Proceedings of the Data for Refugees Challenge Workshop

21 January 2019, Boğaziçi University, Istanbul, Turkey

Scientific Committee:

Albert Ali Salah (Boğaziçi University and Utrecht University)

Alex Pentland (Massachusetts Institute of Technology)

Bruno Lepri (Fondazione Bruno Kessler)

Emmanuel Letouzé (Massachusetts Institute of Technology and Data-Pop Alliance)

Yves-Alexandre de Montjoye (Imperial College London)

Xiaowen Dong (University of Oxford)

Patrick Vinck (Harvard Humanitarian Initiative)

Program

09:00 - 09:30 Opening talks (Boğaziçi University, TUBITAK, Türk Telekom)

09:30 - 10:00 D4R Challenge Award Ceremony

10:00 - 11:00 Session 1 (Oral)

Reducing measles risk in Turkey through social integration of Syrian refugees

Paolo Bosetti, Piero Poletti, Massimo Stella, Bruno Lepri, Stefano Merler and Manlio De Domenico

Data Analytics without Borders: Multi-Layered Insights for Syrian Refugee Crisis

Ozgun Ozan Kılıç, Mehmet Ali Akyol, Oğuz Işık, Banu Günel Kılıç, Arsev Umur Aydınoğlu, Elif Sürer, Hafize Şebnem Düzgün, Sibel Kalaycıoğlu and Tuğba Taşkaya Temizel

UDMIT: An Urban Deep Map for Integration in Turkey

Sedef Turper Alışık, Damla Bayraktar Aksel, Asım Evren Yantaç, Lemi Baruh, Sibel Salman, İlker Kayı, Ahmet İçduygu and Ivon Bensason

11:00 - 12:00 Session 2 (Oral)

AROMA_CoDa: Assessing Refugees' Onward Mobility through the Analysis of Communication Data

Harald Sterly, Benjamin Etzold, Lars Wirkus, Patrick Sakdapolrak, Jacob Schewe, Carl-Friedrich Schleussner and Benjamin Hennig

Measuring fine-grained multidimensional integration using mobile phone metadata: the case of Syrian refugees in Turkey

Michiel Bakker, Daoud Piracha, Patricia Lu, Keis Bejgo, Mohsen Bahrami, Yan Leng, Jose Balsa-Barreiro, Julie Ricard, Alfredo Morales, Vivek Singh, Burcin Bozkaya, Selim Balcisoy and Alex Pentland

Quantified Understanding of Syrian Refugee Integration in Turkey

Wangsu Hu, Ran He, Jin Cao, Lisa Zhang, Huseyin Uzunalioglu, Ahmet Akyamac and Chitra Phadke

12:00 - 14:00 Session 3 (Poster) and lunch

Refugees in undeclared employment - A case study in Turkey

Fabian Bruckschen, Till Koebe, Melina Ludolph, Maria Francesca Marino and Timo Schmid

Mobile Data for Mobility: Travel and Communication Patterns of Syrian Refugees

Eda Beyazıt, Ervin Sezgin, Kerem Arslanli and Mehmet Gencer

Segregation and Sentiment: Estimating Refugee Segregation and its Effects Using Digital Trace Data

Neal Marquez, Emilio Zagheni and Ingmar Weber

Integration of Syrian refugees: insights from D4R, media events and housing market data

Simone Bertoli, Paolo Cintia, Fosca Giannotti, Etienne Madinier, Caglar Ozden, Michael Packard, Dino Pedreschi, Hillel Rapoport, Alina Sirbu and Biagio Speciale

Mobile phone records for exploring spatio-temporal refugee mobility: Links with the Syrian war and socio-economic variations in Turkey

Fatima K. Abu Salem, Al-Abbas Khalil, Ahmad Dhaini, Joachim Diederich, Shady Elbassuoni and Wassim El Hajj

An Overview of Group Behavior on Turkey

Humberto T. M-Neto, Jussara M. Almeida, Artur Ziviani, Virgilio A. F. Almeida, Jaqueline Faria de Oliveira, Douglas C. Teixeira and Haron C. Fantecele

New Approaches to the Study of Spatial Mobility and Economic Integration of Refugees in Turkey

Steven Reece, Franck Duvell, Carlos Vargas-Silva and Zovanga Kone

Syrian Refugee Integration in Turkey: Evidence from Call Detail Records

Tugba Bozcaga, Fotini Christia, Elizabeth Harwood, Constantinos Daskalakis and Christos Papadimitriou

Optimizing the Access to Healthcare Services in Dense Refugee Hosting Urban Areas: A Case for Istanbul

Tarik Altuncu, Nur Sevencan and Ayse Seyyide Kaptaner

Social Integration of Syrian Refugees: Some Insights from Call Detail Record Datasets

Nuran Bayram-Arli, Fatih Cavdur, Mine Aydemir, Fadime Aksoy and Asli Sebatli

Refugee Integration in Turkey: A Study of Mobile Phone Data

Ismail Uluturk, Ismail Uysal and Onur Varol

Measuring Segregation of Syrian Refugees via Mobile Call Detail Records

Fatih Uludağ, Halit Eray Çelik, Serbest Ziyanak, Murat Canayaz and Fikriye Ataman

Reaching all children:A data-driven allocation strategy of educational resources for Syrian refugees.

Suad Aldarra, Lorenzo Lucchini, Elisa Omodei and Laura Alessandretti

Developing Integration Policy for Refugees through Mobile Phone Data Analysis: A Study on Türk Telekom Customers

Ibrahim Zincir, Tohid Ahmed Rana, Ayselin Yıldız and Dilaver Arıkan Açar

14:00 - 15:00 Session 4 (Oral)

Measuring and mitigating behavioural segregation as an optimisation problem: the case of Syrian refugees in Turkey

Daniel Rhoads, Javier Borge-Holthoefer and Albert Solé-Ribalta

Refugee Mobility: Evidence from Phone Data in Turkey

Luisito Bertinelli, Rana Comertpay, Anastasia Litina, Jean-François Maystadt, Benteng Zou and Michel Beine

Mobility and Calling Behavior to Assess the Integration of Syrian Refugees in Turkey

Antonio Luca Alfeo, Mario Giovanni C.A. Cimino, Bruno Lepri and Gigliola Vaglini

15:00 - 16:00 Session 5 (Oral)

Characterizing the Mobile Phone Use Patterns of Refugee Hosting Provinces in Turkey

Ross Gore, Meltem Y. Sener, Christine Boshuijzen-van Burken, Erika Frydenlund, Engin Bozdag and Christa de Kock

Improve Education Opportunities for Better Integration of Syrian Refugees in Turkey

Marco Mamei, Seyit Cilasun, Marco Lippi, Francesca Pancotto and Semih Tumen

Towards an Understanding of Refugee Segregation, Isolation, Homophily and Ultimately Integration in Turkey Using Call Detail Records

Jeremy Boy, David Pastor, Marguerite Nyhan, Rebeca Moreno Jimenez, Daniel Macguire and Miguel Luengo Oroz

Reducing measles risk in Turkey through social integration of Syrian refugees

Paolo Bosetti*, Piero Poletti*, Massimo Stella, Bruno Lepri, Stefano Merler**, and Manlio De Domenico**

Fondazione Bruno Kessler, Via Sommarive 18, 38123 Povo (TN), Italy

Abstract. Turkey hosts almost 3.5M refugees and has to face a humanitarian emergency of unprecedented levels. We use the available Call Detail Records to map the mobility patterns of both Turkish and Syrian refugees, and use these maps to build data-driven computational models for quantifying the risk of epidemics spreading for measles – a disease having a satisfactory immunization coverage in Turkey but not in Syria, due to the recent conflict [de Lima Pereira (2018)] – while accounting for hypothetical policies to integrate the Refugees with the Turkish population. Our results provide quantitative evidence that policies to enhance social integration between refugees and the hosting population might reduce the reproduction number of measles by almost 50%. Moreover, our results suggest that social segregation does not hamper but rather boosts potential outbreaks of measles to a greater extent in Syrian refugees but also in Turkish citizens, although to a lesser extent. This is due to the fact that the high immunization coverage of Turkish citizens can shield Syrian refugees from getting exposed to the infection and this in turn reduces potential sources of infection, in a virtuous cycle reminiscent of herd immunity.

Keywords: Measles · Health · Human mobility · Social Integration

Introduction

Turkey is facing a humanitarian emergency of unprecedented levels. In the last eight years, more than 3.5M Syrians, displaced by the war, have sought refuge in Turkey. This number, through births and new arrivals, is also increasing by approximately 1,000 people per day.

The arrival of a huge amount of people with different economic, health, and living conditions, and from a country where the healthcare system has been almost completely disrupted, may raise serious concerns about the risks of Turkish health systems being overburdened.

For example, Turkish infectious disease specialists are concerned that Syrian refugees' crisis may impose serious risks to their country for infectious diseases previously eliminated or in the process of being eliminated. According to the latest reports from WHO and UNICEF, immunization coverage in Syria dropped from more than 80% before the war to a worrying 41% in 2015 for the most basic vaccines, resulting in millions of unvaccinated children. Direct consequences of this alarming situation are a high risk of epidemics outbreaks (e.g., evidence for polio [1] and measles [2] has been reported) and an increase of mortality due to diseases which could be prevented with vaccines.

Thus, countries, such as Turkey, Lebanon, and Jordan, hosting a great concentration of Syrians perceive the lack of an appropriate immunization coverage as a potential risk of epidemics outbreaks for the local population. This perceived risk ignites a cascade of social dynamics which reinforce: i) segregation of refugees; ii) increase of unemployment and poverty; iii) difficult relationships between healthcare workers and Syrians. It is noteworthy that racial segregation is associated with an increasing of poverty, educational inequalities and increasing of violent crimes. In contrast, boosting social integration leads to societal stability and therefore enhances productivity and individual wealth. In our project, we quantify the risk of observing widespread measles epidemics in Turkey, showing potential public health benefits coming from social integration between Syrian refugees and Turkish citizens.

In particular, measles represents an illustrative case of a highly contagious infectious disease which can be prevented with a safe and effective vaccine. Despite substantial progress towards

* Contributed equally to this work

** Joint senior authors of this work

measles elimination at the global level has been documented, re-emergence of large measles epidemics was observed in the last decade both in low-income and in high-income countries [3]. Measles epidemiology varies widely across different geographical regions, as a consequence of heterogeneous immunity gaps, generated by sub-optimal immunization activities, in different socio-demographic settings [3,4,5,6].

In our project, we propose that the potential spreading of measles in Turkey depends on patterns of human mobility and social mixing among Syrian refugees and Turkish citizens. The crucial role played by both human mobility [7,8,9,10,11,12,13] and mixing patterns [14,15,16,17,18,19,20,21] in shaping the transmission dynamics of infectious disease has been widely documented in the literature and represents a key component of realistic modeling aimed at informing public health policies.

Mobile phone data have been successfully used in the last years as a valuable proxy for human mobility [22,23]. It has been recently shown that such data can be either used to infer socio-demographic information, if missing, or coupled to existing databases to build models of human mobility, mostly based on metapopulation approaches [24,25,26]. Such human mobility models have been used to map the mobility flows between geographical areas at different scales and to improve discrete stochastic modeling of spatial spreading of infectious diseases [24,27,28,29,30]. We capitalize on these works to build, from mobile phone data, a multilayer network [31,32,33] map of human mobility of Turkish citizens and Syrian refugees in Turkey, and we use this knowledge to develop adequate computational models for the potential spreading of measles.

In sum, the contribution of our work is twofold. On the one hand, we quantify the epidemics risk associated with measles in Turkey. On the other hand, we identify an integration policy relating the epidemics risk to policies devised to enhance social integration between Syrian and Turkish populations.

Methods

We developed a simple model of measles spread in realistic epidemiological scenarios. Our model accounts for the current level of immunity to measles in Syrian refugees and Turkish citizens, as inferred from external data sets [34,35] (see Appendix) and, at the same time, for the empirical mobility patterns within Turkey for both populations, as inferred from available mobile phone data in the country [36]. We complement the model by including the effects of social integration of Syrian communities within the Turkish population. Since the current amount of integration is difficult to estimate with available data, we introduce a tunable parameter to account for a variety of scenarios ranging from full segregation to full integration.

The fundamental quantity regulating disease dynamics is the basic reproduction number (R_0), which represents the average number of secondary infections in a fully susceptible population generated by a typical index case during the entire period of infectiousness. Larger R_0, higher the disease transmissibility. If $R_0 > 1$ the infection will be able to spread in a population. Otherwise, the infection will die out. For endemic diseases like measles, R_0 provides insights into the proportion p of immune population (obtained either through vaccination or natural infection) required to prevent large outbreaks; the equation $p = 1 - 1/R_0$ is widely accepted (eg. [37]). For instance, if $R_0 = 20$ at least 95% of the population has to be immune to eliminate the disease. As for measles, typical values of R_0 ranges from 12 to 18 [3,4,37,38,39]. When considering diseases with pre-existing levels of immunity (e.g. childhood diseases like measles), R_0 is a theoretical value representing what could happen in terms of disease transmissibility by removing immunity. In these cases, an appropriate measure of diseases transmissibily is provided by the effective reproduction number (R_e), which represents the average number of secondary infections in a partly immunized population generated by a typical index case during the entire period of infectiousness.

Results are obtained by simulating the spread of measles by assuming different values of R_0 (and thus R_e) and different levels of social integration. Detailed Methods are described in the Appendix.

Results

Two different immunity levels against measles infection are estimated for the Turkish and the Syrian populations. As measles epidemics have not been recently reported in Turkey, we assume

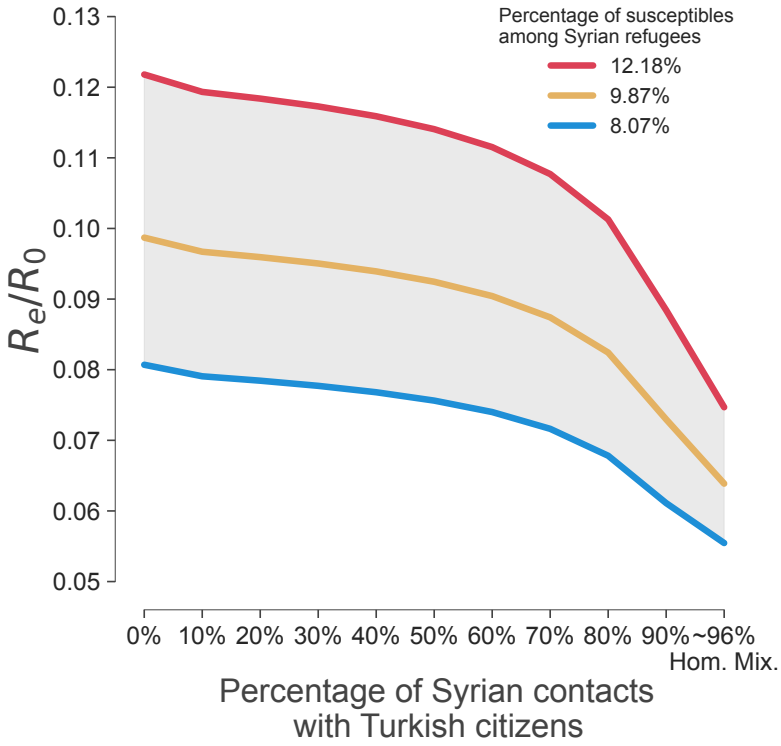


Fig. 1. Effective reproduction number for measles spreading according to our model, rescaled by R_0, as a function of the mixing parameter accounting for social integration between Turkish and Refugees. Coloured lines are associated with the different levels of susceptibility among Syrian refugees within the estimated 95% CI (see Appendix).

the measles immunity level among Turkish citizen reflects the fraction of immunized individuals among birth cohorts between 2006-2016 through 1^{st} and 2^{nd} dose routine vaccination programs (see Appendix). Accordingly, our estimates suggest that only 3.8% of Turkish people might be currently susceptible to measles infection.

Estimates of the immunity level among refugees was instead obtained by inferring the age-specific fraction susceptible individuals in Syria during a recent measles epidemic from the growth rate and age-distribution of cases reported in 2017, and accounting for the age distribution of Syrian refugees in Turkey (see Appendix). We found that the effective reproductive number (R_e) of the recent Syrian measles epidemic was 1.32 (95%CI 1.26–1.38). Consequently, we estimated that the percentage of susceptible individuals in Syria at the beginning of 2017 was 8.92% (95%CI 7.29–10.96). The resulting percentage of susceptible individuals among Syrian refugees in Turkey was estimated to be 9.87% (95%CI 8.07–12.18).

Obtained results suggest that nowadays, in Turkey, 280,000-430,000 out of 3.5M Syrian refugees and about 3M out of 80M Turkish people are measles susceptible.

In Fig. 1 we show the ratio R_e/R_0 as obtained by varying the fraction of Syrian refugees susceptible to measles from 8.07% to 12.18% and by varying the level of social integration from 0% (full segregation of refugees) to 100% (full integration of refugees). We found that pre-existing levels of immunity of the two populations reduce R_e to values lower than 10% of R_0. For example, if R_0 is lower than 10, the probability of observing an epidemic outbreak would be close to 0 because R_e would result lower than 1 as a consequence of pre-existing immunity levels. However, if R_0 is in a more plausible range of values (e.g. 12-18), pre-existing levels of immunity, which are particularly low among Syrian refugees, might not be sufficient to prevent the spread of the disease. Moreover, we found that R_e is maximum when the two populations live socially segregated from each other, whereas it quickly decreases by almost 50% when the two populations are socially well integrated.

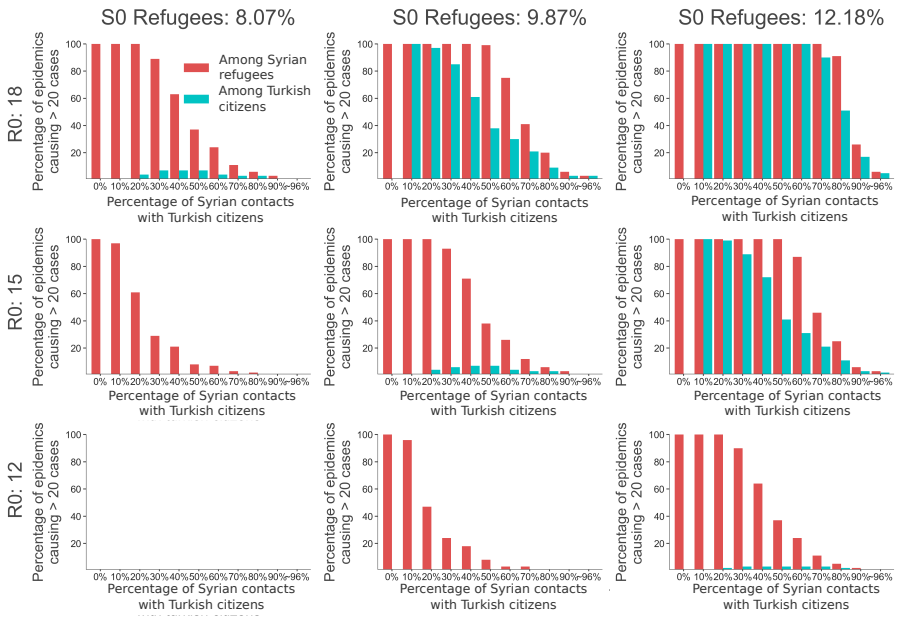


Fig. 2. Percentage of epidemics causing at least 20 cases among Syrian refugees (red bars) or Turkish citizens (azure bars). Three different values of R_0 and immunity levels against measles infection in the Syrian refugees population were considered. For each scenario different proportions of Syrian contacts occurring with Turkish citizens were evaluated. These vary between the total segregation of the Syrian refugees (0%) to the homogeneous mixing between the two population ($\sim 96\%$).

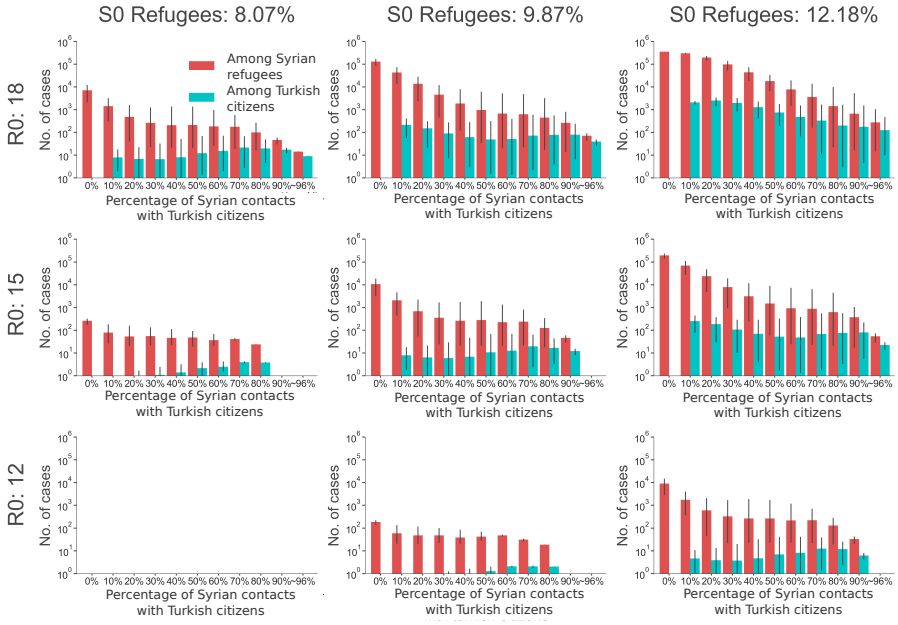


Fig. 3. Cumulative infections by considering epidemics that exceed 20 cases in the entire population. Red and azure bars represent the average number of infections occurring among the Syrian refugees and the Turkish citizens for the model projections, vertical black lines represent 95%CI. Three different values of R_0 and immunity levels against measles infection in the Syrian refugees population were considered. For each scenario different proportions of Syrian contacts occurring with Turkish citizens were evaluated. These vary between the total segregation of the Syrian refugees (0%) to the homogeneous mixing between the two population ($\sim 96\%$).

The immunity level characterizing the Turkish population in 2017 is expected to prevent the spread of future measles epidemic in geographical locations predominantly populated by Turkish

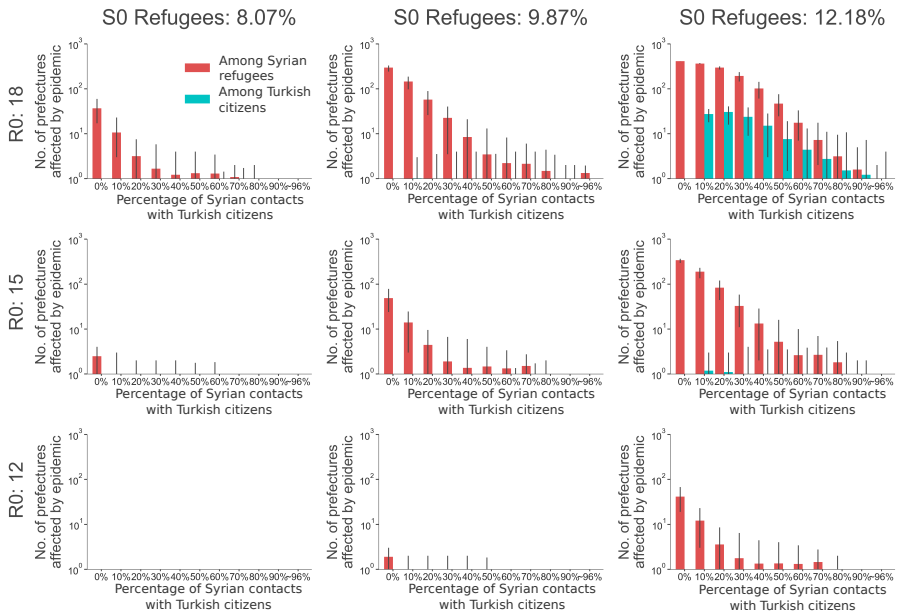


Fig. 4. Estimated number of prefectures affected by the epidemic. Red and azure bars represent the average number of prefectures exceeding 20 cases among the Syrian refugees and the Turkish citizens respectively, vertical black lines represent 95%CI. Three different values of R_0 and immunity levels against measles infection in the Syrian refugees population were considered. For each scenario different proportions of Syrian contacts occurring with Turkish citizens were evaluated. These vary between the total segregation of the Syrian refugees (0%) to the homogeneous mixing between the two population ($\sim$ 96%).

citizens. However, if a measles index case would occur in a population with a sufficiently large proportion of Syrian people, transmission events will be sustained by the lack of adequate immunity levels among refugees. Our modelling analysis show that for any scenario considered the risk of observing large epidemics increases with the basic reproduction number and the proportion of susceptible among the refugees (see Fig. 2,3).

In case of full segregation of refugees (although practically infeasible and therefore unlikely), potential measles epidemics would result in dramatic health consequences among refugees, causing a huge amount of measles cases widespread in the country (Fig. 2,3,4). Specifically, when $R_0 = 15$ is considered and 9.8% of refugees are assumed to be measles susceptible, the probability of observing an epidemic with more than 20 cases is 100% (see Fig. 2) and the final size of potential epidemics is expected to exceed 10,000 cases (mean estimate 10,662 95%CI 3,172–18,414, see Fig. 3). Our results show that the risk of observing sustained transmission in the country is large for any value of R_0 larger than 15 but also for lower values of R_0 (e.g. $R_0 = 12$) and the proportion of refugees susceptible is 9.8% or more (see Fig. 2).

In the case of full segregation, infections would occur only among refugees. However, when assuming high level of segregation (i.e. only a small fraction, yet equal or greater than 10%, of refugees' contacts occur with Turkish citizens), the risk of experiencing large measles outbreak is high (Fig. 2) and measles epidemics could produce non-negligible spillover of cases among Turkish citizens as well (Fig. 3). In particular, in a worst case scenario where $R_0 = 18$, 12.2% of refugees are susceptible and more than 70% of Syrian contacts occur with Syrian people, thousands of measles cases are expected all over the country among the Turkish people as well (see Fig. 3,5).

Obtained results suggest that both the risk of observing measles sustained measles transmission and the final size of potential epidemics are significantly smaller in the presence of high levels of integration of refugees (Fig. 2,3). Specifically, when $R_0 = 15$ is considered, 9.8% of refugees are assumed to be measles susceptible and refugees well mix with the Turkish (e.g. more than 70% of Syrian contacts occur with Turkish people), the propbability of observing epidemic outbreak dramatically decreases to values lower than 10% (Fig. 2). Moreover, in case of outbreak, the expected overall number of cases is no larger than few hundred (Fig. 3), as potentially infectious

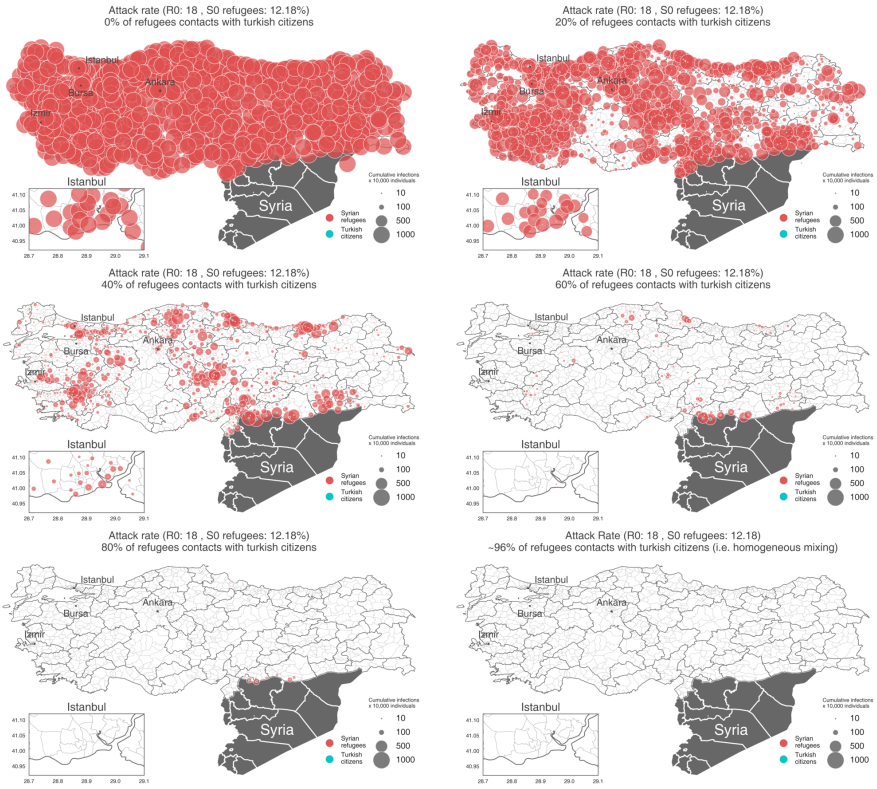


Fig. 5. Estimated measles incidence one year after the first infection, considering the worst case scenario in terms of R_0 and immunity levels against measles infection among Syrian refugees. Red and azure bubbles indicate measles cases among Syrian refugees and the Turkish population respectively. Bubbles size are proportional to the average number of measles cases estimated in Turkish prefectures (per 10,000 individuals). Different proportions of Syrian contacts occurring with Turkish citizens were evaluated. **Inset** displays the Istanbul prefectures.

contacts would more probably occur with Turkish immune individuals, who represent about 90% of individuals currently leaving in Turkey.

Remarkably, larger segregation levels also promote the spatial invasion of the epidemic across the whole country (Fig. 4). In the worst case scenario of $R_0 = 18$, 12.18% of Syrian refugees susceptible, and more than 90% of Syrian contacts occur within the Syrian population, the measles epidemic is expected to affect more than 300 prefectures of Turkey (Fig. 4,6). On the opposite, if more than 70% of contacts of refugees would occur with Turkish people, as a consequence of good integration of refugees with Turkish citizens, for the majority of epidemiological scenarios considered, measles epidemics are expected to remain geographically bounded in less than 10 out of 1021 prefectures of the country (Fig. 4,6).

Figure 7 reports the cumulative incidence of infections over time, in terms of both absolute numbers (left panel) and counts normalized by the average expected population in a prefecture (right panel) in case when no social integration is present. Both the curves of cumulative incidence indicate the presence of some prefectures where more cases of infections are registered at earlier stages. Further, Fig. 7 highlights a cluster of 96 prefectures where the infection spreads earlier and it can reach even up to 10^4 cases in one year. These prefectures, which are strongly affected by the epidemics of measles in our simulations, are mainly urban areas (79 prefectures out of 96) and include metropolitan areas like Istanbul (25 prefectures) and Ankara (5 prefectures). In turn, these results indicate that the absence of social integration in urban, metropolitan areas can boost the incidence of infections.

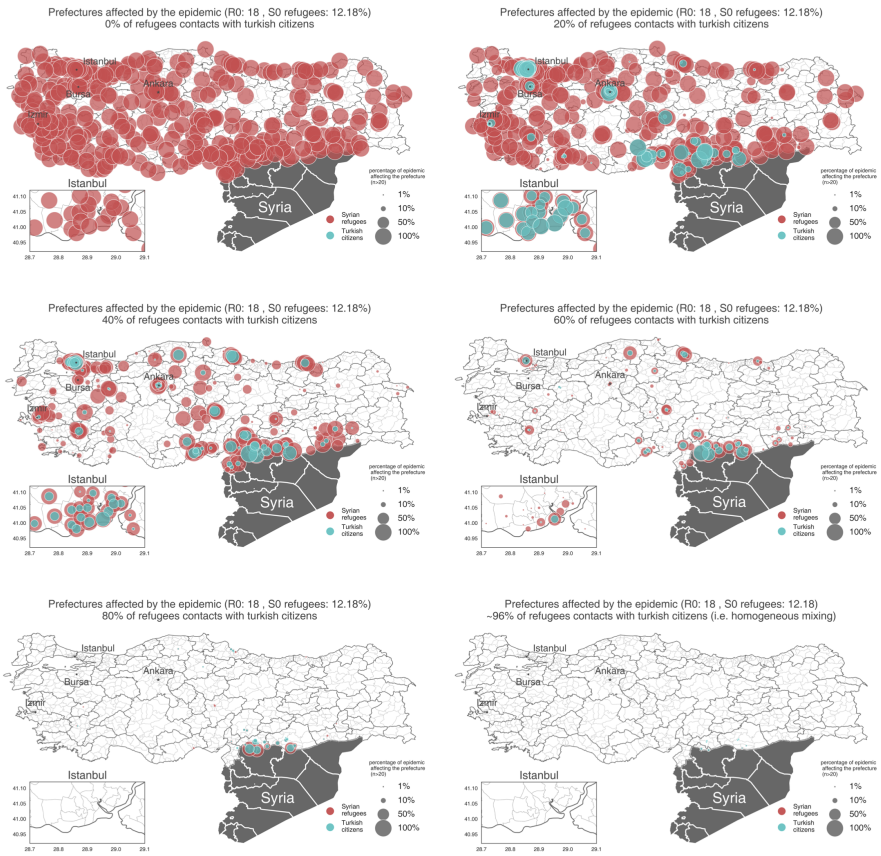


Fig. 6. Spatial invasion. Percentage of epidemic exceeding 20 cases per prefecture one year after the first infection. Red and azure bubbles refer to Syrian refugees and the Turkish population respectively. Different proportions of Syrian contacts occurring with Turkish citizens were evaluated. **Inset** displays the Istanbul prefectures

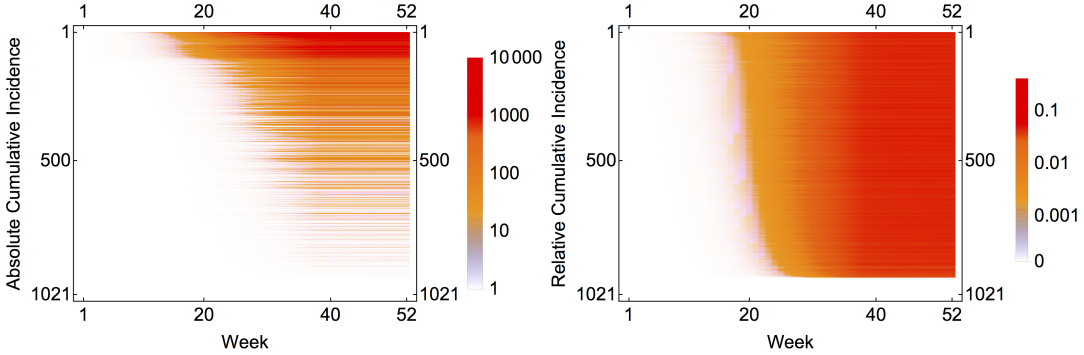


Fig. 7. Left: Raw counts for the cumulative incidence per prefecture over time in the case of full segregation, $R_0 = 15$ and 12.18% of susceptible individuals among Syrian refugees. Right: Relative counts of cumulative incidence over time, normalized by the average population in a prefecture. Prefectures are ranked in decreasing order of cumulative incidence at week 20.

Discussion

The widely accepted critical immunity threshold for measles elimination is 95% of immune individuals. According to our estimates, while Turkish citizens are mostly protected by high vaccine

uptake levels, Syrian refugees display a considerably larger fraction of individuals that is susceptibile to the measles as a consequence of the sub-optimal vaccination during the ongoing war. More specifically, while the level of protection of the Turkish population against the disease is nearly optimal (more than 96% of immune individuals), that of Syrian refugees is far from being acceptable (only about 90% of immune individuals, though highly uncertain).

The strong difference in the immunity levels among the two populations may have deep repercussions on the way society perceives the movement of Syrians within Turkey. As common in Western countries hosting considerable amounts of migrants [40], Turkish citizens might perceive the lower immunization coverage of Syrian refugees as a potential threat to national welfare and health. This perception might be even worsened by the staggering numbers of Syrian refugees registered in Turkey, 3.5M in 2018. This well documented negative perception, in turn, motivates potential segregation mechanisms, aimed at reducing as much as possible interactions and contacts between Syrian refugees and Turkish citizens.

The carried out analysis provides compelling evidence that social segregation does not hamper but rather boosts potential outbreaks of measles to a greater extent in Syrian refugees but also in Turkish citizens, although to a lesser extent. The main result of the current study is the quantitative evidence that social mixing among Syrian refugees and Turkish citizens can be highly beneficial in drastically reducing the incidence and the strength of infection of measles. This is due to the fact that the high immunization coverage of Turkish citizens can shield Syrian refugees from getting exposed to the infection and this reduces potential sources of infection, in a virtuous cycle reminiscent of herd immunity and well documented in many real-world social systems [41]. Our quantitative model combines CDRs data with available epidemiological evidences to estimate the spatial distribution, the immunity profile and the mobility patterns characterizing the two considered populations, allowing the investigation of spatio-temporal patterns of a potential measles epidemic in Turkey.

If social integration is beneficial in terms of reducing the incidence of measles, with possible cost-saving consequences on the economy of the whole country, then the main question, from a policy-making perspective, becomes how to enable and boost social integration itself.

Provided that a full homogeneous mixing of refugees and citizens could prove to be impracticable or rather difficult to achieve, there are several policies that could reduce social segregation. For instance, designing specific housing policies for redistributing refugees across different neighbourhoods of a given metropolitan area could avoid the creation of ghettos, while also increasing the chances of social interactions between refugees and citizens in schools, shops, third places, etc. Although the proposed analysis clearly shows that increasing social mixing between Syrian refugees and Turkish citizens is expected to produce positive public health outcomes, social integration is also expected to provide major societal benefits such as the reduction of violent crimes, economic and educational inequalities.

From a geographic perspective, our analysis confirmed that there are metropolitan areas that are pivotal in diffusing the incidence of the disease over time. These areas are mainly prefectures of Istanbul and Ankara and, unsurprisingly, include also many areas adjacent to the national borders of Turkey with Syria. It is in these areas that the efforts for reducing social segregation should be strategically focused. This poses a great challenge for the future, provided that recent reviews of urban regeneration projects highlighted an important process of social segregation of minorities and non-Turkish ethnicities particularly strong in large cities such as Istanbul [42]. Additionaly, in areas characterized by a large amount of refugees with respect to the Turkish population, as it is the case of many prefectures close to the Syrian border, targeted immunization strategies might critically reduce the chances of measles transmission and prevent the onset of widespread epidemics.

The performed analysis has several limitations that should be considered in interpreting the results. Estimates of immunity levels in Syrian refugees and, to a lower extent, in the Turkish citizens should be considered cautiously as no recent serological surveys are available for the two populations. Immunity levels are inferred from the analysis of vaccine coverage for Turkish citizens and from the analysis of the 2016-2018 measles outbreak in Syria. This last analysis in particular might be affected by under-reporting of cases and does not consider potential spatial heterogeneities that could drastically affect estimates of the overall level of protection against the disease. Also, we assume the same levels of immunity in all municipalities, thus neglecting spatial heterogeneities, for instance induced by differences in vaccine uptake among Turkish citizens. Moreover, no data

on mixing patterns (e.g. by age) are available for either Syrian refugees and Turkish citizens. Consequently, the model neglects potential differences in measles transmissibility by age of individuals and, similarly, potential differences in measles transmissibility for Syrian refugees and Turkish citizens (e.g. induced by different numbers of overall contacts). Finally, CDRs data used in the proposed analysis are associated with only a fraction of the population. Although these data may not perfectly reflect real movements occurring across all the prefectures in the country, they provide valuable evidence to infer a fair approximation of human mobility in the country driving the spatio-temporal spread of an epidemic.

All this considered, the analysis carried out represents a first attempt to quantify the risk of measles outbreak in Turkey and provides striking evidence that, besides policies aimed at increasing vaccination coverage among Syrian refugees, social integration of Refugees within the Turkish population might be an effective countermeasure.

Appendix

Mobility model

To model the mobility of Turkish and Syrian refugees, we assume two populations of individuals, namely population 1 of size N^1 and population 2 of size N^2, living in a territory consisting of K geographically separated patches (i.e., Turkish prefectures) accounting for N_k^1 and N_k^2 individuals, $k = 1, \ldots, K$ with $\sum_{k=1}^{K} N_k^1 = N^1$ and $\sum_{k=1}^{K} N_k^2 = N^2$.

The absolute number of individuals moving between patches is inferred from available Call Detail Records as in Refs. [30,43] and rescaled to adequately represent the volumes corresponding to 80M Turkish individuals and 3.5M Syrian refugees.

Let c_{ik}^1 be the daily number of individuals of population 1 travelling from path i to patch k, with $\sum_{k=1}^{K} c_{ik}^1 = N_i^1$ if we consider that c_{ii}^1 represents non-travelling individuals. Similarly, let c_{ik}^2 be the daily number of individuals of population 2 travelling from path i to path k, with $\sum_{k=1}^{K} c_{ik}^2 = N_i^2$. We show in Fig. 8 the mobility patterns encoded by the two matrices. The number of individuals in any patch i is therefore:

$$P_i(c) = \sum_{k=1}^{K} c_{ki}^1 + \sum_{k=1}^{K} c_{ki}^2. \tag{1}$$

Epidemic transmission model

To account for the effects of social integration and mobility dynamics, simultaneously, we assume that individuals of population 1 mix homogeneously among themselves and with a fraction α (with $0 \leq \alpha \leq 1$) of individuals of population 2, the contact rate of individuals of population 1 in patch i will be proportional to

$$P_i^1(\alpha, c) = \sum_{k=1}^{K} c_{ki}^1 + \alpha \sum_{k=1}^{K} c_{ki}^2. \tag{2}$$

Similarly, the contact rate of individuals of population 2 in patch i will be proportional to:

$$P_i^2(\alpha, c) = \alpha \sum_{k=1}^{K} c_{ki}^1 + \sum_{k=1}^{K} c_{ki}^2. \tag{3}$$

$\alpha = 0$ represents the situation of two completely separated populations; $\alpha = 1$ corresponds to homogeneous mixing among individuals of the two populations.

Let us assume that the contact rate of each individual is equal to a certain value σ for all individuals. Basically, here we assume that the contact rate does not depend on population type, mobility, mixing, and geography. The following equations must be satisfied in any patch i:

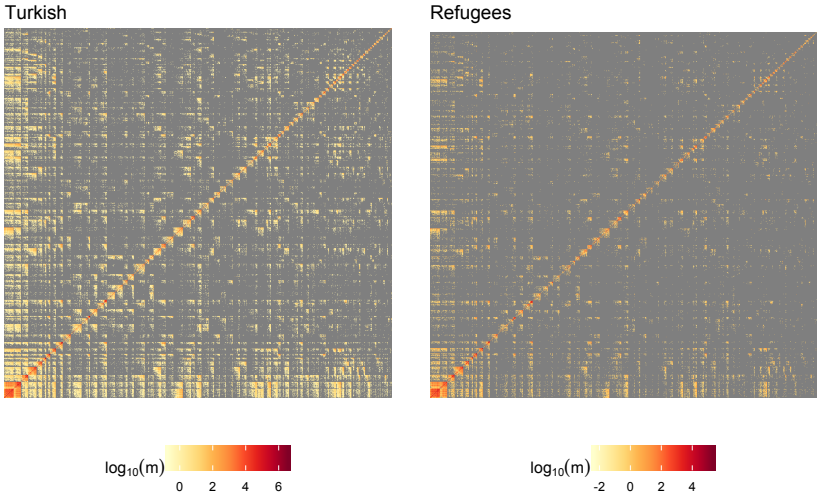


Fig. 8. Mobility patterns of Turkish (left) and Refugee (right) populations, encoded by origin-destination matrices inferred from CDR. Rows are ordered according to the mesoscale structure – estimated by means of Infomap [44,45] – of the underlying mobility network. Both flow direction and volume are taken into account for this estimation. Color encodes the estimated volume of the human flow.

$$\sigma = \sigma^{*1} P_i^1(\alpha, c); \qquad \sigma = \sigma^{*2} P_i^2(\alpha, c), \tag{4}$$

for individuals of population 1 and 2, respectively, which are satisfied by setting:

$$\sigma_i^{*1}(\alpha, c) = \sigma / P_i^1(\alpha, c); \qquad \sigma_i^{*2}(\alpha, c) = \sigma / P_i^2(\alpha, c). \tag{5}$$

The rate of contacts of individuals of population 1 with infected individuals is therefore:

$$Q_i^1(\alpha, c) = \sigma_i^{*1}(\alpha, c) \left[\sum_{k=1}^{N} c_{ki}^1 \frac{I_k^1}{N_k^1} + \alpha \sum_{k=1}^{K} c_{ki}^2 \frac{I_k^2}{N_k^2} \right], \tag{6}$$

where $\frac{I_k^1}{N_k^1}$ and $\frac{I_k^2}{N_k^2}$ represent the fraction of infected individuals of the two populations in patch k. Note that the term between square brackets represents the number of infected individuals among $P_i^1(\alpha, c)$. Similarly, the rate of contacts of individuals of population 2 with infected individuals is:

$$Q_i^2(\alpha, c) = \sigma_i^{*2}(\alpha, c) \left[\alpha \sum_{k=1}^{N} c_{ki}^1 \frac{I_k^1}{N_k^1} + \sum_{k=1}^{K} c_{ki}^2 \frac{I_k^2}{N_k^2} \right], \tag{7}$$

Let p be the probability of infection transmission given a contact and let $\beta = p\sigma$ be the transmission rate. Susceptible individuals of population 1 in any patch i are exposed to the following force of infection:

$$\lambda_i^1(\alpha, c) = \beta \left[\sum_{k=1}^{K} \frac{c_{ki}^1}{P_i^1(\alpha, c)} \frac{I_k^1}{N_k^1} + \alpha \sum_{k=1}^{K} \frac{c_{ki}^2}{P_i^1(\alpha, c)} \frac{I_k^2}{N_k^2} \right] \tag{8}$$

Similarly, the force of infection for individuals of population 2 is given by:

$$\lambda_i^2(\alpha, c) = \beta \left[\alpha \sum_{k=1}^{K} \frac{c_{ki}^1}{P_i^2(\alpha, c)} \frac{I_k^1}{N_k^1} + \sum_{k=1}^{K} \frac{c_{ki}^2}{P_i^2(\alpha, c)} \frac{I_k^2}{N_k^2} \right]. \tag{9}$$

The system of ordinary differential equations regulating the epidemic transmission dynamics is therefore the following:

$$\begin{cases} \dot{S}_i^1 = -\lambda_i^1(\alpha, c) S_i^1 \\ \dot{S}_i^2 = -\lambda_i^2(\alpha, c) S_i^2 \\ \dot{I}_i^1 = \lambda_i^1(\alpha, c) S_i^1 - \gamma I_i^1 \\ \dot{I}_i^2 = \lambda_i^2(\alpha, c) S_i^2 - \gamma I_i^2 \\ \dot{R}_i^1 = \gamma I_i^1 \\ \dot{R}_i^2 = \gamma I_i^2 \end{cases} \tag{10}$$

for $i = 1, ..., K$, where $\gamma^{-1} = 15$ days is the exponentially distributed generation time.

Initial conditions. We assume different levels of protection against the disease f^1 and f^2 ($0 \leq f^j \leq 1$) for individuals of the two populations. We also assume that, at time $t = 0$, the epidemic is seeded by one single index case randomly chosen in a given patch. If the index case is an individual belonging to population 1 and living in patch i^*, the initial conditions in patch i^* are $S_{i^*}^1(0) = (1 - f^1)(N_{i^*}^1 - 1)$; $S_{i^*}^2(0) = (1 - f^2)N_{i^*}^2$; $I_{i^*}^1(0) = 1$; $I_{i^*}^2(0) = 0$; $R_{i^*}^1(0) = f^1(N_{i^*}^1 - 1)$; and $R_{i^*}^2(0) = f^2 N_{i^*}^2$. In patches $i \neq i^*$, the initial conditions are $S_i^1(0) = (1-f^1)N_i^1$; $S_i^2(0) = (1-f^2)N_i^2$; $I_i^1(0) = 0$; $I_i^2(0) = 0$; $R_i^1(0) = f^1 N_i^1$; and $R_i^2(0) = f^2 N_i^2$.

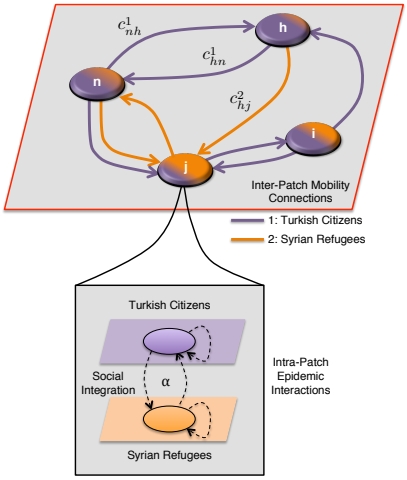


Fig. 9. Schematic illustration of the model considered in this work. Each prefecture of Turkey is considered as a node of a metapopulation network of geographic patches. Two populations, namely Turkish and Syrians, are encoded by different colors and move between patches following the inferred inter-patch mobility pathways. Turkish and Syrian populations encode two different layers of a multilayer system [31,32,33] where social dynamics and epidemics spreading happen simultaneously.

A schematic illustration, summarizing the coupled dynamics of human mobility, social integration and epidemic spreading, is represented in Fig. 9.

Reproduction numbers. Reproduction numbers associated to the epidemic transmission model are computed in a standard way by applying next generation matrix techniques [46,47,48]. If we define $X = (I_1^1, ..., I_K^1, I_1^2, ..., I_K^2)$ and $Y = (S_1^1, ..., S_K^1, S_1^2, ..., S_K^2)$, it is straightforward to observe that equations for X can be written in the form

$$\dot{X}_i = Y_i \sum_{k=1}^{2K} p\sigma m_{ik} \frac{X_k}{N_k} - \gamma X_i \tag{11}$$

for appropriate choices of coefficients m_{ik}, namely:

$$m_{ik} = \frac{c^1_{ki}}{P^1_i(\alpha, c)} \ (i = 1, ..., K; k = 1, ..., K), \tag{12}$$

$$m_{ik} = \frac{\alpha c^2_{ki}}{P^1_i(\alpha, c)} \ (i = 1, ..., K; k = K+1, ..., 2K), \tag{13}$$

$$m_{ik} = \frac{\alpha c^1_{ki}}{P^2_i(\alpha, c)} \ (i = K+1, ..., 2K; k = 1, ..., K), \tag{14}$$

$$m_{ik} = \frac{c^2_{ki}}{P^2_i(\alpha, c)} \ (i = K+1, ..., 2K; k = K+1, ..., 2K). \tag{15}$$

Note that the terms σm_{ik} represent numbers of contacts that individuals in patch i have with individuals of patch k and thus, put in this form, the model resembles a classical age structured SIR model. Let $\mathbf{M}$ be the matrix with entries m_{ik}. It follows that:

$$R_0 = p\rho(\sigma\mathbf{M})\gamma^{-1}, \tag{16}$$

where $\rho(\sigma\mathbf{M})$ indicates the maximal eigenvalue of $\sigma\mathbf{M}$. Since $\mathbf{M}$ is a probability matrix (also termed transition matrix, i.e. all rows sum up to 1), and thus $\rho(\mathbf{M}) = 1$, it follows that $R_0 = \beta\gamma^{-1}$, as for simple homogeneous mixing SIR models.

The effective reproduction number can be computed in a similar way, but accounting for the susceptibility of infectors, that is by defining $\mathbf{M}$ as the matrix with entries:

$$m^*_{ik} = m_{ik}(1 - f^1) \ (i = 1, ..., K; \ k = 1, ..., K) \tag{17}$$

$$m^*_{ik} = m_{ik}(1 - f^2) \ (i = 1, ..., K; \ k = K+1, ..., 2K) \tag{18}$$

$$m^*_{ik} = m_{ik}(1 - f^1) \ (i = K+1, ..., 2K; \ k = 1, ..., K) \tag{19}$$

$$m^*_{ik} = m_{ik}(1 - f^2) \ (i = K+1, ..., 2K \ k = K+1, ..., 2K). \tag{20}$$

Estimating measles immunity levels among Turkish citizens and Syrian refugees. Two different immunity levels against measles infection are assumed in the Turkish and the Syrian populations, hereafter denoted by f^1 and f^2 respectively. As measles epidemics have not been recently reported in Turkey, we assume that f^1 reflects the fraction of immunized individuals among recent birth cohorts through 1^{st} and 2^{nd} dose routine vaccination programs. In particular, by assuming a vaccine efficacy $e = 95\%$ [49] and considering the average coverage levels for the 1^{st} and 2^{nd} doses reported by the WHO for the period 2006-2016, $c_1 = 97\%$ and $c_2 = 88\%$ respectively [34], we estimate $1 - f^1$ as:

$$1 - f^1 = 1 - c_1 + c_1 e(1 - c_2) + c_1 c_2 (1 - e)^2 \tag{21}$$

where $1 - c_1$ denotes the fraction of individuals who have never been vaccinated, $c_1 e(1 - c_2)$ represents the fraction of individuals who have been vaccinated only with 1^{st} dose but have experienced vaccine failure (occurring in a fraction 1-e of vaccinee), and $c_1 c_2 (1 - e)^2$ defines the fraction of individuals who experienced vaccine failure after 2 dose administrations.

Estimates of f^2 were obtained by inferring the fraction of susceptible individuals among different age classes in the Syrian population, on the basis of data on the measles epidemic reported in Syria during 2017 (> 700 cases) [35], and accounting for the age distribution of Syrian refugees in Turkey. More specifically, we estimate the effective reproductive number (R_e) associated with the 2017 epidemic as $R_e = 1 + rT_g$, where $T_g = 15$ days is the measles generation time [37], r is the exponential growth rate in the weekly number of cases reported during 2017 (Fig. 1A,B). Estimates of R_e are used to derive the fraction of susceptible individuals in Syria at the beginning of 2017 as $S_0 = R_e/R_0$, using three values of R_0: 12, 15 and 18 [38,39]. Estimates of S_0 are combined with the age distribution of observed cases (Fig. 1C) and used to estimate the age specific immunity profile of the Syrian population. Specifically, the fraction of immune individuals in each age group a (Fig. 1D) is approximated as:

$$imm(a) = 1 - S_0 \frac{cases(a)}{\sum_j cases(j)}, \tag{22}$$

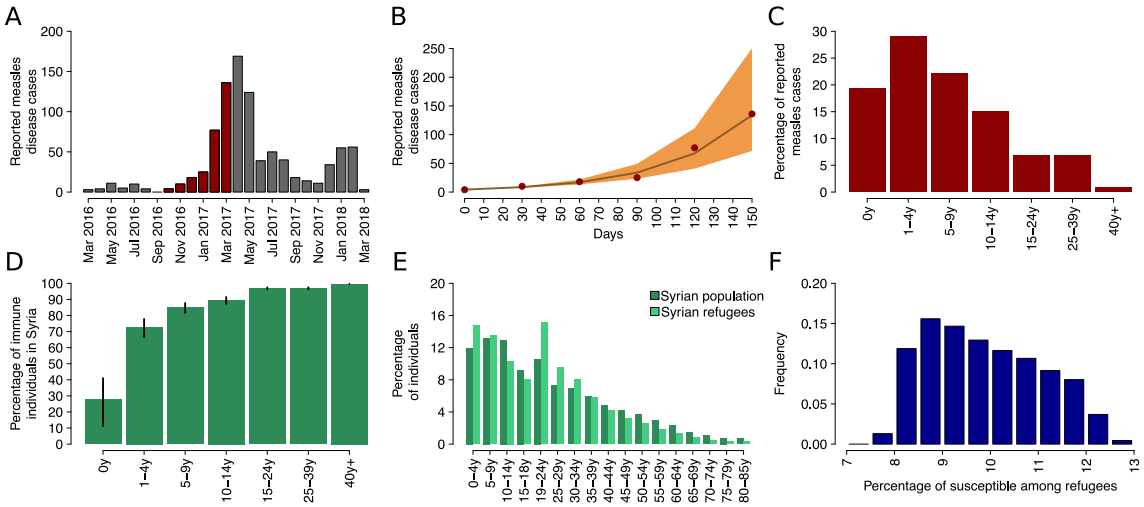


Fig. 10. A Reported number of measles disease cases over time, during the 2016-2018 measles epidemics in Syria as recently reported by the World Heath Organization [35]; red bars correspond to data points used to derive the R_e as a function of the exponential growth rate of the observed epidemic. **B** Obtained fit of the exponential growth of the epidemic between September and February 2017 in Syria: red solid line represents the mean estimate, orange shaded area represents 95%CI. **C** Observed distribution of measles cases across different ages during the 2016-2018 measles epidemics in Syria as recently reported by the World Heath Organization [35]. **D** Estimated age specific serological profile in Syria at the beginning of 2017: green bars represents mean values, vertical black lines represent 95%CI. **E** Observed age distribution of Syrian refugees in Turkey (light green)[50] compared with the population age distribution in Syria (dark green). **F** Estimated percentage of susceptible among Syrian refugees.

where $cases(a)$ denotes the total number of cases observed in age a. Finally, the fraction of susceptible Syrian refugees in Turkey was obtained by combining the age specific immunity profile estiamted for Syria and the age distribution of Syrian refugees in Turkey [50] (Fig. 1E). Estimates obtained on the fraction of measles susceptible refugees (i.e. $1 - f^2$) are shown in Fig. 1F.

The obtained results suggest that the effective reproductive number (R_e) of the recent Syrian measles epidemic was 1.32 (95%CI 1.26–1.38), the percentage of susceptible individuals in Syria at the beginning of the 2017 measles epidemic (S0) was 8.92 (95%CI 7.29–10.96); consequently, the percentage of susceptible individuals among Syrian refugees in Turkey was estimated to be 9.87% (95%CI 8.07–12.18).

Simulating measles epidemics in Turkey Simulations of measles epidemics in Turkey were obtained under three different scenarios of $R_0 = 12, 15, 18$, three different scenarios of $1 - f^2 = 0.0987, 0.0807, 0.1218$ and eleven illustrative values of α. Explored values of α were selected in such a way to reproduce different proportion of Syrian contacts occurring with Turkish citizens: from 0% to 96%; the former representing the full segregation scenario and the latter representing homogeneous mixing between Syrian refugees and Turkish citizens. Starting from Eq. 2 and 3 it is easy to see that, for any given value of α, the fraction of contacts that Syrian refugees have with Turkish citizens in patch i is given by $\alpha B_i/(A_i + \alpha B_i)$ where A_i and B_i are the number of Syrian refugees and Turkish citizens in patch i respectively. It follows that the average fraction of contacts that Syrian refugees have with Turkish citizens in the whole study area is given by $\sum_i w_i \ \alpha B_i/(A_i + \alpha B_i)$, where $w_i = A_i/\sum_i A_i$. Similarly, the average fraction of contacts that Turkish citizens have with Syrian refugees in the whole study area is given by $\sum_i \tilde{w}_i \ \alpha A_i/(B_i + \alpha A_i)$, where $\tilde{w}_i = B_i/\sum_i B_i$. Note that for a given value of α the fraction of Syrian contacts with Turkish citizens is generally different from the fraction of Turkish contacts with the Syrians.

The value of α resulting in a certain fraction x of contacts of Syrian refugees with Turkish citizens can therefore be computed by solving the equation

$$\sum_i w_i \frac{\alpha B_i}{A_i + \alpha B_i} = x. \tag{23}$$

For each considered scenario, 100 measles epidemics were simulated for a year, by seeding each epidemic in one different patch, selected among the 100 prefectures of Turkey with the highest amount of refugees. All simulations were performed under the assumption of $1 - f^1 = 0.038$. An illustrative value of $1 - f^1 = 0.05$ was also considered for sensitivity analysis but it is not shown here for lack of space.

References

1. Organization, W. H. WHO Polio Report, Syria cVDPV2 outbreak Situation Report No. 33. Tech. Rep. (2018).
2. Organization, W. H. WHO Measles Syria, EWARS Weekly Bulletin Week No. 50. Tech. Rep. (2017).
3. Trentini, F., Poletti, P., Merler, S. & Melegaro, A. Measles immunity gaps and the progress towards elimination: a multi-country modelling analysis. Lancet Infect Dis **17**, 1089–1097 (2017).
4. Merler, S. & Ajelli, M. Deciphering the relative weights of demographic transition and vaccination in the decrease of measles incidence in Italy. Proc Biol Sci **281** (2014).
5. Ajelli, M., Merler, S., Fumanelli, L., Bella, A. & Rizzo, C. Estimating measles transmission potential in Italy over the period 2010-2011. Ann Ist Super Sanita **50**, 351–356 (2014).
6. Ajelli, M. & Merler, S. The impact of the unstructured contacts component in influenza pandemic modeling. PLoS One **3**, e1519 (2008).
7. Marziano, V., Pugliese, A., Merler, S. & Ajelli, M. Detecting a Surprisingly Low Transmission Distance in the Early Phase of the 2009 Influenza Pandemic. Sci Rep **7**, 12324 (2017).
8. Zhang, Q. et al. Spread of Zika virus in the Americas. Proc Natl Acad Sci USA **114**, E4334–E4343 (2017).
9. Ajelli, M. et al. Comparing large-scale computational approaches to epidemic modeling: agent-based versus structured metapopulation models. BMC Infect Dis **10**, 190 (2010).
10. Merler, S. & Ajelli, M. The role of population heterogeneity and human mobility in the spread of pandemic influenza. Proc Biol Sci **277**, 557–565 (2010).
11. Merler, S., Ajelli, M., Pugliese, A. & Ferguson, N. M. Determinants of the spatiotemporal dynamics of the 2009 H1N1 pandemic in Europe: implications for real-time modelling. PLoS Comput Biol **7**, e1002205 (2011).
12. Degli Atti, M. C. et al. Modelling scenarios of diffusion and control of pandemic influenza, Italy. Eurosurv **12**, pii=3105 (2007).
13. Ciofi degli Atti, M. L. et al. Mitigation measures for pandemic influenza in Italy: an individual based model considering different scenarios. PLoS One **3**, e1790 (2008).
14. Ajelli, M., Poletti, P., Melegaro, A. & Merler, S. The role of different social contexts in shaping influenza transmission during the 2009 pandemic. Sci Rep **4**, 7218 (2014).
15. Iozzi, F. et al. Little Italy: an agent-based approach to the estimation of contact patterns- fitting predicted matrices to serological data. PLoS Comput Biol **6**, e1001021 (2010).
16. Guzzetta, G. et al. Modeling socio-demography to capture tuberculosis transmission dynamics in a low burden setting. J Theor Biol **289**, 197–205 (2011).
17. Fumanelli, L., Ajelli, M., Manfredi, P., Vespignani, A. & Merler, S. Inferring the structure of social contacts from demographic data in the analysis of infectious diseases spread. PLoS Comput Biol **8**, e1002673 (2012).
18. Merler, S. et al. Spatiotemporal spread of the 2014 outbreak of Ebola virus disease in Liberia and the effectiveness of non-pharmaceutical interventions: a computational modelling analysis. Lancet Infect Dis **15**, 204–211 (2015).
19. Ajelli, M. et al. Spatiotemporal dynamics of the Ebola epidemic in Guinea and implications for vaccination and disease elimination: a computational modeling analysis. BMC Med **14**, 130 (2016).
20. Ajelli, M. et al. The 2014 Ebola virus disease outbreak in Pujehun, Sierra Leone: epidemiology and impact of interventions. BMC Med **13**, 281 (2015).
21. Fumanelli, L., Ajelli, M., Merler, S., Ferguson, N. M. & Cauchemez, S. Model-Based Comprehensive Analysis of School Closure Policies for Mitigating Influenza Epidemics and Pandemics. PLoS Comput Biol **12**, e1004681 (2016).
22. Blondel, V. D., Decuyper, A. & Krings, G. A survey of results on mobile phone datasets analysis. EPJ Data Science **4**, 10 (2015).
23. Barbosa, H. et al. Human mobility: Models and applications. Physics Reports (2018).
24. Balcan, D. et al. Multiscale mobility networks and the spatial spreading of infectious diseases. Proceedings of the National Academy of Sciences pnas–0906910106 (2009).
25. Balcan, D. & Vespignani, A. Phase transitions in contagion processes mediated by recurrent mobility patterns. Nature Physics **7**, 581 (2011).

26. Gómez-Gardeñes, J., Soriano-Paños, D. & Arenas, A. Critical regimes driven by recurrent mobility patterns of reaction–diffusion processes in networks. Nature Physics **14**, 391 (2018).
27. Balcan, D. et al. Modeling the spatial spread of infectious diseases: The global epidemic and mobility computational model. Journal of computational science **1**, 132–145 (2010).
28. Meloni, S. et al. Modeling human mobility responses to the large-scale spreading of infectious diseases. Scientific Reports **1**, 62 (2011).
29. Bajardi, P. et al. Human mobility networks, travel restrictions, and the global spread of 2009 h1n1 pandemic. PloS one **6**, e16591 (2011).
30. Lima, A., De Domenico, M., Pejovic, V. & Musolesi, M. Disease containment strategies based on mobility and information dissemination. Scientific Reports **5**, 10650 (2015).
31. De Domenico, M. et al. Mathematical formulation of multilayer networks. Physical Review X **3**, 041022 (2013).
32. Kivelä, M. et al. Multilayer networks. Journal of complex networks **2**, 203–271 (2014).
33. De Domenico, M., Granell, C., Porter, M. A. & Arenas, A. The physics of spreading processes in multilayer networks. Nature Physics **12**, 901 (2016).
34. Organization, W. H. WHO Immunization, Vaccines and Biologicals, Immunization surveillance, assessment and monitoring. Tech. Rep. (2016).
35. Organization, W. H. Immunization, Vaccines and Biologicals, Immunization surveillance, assessment and monitoring 2016. URL http://www.who.int/immunization/monitoring_surveillance/data/en/. Accessed: 2017-08-01.
36. Salah, A. A. et al. Data for refugees: The d4r challenge on mobility of syrian refugees in turkey. arXiv preprint arXiv:1807.00523 (2018).
37. Anderson, R. M. & May, R. M. Infectious diseases of humans: dynamics and control (Oxford university press, 1992).
38. Grais, R. F. et al. Estimating transmission intensity for a measles epidemic in niamey, niger: lessons for intervention. Transactions of the Royal Society of Tropical Medicine and Hygiene **100**, 867–873 (2006).
39. Lessler, J., Moss, W., Lowther, S. & Cummings, D. Maintaining high rates of measles immunization in africa. Epidemiology & Infection **139**, 1039–1049 (2011).
40. d'Albis, H., Boubtane, E. & Coulibaly, D. Macroeconomic evidence suggests that asylum seekers are not a "burden" for western european countries. Science Advances **4**, eaaq0883 (2018).
41. Fine, P. E. Herd immunity: history, theory, practice. Epidemiologic reviews **15**, 265–302 (1993).
42. Ergun, C. & Gül, H. Urban regeneration and social segregation: The case of istanbul. Toplum ve Demokrasi Dergisi **5** (2014).
43. Matamalas, J. T., De Domenico, M. & Arenas, A. Assessing reliable human mobility patterns from higher order memory in mobile communications. Journal of The Royal Society Interface **13**, 20160203 (2016).
44. Rosvall, M. & Bergstrom, C. T. An information-theoretic framework for resolving community structure in complex networks. Proceedings of the National Academy of Sciences **104**, 7327–7331 (2007).
45. Rosvall, M. & Bergstrom, C. T. Maps of random walks on complex networks reveal community structure. Proceedings of the National Academy of Sciences **105**, 1118–1123 (2008).
46. Diekmann, O., Heesterbeek, J. & Roberts, M. The construction of next-generation matrices for compartmental epidemic models. Journal of the Royal Society Interface rsif20090386 (2009).
47. Diekmann, O., Heesterbeek, J. A. P. & Metz, J. A. On the definition and the computation of the basic reproduction ratio r 0 in models for infectious diseases in heterogeneous populations. Journal of mathematical biology **28**, 365–382 (1990).
48. Melegaro, A. et al. Social Contact Structures and Time Use Patterns in the Manicaland Province of Zimbabwe. PLoS One **12**, e0170459 (2017).
49. Uzicanin, A. & Zimmerman, L. Field effectiveness of live attenuated measles-containing vaccines: a review of published literature. The Journal of infectious diseases **204**, S133–S149 (2011).
50. Republic of Turkey, M. O. I. Distribution by age and gender of registered Syrian refugees. URL http://www.goc.gov.tr/icerik6/temporary-protection_915_1024_4748_icerik. Accessed: 2017-08-01.

Data Analytics without Borders: Multi-Layered Insights for Syrian Refugee Crisis

Özgün Ozan Kılıç[1], Mehmet Ali Akyol[1], Oğuz Işık[1], Banu Günel Kılıç[1], Arsev Umur Aydınoğlu[1], Elif Surer[1], Hafize Şebnem Düzgün[2], Sibel Kalaycıoğlu[1], Tuğba Taşkaya Temizel[1]

[1]Middle East Technical University, 06800, Çankaya, Ankara
[2]Colorado School of Mines, Brown Hall 268, CO 80401, USA
ttemizel@metu.edu.tr

Abstract. This study aims to shed light on various aspects of refugees' lives in Turkey using mobile call data records of Türk Telekom, which is enriched with numerous local data sets. To achieve this, we made use of several techniques in addition to a novel methodology we developed for this particular domain. Our results showed that refugees are highly mobile as a survival strategy, a significant number of whom work as seasonal workers. Most prefer to live in relatively cheap neighborhoods, close to city transport links and fellow refugees. The ones living in low-status neighborhoods appear to be introvert, living in a closed neighborhood. However, the middle and upper class refugees appear to be the opposite. Fatih, İstanbul was found as an important hub for refugees. Finally, the officially registered refugee numbers do not reflect the real refugee population in Turkey. Due to their high mobility, refugees lag behind in keeping up-to-date information about their residential address, resulting in a significant discrepancy between the official numbers and the real numbers. We believe that policy makers can benefit from the proposed methods in this study to develop real-time solutions for the well-being of refugees.

Keywords: health · education · unemployment · social integration · safety and security.

1 Introduction

The civil war in Syria has caused one of the biggest forcibly displaced population in human history [1]. Turkey has become the main destination for Syrian refugees, with around 5 million. Although there are camps built for the refugees with better living conditions than urban areas [2], more than 90% of the Syrian population in Turkey live outside formal camps within host communities, the reasons for which are given as overcrowded camps, illegally entered individuals not being allowed to register to a camp, family ties, and financial independence [3]. The status of Syrian refugees under temporary protection is shaped within the framework of "Temporary Protection Regulation." It is stated that under this regulation, the problems such as education, health, work permit, and access to social services and assistance are solved. They are also treated the same as the Turkish citizens in accessing such rights given that they are

registered with Ministry of Interior Directorate General of Migration Management (DGMM).

At the beginning, Syrian refugees were mainly located in the Southeast Anatolian region bordering Syria. However, over time and with the influx of arriving refugees, they expanded to other regions as well, covering the Mediterranean, Aegean, Central Anatolia, and Marmara regions—İstanbul having the highest number of refugees. So far, Turkey has provided exceptional support to Syrian refugees [4]; however, the problems are mounting. They can be summarized as income, unemployment, education, health, housing, and social tensions [3, 5, 6].

The Syrian refugees have impacted the economy [7]. For instance, "around 1.8 million of the Syrian refugees are of working age" [8]. Although some entrepreneurial efforts have been observed and some of the refugees are skilled, most of the refugees are employed as unqualified labor. Through supplying inexpensive informal labor for labor-intensive sectors, refugees displaced native workers, both formal and informal unemployment rates have increased, and furthermore, it was observed that in these sectors the prices had fallen around 4% [9]. At the beginning of the refugee crisis, the Turkish economy had been experiencing a transition from being a low-wage country to one based on skilled labor. With the arrivals of Syrian refugees, this transition has started to decelerate as they have offered cheap low-skilled labor to the job market, which took advantage of their vulnerabilities. In particular, several refugees found jobs as seasonal workers or in small industrial areas ("sanayi siteleri") [10].

At the time of refugees' arrival, Turkish cities were undergoing a profound transformation in terms of housing. Illegal settlements ("gecekondu") have started to be demolished and TOKI (Governmental Mass Housing Administration) aimed to regulate the housing market [11] which provided partial solutions to the problem. Refugees arrived when Turkey was still struggling with its urbanization problems. Therefore, refugees found safer places in the fragmented cities easily. A good evidence of it is that they settled into still-untransformed poor and environmentally low-quality districts, which are very close to city centers. These districts provided life-saving pockets for refugees where they can survive easily.

Big data have recently started to be used to address big social and environmental challenges in developing countries [12]. With ethical and privacy issues on mind, humanitarian use of private data such as mobile call data records has a great potential in improving society [13]. Data for Refugees, which is a good example of "big data for good", is a research challenge aiming to provide better living conditions to Syrian refugees in Turkey [14]. In this research challenge, we investigated the mobility patterns of refugees from different points of views in order to provide multi-layered insights for the Syrian refugee crisis. We found out that refugees are highly mobile in Turkey as a survival strategy. We also carried out detailed analyses based on three different districts and cities, which we chose according to our previous results. When enriched with our secondary data sets, we saw that those living in low-status neighborhoods are introvert unlike refugees living in middle and high-status neighborhoods.

As a result, the repeatability and the reproducibility of the proposed methods can be beneficial to policy makers to obtain real-time insights about seasonal workers and to arrange services such as mobile health and education services on time. We also put some of our interactive visualization tools online on http://d4r.metu.edu.tr. The project

website also provides detailed information about each step we have carried out in our analyses with several examples.

2 Technical Description

D4R Challenge provided three main data sets (DS1, DS2, and DS3) along with some helper files, which were collected from 992457 customers of Türk Telekom, of which 184,949 are tagged as "refugees", and 807508 as Turkish citizens in 2017. As the paper provides a description of all features, sampling strategies and anonymization methods in depth [15], we will skip those in this paper and describe the other data sets (hereafter called "secondary data sets") we used.

Primary data sets

- Data Set 1 (DS1) comprises annual antenna traffic between each site.
- Data Set 2 (DS2) includes cell-tower identifiers of randomly chosen active users' hourly based two-week call detail records. A user's either incoming or outgoing call traffic is provided but not both.
- Data Set 3 (DS3) consists of randomly chosen refugee and non-refugee annual call traffic but with reduced spatial resolution (district level).
- BTS Locations (BL) comprises cell tower locations in latitude and longitude.
- District Mapping (DM) maps the district IDs used in DS3 to district names.
- City Mapping (CM) maps the city IDs used in DS3 to city names.

Secondary data sets

- Neighborhood-level, district-level, and city-level geospatial data sets indicating the administrative borders in Turkey (GSD) obtained from various official sources and government agencies and used for any information that we needed to filter by a geographic region.
- Neighborhood-level population data and various statistics from 2013 to 2017 for various cities (PD) obtained from Turkish Statistical Institute (TurkStat).
- Coordinates of houses and workplaces for rent and for sale in various neighborhoods in İstanbul (FRE), scraped from Hürriyet Emlak [16].
- Rental fees and other relevant information such as the area of use and building age for various neighborhoods in İstanbul districts (RF), scraped from Hürriyet Emlak [16].
- The results of 2017 Address-based Population Registry System where the education levels of residents are given at neighborhood level (APRS2017).

Some of the terms defined in this study are borrowed from Ahas et al. [17]:

Mobile Operator User (MOU): A subscriber of the mobile operator, which is present in the DS2 and DS3.

Home-Time Anchor Point (HAP): An everyday anchor point, at which the probable home location of a person is identified based on the model.

Work-Time Anchor Point (WAP): An everyday anchor point, at which the probable work-time location of a person is identified based on the model. As the demographics of MOU are not known, it is not possible to determine whether that person is indeed

working, studying or unemployed. Therefore, we called it work-time anchor point to refer to the most probable working time of day of that specific MOU.

Hereafter, we used the abbreviations in the parentheses provided for the data set descriptions in the forthcoming sections.

We mainly used R to manipulate, analyze, query, and visualize data. We also utilized GIS applications such as ArcMap and QGIS to match and enrich the data sets with our secondary data sets, some of which were collected through web scraping and/or Google Places API through Python/R. We benefited from MySQL and NoSQL approaches such as Hive that work inside distributed frameworks that focus on big data, such as Hadoop, to store and manipulate data particularly for organizing and querying data. The network analyses were made in Pajek.

3 Pre-processing

Data quality issues with respect to base transceiver stations (BTS) were handled. Each BTS was assigned to a neighborhood-level, district-level, and city-level administrative units in Turkey by the coordinates, using GSD. The ones that do not fall inside the administrative borders of Turkey or the ones that do not have coordinates in the first place were discarded. This way, a total number of 98854 BTS were matched with city-district pairs. In addition, it was noticed that the city and district columns in BL are not entirely correct. As it can be seen in Fig. 1, there are several BTS coordinates, labeled as located in İstanbul, were not within the administrative district. Those were corrected using GSD.

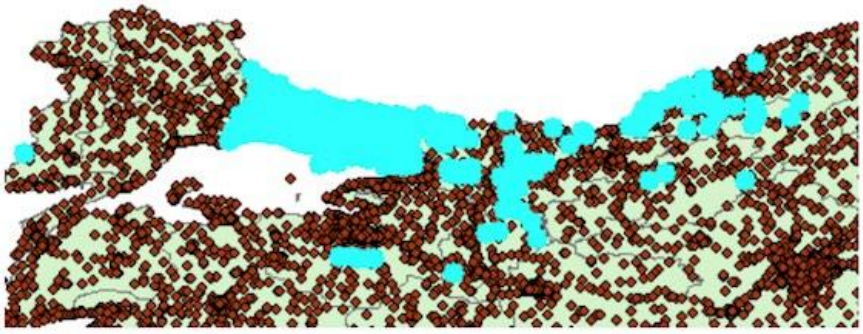


Fig. 1. BTS coordinates provided in BL marked as located in İstanbul but not within the administrative district of İstanbul are highlighted on the map.

DS1 included BTS data, which have zero number of calls or no entries in DS1 in different days or times of days. Fig. 2 shows one of the BTS' call numbers, some of which are not present in DS2. Apart from BTSs that have no data at all (more than half of the whole BTS population), a BTS has 82 missing days at best and 364 missing days at worst in total while the median number of the missing days is 92. Considering that some of the most missing ones belong to urban and dense urban areas (as also noted by the data set itself), it suggests that at least some of the missing days might be related to data quality issues rather than the lack of mobile traffic. This led us to work with

average call numbers per BTS but not with cumulative sums, where the days with no data are excluded from calculations.

Fig. 2. The number of calls per day in July, 2017 in Çaykara, Trabzon measured in BTS number 5066930. The red line shows the total outgoing refugee call and SMS traffic and the blue line shows the total outgoing non-refugee call and SMS traffic. There are no instances in the data set for a number of days.

4 Methodology

The mobility of refugees is analyzed in five folds. In the first approach, we aim to understand how comparable the registered Syrian population in each city of Turkey is with that of call records in city level obtained from BTS call statistics. With the second approach, we analyze city networks connected by refugee mobility. With the third approach, we investigate the monthly refugee movement per district to be able to understand the influx of refugees over time. In particular, we address which districts are increasingly attracting refugees and whether there are any specific districts attracting refugees at a specific time of the year. In the fourth approach, we identify the most probable work and home-time anchor points from two-week data on selected rural and urban locations using mobility statistics and clustering algorithms. This output enabled us to identify the popular places for work and home of refugees. Here, we developed a new method to identify possible work and home time locations. Finally, we explore three different locations to understand the mobility patterns of refugees in depth using several statistics.

4.1 Background: Approaches for Determining Meaningful Places

Meaningful places or meaningful locations are defined as regularly visited places, which have a meaning for a person [18]. Mobile positioning data have been increasingly used for determining meaningful locations. In particular, finding work-time and home-time anchor points have been extensively studied in the literature [17, 19]. The common assumption in finding these places relies on a frequentist approach. First, a specific time interval is defined for work and home time periods. Second, the number of call days and the total number of calls are considered within these time intervals to identify significant anchor points. For example, if mobile calls are made from a specific BTS repeatedly during the home time, that BTS is marked as a potential home-time anchor

point. In urban areas, several BTS can be located at the same site. Such BTS locations can be clustered using Hartigan's algorithm or k-medoid, and site information can be utilized when calculating mobile call metrics. Networks are also constructed including these frequent nodes (a node corresponds to a cluster comprising one or more BTS locations) to correctly identify home and work locations [19]. Finally, several metrics such as regularity, entropy, or radius of gyration (RoG) can be computed to understand the behavior of citizens [20].

Mobile traffic signatures, defined as the typical activity pattern of the mobile demand at one specific geographic zone, have been recently used to investigate the relationship between urban fabrics such as touristic and leisure places, and mobile network usage [21]. Different metrics were proposed based on voice and text traffic volume over weekends and weekdays some of which take into consideration the seasonality.

4.2 Comparison of Registered Syrian Refugee Population with the Number of Calls

Directorate General of Migration Management of Turkey published the registered Syrian population in each city of Turkey for 2017 [22]. According to the statistics, the highest numbers of refugees are located in İstanbul, Şanlıurfa, Hatay, and Gaziantep with 479555, 420532, 384024, and 329670, respectively. In this part of the study, we aggregated BTS call statistics at city level for refugees as follows: First, we calculated the total number of refugee calls and such for each day per BTS. Then, we calculated the monthly average from the daily totals per BTS in each month while discarding the days with no calls due to the aforementioned data quality problem. Then, we computed the sum of each BTS over a year to obtain the cumulative annual use of each BTS, which is denoted as BR_i. Each BTS is geospatially associated with a district and city using GSD. Finally, for each city, we summed up all BTS aggregated call data, which is denoted as C_i. The vertical percentage of refugee calls for each city is found using $VPCR_i = C_i / (\sum_{i=1}^{81} Ci)$, where 81 is the total number of cities in Turkey. Then, we made use of the registered Syrian population PR_i and the total population of each city in Turkey P to calculate the vertical percentage of the registered refugee population with $VPR_i = PR_i / P_i$. Finally, we divided $VPCR_i$ by VPR_i to obtain the magnitude of the difference between the two statistics, which is denoted as MD. The map plotted in Fig. 3 shows the results. Although the registered Syrian refugees are officially reported low in Antalya, the total number of refugees' calls in Antalya is quite higher than expected. The second highest city is Kilis, which shares a border with Syria. There also seems to be an undocumented influx to East and West Black Sea regions, although it is not as significant as the previously mentioned ones, which will be examined closely in the following sections.

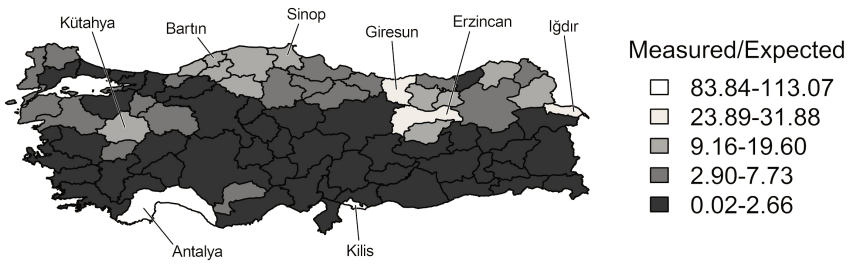


Fig. 3. The ratio of the total refugee calls per city in 2017 to the official residency records, higher numbers (represented with lighter colors) indicating a higher refugee influx compared to official figures (higher than expected)

The results show that there is a discrepancy between the number of registered users and the call data in certain cities. They indicate that although some refugees change their residency addresses, they do not inform the state agencies about this change on time. An expert working in Refugees and Migrants Solidarity Association we consulted also stated that they were suspecting of numerous undocumented refugees living in Antalya but they are not sure about how big the discrepancy is.

4.3 Analysis of City Networks Connected by Refugee Mobility

In order to understand how Syrian refugees use space in Turkey, incoming and outgoing calls data in DS3 were used to form 1-mode and 2-mode networks of refugees and the cities in which they have made phone calls. Initial basic statistics showed that refugees are highly mobile. Out of 37300 refugee MOUs, 53.8% visited only one, 19.9% two, 9.3% three, 5.2% four and 11.8% five or more cities.

In network analysis, initially multiple lines were summed in the 2-mode network of refugees and the cities to obtain the total number of calls a refugee has made in a city. In order to determine the cities where refugees reside in, rather than visit briefly during a trip, the ties that indicate less than 100 phone calls in a city in one year were removed. Then, all line values were replaced with the value 1, since the focus of attention is the presence of a refugee in a city, not how many phone calls they have made. This 2-mode network was then used to obtain the 1-mode network of cities. In this network, a pair of cities is connected by refugees who have been to both cities. In the 1-mode network, the value of the line that connects any two cities is the total number of refugees who have been in those cities (aka "network traffic" in Figure 4).

In order to quantify the most important cities for refugee mobility, weighted degree centralities were calculated. Top 10 cities that receive the highest refugee traffic in descending order were found as İstanbul, Gaziantep, Ankara, İzmir, Mersin, Adana, Hatay, Antalya, Şanlıurfa, and Kocaeli. Then, the lines with values lower than 50 were removed to simplify the graph and the largest connected component was determined, which yielded 33 cities. Fig. 4 shows the resulting ties between these cities. The sizes of the vertices show the registered Syrian refugee population [22] in the corresponding cities with a minimum of 155 (Giresun), a maximum of 479,555 (İstanbul), and a median of 8120. The widths of the lines indicate the total number of refugees linking

the cities, with a minimum of 50 (between İstanbul and Kayseri), a maximum of 318 (between İstanbul and Kocaeli), and a median of 79.5. Some cities, such as Edirne, Tekirdağ, Çanakkale, Aydın, Muğla, Antalya, Kastamonu, Samsun, Trabzon, Ordu, Giresun, Tokat, and Sivas have a small number of registered refugees. Yet, the analysis shows that refugees visit and stay in these cities. The majority of these travels are to/from İstanbul only. The strongest single link a city has (compared to its registered refugee population) belongs to Antalya, tied to İstanbul.

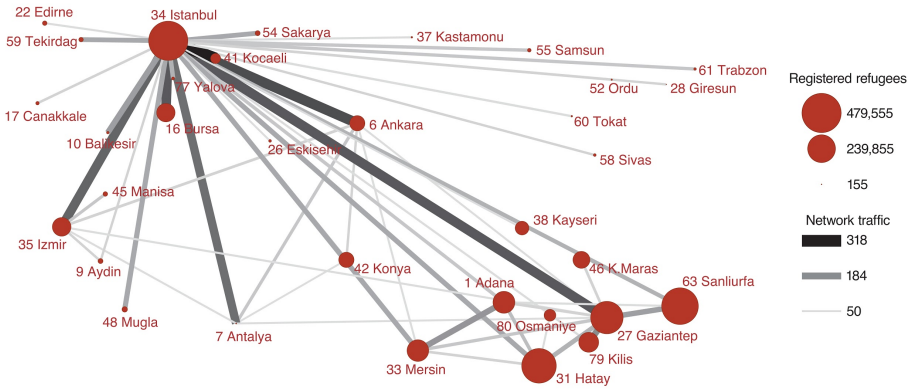


Fig. 4. Network map indicating the refugee links between cities (lines) along with their registered Syrian refugee populations (vertices)

4.4 Investigating monthly refugee movement per district

In this section, monthly refugee mobility is studied. Some refugees are known to be highly mobile and work as seasonal agricultural workers in Turkey. To be able to identify districts receiving the highest refugee influx, we calculated the sum of BR_i for each district, which is denoted as BRD_i. As we will study the density and distribution of the outgoing refugee calls in DS1, we need to consider districts that have a significant number of call data. This is due to the fact that the mobility pattern of subscribers without heavy phone usage can be properly characterized, is indeed questionable [23]. The study reported that 17% and 38% of subscribers in their CDR data set had two or fewer records and fewer than seven CDR, respectively. In addition, some BTS records are not present in DS1. As it is not possible to understand the reason of omissions (whether it is a measurement problem or indicating zero call entries), we performed filtering, which resulted in discarding two-thirds of district data. As a rough cut off point, the districts with total outgoing refugee calls in 2017 lower than 3650 were filtered out to improve calculation time and reliability of the analysis results. This is due to the fact that to make an inference with a few calls can produce biased estimations. This threshold is determined using the mobile phone usage distribution statistics provided in [23].

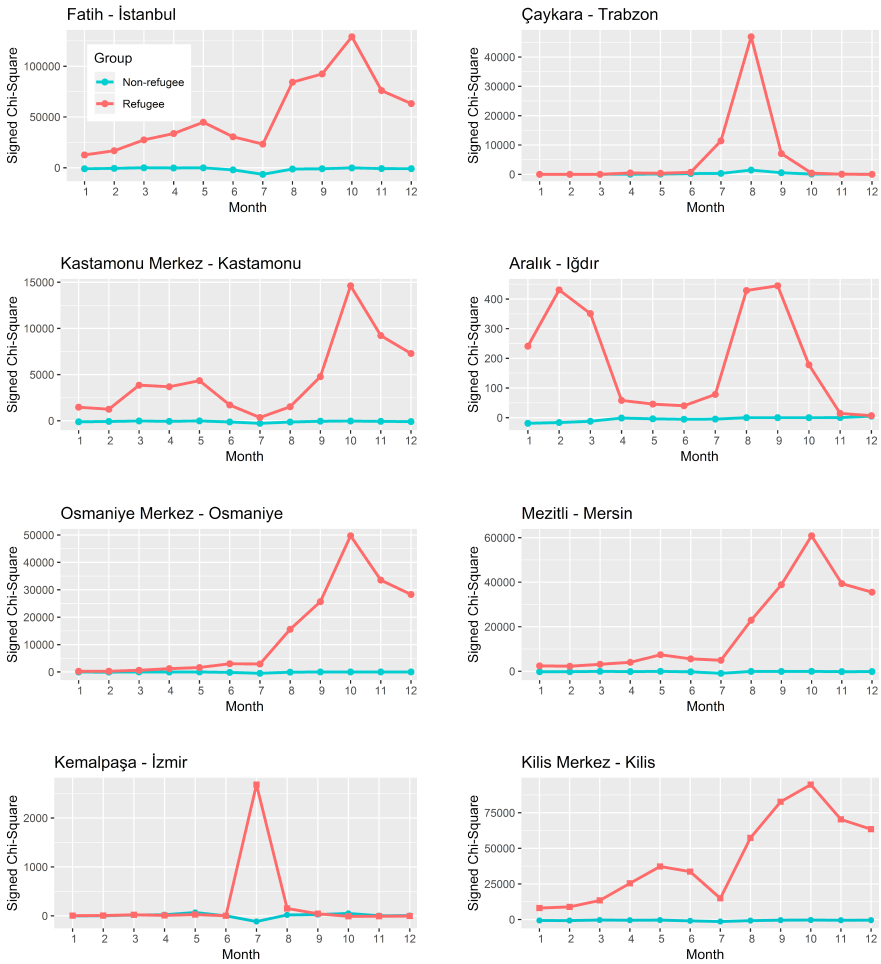


Fig. 5. Some of the districts with significantly higher outgoing refugee calls and/or monthly changes that can be explained

By looking at the monthly SCS values and their first order differences, we highlighted some of the districts that feature significant changes as seen in Fig. 5. Later, we consulted experts to understand the reason for the influx of patterns. Fatih-İstanbul was reported as an attraction center for refugees, the exact reasons for which are not clear. However, it is advocated that the low prices in accommodation, Fatih's easy access to other districts and the existing local community's religious background might have attracted them. This district is studied in depth in the forthcoming section. The agriculture expert we consulted informed us of many changes in the districts coincide with the harvest times. To be specific, the harvests of Çaykara-Trabzon in July and August are hazelnut and tea, Kastamonu in September and October are sugar beet, garlic, and paddy. In addition to that, the harvests of Iğdır in August/September are

sugar beet and cotton while the harvest of Osmaniye in October is peanut. Wealthy refugees visit summer locations as frequently as non-refugees. We see an interesting peak in Kemalpaşa-İzmir in July and a smaller peak in August. In addition, for February/March in Iğdır, there is another one we observed. However, at the time of writing this report, we still could not reach an authority who could explain the exact reason for this refugee influx in July for Kemalpaşa and in February/March for Iğdır. On the other hand, for the smaller peak in Kemalpaşa in August, we were told by the expert that the reason of the increase in August could be attributed to the religious festival for Alevis, which is called Hamzababa Anma Törenleri (Hamzababa Commemoration) taking place between 28 and 29th of August each year. For the peak in July for Kemalpaşa, the experts from numerous local municipality services we consulted, speculated that it is a high season for harvesting and refugees might have stayed in this district for temporary accommodation. The agriculture expert also suggests the refugee influx in Mezitli is based on tourism and the fact that the peak happens in October does not contradict with it since Mersin is known for its scorching heat and humidity which shifts the tourism season towards fall. Kilis has a border with Syria, constantly receives refugees. We have provided the results of some districts in Fig. 3. Many more others can be viewed on our web site using an interactive tool.

4.5 A New Method to Understand Possible Work-Time and Home-Time Locations

In this section, we investigate refugees' meaningful places, specifically work-time and home-time mobility patterns, which are quantified using well-known measures used in the literature [24]. In particular, different aspects of irregularity are studied. We extended a well-known method in the literature, which was described in Section 4.1. The main contribution of our proposed method is rather than using a pre-determined time to identify work and home time locations, we identified important anchor points automatically using an algorithm. Many people such as white-collar workers have very structured daily lives. Their work hours are usually between 9:00 A.M. to 6:00 P.M. However, a garbage collector visits several districts to collect recycled materials such as glass, paper or plastics, while a seasonal agricultural worker might have changed his location within a week and moved to another district. Some people might have been working at different times of days, which we came across in the data sets considerably. Therefore, it may not be convenient to use a pre-determined time to find WAP and HAP locations. In addition, a static threshold for number of calls or number of call days is generally used to filter out the MOUs as their HAP and WAP information cannot be obtained due to limited number of call data. Instead of it, we used a clustering algorithm in this study. Later, we made use of DS2 to test our methodology.

The pseudocode of the algorithm is provided in Table 1. The first step of the algorithm involves finding the idle hours of a given user. As there might be idle hours during the working-time, we made use of a median filter, where each hour is replaced with the median value of a moving filter. Finding the first quartile enables us to identify the start and end points of a continuous time block. For the home-time period, we assume that MOU can be highly likely to be at home just before the idle time period so we chose a time interval starting three hours before the starting point of the idle time and ending at the end point of the idle time. Likewise, we considered four hours after

the idle time as a starting time of a work-time period as we assumed that these four hours will highly likely to include home and commute locations. It is not meant the real work-time will start at that point rather we aim to discard noisy data. These numbers are obtained empirically based on DS2. Finally, the closer BTS locations are clustered. As mentioned by Isaacman et al. [25], in urban areas the BTS can be as dense as 200 meters and in suburban areas, they can be 3-5 kilometers apart. Therefore, we have tried different values for the radius, such as 200 meters, 500 meters, and 1 kilometer and the radius of 1 km gave more meaningful results.

Table 1. Algorithm

Algorithm to detect work-time and home-time patterns
input: caller_id *CID* in DS2 output: *WAP* and *HAP* 1. Retrieve *CDR* of *CID* 2. Calculate hourly call counts $hourly\_calls_i$ from CDR, where i>=0 and i<24 3. If there is no CDR in any $hourly\_calls_i$, assign it to zero. 4. Apply median filter with a window size three on $hourly\_calls_i$, to obtain filtered data, denoted as $filtered\_hourly\_calls_i$. 5. Sort the $filtered\_hourly\_calls_i$ in descending order. 6. Obtain the first quartile of $filtered\_hourly\_calls_i$, which is denoted as fq. 7. Find the minimum and maximum hours in fq, denoted respectively as fq_{min} and fq_{max} respectively 8. Determine the start and end points of WAP and HAP as follows: *Let wtp_start* be the work-time period start time, where $wtp\_start = fq_{max}+4$ Let *wtp_end* be the work-time period end time, where $wtp\_end=wtp\_start + 6$ Let *h_start* be the home-time period start time, where $h\_start = fq_{min} -3$ Let *h_end* be the home-time period end time, where $h\_end = fq_{max}$ 9. Find the most used BTS in terms of calls days between *wtp_end* and *wtp_start,* which is denoted as $BTSW_{max}$ 10. Apply Hartigan's leader algorithm [26] to all BTS between *wtp_end* and *wtp_start*. 11. Select the cluster in which the most used BTS resides, which is denoted as *WAP*. 12. Find all BTS on the same calls days with *WAP* between *h_end* and *h_start,* denoted as $BTS_{HomeAll}$ 13. Find the most used BTS in $BTS_{HomeAll}$, denoted as $BTSH_{max}$ 14. Apply Hartigan's leader algorithm to all BTS between *h_end* and *h_start* 15. Select the cluster in which the most used BTS resides, which is denoted as *HAP*

Then, we have extracted 31 number features from two-week CDR data of each MOU. Some of these features are borrowed from Soto et al. [27]. The complete list of features can be found on our website. The significant features selected by the process described below and our consequent statistical analyses are as follows:

- N_call_days/N_call_days_home_time/N_call_days_work_time: The number of unique days, unique days recorded within HAP, unique days recorded within WAP respectively
- N_calls: The number of calls
- N_city: The total number of cities the MOU is seen

- N_district: The total number of districts the MOU is seen
- RoG [27]
- Entropy_bts: Entropy of MOU based on the BTS footprint
- Entropy_district: Entropy based on the district footprint
- Entropy_district_home_time: Entropy calculated within HAP (based on district information)
- Entropy_district_work_time: Entropy calculated within WAP (based on district information)
- Entropy_cluster_home_time: Entropy calculated within HAP (based on the clusters formed after Hartigan's leader algorithm)
- Entropy_cluster_work_time: Entropy calculated within WAP (based on the clusters formed after Hartigan's leader algorithm)
- Ref_nonref_ratio: Ratio of calls made/received to/from refugees to non-refugees
- Total_movement: Total distance of travel made by MOU, calculated by summing the distance between each following BTS used by the MOU
- Work_home_dist: Haversine distance between the WAP and HAP

After deriving these features, we first applied sparse K-means (SK-means) algorithm using R's RSKC package [28] with different parameters to obtain the most relevant, uncorrelated and non-redundant features. As a result, we ended up with the following features (sorted in descending order according to their significance values): n_call_days, n_calls, n_call_days_work_time, n_call_days_home_time, entropy_cluster_home_time, entropy_cluster_work_time, entropy_district, entropy_district_home_time, rog_work_time, rog_home_time, n_city, and work_home_dist to cluster the MOUs. Finally, these features are used as inputs to Self Organizing Map (SOM) [29], which is a type of artificial neural network to visualize high dimensional data in a low-dimensional grid. SOM produced different mobility clusters and we selected the instances falling into the cluster nodes, where there is sufficient number of call days and number of calls, which are important to detect HAP and WAP more accurately. Finally, we have identified regions mostly preferred for work and living purposes by refugees. Note that two week data is quite limited to estimate WAP and HAP of an individual accurately. Hence, if there are not sufficient numbers of data points for a MOU, the algorithm cannot successfully identify these locations. The detailed steps of the algorithm are provided in the Appendix section including how to interpret clusters with examples as well.

Since we did not have a ground truth in hand for evaluation, we scraped seemingly the biggest online real estate website in Turkey, Hürriyet Emlak (2018), and obtained a total of 701 house/workplace ads for rent/for sale that correspond to the neighborhoods in Fatih, İstanbul (FRE). The sampling was neighborhood-wise stratified (to ensure each neighborhood is represented) and ad-wise systematic (to ensure that the price range and distribution are reflected) as every n^{th} ad was recorded from the list of ads ordered by price while while $n = \lfloor ad\_size/sample\_size \rfloor$. The sampling of the ads for a specific neighborhood was only applied if the neighborhood had more than 10 ads for houses or workplaces, separately. After collecting 338 houses and 363 workplaces, we sampled the house and workplace locations that fall inside Fatih, İstanbul (by also using GSD) from the cluster of refugees that we are most certain. Again, we used stratified sampling to sample exactly 15 (house & workplaces in total) coordinates from each neighborhood and we obtained 338 houses and 517

workplace locations (855 points in total). The difference in numbers of points (701 vs. 855) is due to different sample sizes, Hürriyet Emlak not having all the neighborhoods available for filtering, and lack of ads for certain neighborhoods. Then, we compared the locations of the ads with refugees' predicted work/home locations. As can be seen in Fig. 6, some locations such as İskenderpaşa neighborhood (central region) are both residential and working places. The shops are mainly located on the sideways of the main street. However, like in Tahtakale neighborhood (the northeastern region of Fatih), some are mostly business related locations while some are mostly residential places as in Şehremini or Silivrikapı neighborhoods (southwestern region). Our results coincide with the ad types.

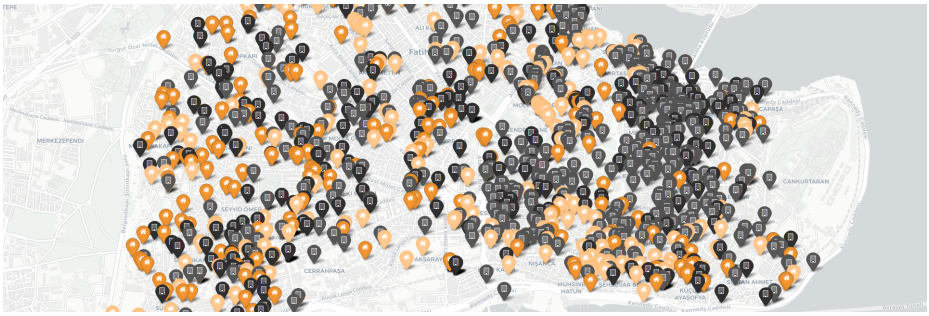


Fig. 6. HAP and WAP locations obtained from refugees' call data (indicated with light orange and light gray) along with the houses and workplaces obtained from Hürriyet Emlak (indicated with orange and dark gray)

5 Investigating Regions in Depth: Case Studies for Çaykara/Trabzon, Fatih/İstanbul, and Mezitli/Mersin

In this section, based on our previous findings, we explore three districts and cities in detail. We look into the characteristics of these locations and attempt to understand refugees' living conditions.

5.1 Findings for Çaykara/Trabzon

Our initial analysis showed that Çaykara, Trabzon has a significant peak in August. An agriculture expert we consulted suggested that it is most probably related to the harvest of tea (which also gives its name to the district) and hazelnut. We also confirmed it with "Development Workshop" reports [10]. The report states that the Black Sea region in Turkey receives seasonal Syrian agricultural workers between August and September, first starting to work in gardens in coastal regions then move innermost.

We analyzed DS2 and DS3 to find the refugees that were present around the specific BTS (#5066930) in August and understand where they come from. We found that people present in the area were also highly present in İstanbul and Mersin. The coordinates where the phone calls occurred also come right on top of the usual intercity bus route between İstanbul and the Black Sea region. The seasonal workers here might be later switching to Mezitli, Mersin region (which can explain its peaks in October)

once the harvest is done. When we looked at DS3 to find the refugees that were in Çaykara in August, we found 306 callers (people with outgoing call data in DS3) and 275 callees (people with incoming call data in DS3). Interestingly, the most common district that these callers and callees present in 2017 was Fatih, İstanbul (68.4% and 60%, respectively). However, based on their most frequent call locations, it seems like these refugees do not live in Fatih, İstanbul. They either live in Ortahisar, Trabzon (a coastal district) or various districts in İstanbul, such as Kağıthane (a district known until recently for its low living standards but undergoing a rapid transformation in some parts).

5.2 Findings for Fatih/İstanbul

İstanbul appeared to be the top refugee location according to Directorate General of Migration Management of Turkey (DGMM 2017) and in DS1. The first thing to be noted as for the distribution of refugee calls in İstanbul where almost 40% of registered refugees live, is the high concentration of such calls on the European side (see Fig. 7). Given the fact that most of the city's population, central business activities, job opportunities, and urban amenities are on the European side, this concentration of refugee calls is in harmony with the basic characteristics of the human geography of the city.

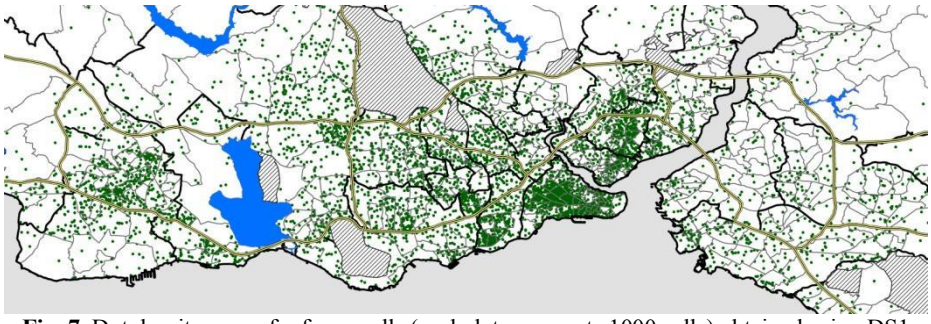


Fig. 7. Dot density map of refugee calls (each dot represents 1000 calls) obtained using DS1

On the European side, refugees seem to prefer such districts as (from west to east) Esenyurt, parts of Güngören, Bağcılar, Zeytinburnu, Fatih, Beyoğlu, Şişli, and Beşiktaş districts. Of these districts, Şişli and Beşiktaş are known to be the parts of the modern city center. The neighborhoods preferred by the refugees are close to the main transport routes and provide easy access to some important urban amenities such as city center, business districts, and entertainment facilities. These locations are also mostly low-income areas of the city. In an attempt to prove this claim, we use APRS2017 data set. The map in Fig. 8 shows the distribution of university graduates by neighborhoods as a percentage of total population above 6 years of age together with the dot density map of refugee calls on the European part of the city. We know from previous studies that education level is almost perfectly correlated with income level, meaning that the higher the education level in a given neighborhood, the higher the income of its

residents [30]. A comparison of the maps in Fig. 6 makes it clear that refugees tend to conglomerate in areas where the residents are low-income groups.

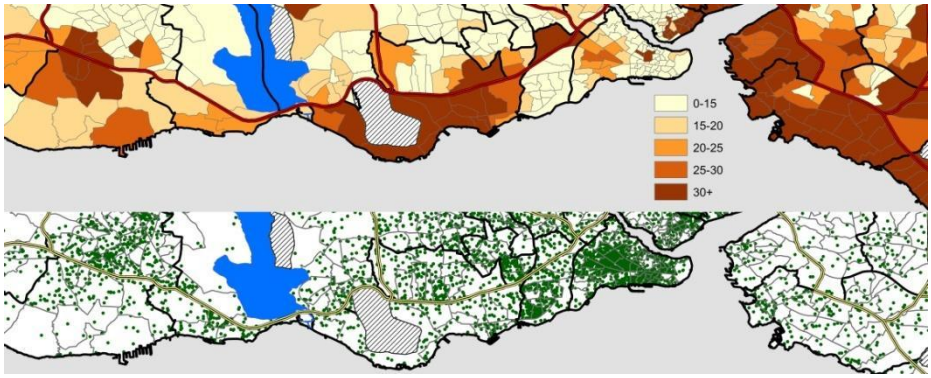


Fig. 8. The percentage of university graduates compared to the dot density map

Table 2. Monthly rents for each square meter of housing units in selected neighborhoods

District	Neighborhood	Monthly rent per m^2 (TL)	District	Neighborhood	Monthly rent per m^2 (TL)
Selected neighborhoods with high concentration of refugees					
Bağcılar	100. Yıl	11.9	Güngören	Abdurrahman Nafiz Gürman	17.7
Beyoğlu	Hüseyinağa	19.2	Güngören	Mehmet Nesih Özmen	17.2
Beyoğlu	Katip Çelebi	32.1	Güngören	Tozkoparan	15.4
Beyoğlu	Kocatepe	24.1	Güngören	Güven	14.8
Beyoğlu	Kuloğlu	30.5	Zeytinburnu	Beş Telsiz	27.6
Beyoğlu	Şehit Muhtar	28.4	Zeytinburnu	Çırpıcı	27.3
Esenyurt	Barbaros Hayrettin	15.0	Zeytinburnu	Merkezefendi	26.7
Esenyurt	Cumhuriyet	14.4	Zeytinburnu	Seyitnizam	25.4
Esenyurt	Fatih	8.1	Zeytinburnu	Sümer	16.2
Esenyurt	Mehterçeşme	9.8	Zeytinburnu	Nuri Paşa	18.5
Esenyurt	Mevlana	12.2	Fatih	Mevlanakapı	19.0
Esenyurt	Talatpaşa	9.4	Fatih	Şehremini	20.1
Esenyurt	Yenikent	10.9	Başakşehir	Başakşehir	17.7
Selected neighborhoods with low concentration of refugees					
Beşiktaş	Akatlar	77.9	Beşiktaş	Levent	92.1
Beşiktaş	Arnavutköy	126.2	Beşiktaş	Ulus	87.8
Beşiktaş	Gayrettepe	58.9	Şişli	Şevketpaşa	66.7

The zones in İstanbul where refugees are concentrated offer relatively cheap housing opportunities for its residents, where the housing rents are generally low. In order to reach a better understanding of refugee concentration areas in İstanbul, we compiled the monthly rents for 3186 housing units advertised on 6 September 2018 in Hürriyet Emlak (RF), for selected neighborhoods of İstanbul (see Table 2). The results show that the areas preferred by refugees for housing purposes are low-housing cost areas, with rents in some cases 4 to 5 times lower than those in middle and upper class areas characterized by the absence of refugees. We can also add from the existing literature on Syrian refugees in Turkey [31] that even this is a partial and misleading picture of the housing conditions of refugees both in İstanbul and Turkey as a whole, as the figures published in real-estate agents' websites are only for those housing units on the "formal" housing market and refugees do have to live in sub-standard housing units not preferred by the local population.

To complement this study, we also investigated Fatih in street level using the proposed algorithm to understand in which parts of Fatih they usually live and work. The results show that HAP of non-refugees is clustered in the more expensive areas of Fatih district. However, HAP of refugees is clustered relatively in poorer areas.

As for the distribution of refugees in İstanbul, the following conclusions can be drawn:

- Refugees tend to live close to fellow refugees (evidenced by the exceptionally high concentrations of refugee calls in some parts of İstanbul);
- Refugees prefer those areas of the city where the poor and low-income groups live (evidenced by the comparatively low-education levels of refugee concentration zones in İstanbul);
- Refugees tend to concentrate in low property value areas of the city (evidenced by housing rent values of refugee concentration zones in İstanbul);
- Refugees live in inner city areas, close proximity to main transport lines for easy access to job opportunities and urban amenities (evidenced by the general distribution of refugees in İstanbul).

Fatih, known for its rather traditional and religious residents since the Early Republican Period, is the most refugee-saturated district not only in İstanbul but in the whole country. It is believed that many Syrian refugees started their new life in here, thanks to relatively cheap rental prices of basement floor flats, with hopes to hold on to İstanbul where there are more jobs and the city life is dominant.

5.3 Findings for Mezitli/Mersin

Mezitli, a district of Mersin, appeared to have received a significant refugee influx according to our chi-square analysis results. Therefore, in this section, we aim to study Mersin in depth. Using our algorithm, we obtained HAP and WAP locations for each MOU who has been seen at least two times according to their calls records in Mersin. We have selected the MOUs after we have clustered them using SOM. The results of the SOM clustering can be seen in Fig. 9.

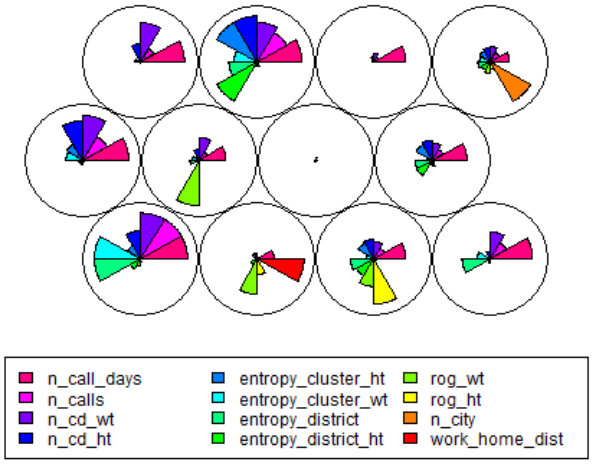


Fig. 9. Codes plot of the SOM clustering of refugees in Mersin

While selecting the nodes, we have considered the ones having sufficient number of call days and number of calls, as we have discussed before. As a result, we have selected MOUs contained by the 1st, 5th, and 10th nodes (enumeration starts from the bottom left and ends at the top right).

Using the GSD data set, we mapped each HAP and WAP locations to neighborhoods in Mersin. Furthermore, using the PD data set, we labeled each quarter as low-status (LSN), middle-status (MSN), and high-status neighborhoods (HSN) in seaside districts of Mersin, which are Mezitli, Akdeniz, Yenişehir, and Toroslar, according to their education level and youth population ratio. Among these districts, Mezitli is one of the high-status neighborhoods in Mersin. The locations of the HAP and WAP of the selected MOUs can be seen in Fig. 10.

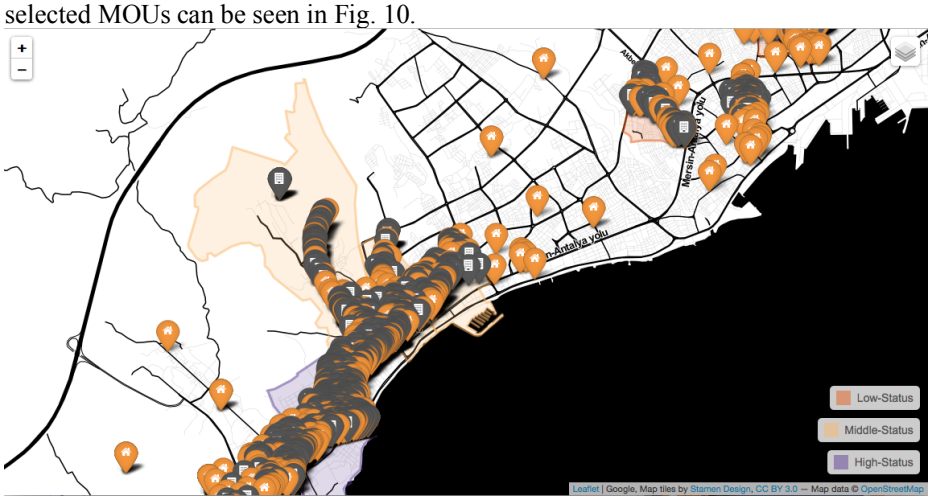


Fig. 10. HAP (marked with orange color) and WAP (marked with gray color) locations of selected MOUs in LSN (marked with light orange), MSN (marked with peach color), and HSN (marked with purple)

Considering all 31 features, we investigated whether there is a statistically significant difference between the neighborhoods using a pairwise Mann-Whitney U test, as it is a non-parametric test that can be used for non-gaussian distributions. We have presented the number of MOUs in each neighborhood (N), followed by the mean and median values of each feature per neighborhood in Table 3.

The results presented in Table 3 only shows the features that have been found as significant in comparisons, which are presented in Table 4, with p-value < 0.05. Additionally, when Bonferroni adjustment is applied, significance level drops to 1.66e-2 since we have made 3 different comparisons for each feature. Table 4 shows the comparison results, which were found significant. We make 2-sided comparisons with Mann-Whitney U Test and p-values implies the 2-tailed exact significance.

As a result, the test indicated that the distance between HAP and WAP locations is the highest for HSN indicating that MOUs living in HSN probably work in farther neighborhoods. Additionally, the entropy calculated based on different district visits (Entropy_district) for LSN is the lowest, which is also confirmed by the entropy based on the BTS visits (Entropy_bts), and for MSN it is the highest among others. Also, MOUs living in HSN appear to be more regular and MOUs living in the LSN appear to

be more irregular than the others in their home-time periods based on their entropy based on BTS visits (Entropy_bts_home_time). One feature that can shed light on the introversion of the MOUs is the refugee to non-refugee call ratio (Ref_nonref_ratio) and in our tests, we have seen that MOUs in LSN have the lowest ratio whereas MOUs in HSN have the highest ratio, showing that refugees in LSN have the least interaction with the non-refugees and refugees in HSN have the most interaction with the non-refugees. Lastly, we have seen that the total movement (Total_movement) and RoG is the lowest for LSN indicating that the total movement and RoG increase as the level of education and youth population level increases.

Table 3. Descriptive statistics of LSN, MSN, and HSN for each feature presenting the number of MOUs (n), mean (μ), standard deviation (σ), min, max, and percentiles of the distributions.

		n	μ	σ	Min	Max	Percentiles		
							25th	50th (Median)	75th
Work_home_dist	LSN	176	0.57	1.6	0	14.20	0.09	0.15	0.35
	MSN	342	1.46	4.0	0	34.77	0.07	0.18	0.96
	HSN	619	1.53	3.4	0	26.68	0.14	0.36	1.40
Entropy_district	LSN	176	10.1	9.2	0	56.9	3.5	7.9	13.9
	MSN	342	17.7	15.0	0	71.3	5.9	13.0	26.2
	HSN	619	13.8	13.3	0	100.4	4.5	9.5	19.0
Entropy_bts_home	LSN	176	5.90	3.0	0.4	19.2	3.9	5.4	7.2
	MSN	342	5.20	3.2	0	22.4	3.0	4.7	6.5
	HSN	619	4.60	3.0	0	23.5	2.8	4.1	5.6
Ref_nonref_ratio	LSN	176	0.14	0.2	0	1.54	0.00	0.04	0.16
	MSN	342	0.33	0.8	0	6.75	0.02	0.09	0.28
	HSN	619	0.33	0.4	0	2.80	0.05	0.18	0.42
Rog	LSN	176	5.70	5.1	0.84	42.74	2.72	4.32	6.77
	MSN	342	6.30	4.6	0.41	36.52	3.54	5.04	7.80
	HSN	619	6.40	3.8	0.80	47.62	4.17	5.39	7.95
Total_movement	LSN	176	278	256	26	2161	118	211	373
	MSN	342	355	310	20	1604	134	277	462
	HSN	619	344	266	15	1970	172	271	421
Entropy_bts	LSN	176	35.3	21.8	4.7	166.4	20.2	30.6	44.7
	MSN	342	41.7	24.7	6.0	135.9	23.4	36.6	55.3
	HSN	619	43.0	25.3	3.3	158.5	25.2	37.5	53.8

Table 4. Mann-Whitney U Test results for the comparisons of LSN, MSN, and HSN for each feature (Comparisons read as follows; for all comparisons like "LSN vs HSN", the one with the lower median value written on the left hand size, that is, the median of LSN is less than the median of HSN for this example).

		Mann-Whitney U Test	
		U	**p**
Work_home_dist	LSN vs HSN	71765	6.27e-11
	MSN vs HSN	126860	3.42e-07
Entropy_district	LSN vs MSN	21408	7.28e-08
	LSN vs HSN	46842	2.27e-03
	HSN vs MSN	38784	7.28e-08
Entropy_bts_home	MSN vs LSN	35412	9.85e-04
	LSN vs HSN	71254	4.30e-10
	MSN vs HSN	92768	1.50e-03
Ref_nonref_ratio	LSN vs MSN	24361	3.56e-04
	LSN vs HSN	32263	4.40e-16
	MSN vs HSN	128090	6.58e-08
Rog	LSN vs MSN	25286	2.88e-03
	LSN vs HSN	41825	2.55e-06
Total_movement	LSN vs MSN	25756	7.20e-03
	LSN vs HSN	44189	1.31e-04
Entropy_bts	LSN vs MSN	25317	3.06e-03
	LSN vs HSN	43739	6.54e-05

6 Summary of Findings (SF)

SF1. A place in the city—Place as a means of survival: Finding an adequate place to live in is a key to success in the urban jungle. A place where they can interact with their peers, have easy access to urban facilities such as work and leisure is vital for their survival in the city. Using the established networks of solidarity among the refugee community, newcomers maximize their access to flows of information, which may from time to time play a life-saving role. Almost as a rule, the newcomers to a city tend to concentrate in particular parts of urban areas [32].

We have discovered in the second layer of our findings that Syrian refugees in Turkish cities tend to live in areas where [a] they can have maximum interaction with

Syrian community, which is crucial in access to information flows; [b] the rents are lower; [c] they can have easy access to urban facilities, namely in inner city areas.

SF2. Networks—Sine qua non for survival: All of the above—i.e. joining the complex web of relations characterizing seasonal agricultural work and finding the best place to live in a city—could not have been achieved without networks. Some of the hotspots we have discovered seem to function like hubs in a network. For example, Fatih district definitely plays a high-level hub role, as we have detected a non-negligible number of refugees (whose speculated homes are outside Fatih) coming to Fatih and also heading to Çaykara where they work for tea harvest.

SF3. Better off Syrians—Internal divisions: We know from other studies that an important portion of Syrians came to Turkey without a chance to turn their savings in their homeland into cash [31]. There are also some cases, though not many, where some Syrians arrive in Turkey with some accumulated wealth. This we believe creates a division within the Syrian refugee population, with better off Syrians living in comparatively higher status neighborhoods and having different mobility patterns, compared to the rest of the refugees. Mersin appeared to be a very interesting city where refugees from different socioeconomic levels are living. Our analyses showed that middle and high-status refugees use a very large space, travel long distances, have regularity in their mobility (regular work home patterns), whereas low-status refugees appear to be trapped in a small neighborhood meaning that traveling not very distant districts, having high irregularity in their mobility.

SF4. Seasonal work—Geographical mobility as a survival strategy: One of the most important findings of our study is the one that shows the movement of Syrian refugees among various districts of Turkey. The evidence to this fact is the unusually high number of calls made by refugees in certain parts of the country in certain months of the year. What seems to be an anomaly of the data set is, in fact, a practice that is very common in Turkish agriculture: the prevalence of seasonal work in agriculture. We know from a recent study that Syrian refugees have replaced seasonal Kurdish workers in the last few years especially in cotton and hazelnut harvests, two most common products utilizing seasonal labor [33]. They have also replaced Georgian migrants in tea harvest which, compared to other products, requires highly skilled labor.

The unusual peaks in the number of outgoing refugee calls in some regions in certain months attest to the fact that a significant portion of Syrian refugees are on the move in search of an adequate job and are in harmony with the harvest seasons of agricultural products. Their enhanced geographical mobility as a survival strategy shows the capability of Syrian refugees to adapt to new conditions. Furthermore, the fact that they have replaced (or are in the process of replacing) Georgian migrants in tea harvest is a testimony to their ability to alter and penetrate the existing networks.

7 Policy Recommendations

In this section, after a thorough analysis of D4R data, available literature reviews and investigation of available social protection mechanisms, we list and group the main vulnerabilities of Syrian refugees. In addition, we propose short-term and long-term solutions and policy recommendations. We investigate the problems followed by

proposed recommendations under the subheadings of 1) Education & Employment 2) Safety & Security 3) Work, 4) Healthcare, and 5) Integration.

Education & Employment: As our findings demonstrate, Syrian refugees use geographical mobility as a survival strategy (SF4) which means spending short amounts of time in certain regions. Policies are required for both children and agricultural laborers. For children, the main problem is the access and participation to education. Harvesting times often coincide with the school period for some districts. Although we are not sure whether they are moving with their families (demographics of MOUs and their call networks are not provided in detail in the data sets), it is highly probable that there may be Syrian refugee children who are unable to benefit from education services. Furthermore, having one parent moving all year long is not a healthy environment to raise children. Studies also show that language barriers, insufficient salary, invisible costs of education (i.e. costs for course materials, transportation, and lunch costs), and distance from urban centers may result in education problems. As a remedy, introductory programs can be expanded in order to take language and cultural differences into account. Social assistance policies should specifically focus on the involvement of children in education. In line with our finding (SF4), mobile school programs such as TÜBİTAK 4007's science buses and science fairs that bring science to rural areas, can visit farm areas to help sustain continuity in children's education. As for adults, who are mostly seasonal workers in agriculture, problems are many: long working hours, low wages, health and hygiene problems, makeshift tents they have to live in, the lack of basic amenities, and so on. From a wider perspective, seasonal labor cannot be a long-term solution in the sense that it guarantees only (or even falls short of) a minimum level of subsistence. In addition, seasonal work is a strategy that may lead to the failure of coming generations, since it almost requires the movement of young members of the family who cannot attend their school during seasonal work, a fact that may lead to persistent long-term problems. It was also reported that in numerous stages of agriculture, different skills are required but refugees do not have the necessary background skills [10], which can be solved with specialized training programs. Generations may stick in seasonal work due to a lock-in network. As networks are critical (SF2), conditions of the agricultural work network should be improved and new occupational networks should be introduced. Agricultural occupations can be made more structured with better pay. For the latter, new job opportunities aiming to integrate Syrians into the labor force should be created. This may be achieved by offering vocational and entrepreneurship trainings.

Safety & Security: The results of our analyses showed that Syrian refugees tend to live close to other fellow refugees (SF2) and they live in poor housing conditions in underdeveloped areas of the cities (SF1). This creates ghetto-like communities and becomes a handicap on overall integration in Turkey. To address these vulnerabilities, "urban transformation" policies should be redesigned taking Syrian refugees into account and integration to the society should be aimed. Housing is a problem for seasonal workers. Recently, a project called METİP (Project for Improvement of the Working and Living Conditions of Seasonal Migratory Agricultural Workers) [34] has been put into effect by the Ministry of Labor and Social Security with the goal of improvement of the current living, shelter, transportation, education, health, security, social relations and social security status of the seasonal migratory agricultural workers

who migrate to other cities with their families to be employed as agricultural workers. It is important that the continuity of the METİP project will be ensured in such districts.

Healthcare: As mentioned above, most of the Syrians are employed in seasonal work (SF4). Seasonal workers and women are the most vulnerable group in terms of accessing healthcare due to remoteness and language barriers. Seasonal changes in welfare, difficulties in access to childcare services, obligation to work during pregnancy and concern about nutrition quality are the main problems of women. Hygiene is a problem for everyone. Mobile health units that travel farms can be established as a remedy to provide the necessary healthcare services.

Integration: All of our findings are in fact intertwined and part of the bigger challenge that is going to take a long time for everyone involved to deal with – **integration**. For instance, just with being employed in seasonal work will not be enough for Syrians to integrate with the Turkish society. It also brings out the competition on limited resources with Turkish seasonal workers. Another example may be that if Syrian children do not receive the necessary education, their destiny will be low-paid jobs and no integration. Ghettos would be another obstacle for integration. Moreover, there are even fractions within the Syrian communities itself (SF3). Thus, even with the limited data provided to researchers in D4R, it is obvious that addressing the integration is the real challenge ahead of Turkey in the long run. In our opinion, in addition to NGOs and academics, government bodies and international organizations should work together to establish a big coalition to come up with an integration strategy and obtain a buy-in from the public to tackle the integration challenge, because the last challenge will require not only huge amounts of resources and wisdom but also empathy, compassion, and tolerance.

8 Acknowledgement

We would like to thank Türk Telekomünikasyon A.Ş. for the one-year anonymized mobile communication data they provided within the D4R Challenge.

References

1. United Nations High Commissioner for Refugees (UNHCR): Global trends 2017: Forced displacement in 2017. http://www.unhcr.org/5943e8a34.pdf (2017). Accessed 31 August 2018
2. ORSAM (Ortadoğu Stratejik Araştırmalar Merkezi): Suriye'ye komşu ülkelerde Suriyeli mültecilerin durumu: Bulgular, sonuçlar ve öneriler. http://www.madde14.org/images/e/e5/OrsamSuriyeKomsu2014.pdf (2014). Accessed 31 August 2018
3. İçduygu, A.: Syrian refugees in Turkey: The long road ahead. Washington DC: Migration Policy Institute (2015)
4. Kaya, A.: Istanbul as a space of cultural affinity for Syrian refugees: "Istanbul is safe despite everything!" Southeastern Europe 41(3):333–58 (2017)
5. World Bank: Turkey's response to the Syrian refugee crisis and the road ahead. Washington DC: World Bank (2015)

6. Stock, I., Aslan, M., Paul, J., Volmer, V., Faist, T.: Beyond humanitarianism: Addressing the urban, self-settled refugees in Turkey. Bielefeld: COMCAD (2016)
7. Cagaptay, S., Menekse, B.: The impact of Syria's refugees on Southern Turkey. Policy Focus 130. Washington, DC: The Washington Institute For Near East Policy. http://www.washingtoninstitute.org/uploads/Documents/pubs/PolicyFocus130_Cagaptay_Revised3s.pdf (2014). Accessed 31 August 2018
8. Asik, G.A.: Turkey badly needs a long-term plan for Syrian refugees. Harvard Business Review. https://hbr.org/2017/04/turkey-badly-needs-a-long-term-plan-for-syrian-refugees (2017). Accessed 14 September 2018
9. Balkan, B., Tumen, S.: Immigration and prices: Quasi-experimental evidence from Syrian refugees in Turkey. Journal of Population Economics 29:3:657–686 (2016)
10. Dedeoğlu, S.: Yoksulluk nöbetinden yoksulların rekabetine: Türkiye'de mevsimlik tarımsal üretimde yabancı göçmen işçiler mevcut durum raporu. http://www.ka.org.tr/TumYayinlar (2016). Accessed 12 September 2018
11. Isikkaya, A.D.: Housing policies in Turkey: Evolution of TOKI (Governmental Mass Housing Administration) as an urban design tool. Journal of Civil Engineering and Architecture 10:316–326 (2016)
12. Chandy, R., Hassan, M., Mukherji, P.: Big data for good: Insights from emerging markets. Journal of Product Innovation Management 34(5):703–713 (2017)
13. Alemanno, A.: Big data for good: Unlocking privately-held data to the benefit of the many. European Journal of Risk Regulation 9(2):183–191 (2018)
14. Türk Telekom: Data for refugees. http://d4r.turktelekom.com.tr/ (2018). Accessed: 31 August 2018
15. Salah, A.A., Pentland, A., Lepri, B., Letouzé, E., Vinck, P., de Montjoye, Y.A., Dong, X., Dağdelen, Ö.: Data for Refugees: The D4R Challenge on mobility of Syrian refugees in Turkey. arXiv preprint arXiv:1807.00523 (2018)
16. Hürriyet Emlak: Online real estate ads. https://www.hurriyetemlak.com (2018). Accessed 8 August 2018
17. Ahas, R., Silm, S., Järv, O., Saluveer, E., Tirui M.: Using mobile positioning data to model locations meaningful to users of mobile phones. Journal of Urban Technology 17(1):3–27 (2010)
18. Nurmi, P., Koolwaaij, J.: Identifying meaningful locations. In: Mobile and Ubiquitous Systems: Networking & Services, 2006 Third Annual International Conference, pp. 1–8. IEEE, San Jose (2006)
19. Jiang, S., Ferreira, J., González, M.C.: Activity-based human mobility patterns inferred from mobile phone data: A case study of Singapore. IEEE Transactions on Big Data 3(2):208–219 (2017)
20. Graells-Garrido, E., Peredo, O., García, J.: Sensing urban patterns with antenna mappings: the case of Santiago, Chile. Sensors 16(7):1098 (2016)
21. Furno, A., Fiore, M., Stanica, R., Ziemlicki, C., Smoreda, Z.: A tale of ten cities: Characterizing signatures of mobile traffic in urban areas. IEEE Transactions on Mobile Computing 16(10):2682–2696 (2017)
22. Directorate General of Migration Management of Turkey (DGMM): Hangi ilde ne kadar Suriyeli var? İşte il il liste. http://www.bik.gov.tr/hangi-ilde-ne-kadar-suriyeli-var-iste-il-il-liste/ (2017). Accessed 31 August 2018
23. Zhao, Z., Shaw, S.L., Xu, Y., Lu, F., Chen, J., Yin, L.: Understanding the bias of call detail records in human mobility research. International Journal of Geographical Information Science 30(9):1738–1762 (2016)
24. Yang, P., Zhu, T., Wan, X., Wang, X.: Identifying significant places using multi-day call detail records. In: 2014 IEEE 26th International Conference on Tools with Artificial Intelligence (ICTAI), pp. 360-366. IEEE, Limassol (2014)

25. Isaacman, S., Becker, R., Cáceres, R., Kobourov, S., Martonosi, M., Rowland, J., Varshavsky, A.: Identifying important places in people's lives from cellular network data. In: International Conference on Pervasive Computing, pp. 133–151. Springer, Berlin, Heidelberg (2011)
26. Hartigan, J.A.: Clustering algorithms. New York, NY, USA (1975)
27. Soto, V., Frias-Martinez, V., Virseda, J., Frias-Martinez, E.: Prediction of socioeconomic levels using cell phone records. In: International Conference on User Modeling, Adaptation, and Personalization, pp. 377–388. Springer, Berlin, Heidelberg (2011).
28. Kondo, Y., Salibian-Barrera, M., Zamar, R.: RSKC: An R package for a robust and sparse k-means clustering algorithm. Journal of Statistical Software 72(5):1–26 (2016)
29. Kohonen, T.: The self-organizing map. Neurocomputing 21(1–3):1–6 (1998)
30. Işık, O., Ataç, E.: Yoksulluğa dair: Bildiklerimiz, az bildiklerimiz, bilmediklerimiz. Birikim 269(268):66–86 (2011)
31. Eraydın, G.: Migration, settlement and daily life patterns of Syrian urban refugees through time geography: A case of Önder neighborhood. Unpublished PhD Thesis, Middle East Technical University (2017)
32. National Academies of Sciences, Engineering, and Medicine (NASEM): The integration of immigrants into American society. Washington, DC: The National Academies Press (2015). doi: https://doi.org/10.17226/21746
33. Kalkınma Atölyesi: Mevsimlik gezici tarım işçiliği izleme: Mevcut durum haritası (2012–2013) http://www.ka.org.tr/TumYayinlar (2013). Accessed 12 September 2018
34. Prime Minister's Office: Memorandum Circular n. 2017/6. Official Gazette of Turkey, 30043 (2017)

9 Appendix: Finding HAP and WAP

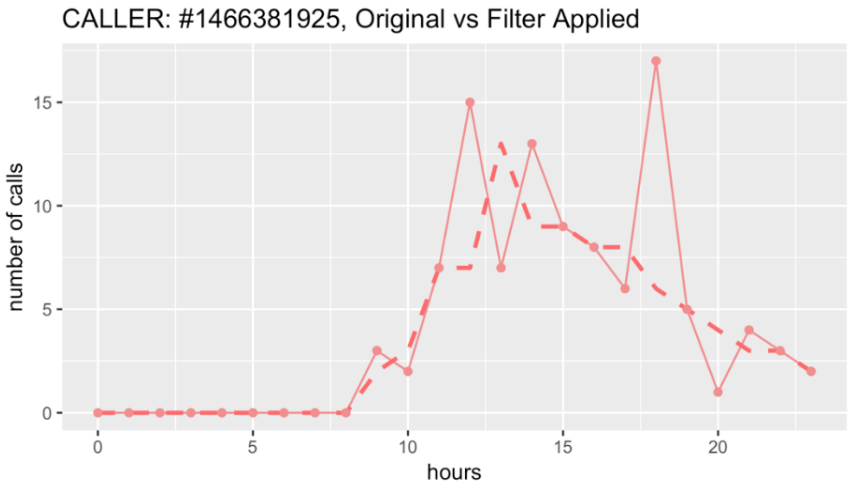


Fig. a1. The number of calls per hour for a MOU, original is shown with straight line and the median filter applied is shown with dashed line.

In order to extract the important places of individuals, we study their fine grained mobility using DS2. More specifically, we try to find WAP and HAP for each MOU. To be able to find those points, first we aggregate the hourly calls counts for each MOU

and we apply a median filter with the bandwidth of three to hourly calls signal to smooth unexpected low and high values of call counts out. As depicted in Fig. a1. after the median filter is applied, the hourly call counts signal is smoothed.

Then, we sort the hours according to the number of calls made during that hour in ascending order. We select the hours in the first quartile and three hours before that as the Home-Time Period (HTP) assuming that this period is spent during the most probable HAP. After that, we find the Work-Time Period (WTP) by first adding four hours, which is allocated for preparation and commute, to the end of HTP and selecting the next six hours, which is assumed to be a safe period for work activities for most of the people, as the WTP. In Fig. a2, HTP and WTP have been marked on the filtered data.

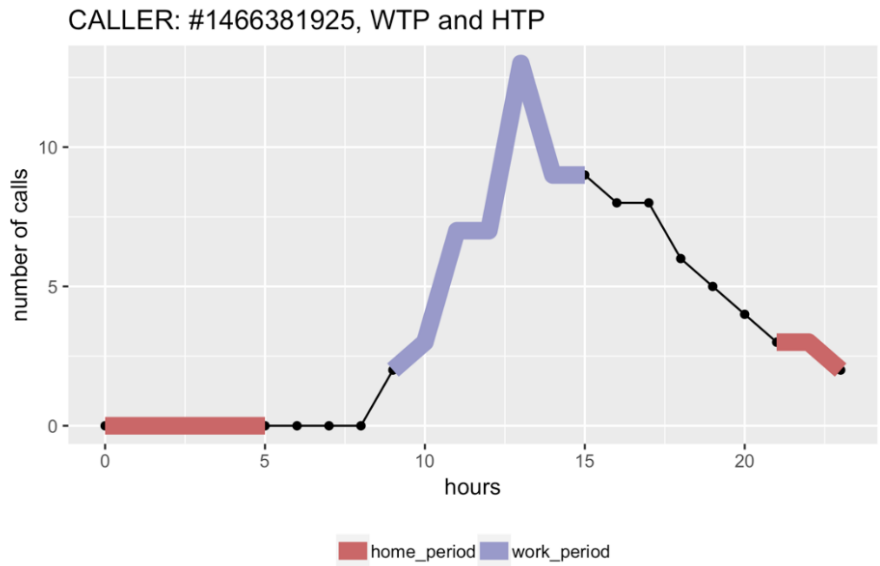


Fig. a2. The graph showing the Work-Time Period (WTP) and Home-Time Period (HTP) of a MOU. HTP included the hours in the first quartile (between 00:00 and 5:00) and three hours before that (between 21:00 and 23:59). WTP is found by first adding four hours to the end of HTP (9:00) and selecting the next six hours (end point is 15:00).

After finding the HTP and WTP, we derive the BTS used in those periods to find HAP and WAP. First, we find the HAP by sorting BTS by the number of unique days they used and we select the one with the highest number of unique day usage. In the next step, we cluster the BTS using Hartigan's leader clustering algorithm. The advantage of the Hartigan's leader algorithm, unlike clustering algorithms like K-means, we do not need to set the number of clusters at the beginning. We only need to set the radius to cluster the BTS based on their proximity. We set the radius as 1 km. After clustering the BTS, we select the cluster, in which the BTS with the most call days resides and then set the centroid of the cluster as HAP. In order to find the WAP, on the list of possible BTS for WAP, we apply the same steps by applying the Hartigan's leader algorithm and then selecting the centroid of the cluster in which the BTS has the most number of call days in weekday usage. Aside from the hour differences between the

HTP and WTP, to specify the WTP we only look at the calls made or received on the weekdays within the designated work time hours. In Fig. a3, we present the possible WAP, which is represented by the green circle, and HAP, which is represented by the red circle, locations for a MOU living around Siteler, Ankara.

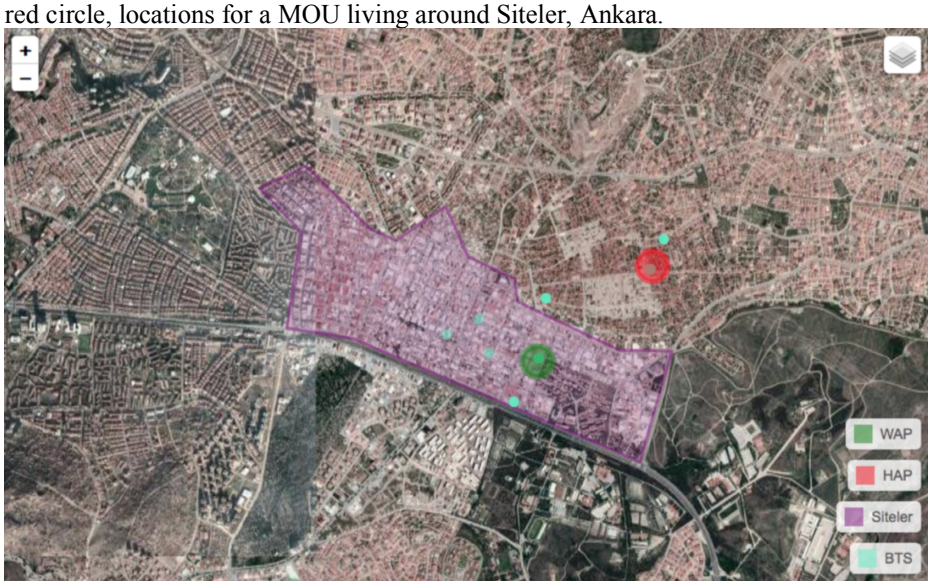


Fig. a3. WAP and HAP for a person living around Siteler, Ankara

10 Appendix: SOM Clustering

In SOM clustering, we trained the neural network until the average distance between the data points and the node centroids converged to a relatively small distance and then, we get the codes plots of the clusters, an example can be seen in Fig. a4. If we are to analyze some of the clusters, for instance, in Node 1 (enumeration starts from the bottom left and ends at the top right), we observe that MOUs in this node have higher entropy while having close to zero radius of gyration and visited very few numbers of different cities. As their HAP and WAP locations are very close, their daily commute is expected to be small. Even though they visited different numbers of BTS and districts, their travel distance is relatively small and they wondered in a very small area by visiting a lot of places, which are very close to each other. On the other hand, for Node 6, we see that MOUs in this node have low entropy and high radius of gyration indicating that these MOUs are making longer commutes, especially in the work-time periods, while visiting a fewer number of different places. Additionally, larger commutes were also confirmed by the larger distance between HAP and WAP locations. Lastly, we observe MOUs in Node 10 having only higher values of radius of gyration in the home-time period while other features like entropy, radius of gyration in work-time period, number of different cities visited, and WAP-HAP distance are small. Hence, we can deduce that these MOUs are making larger commutes in their home-time period but they are not visiting different places in terms of BTS, districts, or cities and their HAP and WAP locations are very close to each other.

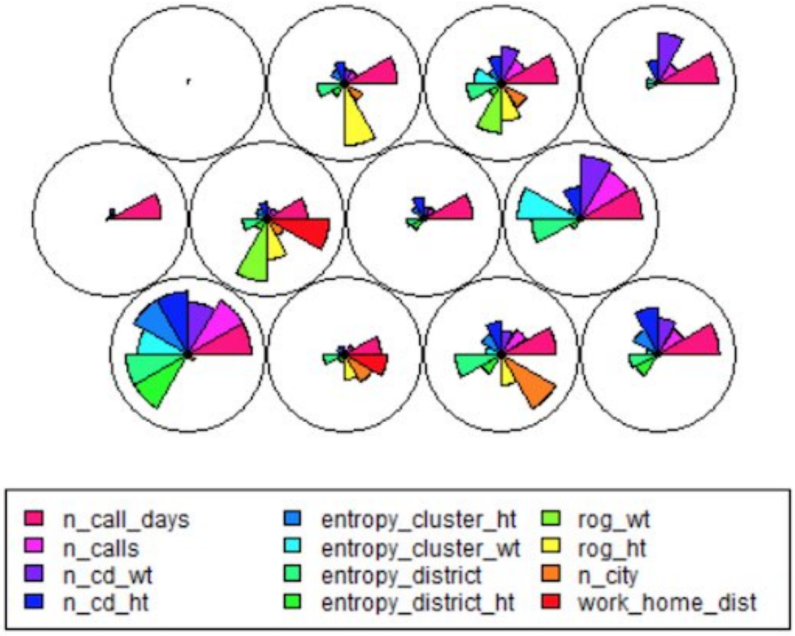


Fig. a4. Codes plot of a SOM clustering

UDMIT: An Urban Deep Map for Integration in Turkey

Sedef Turper Alışık[1[0000-0001-5659-3181]] Damla Bayraktar Aksel[1[0000-0002-4157-812X]] Asım Evren Yantaç[1[0000-0002-3610-4712]] Lemi Baruh[1[0000-0002-2797-242X]] İlker Kayı[1[0000-0002-4115-6613]] Sibel Salman[1[0000-0001-6833-2552]] Ahmet İçduygu[1[0000-0002-8145-5888]] Ivon Bensason[2]

[1] Koç University, 34450 İstanbul, Turkey
[2] Attorney at Law, İstanbul, Turkey
lncs@springer.com

Abstract. This report provides an overview of the data analysis and visualization steered under "An Urban Deep Map for Integration in Turkey" (UDMIT) project, which uses mobile call data records of Syrian refugees under temporary protection provided by Data for Refugees: The D4R Challenge on Mobility of Syrian Refugees in Turkey (Salah et al. 2018). First, in an attempt to examine Syrian refugees' temporal and spatial dimensions of mobility, the study concentrates on their interprovincial migration patterns within Turkey. Based on an analysis on these patterns, the report offers assumptions on the potential motivations for regular and seasonal internal mobility, especially regarding access to services and employment opportunities in the formal and informal labor market. The findings are also complemented by policy recommendations on how the D4R data can be of use to central and local authorities on providing occupational health and safety services and on improving refugees' access to information. Second, the study traces the host and refugee community interaction patterns, interpreting their temporal and spatial distributions. The findings of the analysis on the interaction patterns are expected to guide the ongoing empirical work undertaken by project members in creating an interactive integration governance model. Finally, the current project delivers a web based deep mapping platform that allows generating and reporting visual representation of refugee population densities and mobility across Turkey on a real-time basis. The interface enables examining the spatio-temporal D4R data at three scales (country, province and district level) together with other layers of data, including (a) demographic information at the province and district levels, (b) service providers (nongovernmental organizations, schools and healthcare services), (c) media analytics and (d) public discussion. Within the scope of this limited study, the deep mapping platform has been developed as an early-version prototype to demonstrate the potential of opening the data to the use of experts and public with a multilayered, visual and interactive tool.

Keywords: Interprovincial Migration, Seasonal Labor Migration, Health Services for Seasonal Labor Migrants, Host- Refugee Interaction, Deep Mapping.

1 Introduction

The ongoing civil war in Syria in its seventh year has displaced more than 10 million people, including 5.6 million people seeking safety in third countries. Since the onset of the Syrian conflict, Turkey adopted a generous open-door policy towards those Syrians fleeing conflict. While initially the number of Syrian refugees entering Turkey was relatively small, the inflow of Syrian refugees gained significant momentum after the breakdown of the ceasefire negotiations in mid-2012. According to the data provided by Directorate General of Migration Management (DGMM) as of 2018, Turkey hosts more than 3.5 million registered Syrian refugees under the extended status of temporary protection with access to certain welfare provisions such as health and education [1]. While more than 90 percent of the Syrian refugees live outside camps and engage in daily interaction with the local communities [1], the registration statistics provided by DGMM suggests that more than half of the Syrian population concentrated in five provinces in Turkey, namely, Istanbul, Gaziantep, Şanlıurfa, Hatay and Mersin [1].

Considering that majority of the refugee population is living in an urban setting in close interaction with the local communities, the prolongation of their stay due to the multifaceted conflict in Syria and unpredictability of their return, integration policies have become even more prominent. In addition to the access to health and education, in January 2016, "The Regulation on Work Permit of Refugees under Temporary Protection' took effect, that allows refugees to apply for work permits six months after their registration under the status of temporary protection. However, the number of refugees who received work permits has been only limited to 20,970 people, which is no more than 5 percent of the estimated labor force [2]. As far as the sectors that Syrian refugees work are concerned, Syrians are found to be working in those sectors characterized by informal work force such as seasonal agriculture and construction works [2]. Working in those sectors, especially in seasonal agriculture work [3], indicates not only the problems of exploitation, precarity, chronic poverty, insecurity, health issues experienced by the refugees, but also their interactions with local communities within the context of tensions derived from economic competition over resources and jobs. Thus, drawing from D4R [4] data, this study attempts to offer a framework enabling us to examine the temporal and spatial dimension of seasonal work, which is characterized by informal and irregular migration movements. After underlining the problems, this study will attempt to offer policy recommendations for overcoming the problems in the realms of integration and health.

2 The Internal Migration Patterns of the Syrians under Temporary Protection

As the statistics by the DGMM are compiled based on the premise that refugees continue to reside in the provinces that they are registered, it becomes a vital necessity to analyze refugees' migratory movements within Turkey in order to better estimate the service needs of refugees that would inform the drafting and implementation of well-tuned integration policies in Turkey. Thus, this section analyzes internal migratory

movements of refugees and traces the signs of seasonal work motives underlying migratory movements of refugees within Turkey.

To investigate the internal migration patterns of the Syrian refugees we utilize Dataset 3 that provides coarse grained trajectories of 37,300 randomly selected refugees for the entire observation period of 2017. To identify the migratory movements of refugees, we aggregated the daily user data into monthly data in which each refugee user is assigned to a province for each month. The assignment procedure is done by considering the number of days that each refugee user had been active in each province. Accordingly, those refugees who had call activity in a single province throughout the month are assigned to that province, whereas those refugees who had call activities in more than one province throughout month are assigned to the province where they had been active for highest number days. We further cleaned the dataset to include only those users with call activities in at least four months of the year, and this procedure yielded a total number of 24,233 unique refugee users to be included in the final dataset.

Findings of our analysis reveal that among those 24,233 refugee users included in the analysis, 19,835 refugees (81.9 percent) had call activities logged in a single province for the entire observation period. The remaining 4398 refugee users, corresponding to 18.1 percent of all the refugees in our sample, are on the other hand, found to have stayed in at least two different provinces throughout the year. Among these 4398 mobile refugees, 3852 of them, constituting the 15.9 percent of all the refugees in our sample, found to have migrated only to one province whereas the share of those refugees who stayed in three different provinces during the observation period constituted 2 percent of the whole refugee population.

	N	%
Resided in one province	19835	81.9
Resided in two provinces	3852	15.9
Resided in three provinces	489	2.0
Resided in four provinces	47	0.2
Resided in five provinces	9	0.0
Resided in six provinces	1	0.0
Total	24233	

Table 1. The Number of Provinces Refugees Resided in 2017 (raw numbers, percentages)

According to the population statistics provided by the DGMM, Istanbul (538,000 people), Şanlıurfa (463,000 people), Hatay (457,000 people), Gaziantep (350,000 people) and Mersin (192,000 people) had the highest number of registered Syrian nationals at the end of 2017. The interprovincial mobility information that we have analyzed by using the D4R Data provides some insights on the possible motivations of mobility among the Turkish provinces, and especially among the main host provinces. An inspection of the migratory flows of refugees reveals that the migratory flows between ten provinces, namely, Adana, Ankara, Antalya, Bursa, Gaziantep, Hatay, İçel, İstanbul, İzmir, Konya, Şanlıurfa account for 53.6% of all the interprovincial migration of refugees in Turkey. Istanbul alone receives 14.3 percent of total interprovincial mobility as a destination province, and it is followed by Gaziantep (4.6%), Hatay (3.6%),

Ankara (3.6%), and İzmir (3.5%) as mostly preferred destination provinces. As far as the sending provinces that are concerned, Istanbul is observed also to be the major origin province for refugees moving to other provinces within Turkey. Accordingly, 11.9 percent of total mobility is observed to be from Istanbul to other provinces, and Istanbul is followed by Gaziantep (5.0 %), Hatay (3.6%), Ankara (3.2%), and İzmir (3.0%) as other major departure provinces. Furthermore, an inspection of the most frequently observed pairs of origin and destination provinces reveal that migration from Gaziantep to Istanbul (106 people), Ankara to Istanbul (94 people), Istanbul to Izmir (85 people), and Istanbul to Gaziantep (82 people) are the most frequently observed routes for interprovincially mobile refugees.

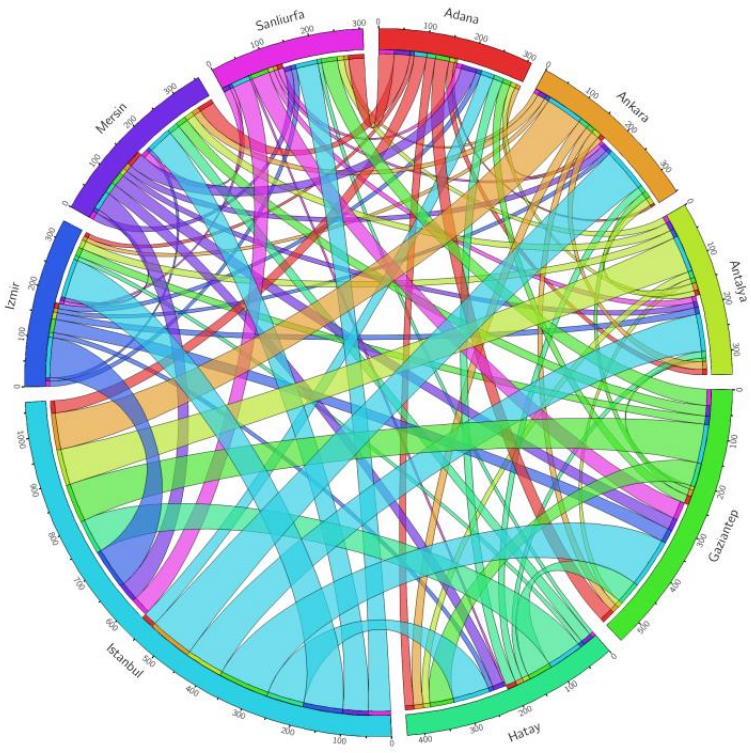


Fig. 1. Internal Migration Flow of Syrian Refugee Users in 2017[1]

[1] Source and Legend: D4R Data, visualized using Circos Table Viewer. This data provides the internal migration flow of Syrian refugee users from the main provinces of origin in Turkey. Origin and destination provinces are represented by segments around the circle. Each province is assigned with a colour (Istanbul: light blue) and outgoing flows have the same color as the origin.

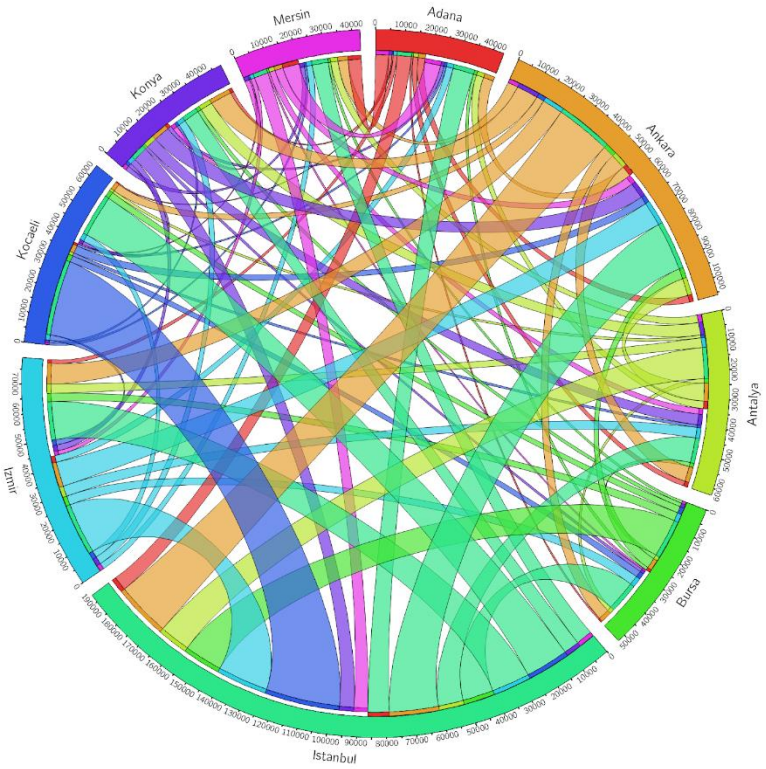


Fig.2. Internal Migration in Turkey (2017)[2]

The higher rates of interprovincial mobility to and from İstanbul meet the a priori expectations, since the province has traditionally been the main hub for both internal and international migratory flows in Turkey (See Fig. 1 and Fig. 2), due to a number of reasons including the employment opportunities and existing migrant networks. Moreover, the province alone is hosting nearly 16% of all the Syrian population in Turkey. In a parallel vein, the mobility to and from the bordering provinces such as Gaziantep and Hatay are also expected, as they are the first provinces of entry to Turkey for many Syrians, have higher registration rates and population sizes. Although the population

[2] Source and Legend: TURKSTAT data on Internal Migration in Turkey, visualized using Circos Table Viewer. This data provides the internal migration flow of Syrian refugee users from the main provinces of origin in Turkey. Origin and destination provinces are represented by segments around the circle. Each province is assigned with a colour (Istanbul: light green) and outgoing flows have the same color as the origin.

size in İzmir and Ankara are comparably lower, the mobility to and from these provinces can be interpreted by other factors. In the case of Ankara, the existence of the main headquarters of refugee management, status determination and resettlement institutions such as the DGMM, United Nations High Commissioner for Refugees (UNHCR) and the International Organization for Migration (IOM) may account for the higher interprovincial mobility of Syrians who are concerned about further mobility from Turkey. Positioned on the western coast of Turkey, İzmir may witness interprovincial mobility as a transit center on the path to Greece or due to opportunities of seasonal work in the tourism sector and citrus farming [5].

2.1 Seasonal Agricultural Migration

The delay in the work permit arrangements for Syrians under temporary protection in Turkey has resulted in their participation to informal labour market and in low-skilled jobs in seasonal agriculture, textile, construction and manufacture [6]. According to İçduygu and Diker [7], the relatively high number of Syrian farmers who have migrated from rural areas to Turkey, in comparison to other host countries in the region, accounts for a crucial reason for their participation to agricultural work. A report by Development Workshop on the migrant workers in the seasonal agriculture in Adana indicates that the inability to find work in their primary occupations is also a crucial reason for Syrians' participation to this sector [8]. Despite the arrangements for work permits in other sectors during 2016, seasonal agricultural work has been positioned in the scope of work permit exemptions. Under current conditions many Syrians working in the seasonal agriculture are faced with the unhealthy and unsafe conditions with little access to education, health and other social services [7].

Previous studies hint that refugees working in seasonal agricultural jobs in Turkey migrate between provinces to work in harvesting of agricultural products such as hazelnut, apple and cotton [9], however to the best of our knowledge, there are no earlier studies documenting interprovincial migratory movement patterns of Syrians under temporary protection. In order to trace internal migratory movements of refugees for seasonal agricultural work purposes, we first investigate in which provinces refugee activities are seasonally increasing in areas categorized as rural areas. To this end, we utilize Dataset 1 which provides activity logs for each base station on a daily basis between January 2017 and December 2017. Having a specific focus on the area type that these base towers are located, we first investigate how the share of the refugee activities within urban, suburban, rural, industrial and seasonal areas of in those provinces where seasonal agricultural works are expected, has varied across months. This analysis informs us about over-time changes in the concentration of refugee activities in urban, suburban, rural, industrial and seasonal parts the provinces in regard, and hence, provides us with suggestive evidence for seasonal agricultural migratory movements of refugees. However, inspection of changes in the refugee activity volume in rural settings cannot allow us to infer whether the increased refugee activity in a rural setting is induced by increasing number of users or increasing number of call activity per user. To overcome this shortcoming, we conduct a complementary analysis by utilizing Dataset 2 that consists of fine grained mobility data for randomly chosen refugee

samples covering 15-day periods. As the refugee samples in this dataset are freshly chosen in every 15-day period, they do not allow us to trace the same users throughout the year, but, the random sampling method utilized in the selection of the samples allows us to trace the over-time changes in the total number of users in each province. In our complementary analysis utilizing Dataset 2, we identify the number of unique refugee users in the rural and non-rural areas of the provinces in regard.

We start our analysis by focusing on the seasonal agricultural migratory flows to the Black Sea region. The Black Sea region is known to host a considerable number of internal and international migrant workers (from Syria, Georgia, Armenia, Russia and Azerbaijan [8]) seasonally migrating to the region to meet the agricultural labor needs in harvest seasons. While Rize attracts seasonal agricultural workers for harvesting tea leaves, Giresun and Ordu are the two provinces where the seasonal agricultural workers mainly work in harvesting hazelnut.

Rize is a coastal province located in the eastern part of the region mostly famous for tea agriculture. According to statistics provided by the Turkish Statistical Institute, approximately 50,000 hectare land is used for cultivating tea plant in Rize, and the tea produced in Rize constitutes approximately 65 percent of all the tea production in Turkey. The harvesting time of the tea leaves usually start at the end of March and the harvesting season can last until November. Figure 3a presents the distribution of refugee call activities across urban, suburban, rural and seasonal parts of Rize province for each month. An inspection of the Figure illustrates that the share of refugee activities in rural and seasonal categories significantly increases between May and November. While the mobile activity of refugees in rural parts of Rize constituted 22.8 percent of all refugee activity logged in the Rize province in May, the share of rural refugee activities within all refugee activity in Rize increased to 25% in June, 40.3% in July, 42.0% in August and gradually decreased to 15.4% in November. In a similar vein, seasonal refugee activity share sharply increased from 1.2% in June to 19.3% in July and to 29.6 in August and gradually decreased to pre-July levels in October.

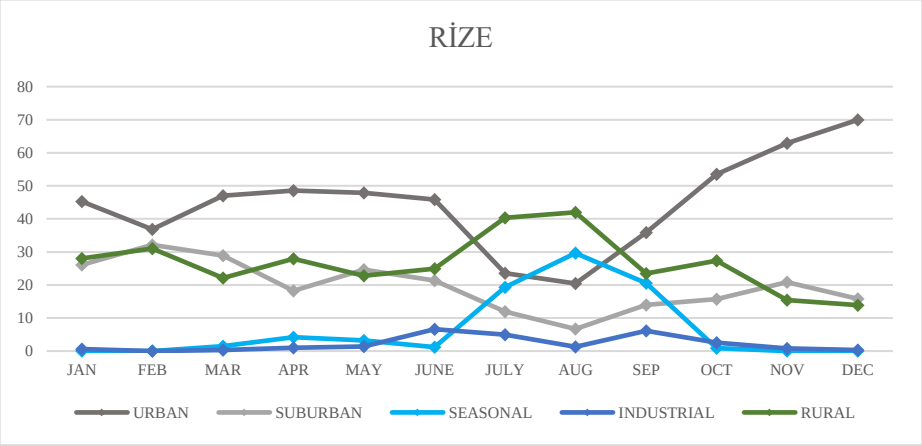


Fig. 3a. Distribution of Refugee Activities Across Area Types in Rize (percentages)

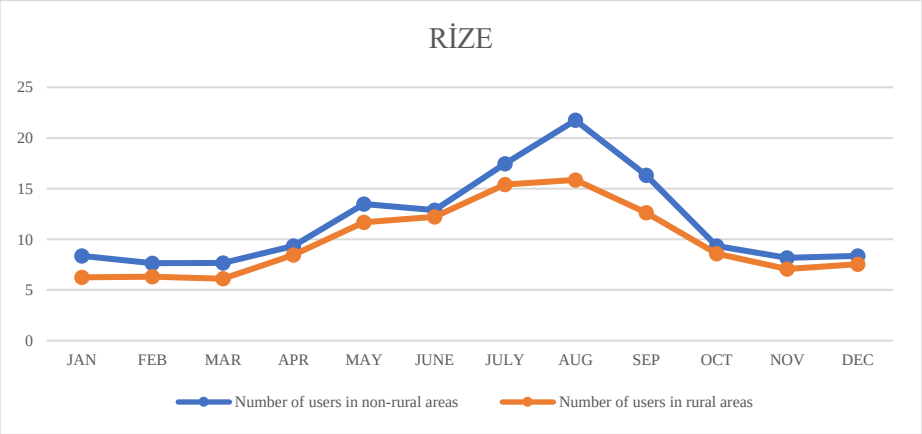


Fig. 3b. The number of Unique Refugee Users in Rural and Non-Rural Areas in Rize (raw numbers)

Figure 3b presents the number of unique refugee users observed in rural and non-rural (including urban, suburban, seasonal and industrial) areas of Rize province based on the analysis of the refugee samples constituting Dataset 2. An inspection of the graph lends support for the existence of seasonal migratory movements towards Rize between March and November. Our findings demonstrate that the number of refugees also increases during the harvest season from March to August, and gradually decreases to pre-harvest season levels in November. All in all, our findings from the complementary analysis suggests that the increasing refugee activity in rural areas can at least be partially accounted by the increasing number of refugee users in rural parts of the Rize province.

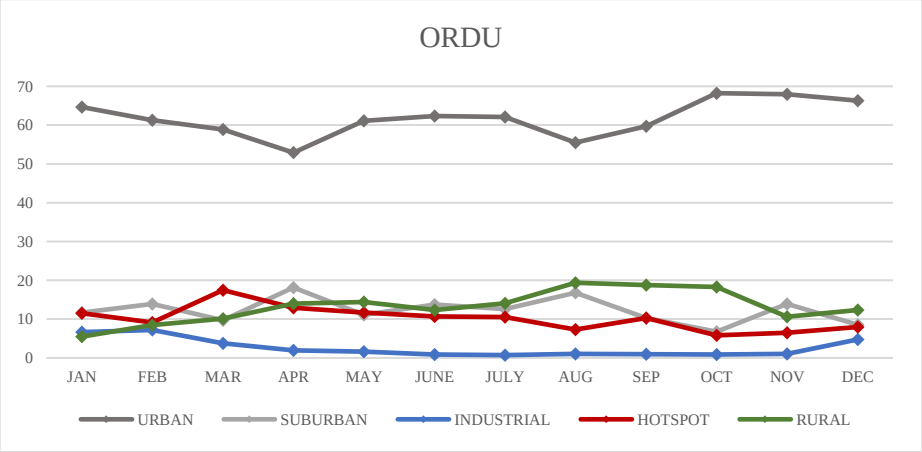


Fig. 4a. Distribution of Refugee Activities across Area Types in Ordu (percentages)

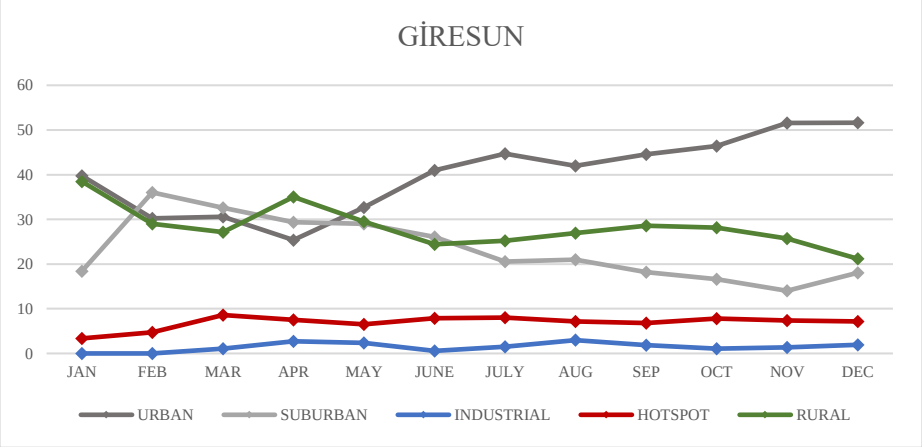


Fig. 4b. Distribution of Refugee Activities across Area Types in Giresun (percentages)

Turkey is also one of the biggest producers of hazelnut globally and Ordu and Giresun provinces constitute the main agricultural sites for hazelnut production in Turkey. According to 2015 Turkish Statistical Institute statistics, the amount of hazelnut produced in these two provinces comprises approximately half of the all the hazelnut production in Turkey. The harvesting of the hazelnut usually starts at the end of July and the harvesting season lasts until September. Figure 4a and Figure 4b presents the distribution of refugee call activities across urban, sub-urban, rural and industrial parts of Ordu and Giresun over the year. An inspection of the Figures reveals that the share of refugee activities in rural areas significantly increases between July and November in Ordu, while the share of refugee activities in rural areas of Giresun are observed to slightly increase during harvesting season of hazelnut but not exceeding the January and April levels.

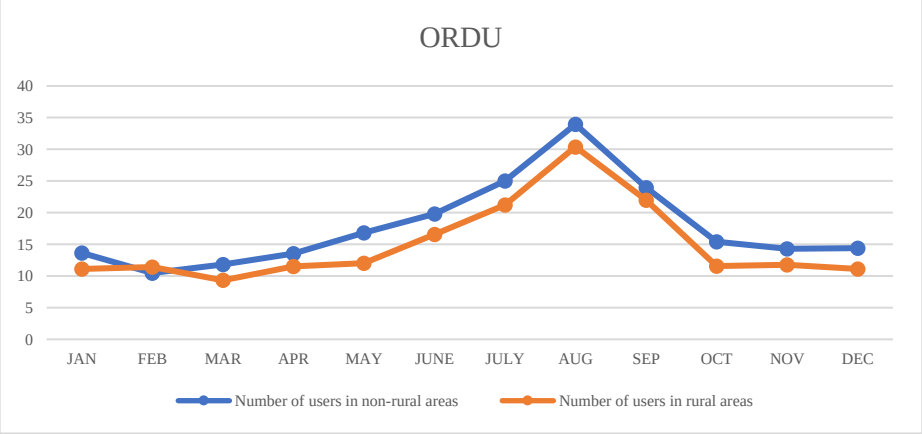


Fig. 5a. The number of Refugee Users in Rural and Non-Rural Areas in Ordu (raw numbers)

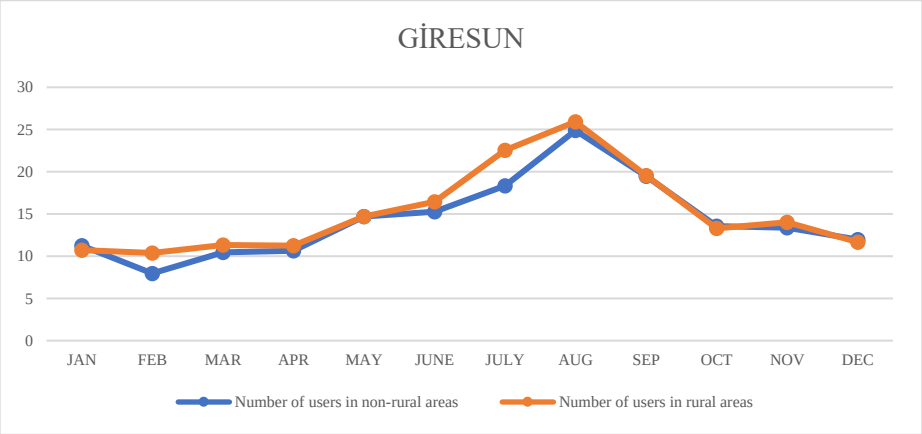


Fig. 5b. The number of Refugee Users in Rural and Non-Rural Areas in Giresun (raw numbers)

An inspection of Figure 5a and Figure 5b presenting the number of unique refugee users observed in rural and non-rural areas of Ordu and Giresun provinces, lends further support for the seasonal migratory movements towards Ordu and Giresun during the harvesting time of hazelnut in these provinces. Our findings demonstrate that the number of refugees gradually increases starting from March onwards and reaches a peak in August and returns to pre-harvest season levels in October. All in all, our findings from the refugee activity shares in rural areas and complementary analysis indicate seasonal agricultural migratory flows also to Ordu and Giresun provinces.

When we move on to the Central Anatolian region, Niğde, a province that accounts for approximately 30 percent of all potato production in Turkey stands out as one of the agricultural sites that provide refugees seasonal work throughout the harvesting season. The harvesting season of potato plant in Niğde typically starts at the end of July and lasts until September or October depending on the type of potato plant. As illustrated in Figure 6a, while the share of rural refugee activity within all refugee mobile activities drops during the harvest season, our findings from the analysis of the number of unique refugee users in rural and non-rural settings of Niğde indicates a significant increase in the number of refugees residing in Niğde during the harvesting season of potato as shown in Figure 6b. Accordingly, the number of refugees actively using their mobile phones in rural areas is found to increase approximately by 150% in August when compared to the number of refugees in thepre-harvest season.

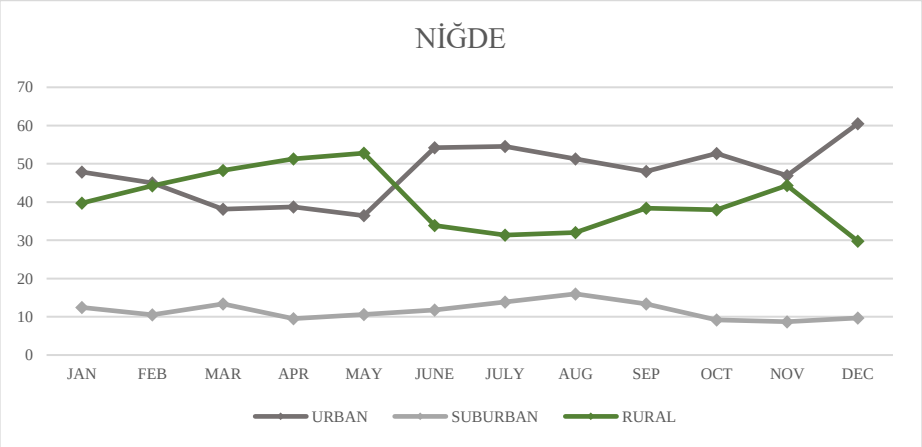


Fig. 6a. Distribution of Refugee Activities across Area Types in Niğde (percentages)

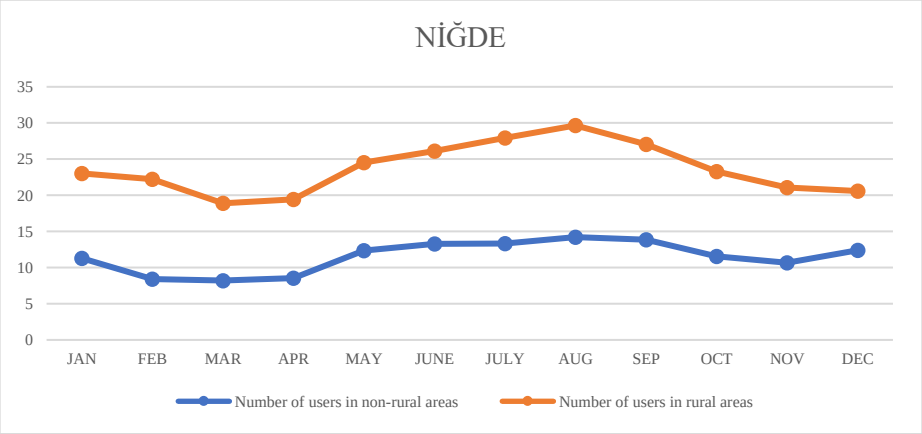


Fig. 6b. The number of Refugee Users in Rural and Non-Rural Areas in Niğde (raw numbers)

When we focus on the Eastern Anatolian region, Malatya stands out as one of the biggest provinces whose economy is mainly centered on agriculture. Apricot production lies at the core of its agricultural production and the apricot harvesting season in Malatya typically starts in mid-June and lasts until early August. As illustrated in Figure 7a the share of mobile activities of refugees in rural areas within the refugee activities in all area types significantly increases from June to July and drops back to pre-harvest season levels shortly after August. An inspection of the number of refugee users in rural and non-rural settings of Malatya also indicates a gradual increase in the number of refugees between June and September, which can be interpreted as a sign of seasonal agricultural migratory flows of refugees.

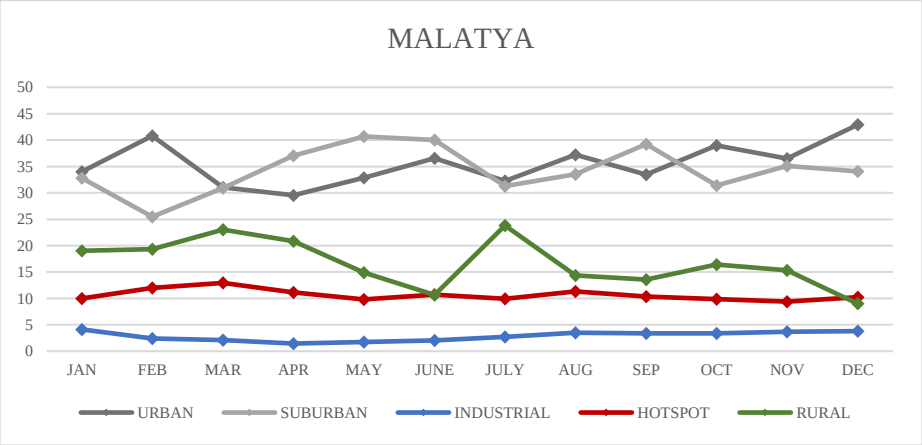


Fig. 7a. Distribution of Refugee Activities across Area Types in Malatya (percentages)

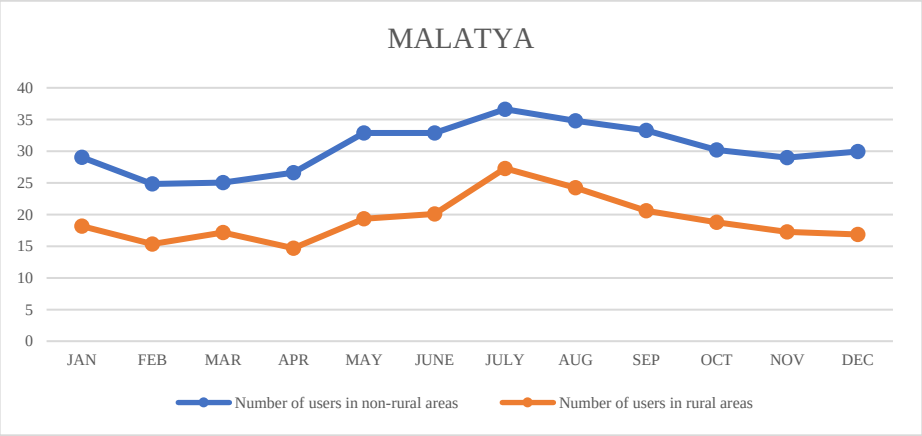


Fig. 7b. The number of Refugee Users in Rural and Non-rural Areas in Malatya (raw numbers)

Adana, a well-documented destination for the seasonal agricultural workers, on the other hand, offers opportunities for work in almost every season of the year with its appropriate climate for diverse range of agricultural products including cotton, oranges, lemons, tangerines and onions. Adana, especially the districts of Yüreğir, Karataş, Yumurtalık, Ceyhan, and Seyhan, hosts more than ten thousand seasonal workers every year, and our findings suggest that Adana is one of the provinces that agricultural refugee workers also work in the harvesting of a diverse set of agricultural products in Adana throughout the year. An inspection of Figure 8a demonstrates that the share of refugee call activities placed in rural areas within all refugee activity fluctuates at a much lower rate than the other provinces we have so far analyzed and slightly increases between June and October. In a similar vein, Figure 8b demonstrates that the number of refugees in the rural regions of Adana is observed to be significantly higher than the number of refugees in the rural parts of other provinces throughout the year. On an

additional note our analysis with Dataset 3 exploring the interprovincial migration rates further demonstrates that the migratory flow from and to Adana makes up of 7.2 % of all the refugee migratory movements in our sample. When all taken into account, our findings suggest Adana to be one of the agricultural sites that attracts seasonal refugee migration almost in every season of the year. The findings on the migratory flows through refugee call activities also comply with the existing empirical research on the Syrians' access to seasonal agricultural work in Adana, which suggests that they participate in a wide range of agricultural activities in the region, including the harvesting of vegetables, cotton, citrus, peanuts and greenhouse cultivation [8].

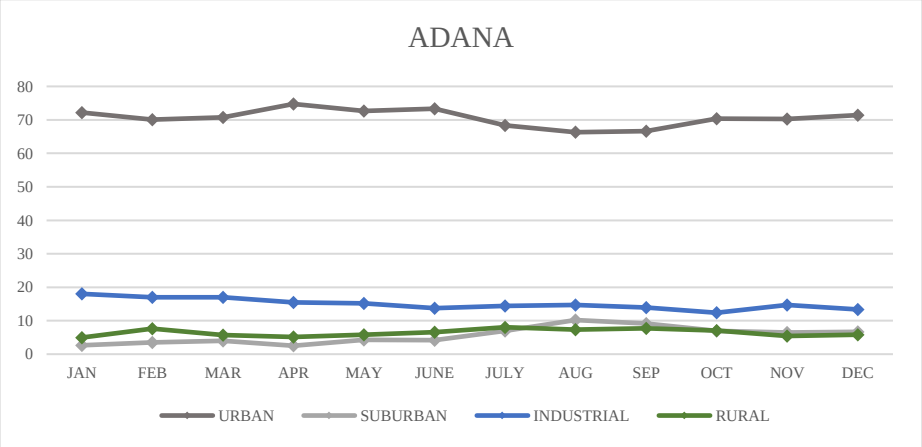


Fig. 8a. Distribution of Refugee Activities across Area Types in Adana (percentages)

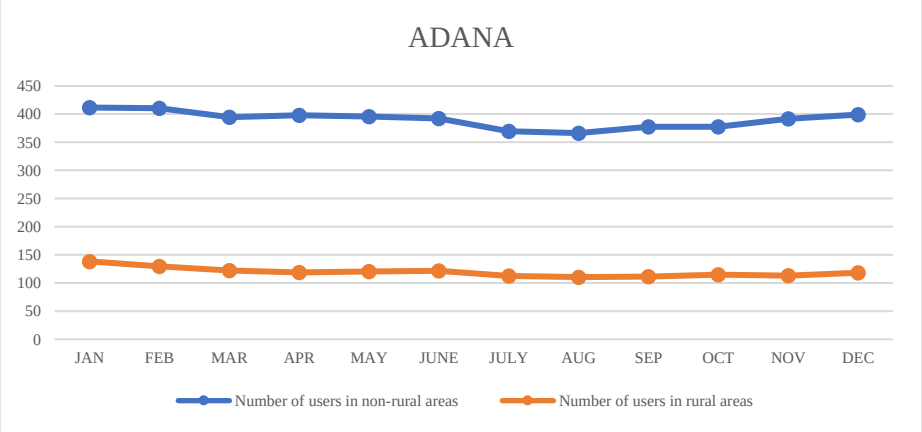


Fig. 8b. The number of Refugee Users in Rural and Non-Rural Areas in Adana (raw numbers)

2.2 Seasonal Tourism Worker Migration

While seasonal agricultural work is found to be one of the important drivers of internal migration of refugees in Turkey, seasonal needs for workers in the tourism sector also seems to be one of the crucial drivers of internal migratory movements of refugees in Turkey. The inspection of the interprovincial migration patterns of the refugees reveal that the two provinces with highest rates of tourism sector activities, namely Antalya and Muğla, received a considerable share of their internal refugee migrants during the high tourism seasons. While initially Syrians were not officially allowed to reside and take up paid employment in Antalya [10] as evinced by the "Open Hearts and Welcoming Communities for Immigrants and Refugees" project carried out by Akdeniz University Department of Tourism in collaboration with Konyaaltı municipality and other local partners, public attitudes and policies towards Syrian refugees in Antalya has significantly became less restrictive by the year 2017[11]. Reflecting this change, approximate 50 percent of all refugee mobility towards Antalya took place during a four-month period between June and September. In a similar vein, approximately 60 percent of all Muğla's internal refugee migrant intake is observed during the period between May and August. The provinces of Antalya and Muğla also stand out as the two provinces where the share of refugee activities in seasonal areas significantly increases during the summer holiday seasons between May and September, indicating that migratory movements of refugees to these provinces are driven by labor needs in the tourism service sector. As illustrated in *Figure 9a*, the share of refugee activities in seasonal areas of Antalya steadily increases from 15.2 percent in April to 22.6 percent in July and slowly drops back to 13 percent levels in October. In a similar vein, as illustrated in Figure 9b, the share of refugee activities in seasonal areas of Muğla also increases from 14.4 percent in April to 21.6 percent in May, and it further increases 16 percentage points in July reaching at its peak levels of 38.4 percent.

The high levels of refugee intake during the high season for tourism and increasing levels of refugee activity in the seasonal areas of Antalya and Muğla suggest interprovincial mobility to these provinces to be driven by seasonal labor opportunities in the tourism sector. However, we further test our assumptions of seasonal work migration to Muğla, as the May-September period is also the high season for irregular border passages from Turkey to Greece, which in the case of Muğla, takes place from Bodrum to Kos island. To this end, we turn to our analysis of interprovincial migration utilizing Dataset 3 that provides us with the coarse-grained call activity data for a sample of 24,233 refugees and trace the mobility patterns of those refugees who had been in Muğla between May and September. In this dataset, we identified 169 refugees for whom Muğla was spotted as the province of residence for at least one month during the period between May and September. Findings from our analysis demonstrate that while approximately 40 percent of these refugees stayed in Muğla for the rest of the observation period, almost half of these refugees (47.6 percent) has moved to other provinces after their stay in Muğla. The percentage of those refugees who were last spotted in Muğla and had no call activity within the last three months of the observation period, on the other hand, is found to be lower than the percentage of those inactive refugees for the same period in the larger dataset (7.1% in Muğla as oppose to 11% in the whole

dataset), which suggests that observed inactivity trends in Muğla are more likely to be stemming from missing data patterns of the dataset itself rather than being indicative of irregular passages from Turkey to Greece. In summary, our findings reveal the existence of high rates of seasonal migration to Muğla and lend support for the interpretation that these migratory movements of refugees are more likely to be driven by seasonal labor needs of the province rather than irregular border crossing motivations.

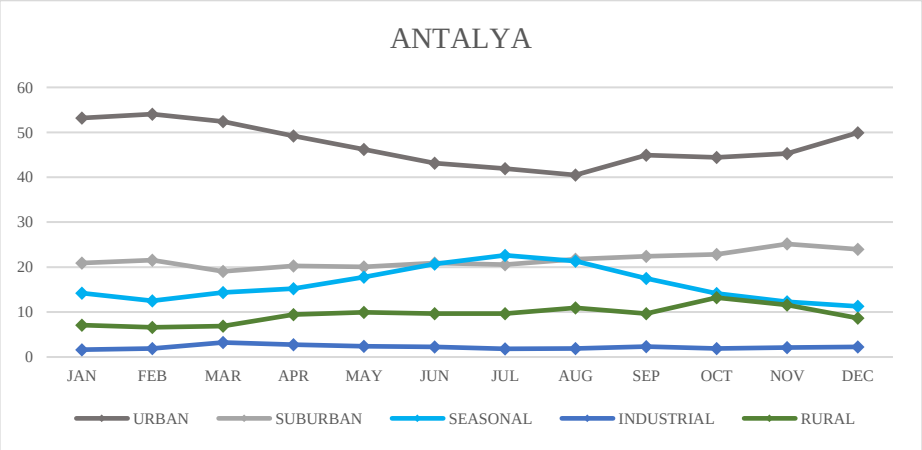


Fig. 9a. Distribution of Refugee Activities across Area Types in Antalya (percentages)

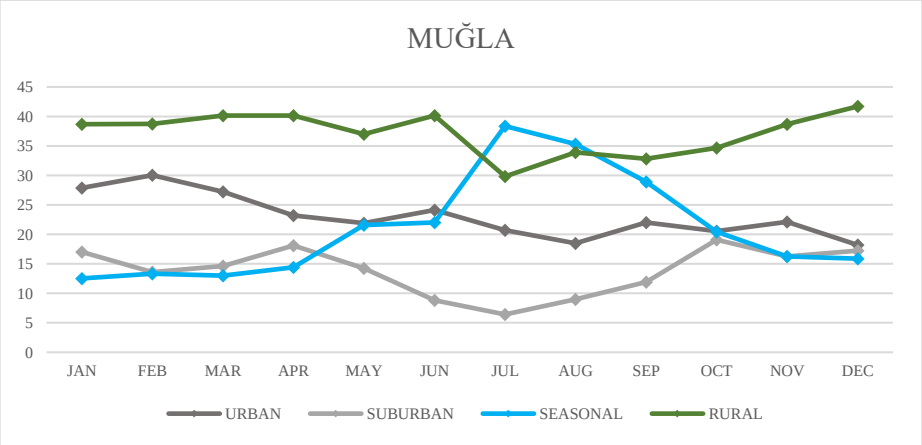


Fig. 9b. Distribution of Refugee Activities across Area Types in Antalya (percentages)

As in the case of other sectors, the employment of Syrian nationals in the seasonal tourism sector remains within informality, as it can be captured from the data on the work permits to foreign nationals provided by the former Ministry of Labour and Social Security. Among the 110 work permit applications (96 of which were granted) made to MLSS in Antalya in 2017, 14 applications were in areas that are related to tourism (i.e.

travel agency official, receptionist, cook, and animator). In Muğla, 52 of the 78 applications in 2017 were granted with a work permit and only four applications (i.e. translator and waiter) can be attributed with the tourism sector [12]. As far as the interprovincial mobility trends of the refugees in Antalya and Muğla are taken into consideration along with the relatively low number of work permits issued in the field of tourism in these provinces, our findings support the argument that the employment of Syrian nationals in the seasonal tourism sector remains within informality.

3 Host and Syrian Community Interaction Patterns

In order to trace the host and refugee community interactions, we utilized Dataset 2 and identified the number of calls made by refugees and non-refugees to users registered as refugees, non-refugees or unidentified users. Figure 10 illustrates the percentage of calls exchanged between refugee and non-refugee users within all the calls placed by refugee and non-refugee users between January 2017 and December 2017. The inspection of the data reveals that the calls exchanged between refugees and non-refugees constituted approximately 8% of all the calls placed by refugees and non-refugees in January 2017, and the share of refugee and host community mobile interactions has significantly increased throughout the year. The percentage of refugee and Turkish citizen interactions steadily increased in the first three quarters of 2017, reaching a peak in October (15.9 %), and slightly decreasing by the end of 2017. In December 2017 the refugee and Turkish citizen interactions were recorded as constituting 14.5 percent of all the calls placed by refugees and Turkish citizens.

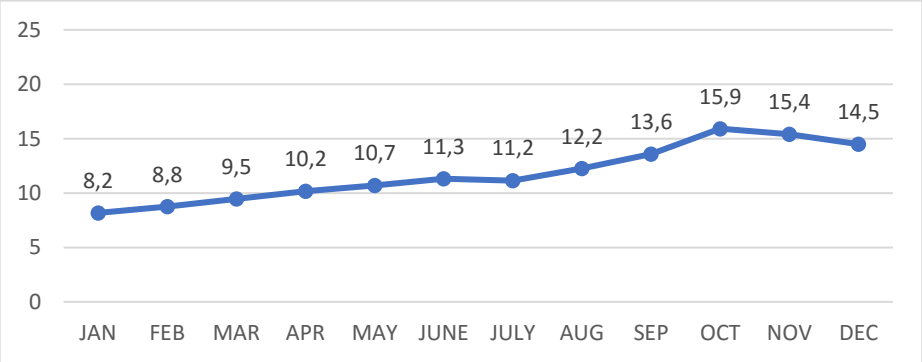


Fig.10. The Share of the Calls between Refugees and Non-Refugees within All Calls Placed by Refugees and Non-Refugees

The inspection of the calls placed by refugee users also reveal that refugees place calls to non-refugees more often than they place calls to other refugees. Accordingly, the percentage of calls to users registered as non-refugees constituted 87.4% of the all calls placed by refugees in Turkey throughout the year, and the percentage of calls from

refugees to non-refugees constituted 74 to 98 percent of all refugee calls in each province over a year's period. While the lowest shares of refugee-Turkish citizen interactions are recorded in Muş, Ağrı, Trabzon and Tunceli (74 to 76% of all refugee activity), the highest shares of refugee-Turkish citizen interaction are recorded in Iğdır, Kırşehir, Ardahan, Bitlis and Bingöl where the refugee calls to Turkish citizens constituted 95 to 98 percent of all refugee activity in those provinces.

While the shares of refugee calls to Turkish citizens within all refugee call activity allow us to make inferences about the ratio of refugee to refugee and refugee to Turkish citizen interactions, they still fall short of providing us with a complete picture as they cannot reveal how frequently refugees and Turkish citizens are interacting with one another. Therefore, we further analyze the mean number of calls placed by refugees to non-refugees, to refugees and to unidentified users. Figure 11 presents the mean number of calls per refugee by call type and province over a year's period. An inspection of the figure reveals that the highest number of calls per refugee to non-refugees is observed in Gaziantep (30.8 calls per refugee), it is closely followed by İçel (29.9 calls per refugee), Samsun (29.3 calls per refugee), İstanbul (29.2 calls per refugee) and İzmir (29.1 calls per refugee). A detailed analysis of the data also illustrates that the number of refugee to non-refugee calls reached a peak in Gaziantep and İçel in August, whereas the highest number of refugee calls to non-refugees are observed in January in İstanbul and Samsun, and in September in İzmir. Amasya, Bolu, Bayburt, Artvin and Hakkari, on the other hand, are observed to be the provinces with the lowest number of calls per refugee to non-refugees in 2017.

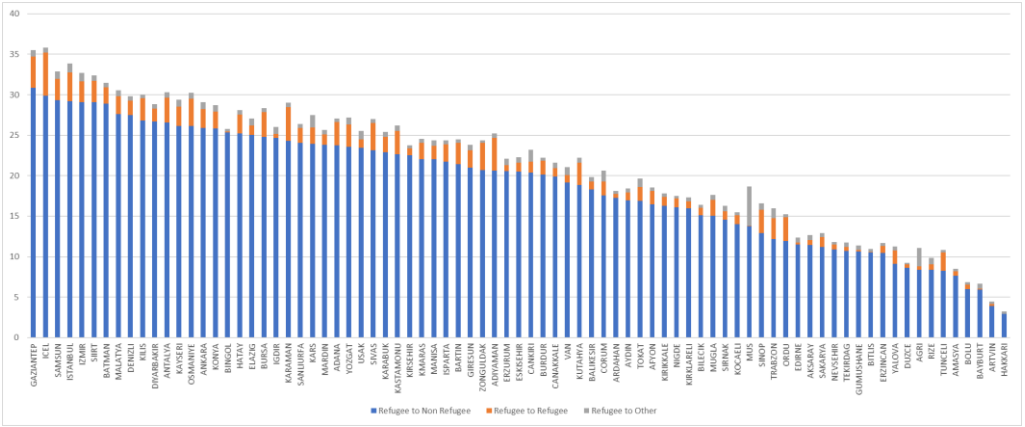


Fig.11. The Number of Calls Per Refugees by Call Types and Provinces

While our analysis so far revealed that refugee calls to non-refugees constituted 87.4 % of all the calls placed by refugees throughout the year, the non-refugee calls to refugees make only the 0.05% of all the calls placed by non-refugee users. The average number of calls to refugees placed by non-refugees are also observed to be significantly lower when compared to average number of refugee calls to non-refugees. While the

average number of non-refugee calls to refugees is observed to be the highest in Gaziantep (0.034 calls per non-refugee), Batman (0.031 calls per non-refugee), İstanbul (0.029 calls per non-refugee), the lowest averages are recorded in Sivas, Ardahan and Gümüşhane (less than 0.001 calls per non-refugee).

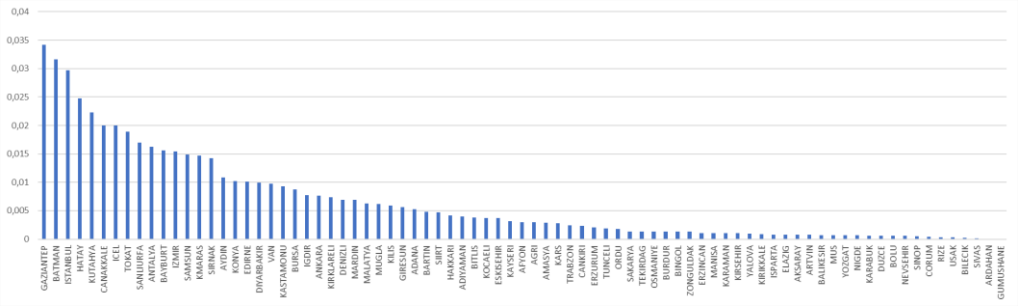


Fig. 12. The Number of Calls to Refugees Per Non-Refugee User

All in all, our analysis illustrates that the percentage of calls exchanged between refugee and non-refugee users within all the calls placed by refugee and non-refugee users between January and December 2017 has steadily increased to a level where refugee and non-refugee mobile interactions constituted almost 15% of all the call activity recorded by the end of 2017. The closer inspection of calls logs of refugee and non-refugee users, however, suggest that these interactions are mainly initiated by refugees as approximately 8.5 million calls were placed by refugees to non-refugees as opposed to a total of 32,663 non-refugees calls to refugees. Moreover, the data reveals that contrary to our earlier assumptions, refugee to non-refugee interactions constituted a higher share in the overall mobile interactions done by refugees. A further analysis done by controlling for the number of calls per refugees has also provided insights on the fact that frequent interactions have taken place between refugees and non-refugees especially in some of the main host provinces, including Gaziantep, İçel, İstanbul and İzmir. While these findings shed light on the refugee and host community relationships, we believe that they can be supported with further evidence that takes into consideration other intervening factors, such as demographics and socio-political context. This project offers the prototype of a web based deep mapping platform that incorporates spatio-temporal D4R data with other layers of data on a real-time basis, which could be used as an open source for further research and development by academic researchers and policy makers to further investigate the patterns of refugee-host community interactions along with various other research questions not only in the field of integration but also in the fields of healthcare, education, safety, employment.

4 Deep Mapping D4R Content

Maps have always been strong tools for archiving, representing and sharing location or place-based data. For the last 50 years, Geographical Information Systems (GIS)

made use of satellites, imaging technologies and the internet to enrich the abilities of maps especially for researchers working with quantitative methods. With the further advances in technologies such as ubiquitous computing, mobile devices, and sensors, we have been facing a new turn in the Digital Humanities (DH) field; Deep Mapping. The term Deep Mapping is used for online and interactive mapping platforms, which represent spatially contextualized, multilayered, complex, spatio-temporal data in a visual, open-ended, experiential and flexible approach unlike the structured and precise approaches of the GIS. With this inclusive approach deep maps bring together researchers, experts working with both quantitative and qualitative data, in addition to the public, to create deeper discussions and narratives.

With this perspective, we conceptualized, designed and prototyped a deep mapping platform for the D4R data, namely "Datascapes4Refugees" (see Fig. 13). Aim of this platform is to overlay data from different sources (D4R, NGOs, demography, news etc.) and provide tools for spatial (Fig. 14.a and 14.c) and temporal (see Fig. 14.b) navigation of these multilayered data layers structure (see Fig. 14.d). One can examine the data with different horizontal and vertical scales and finally discuss for the cause and effect, possible actions for each data item relating data from different sources (see Fig. 15). We organized expert meetings to discuss the data, how we can ask questions to this data, create narratives using it which was later on used to build the structure of this tool. But this has been mostly an early stage for the design and development of a well-established community platform. More work is needed on how to collect census data, news and events from media and social media sources and what kind of other data types can be employed. The early prototype of the D4R Deep Mapping platform can be accessed via the following link: http://d4r.kakare.net.

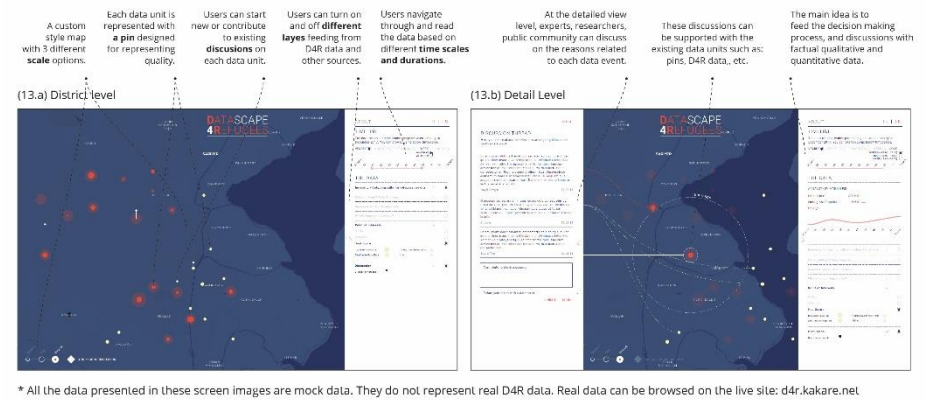


Fig. 13. Two of the Datascape4Refugees Deep Mapping Platform interfaces. (13.a) The District level interface (left) and (13.b) Detail Level interface (right). These screenshots from the design phase makes use of mock data and they do not represent any real data.

Fig. 14. Interface elements of the deep mapping interface: map, timeline, spatial scaling and data layers.

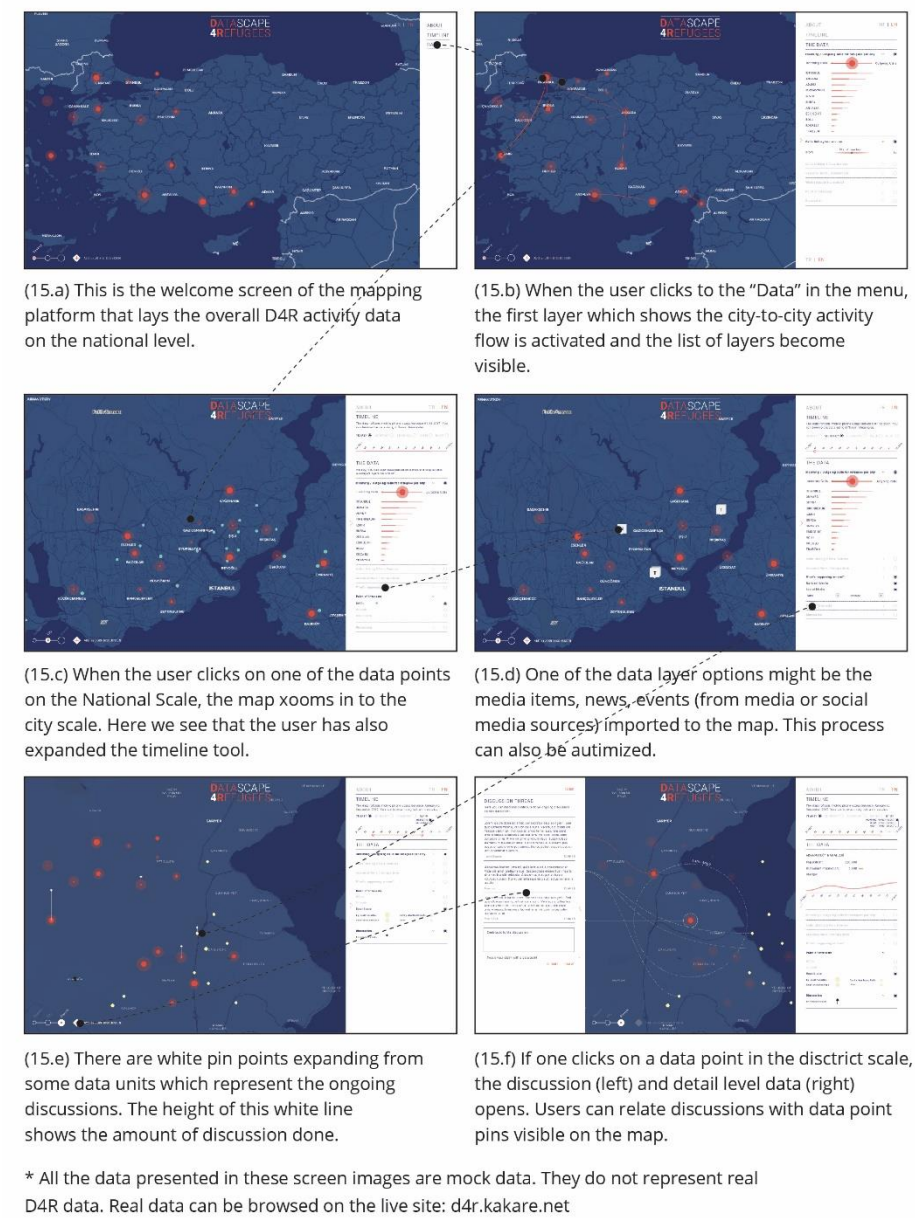


Fig. 15. A typical browsing scenario for a user visiting the deep mapping platform. In this typical scenario, user start from the home screen, and browses through different levels and layers of the D4R data ending with a detail level discussion screen.

5 Discussion and Conclusion

The official statistics on Syrian refugees in Turkey are compiled mainly on the basis of registration data and these statistics can offer only a limited capacity to inform well-tuned integration policies that would meet the needs of a dynamically changing target population. Our findings demonstrate that nearly 20 percent of Syrian refugees, in other words, one in every five Syrian refugees, has moved to at least one other province than they initially were in within a single year's time. We can claim that the Syrian population in Turkey is highly mobile, and our findings further suggests interprovincial migratory movements of refugees to be driven by seasonal labor needs in the destination provinces.

While Antalya and Muğla identified as the provinces where the seasonal migratory movements of the refugees to these provinces are found most likely to be driven by the seasonal work opportunities in the tourism sector, migratory movements towards Rize, Ordu, Giresun, Niğde and Malatya are found to be driven by agricultural worker needs. Adana, a well-documented destination for the seasonal agricultural workers, on the other hand, is observed to accommodate a large number of agricultural refugee workers throughout the year as it offers opportunities for work in almost every season of the year with its appropriate climate for diverse range of agricultural products including cotton, oranges, lemons, tangerines and onions.

It is a well-known fact that agriculture is among the most hazardous occupations in the world. Farmworkers are exposed to various occupational and environmental hazards including pesticides, dust, unfavorable weather conditions, transport, long working hours, food safety and security, infectious diseases, lack of sanitation and hygiene in living and working environment, injuries due to accidents and wild animals [13].

There are already many existing problems for people working in agriculture in Turkey. To name a few, firstly substandard housing is an important issue. 80% of seasonal migrant farm workers (SMFWs) accommodate in tents and more than 75% have no access to clean water and sanitation services. SMFWs usually travel and work as a family with their elderly and children. Child labor is not uncommon among SMFWs due to their economic returns. One out of four persons does not utilize the health services when they feel sick. Moreover, it is crucial for children to get the vaccines on time and while working on the field there is a barrier to access to immunization services, as it is not always easy to leave work or commute from rural areas to a close by health center. When compared to the general population, infant mortality is four times and maternal mortality is 10 times higher among SMFWs in Turkey. Access to information is another asset, which SMFWs have limited opportunities. For example, 90% of women and 80% of men has reported that they do not know the health impacts of pesticides [14].

It is, therefore, necessary to provide occupational health and safety services to SMFWs along with ensuring access to information, decent housing or accommodation during work, social and educational services for women and children, and access to health care when needed. The Prime Minister Circular no.2017/6 mandates the governors to ensure that these needs are met via coordinated efforts of different governmental institutions including directorates of health, labor and social security, education, family

and social services, and security forces [15]. However, the implementation phase of the regulation cannot be performed in an effective way, leading to informal and unhealthy working in agriculture.

CDR data provided by Turk Telecom and the DataScape4Refugees Mapping Platform can be of use in supporting the activities defined in the circular. Our analysis supports the previous reports that Syrian Refugees are engaging in agricultural activities, illustrated by the increase in the rural activity logs in certain months in line with relevant harvest time of certain products. This means Syrian Refugees require occupational health and safety services to say the least. Outreach activities play a vital role in meeting the services available to the workers. These include inspectors from various directorates and mobile services such as health care [16]. Besides, as SMFWs travel as a family, preventing child labor during the harvest time, providing transport for the children to schools nearby are needed to support the healthy development of the child. Outreach can also increase the awareness among Syrian SMFWs by providing information on how to work safely in agriculture, and services provided by the government to SMFWs. Governors can benefit from the data that shows the movements of Syrian Refugees from one province to another and from urban to suburban or rural especially in certain times such as harvest. The call activity logs of Syrian Refugees and their location as a generalized and anonymous dataset can be used to see the changes in the number of Syrian Refugees in certain locations during certain times. Such a surveillance will allow governors to assign outreach teams from different institutions to detect the location of Syrian MSFWs that would lead the way to meet them with the services they required. For example, some Provincial Health Directorates have purchased specially built vans to be used as mobile health clinics to provide environmental health services, vaccination for children, reproductive health and screening services to the MSFWs [16]. However, temporary accommodation facilities are not known to these outreach teams. As a consequence, the success of the program is limited with the chance to encounter these settlements or with the calls they get from SMFWs to ask for the provision of such services. The DataScape4Refugees Mapping Platform would also allow developing a surveillance system for outreach teams that might improve the quality of living and working conditions of Syrian SMFWs as well as their access to services such as health care, education and social services.

Acknowledgements:
The authors would like to thank Eleni Diker for the provision of the dataset on non-governmental organizations, and to student assistants Abdullah Coşkun, Hüseyin Kuşçu, Damla Çay and Salih Tosun for their efforts in research and reporting of findings.

References

1. DGMM Homepage, http://www.goc.gov.tr/icerik6/temporary-protection_915_1024_4748_icerik, last accessed 2018/09/10

2. Kaygısız, I. Suriyeli Mültecilerin Türkiye İşgücü Piyasasına Etkileri. Report, Friedrich Ebert Stiftung, (2017)., http://www.fes-tuerkei.org/media/pdf/Dünyadan/2017/Du308nya-dan%20-%20Suriyeli%20Mu308ltecil-erin%20Tu308rkiye%20I307s327gu308cu308%20Piyasasina%20Etkileri%20.pdf, last accessed 2018/09/10
3. For an extended research on refugees working at seasonal agricultural jobs, see Kavak, S.: Syrian refugees in seasonal agricultural work: a case of adverse incorporation in Turkey. New Perspectives on Turkey 54, 33-53 (2016).
4. Salah, A.A., Pentland, A., Lepri, B., Letouzé, E., Vinck, P., de Montjoye, Y.A., Dong, X. and Dağdelen, Ö., 2018. Data for Refugees: The D4R Challenge on Mobility of Syrian Refugees in Turkey. arXiv preprint arXiv:1807.00523.
5. Kalkınma Atölyesi, Türkiye'de Mevsimlik Tarım İşçisi Olarak Çalışan Yabancı Göçmen İşçilerin Geldikleri Ülkeler, http://www.ka.org.tr/dosyalar/file/Yayinlar/Raporlar/TURKCE/03/G%C3%96%C3%87MEN%20%C4%B0%C5%9E%C3%87%C4%B0LER%20MEVCUT%20DURUM%20HAR%C4%B0TASI.pdf, last accessed 2018/09/10
6. İçduygu, A.: Turkey: Labour Market Integration and Social Inclusion of Refugees, Study for the EMPL Committee, (2016) http://www.europarl.europa.eu/RegData/etudes/STUD/2016/595328/IPOL_STU(2016)595328_EN.pdf., last accessed 2018/09/10
7. İçduygu, A. and Diker, E.: Labor Market Integration of Syrian Refugees in Turkey:From Refugees to Settlers. Göç Araştırmaları Dergisi. 3(1), 12-35 (2017).
8. Kalkınma Atölyesi, Adana Ovası'nda Suriyeli Göçmen Mevsimlik Tarım İşçileri Haritası, http://www.ka.org.tr/dosyalar/file/Yayinlar/Raporlar/TURKCE/04/MIT%20HARITA.pdf., last accessed 2018/09/10
9. Aygül, H. H.: Mülteci Emeğinin Türkiye Piyasalarındaki Görünümü ve Etkileri. Süleyman Demirel Üniversitesi Vizyoner Degisi. 9(20), 68-82 (2018).
10. Uzun, S.: Suriyeli göçmenlere Antalya yasağı. Hürriyet, 07 November 2014, http://www.hurriyet.com.tr/gundem/suriyeli-gocmenlere-antalya-yasagi-27534479, last accessed 2018/09/12
11. Akdeniz University Website, http://bhim.akdeniz.edu.tr/akdeniz-universitesi-multeci-ve-gocmen-topluluklari-icin-hosgoru-projesi-ile-hibe-almaya-hak-kazandi/, last accessed 2018/09/12
12. Çalışma İstatistikleri Bilgi Sistemi, http://cibs.csgb.gov.tr/RaporOlusturma-Sihirbazi.aspx?kullanicisiz=1., last accessed 2018/09/13
13. ILO. Safety and health in agriculture. ILO code of practice. International Labour Office - Geneva, (2011)
14. Şimşek Z. Needs assessment of seasonal migrant farm workers and their families 2011. Harran University Faculty of Medicine & UNFPA (2012) Ankara. [Turkish]
15. TC Resmi Gazete. Mevsimlik Tarım İşçileri ile İlgili 2017/6 Sayılı Başbakanlık Genelgesi 19 April 2017 No: 30043 http://www.resmigazete.gov.tr/eskiler/2017/04/20170419.htm, last accessed 2018/09/13
16. Denizli Saglik Mudurlugu Homepage, http://denizlism.gov.tr/TR,73486/quotmevsimlik-tarim-iscilerine-sunulan-saglik-hizmetlerini-degerlendirme-ve-gelistirme-egitimiquot.html last accessed 2018/09/14

AROMA_CoDa: Assessing Refugees' Onward Mobility through the Analysis of Communication Data

Harald Sterly[1] [0000−0001−8819−1638], Benjamin Etzold[2][0000−0002−1109−7640], Lars Wirkus[2][0000−0002−9795−3187], Patrick Sakdapolrak[3][0000−0001−7137−1552], Jacob Schewe[4][0000−0001−9455−4159], Carl-Friedrich Schleussner[5][0000−0001−8471−848X], Benjamin Hennig[6][0000−0002−5754−2455]

[1] University of Bonn, 53115 Bonn, Germany, sterly@giub.uni-bonn.de
[2] Bonn International Center for Conversion, Pfarrer- Byns-Straße 1, 53121 Bonn, Germany
[3] University of Vienna, Universitätsstraße 7/5, 1010 Vienna, Austria
[4] Potsdam Institute for Climate Impact Research, Telegrafenbg A56, 14473 Potsdam, Germany
[5] Climate Analytics gGmbH, Ritterstraße 3, 10969 Berlin, Germany
[6] University of Iceland, Askja, Sturlugata 7, 101 Reykjavík, Iceland

Abstract. Secondary or onward mobility of refugees can pose considerable challenges for targeted and timely humanitarian assistance, and for long-term integration. There is very little systematic knowledge on the onward migration of refugees after their initial flight to a country of reception in general, and specifically in Turkey. In the paper we describe how the analysis of mobile phone Call Details Records can help to better understand spatio-temporal patterns of refugees' onwards mobility. The analysis reveals some clear, large-scale mobility patterns (from South to North, from East to West, from Centre to the Coast, to large urban areas), and also some temporal patterns, but also shows that human mobility is complex and accordingly requires more advanced analytical tools. We conclude that it might be worth of reframing registration policies for refugees, given the highly mobile share of refugee population, and the important role that this mobility probably plays for livelihoods.
Keywords: Call data record, mobility, social integration, Turkey, refugees movements

1 Introduction: Why is secondary or onward mobility important?

By the end of the year 2017, 68.5 million people were forcibly displaced globally due to conflict, persecution or violence — an increase of almost 3 million compared to 2016 [1]. While many of the displaced find refuge at the places to where they flee, internationally or internally, many continue their journey — in order to reunite with family or kin, to escape poverty or to improve their livelihoods,

to flee further persecution, or due to other reasons. This *secondary* or *onward mobility* is an important issue due to several reasons.

First, and apart from often violating regulations for refugees or asylum seekers, it poses often practical challenges for humanitarian assistance and long-term integration of refugees: it is difficult to effectively address mobile persons with targeted and timely support for immediate needs (such as shelter, food or health), but even more so to socially and economically integrate mobile populations into the receiving society. On the other hand, unknown mobility patterns also challenge efficient planning and management of the assisting communities and institutions: investments of (scarce) resources in infrastructure and measures for support are likely to be misplaced if a target group moves on.

Second, there is very little knowledge about the secondary mobility of refugees – the size of moving populations, their routes and central nodes in mobility networks, the specific trajectories and timings of movements, as well as interim and final destinations. In general, however, it seems like refugees' journeys – even from the same regions of origin – have become not only more diverse, multi-directional and longer, but also fragmented as periods of mobility interchange with longer phases of immobility. With more than 3.5 million displaced persons [1], Turkey is not only a particularly important host country for a highly vulnerable, yet mobile population group, but also the most significant mobility hub for onward mobility to Europe and for the return to countries of origin.

2 Objectives, Data and Methods

Knowing larger scale patterns of onward mobility of refugees within Turkey helps to better assess where and how support (e.g. emergency shelter, health services) should be provided, and where sustainable investments in the physical, economic and social infrastructure (e.g. employment opportunities, educational facilities) are best allocated. Knowing key drivers of secondary mobility, moreover, helps to anticipate future mobility, to make use of refugees' flexibility and to enhance incentives for refugees to stay at places where their needs can be catered for adequately.

We aim to show how the analysis of mobile phone Call Detail Records (CDR) can yield valuable information on the spatial and temporal patterns of secondary mobility of refugees, including flows over time and spatial trajectories between important points of origin and destination. In addition, and with additional information from secondary data, CDR analysis can also help to better understand the drivers of spatio-temporal mobility patterns of refugees.

2.1 Data

We were granted access to a set of CDR from the Turkish mobile operator Türk Telekom. Access was granted in the context of a call for projects with the aim of investigating the potential of analysis of mobile call data for the improvement of the situation of Syrian refugees in Turkey ("Data for Refugees: The D4R

Challenge on Mobility of Syrian Refugees in Turkey", see [2]). The data was sampled from CDR over the period of one year, from 01.01.-31.12.2017 from 992.457 Türk Telekom customers, thereof 184.949 that were registered in the customer database as "refugees"; 75% of these "refugees" were registered as male [3], however gender is not stated in the dataset. It has to be noted that the flag "refugee" in the datasets includes migrants, asylum seekers and foreigners with a temporary protection status; and while the individual attribution of the refugee status to an individual caller ID might not be possible with certainty, it should however be possible to deduce general patterns from aggregate analyses [3].

The data consisted of three datasets: Dataset 1 "Antenna traffic between cell tower locations", consisting of the total exchange of calls and text messages between cell towers; Dataset 2 "Fine Grained Mobility" of about 65.000 users that were newly sampled and assigned random ID numbers for time periods of two weeks over the course of the year; and Dataset 3 "Coarse Grained Mobility", containing of the CDR of a subsample of users over the course of the whole 12 months but spatially aggregated on prefecture / district level [3].

We concentrated for the present (first) analysis on the analysis of Dataset 3, which consisted, after consolidation, of 56,433,358 entries of the form 'Caller ID', 'Timestamp (DD-MM-YYYY HH:MM)', 'District ID', 'City ID'. All steps of data analysis were carried out using the open source statistical software R (3.5.0), on a standard desktop computer. Visualization was done in R, MSExcel and in Adobe Illustrator.

2.2 Methods

Dataset 3 was of special interest for our analysis, as it allowed for a longer temporal overview of mobility patterns, and thus enabled us to differentiate between short-term movements (e.g. for visits), circular, seasonal or more longer-term migratory movements.

In a first step, the data was checked for consistency and was consolidated. This consisted of the removal of duplicate data (in total almost 9,6 million rows) and the combination of data for outgoing and incoming voice calls to improve temporal coverage, as the dataset showed considerable gaps, especially in February and March (see Figure 1).

In a next step, we aggregated the DS3 dataset over individual callers and days, resulting in a dataset with a combination of all callers and the days on which they actively called or were called, their first and last districts for each given day (dayn), as well as the first and last districts where the callers had been on the preceding day (dayn-1) and on the subsequent day (dayn+1). We take the location of the last districts where callers have been on these days as a proxy of their places of residence, as this usually refers to places in the evening hours.

We then calculated the distance between the first district where callers were using their phone, and the last district, i.e. where they "entered" and where they "exited" the dataset. Thus a caller who started at district A, moved to B, to C, etc., and finally returned to A would be showing a zero value here, whereas for a caller moving from A to B, to C, etc. and ending in X we would obtain

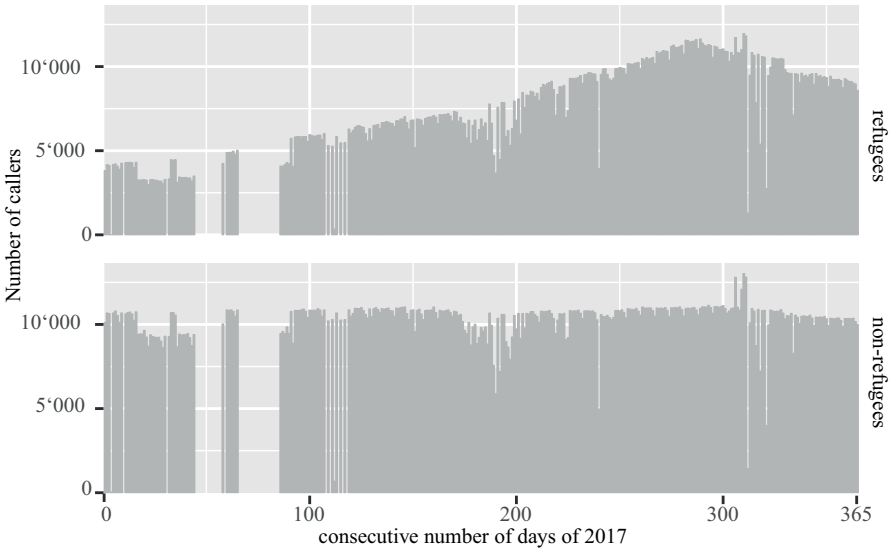


Fig. 1. Number of callers flagged as refugees and as non-refugees per day, over the whole of 2017. Note the large gaps in February and March, as well as also on single days later in the year.

the distance between A and X. Figure 2 shows the distribution of the callers according to this distance.

In a next step, a mobility (origin-source) matrix was created, containing the number of refugees moving between districts. From this, information on net migration flows between districts was calculated.

Based on this, a subset of refugees was selected who moved more than 100km between their districts of first and last appearance in the dataset. From an analysis of the mobility patterns of a small sample of callers we assume that a movement of more than 100km can be interpreted as a proxy for a temporary or permanent shift of residence, hence as migration as opposed to commuting.

Then we determined when/on which days these “migrating” refugees made movements of more than 100km, and for each day we summarized the number of the callers moving more than 100km on that day. Because of the high temporal variability of callers covered in the dataset (see Figure 1), we normalized the number of callers moving with the total number of callers on each particular day, in order to ensure comparability over time.

3 Results

Figure 3 shows an overview of “mobile” and “less mobile” (defined as moving more, respective less, than 100km between their first and their last appearance in the dataset) callers. If we consider this a relevant proxy for migration, then

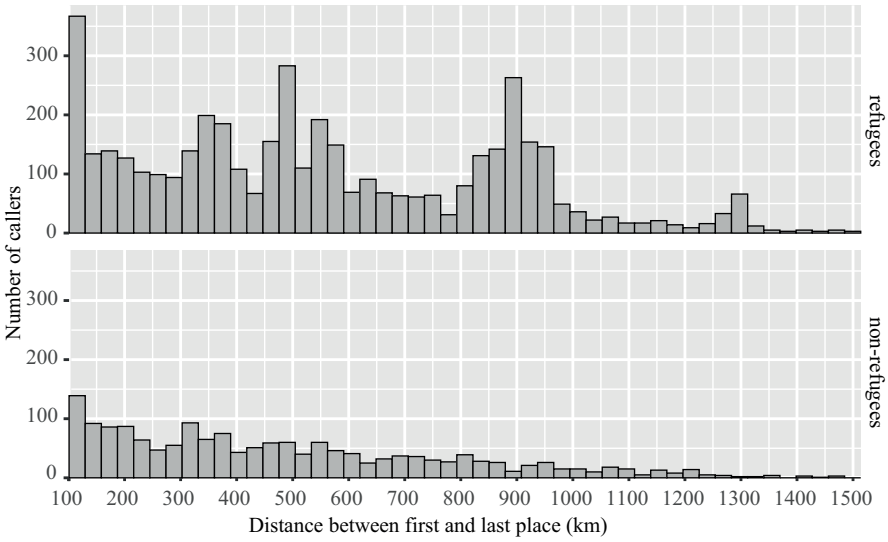


Fig. 2. Distances between first and last place of callers (1=refugees, 2=non-refugees, only dis-tances >100km shown)

about 14% of refugee callers in the whole of Turkey can be considered as onward migrating. In the border provinces to Syria (Hatay, Gaziantep, Sanliurfa, Kilis and Mardin) this ratio is about 17.7%.

If all movements of refugees of more than 100km are considered, regardless of the overall distance between their first and last district, a large number of bidirectional movements (between districts and back) becomes apparent, as Figure 4 shows. This indicates that refugees are highly mobile, even if they do not permanently (or at least over a longer period of time) change their place of residence.

This is also reflected in the high total cumulative distance (i.e. the sum of any movement, adding up also cyclical mobility, both migration and/or commuting) that many refugees travel over the time of their coverage in the dataset, compared to the (directed) total distance between the first and the last place when they are registered in the dataset (Figure 5).

When the total distance between the first and last districts of appearance in the dataset is taken into account (i.e. when "mobile persons" are defined as having a distance between the first and last place of more than 100km), and when the net flows of these "mobile persons" between places are calculated, a clearer pattern of origins and destinations emerges (see Figures 6 and 7). Generally, three larger migration systems seem to dominate: a) larger urban centres as destinations (notably Istanbul, Ankara, Adana, Antalya and others), b) a general direction from South to North and from East to West, and c) the movements within the Western Turkey-Syria border region, including movements back to border towns indicating return migration.

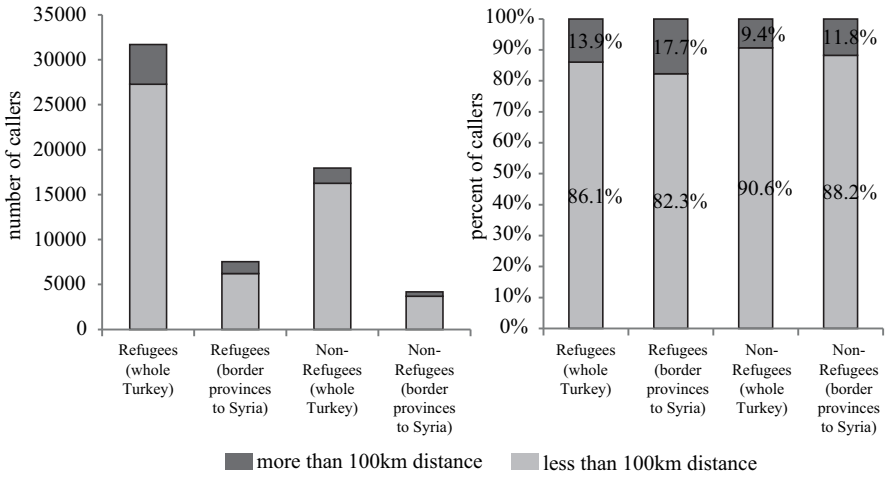


Fig. 3. "Mobile" and "less mobile" callers in the dataset (mobile = more than 100km between first and last district, less mobile = less than 100km), left: absolute numbers, right: percent; with "border provinces" we refer to Hatay, Gaziantep, Sanliurfa, Kilis and Mardin

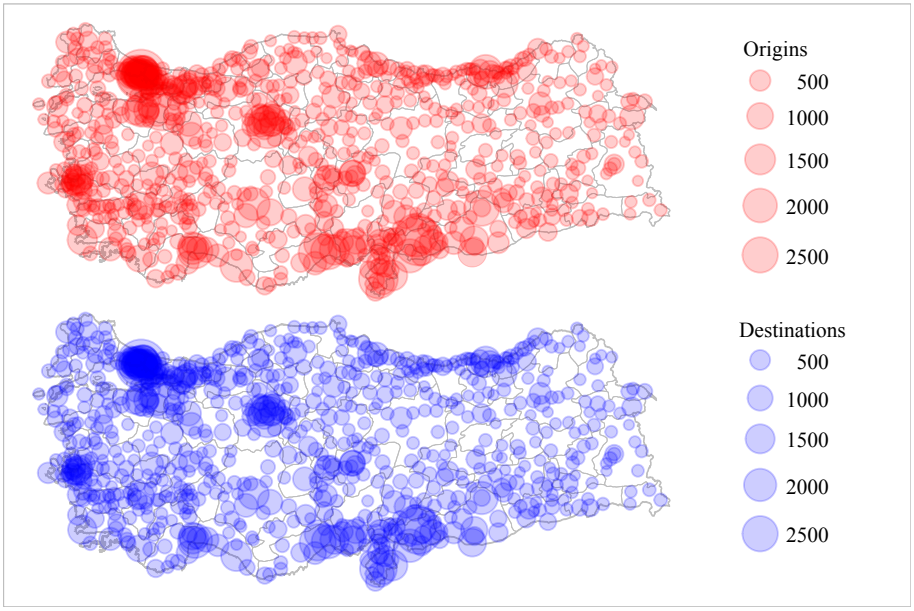


Fig. 4. Source and destinations of refugee movements (>100km) of all refugees over the total time of 2017. The similarity of the origin and destination maps (i.e. destinations are at the same time sources) indicate the high number of circular or back-and-forth movements of refugees.

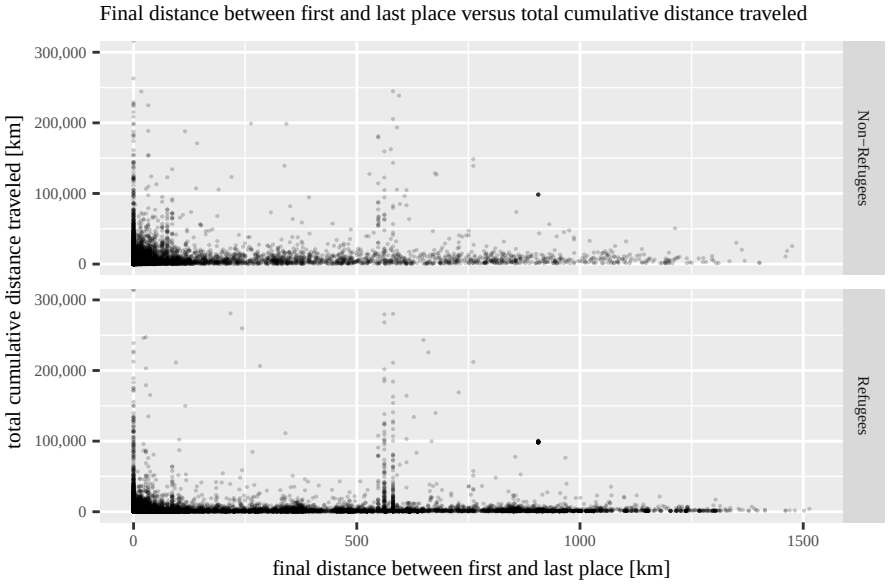


Fig. 5. Final distance between first and last place versus total cumulative distance traveled (1=refugees, 2=non-refugees), each dot equals one caller. The final distance refers to the distance between the first and the last district where a caller was registered in the dataset, the total cumulative distance includes all movements (one-time-migration, cyclical migration, commuting, visits, etc.).

Top 10 Provinces (in-migration)		Top 10 Provinces (out-migration)	
Istanbul	929	Mersin	262
Hatay	177	Mugla	175
Bursa	113	Izmir	166
Sanliurfa	64	Trabzon	157
Kayseri	28	Konya	143
Agri	22	Antalya	136
Kilis	14	Adana	109
Usak	7	Kocaeli	83
Karabük	3	K. Maras	81
Kinkkale	3	Sakarya	62

Table 1. Top 10 in-migration and out-migration provinces (number of refugees, according to their first and last appearance in the dataset)

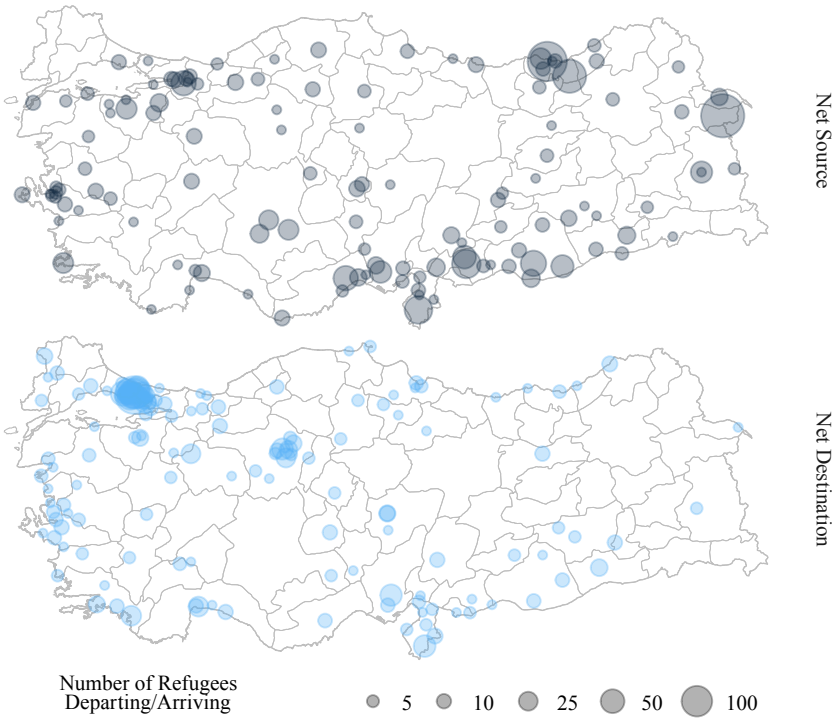


Fig. 6. Net sources and destinations of refugee movements (only refugee callers who moved more than 100km between their first and last district) over the total time of 2017 (by districts)

To determine the "migration intensity" over time, the days with movements of more than 100km of all refugees were identified. The number of refugees with such movements was summarized per day and normalized with the number of refugees appearing in the dataset on every day (see Figure 8). This data shows distinct patterns, for example the weekly mobility of the non-refugees, or the increased mobility during the two Eid festivals (both refugees, and especially non-refugees).

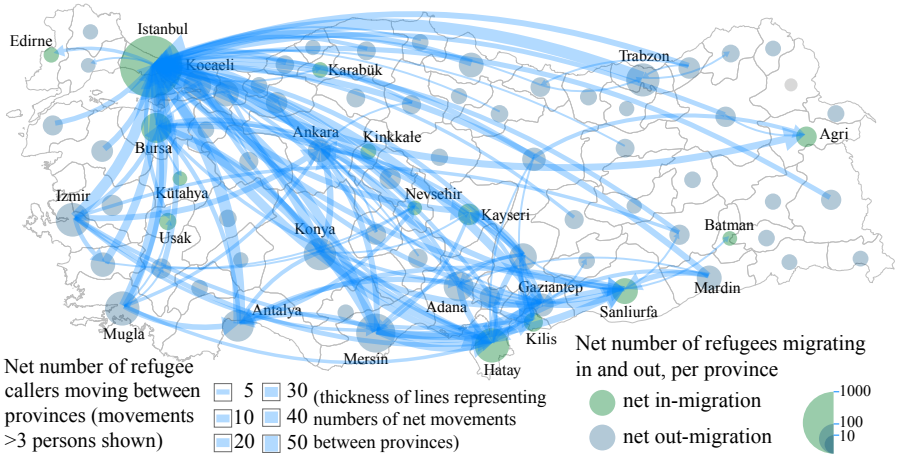


Fig. 7. Net flows (directed) of refugees, aggregated to flows between provinces of origin and destination (only flows of more than 3 individuals shown, for purposes of clarity)

4 Discussion

First it seems noticeable that the larger share of refugees do not *permanently move* more than 100km. This is in line with findings of existing research that many Syrian refugees stay either close to the border (in order to easily return when this is possible) or in areas where they have social networks and can find accommodation and jobs [4][5]. Also, refugees registered as temporary protection beneficiaries are required to stay in their assigned province and have to comply with reporting requirements [6].

At the same time, even refugees who cannot be considered as shifting their residence permanently do show a remarkably high degree of mobility. This is noteworthy especially insofar, as refugees would presumably have less financial resources that are necessary to travel and cover large distances, and it would be very interesting to investigate this issue further.

What becomes also obvious is that among those callers who *do move* more than 100km between their first and last place, refugees are *relatively more mobile*

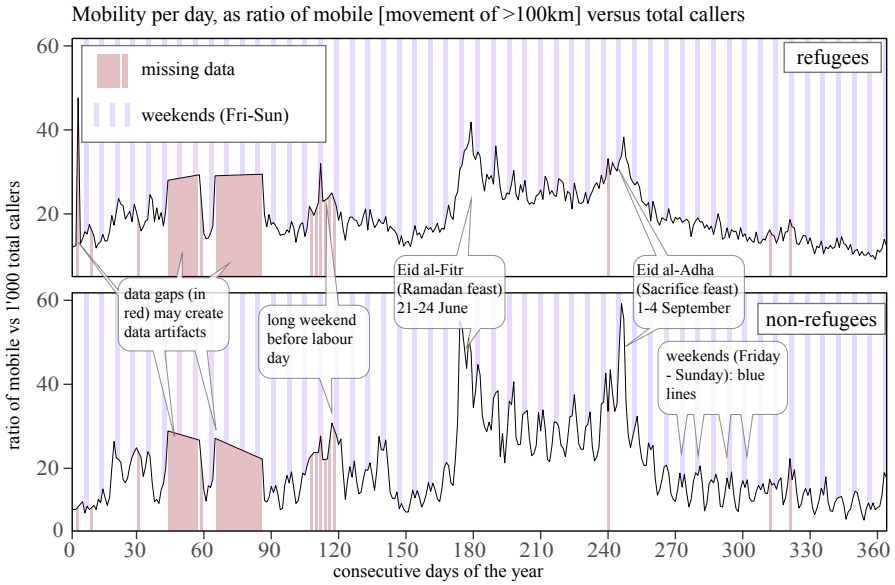


Fig. 8. Mobility per day, expressed as the ratio of mobile callers (moving more than 100km on that particular day) by total refugee / non-refugee callers on that day; note the data gaps in February/March.

than the comparable non-refugee population in the dataset (Fig 2 and Fig 3). While non-refugees cover almost twice the mileage than refugees (mostly through commuting and visits), refugees do move further, when the distance between their first and last place of appearance in the dataset is considered (Table 2).

Refugee Status	Total cumulative distance traveled	Absolute distance between first and last place
Refugees	6'365 km	588 km
Non-Refugees	12'195 km	475 km

Table 2. Total cumulative distance traveled, and absolute distance between first and last place, by refugee status

The dataset also shows clear patterns of larger distance movements to particular destinations, e.g. between places at the Syrian border, the Black Sea, and Istanbul, Ankara or Izmir.

Regarding mobility patterns in general, there seems to be a considerable back-and-forth movement, reflected in the differences of the total vs. the net flows (Fig 4 vs 6). The net flows (Fig 6 and 7) do reflect general movement patterns from the South to the North and from the East to the West, predominantly to Istanbul, Ankara and the Coast. There is also some movement to other urban areas, and

also bidirectional mobility to and from more rural places in central Turkey, indicating agricultural labour migration. There is also movement towards the Syrian border region, indicating possibly some return mobility.

The movement over time (Fig 8) clearly shows a growing mobility towards the two Eid festivals (Eid al-Fitr on 21st - 25th June and Eid al-Adha on 1st – 4th September 2017), which is then declining again. Mobility around the Eid festivals is much more pronounced among the non-refugee population, which might be explained by the better economic situation (allowing for leisure travel and family visits). Also visible is the stark contrast for weekend-mobility, which is strongly appearing in the non-refugee data and presumably related to commuting, and only very weak in the refugee-data. Although visible to some extent, there seems to be less seasonal mobility appearing in the data than we initially expected (with regard to temporal agricultural labour demand in central Turkey). This, and also other, smaller variations in mobility, especially for the refugees, need more explanation and require a more in-depth look into the data.

5 Conclusion and outlook

Methodologically: we can conclude that the analysis of CRD enables the unveiling of mobility and migration patterns to a hitherto unprecedented level of detail, both temporally as well as spatially. However, the complexity of human mobility requires also more advanced and in-depth analyses of mobility patters, including approaches such as most frequented locations or full temporal origin-destination-matrices [7][8]. Due to resource restrictions, we have applied a very simple approach of delineating migration, by assuming that those individuals can be regarded as migrants (permanently or semi-permanently shifting their residence), whose first and last districts of appearance in the dataset are located more than 100km apart from each other.

Knowledge on refugees' mobility: general patterns of mobility (South to North, East to West, to urban centers and to the coast, to and from agricultural areas) are clearly reflected in the data; however it becomes also evident that mobility is more complex, and that boundaries between commuting, visiting, temporary, seasonal and permanent migration might be more fluid than often conceptualized in migration research. Cultural motifs for movement (i.e. the Eid festivals) are more important than initially expected, and seasonal mobility due to agricultural labour demand seems to be less important than initially expected. However, the underlying motivations for mobility can only be inferred indirectly—either through additional data, or through making assumptions (i.e. that more intense connectivity as represented in dataset 1 might be a reason for mobility between persons).

Implications for refugees' wellbeing: although the majority of refugee callers is less mobile or at least appears not to permanently change location between the first and last appearance in the dataset, still a significant share of refugees (about 14%) does so. Within the existing registration system this implies limited or difficult, or even suspended, access to social services such as

healthcare, education, housing, and so on for these mobile refugees. Thus, it could be an important contribution to these refugees' wellbeing if changing the place of registration would be possible (more) easily and quickly — reflecting the mobile reality of their lives and livelihoods.

Open questions: As the analysis work was done with limited time resources, there still remains some work to do in a series of next steps:

- More in-depth analysis of general mobility patterns: it would be necessary to better differentiate between permanent and temporary migration (as a permanent or temporary shift of residence to another place), visiting mobility (as short-term change of place) and labour mobility (as longer term but onwards and finally returning mobility, e.g. following the changing places of agricultural labour demand), as these different kinds of mobility are resulting in different needs of the refugees in terms of support and service provision;
- A spatially more fine grained analysis of mobility would be beneficial, differentiating between mobility patterns from and to rural areas, smaller towns and larger urban centers, as this will also allow to link mobility with structural drivers (labour markets, existing social networks to refugees in the destination places, etc.);
- A spatially explicit analysis of mobility over time would help to disentangle different types of mobility (on the basis of the differentiation between the four mobility patterns mentioned above), and would allow for associating these types of movements with the spatio-temporality of seasonality, of weather and climate phenomena (precipitation, land surface temperature, drought indices, cf. [9] and [10]), conflict and political events (we have started working on [11] and [12]), and the important role of translocal connectivity of refugees to others (derived from the communication dataset DS1 provided through the D4R challenge).

On a more general level, it seems important to us to remark two issues: first, the inference of mobility patterns from CDR obviously poses challenges of data protection and privacy. In the context of this study, the organizers of the D4R challenge put a special emphasis on these issues, including anonymizing the datasets, a review process safeguarding the interests of refugees and setting up clear contractual agreements for the research teams using the data. The analysis of such data in other contexts would require similar standards. However, and secondly, the analysis of CDR can yield a wealth of fine-grained information on human mobility that is almost impossible to achieve with traditional means (e.g. surveys, registration or census data). Thus, given the safeguarding of privacy and data protection, researchers' access to CDR and similar data could benefit both science and development practice.

6 Acknowledgement

We would like to express our sincere gratitude to the D4R challenge organizers and Türk Telekom for setting up the challenge and for providing us with the datasets.

7 References

1. UNHCR: Global Trends – Forced Displacement in 2017, http://www.unhcr.org/5b27be547.pdf, last accessed 10.09.2018
2. D4R Homepage: http://d4r.turktelekom.com.tr/, last accessed 15.09.2018
3. Salah, A.A., Pentland, A., Lepri, B., Letouzé, E., Vinck, P., de Montjoye, Y.A., Dong, X. and Dağdelen, Ö., 2018. Data for Refugees: The D4R Challenge on Mobility of Syrian Refugees in Turkey. arXiv preprint arXiv:1807.00523.
4. Istanbul Policy Center (Ed) 2015, Urban Refugees: The Experiences of Syrians in Istanbul. Istanbul Policy Center. Instanbul.
5. Tuzcu, N., 2014, Syrian Urban Refugees in Turkey: Spatial & Social Segregation. Pub-lished by Displacement Research & Action Network, MIT Department of Urban Studies and Planning. Online available http://mitdisplacement.org/new-page-43/, last accessed on 05.09.2018.
6. AIDA - Asylum Information Database: Freedom of Movement, Turkey, http://www.asylumineurope.org/ movement-1, last ac-cessed 12.09.2018
7. González, M.A., Hidalgo, C.A., Barabási, A.-L., 2008, Understanding individual human mobility patterns, Nature Vol 453(5), pp. 779-782, doi: 10.1038/nature06958
8. Barbosa, H., Barthelemy, M., Ghoshal, G., James, C.R., Lenormand, M., Louail, T., Menezes, R., Ramasco, J.J., Simini, F., Tomasini, M., 2018, Human mobility: Models and applications, Physics Reports 734 (2018) 1–74, doi:10.1016/j.physrep.2018.01.001
9. PSD Gridded Climate Datasets, https://www.esrl.noaa.gov/psd/data/gridded/, accessed 12.09.2018
10. GHCN Gridded Products – Temperature, Precipitation and Drought, https://www.ncdc.noaa.gov/temp and-precip/ghcn-gridded-products/, last accessed 12.09.2018
11. UCDP Georeferenced Event Dataset (GED) Global version 18.1 (2017), http://ucdp.uu.se/downloads/, last accessed 17.08.2018
12. ACLED – Armed Conflict Location and Event Data Project, https://www.acleddata.com/, last accessed 17.08.2018

Measuring fine-grained multidimensional integration using mobile phone metadata: the case of Syrian refugees in Turkey

Michiel A. Bakker[1], Daoud A. Piracha[1], Patricia J. Lu[1], Keis Bejgo[1], Mohsen Bahrami[1,2], Yan Leng[1], Jose Balsa-Barreiro[1], Julie Ricard[3], Alfredo J. Morales[1], Vivek K. Singh[4], Burcin Bozkaya[2], Selim Balcisoy[2], and Alex 'Sandy' Pentland[1]

[1] MIT Media Lab, Cambridge, Massachusetts, USA {bakker,pentland}@mit.edu
[2] Sabanci University, Istanbul, Turkey
[3] DataPop Alliance, New York, USA
[4] Rutgers University, New Jersey, USA

Abstract. The current Syrian civil war has led to a mass migration of Syrian refugees into Turkey. As the Syrian conflict has intensified and lengthened, many refugees have faced challenges integrating into their host societies. Here we introduce and evaluate different measures extracted from mobile phone metadata to study integration of refugees along three dimensions: (1) social integration (2) spatial integration and (3) economic integration through signatures of employment activity. We use these measures to compare integration across different regions in Turkey and find striking differences both in the distributions of these dimensions and the relations between them. Finally, leveraging the results from two general elections in Turkey in 2015 and 2018, we confirm earlier findings concerning the impact of refugee presence on voting behavior, and demonstrate that we can better explain voting behavior by incorporating integration metrics.

Keywords: Local integration · Employment

1 Introduction

The Syrian civil war that began in 2011 has had an enormous human cost and impact on the region. The United Nations Refugee Agency (UNHCR) has estimated that over 12 million people have fled their homes since the war started. Around 6.6 million Syrians are internally displaced while 5.6 million people fled Syria, seeking safety in Turkey, Lebanon, Jordan and beyond [1]. The vast majority of internationally hosted Syrian refugees live in urban areas, while around 8% are accommodated in refugee camps. Meanwhile, the crisis is in its seventh year with no clear end in sight.

Despite reports of some tens of thousands of Syrians sporadically returning to safer parts of their home country, it remains unclear when and if the majority

of the refugees will return to Syria [13, 3]. Settlements in host countries have transformed from temporary to permanent, refugees have established social ties with their host communities and many have found jobs, predominantly in the informal sector [21].

This work studies local integration of Syrian refugees in Turkey quantitatively using call data records (CDR) from mobile phones. While aspects like education and establishment of appropriate legal processes [22, 14] are instrumental to local integration, we focus on social, spatial and economic integration. We expect our work to shed light on new ways to measure integration and feed the discussion on which types of interventions could and should be adopted.

1.1 Local integration

Historically, local integration has been a guiding principle of refugee programs. According to the 1951 UN Refugee Convention, restoring refugee dignity and ensuring the provision of human rights includes an approach that would facilitate integration into the host society [2]. This Convention uses the word 'assimilation' which implies the disappearance of differences between refugees and the local population. Most authors, however, emphasize the importance of maintaining individual identities with the purpose of *integrating* people, instead of *assimilating* them to the national culture. According to a more recent UNHCR report, *local integration* is commonly referred to as one of the three 'durable solutions' for refugees, in addition to the voluntary repatriation to the home country and the resettlement in a third country [14].

We study local integration as a multidimensional process of developing social and economic ties with the host country and community, and becoming increasingly self-reliant. The first dimension is *social integration*, i.e. the formation of social ties between refugees and the host country. A strong social network has proven to be instrumental in finding housing, employment and healthcare in a longitudinal study of Syrian refugees in Canada [20], a longitudinal study of immigrants of refugees in the UK [11, 12] and a large study across European countries [24]. Within social networks among refugees and immigrants, most prior work makes a distinction between *bonding capital*, encompassing the interaction within the refugee group, and *bridging capital*, describing the interactions between refugees and the local citizens. While Milgram in his classic study [18] has argued the importance of 'weak ties' (connected with bridging capital) for finding employment, recent immigration-based studies from the UK and Canada show that bonding capital is more important for finding employment and housing, especially in the short term. More generally, the literature has reported both bridging and bonding capital to be relevant for finding jobs and employment, and that both provide immigrants with access to different and unique information and opportunities. Our work focuses on the bridging capital and the social integration with local citizens. While prior work on the subject relies predominantly on self-reported survey data, we measure social integration as the relative

refugee to local call volume, which has shown to be more informative [29].

The second dimension, *spatial integration*, has been well studied in the context of integration of immigrant and minority populations using census data [26, 31] but also more recently using Twitter data in cities globally [23]. *Spatial integration* is also often called *urban* or *residential integration*. Prior work has shown the importance of spatial integration and its impact on academic performance [8] and health [35], among other factors. We adapt methods for the study of spatial segregation in census data from prior work and use them to study spatial integration using the geolocations of cell phone towers in the CDRs. Additionally, we introduce a new method to study spatial integration that captures the likelihood of a refugee encountering a non-refugee, not only in their residential area but dynamically throughout the day.

The last dimension along which we study integration is *economic integration* or employment. Most previous studies focus on the impact of immigrants and refugees on the labor market of the host country [16, 9]. However, in this study we will focus on the impact of employment on other dimensions such as social and spatial integration, as well as the effects of economic integration on the local environment of the refugee. Our estimation of employment is inspired by prior work on human behavior during crises using CDRs in combination with complementary datasets [33, 4, 34].

1.2 Syrian refugee immigration in Turkey

Officially, the Syrian refugees in Turkey are recognized as guests in Turkey rather than as asylum seekers [28]. Unlike the *refugee status*, the *guest status* entails that a refugee can technically be relocated at any time without notice. To limit this uncertainty for the refugees, the country enacted a temporary protection status that ensures no forced exits. This is important for the analysis because refugees are not counted in official statistics on internal migration and housing.

Moreover official statistics lack refugee employment data. Although formal employment mechanisms for Syrian refugees in Turkey have existed since 2016, the number of officially employed refugees has not increased since the law's passage, and the number of informal workers has been estimated to exceed the formal ones by a factor of 50 [17]. Within the informal sector refugees are employed mainly in low-skilled jobs such as construction and the service sectors, since the language barrier is likely to limit their access to high-skilled jobs [15].

1.3 Outline

The remainder of the paper is organized as follows. Section 2 provides a brief description of the datasets. Section 3 introduces our measures of integration. Section 4 discusses the measures of integration, how they affect each other and

how they affect the host society. Section 5 concludes and Section 6 provides recommendations and suggestions for future work.

2 Data

For this study, we use anonymized mobile phone metadata, known as call data records (CDRs), and official census data.

2.1 Call data records

We use two different types of CDRs provided by a single telecommunications service provider [30] for the whole year 2017. The dataset is collected from 807K Turkish customers and 185K 'refugees' - customers with a 'temporarily protected foreign individual' status. Though most of these customers are in fact refugees, these also include some migrants, asylum seekers and even other foreigners. Moreover, the users are not uniformly sampled across the population. About 45% of refugees and locals in our dataset are in Istanbul, while only 18.5% live there officially.

Fine-grained mobility and communication The first dataset tracks the calls or text messages of a randomly chosen subset of users during a two-week interval. There are in total 26 intervals throughout the year, each of which is partitioned twice - first into SMS and voice calls and second into incoming and outgoing calls. New random identifiers are created for each interval and each partition so that no single user can be tracked throughout the year or across partitions. On average, each partitioned dataset contains records of 61K unique user IDs while each user on average has 31 records during the two-week period.

There are four fields for each record. The random *user ID* generated for the two-week period identifies the user as a local or refugee. The *timestamp* specifies the specific day and hour. The *user2 ID* denotes whether the other person is a refugee, local or unknown. Note that this second person is sometimes the caller and other times the callee, depending on whether it is an outgoing or incoming partition. The *site ID* is a unique ID for each cell tower for which we know the exact latitude and longitude.

Antenna tower traffic The second dataset includes all site-to-site calls and SMS traffic between cell towers on an hourly basis for the year 2017. Calls with other operators other than the operator providing the data only have information from one side.

There are seven fields for each record: the timestamp denoting the day and hour, the outgoing and incoming cell tower ID, the number of calls during the day and hour, the number of calls with a refugee-labeled user on either side, and the total duration of the calls.

2.2 Complementary datasets

We combine the CDRs with two complementary datasets. The first is from the Turkish Ministry of Culture and Tourism and contains the 2017 number of arriving and departing foreign visitors by district [5]. The second is the votes per polling station during 2015 and 2018 general elections from the Ballot Result Sharing System from the Turkish Supreme Electoral Council [6].

3 Measuring integration

We have developed and adopted a number of methods that can be used to locally probe all three dimensions of integration using the datasets described in Section 2 but also more generally with other CDR datasets. Each measure can be used to estimate the level of integration for a single user, but also on the neighborhood, district and province level.

3.1 Social integration

Social integration is measured using the fine-grained CDR dataset (see Section 2). For each minority user, social integration is defined as the number of calls that are made to majority users relative to the total number of calls made to all users.

$$\textit{Social integration} = \frac{Calls_{minority \to majority}}{Calls_{minority \to majority} + Calls_{minority \to minority}} \quad (1)$$

When looking at aggregate estimates for multiple users in a single region, we measure social integration for the region as the relative number of calls that are made to majority users by minority users while being in the region. In our dataset, remarkably, 91% of all calls made by refugees are made to non-refugees, much higher than one would expect from prior studies on social networks of refugees [20]. We hypothesize that this is either because of noise in the labels, which is known to be present in the CDR data, or because the users that receive the call are with a different operator and thus the refugee status is not known.

3.2 Spatial integration

We focus on two measures of spatial integration: (a) *evenness*, which involves the differential distribution of the minority population and (b) *exposure*, measuring the potential for contact between the minority and majority population. For a comprehensive survey of dimensions and measures of spatial integration we refer to [25].

[5] `http://www.kultur.gov.tr/EN,153018/number-of-arriving-departing-visitors-foreigners-and-ci-.html`

[6] `https://sonuc.ysk.gov.tr/`

Evenness - Gini coefficient The most commonly used measure of integration is evenness, measured either by dissimilarity or the Gini coefficient. In our study both gave highly similar results, and we report only the inverse Gini coefficient defined as

$$\textit{Spatial integration Gini} = 1 - \frac{\sum_i \sum_j t_i t_j \left|(p_i - p_j)\right|}{2T^2 P(1-P)} \tag{2}$$

where t_i is the population in area i, p_i is the proportion of minority group members in area i, T is the total population across all areas and P is the proportion of minority group members across all areas. The $1-$ ensures that all our measures are defined in the same way, varying between 0 and 1, with, 1 indicating maximum integration and 0 indicating maximum segregation.

Exposure - Encounter index Exposure measures the possibility of interaction between the minority and majority group members. The most commonly used measure of integration is *interaction*, reflecting the probability that a minority person shares the area with a majority person. This measure is static and based on the home location of an individual. We leverage the richness of the CDR dataset and compute an encounter index for each refugee defined as

$$\textit{encounter index} = \frac{1}{N} \sum_j p_{i_j, t_j} \tag{3}$$

where N is the total number of phone calls j a minority person made and p_{i_j,t_j} is the proportion of calls that were made by majority persons in the area i at the time t the refugee was making call j. Especially for our work this measure is powerful since we know from the fine-grained dataset where and when the refugee made each call, while at the same time we can access the exact value of p_{i_j,t_j} using the antenna traffic dataset which describes all call traffic at time t for tower i in aggregate without sampling. Our encounter measure thus not only measures the probability of encountering a local near home but also measures the probability of encountering a local throughout the day.

3.3 Economic integration

We measure economic integration by computing the regularity of individual commuting patterns. For each individual in the fine-grained CDR dataset, we compute a week and weekend mobility similarity matrix S. To compute each element $S_{i,j}$ in the matrix S, we calculate the cosine similarity between the cell tower IDs for all calls made during hour i and the cell tower IDs for all calls made during hour j. If an individual is consistently at the same cell tower at hour i and j, $S_{i,j} = 1$. Otherwise, if the user has erratic calling patterns during i and j, $S_{i,j} = 0$.

The 'ideal' commute mobility pattern describes an individual who is always at the same home location during the evening, and at the same work location,

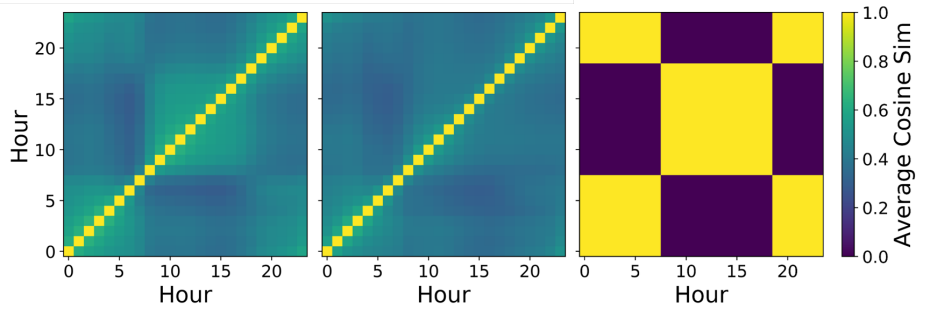


Fig. 1: Cosine similarity matrices to compute the employment score for week (left) and weekend (middle). On the right we see the ideal matrix for someone with employment score 1. In the weekday matrix on the left, we see how the metric reveals employment. Users, on average, are at the same location during office hours and at the same location during the night resulting in two high similarity yellow/green blocks around the diagonal. At the same time, the off-diagonal low similarity regions in blue show clearly that these office and house locations are two different ones.

which is different from the home location, during office hours. In the rightmost plot in Fig 1, we see the similarity matrix for an individual with this ideal pattern. To measure the employment score we now use the *Frobenius norm* to measure how similar a mobility pattern is to the ideal mobility pattern

$$employment\ score = 1 - \|S^* - S\| = 1 - \sqrt{\sum_{i,j}(S^*_{i,j} - S_{i,j})} \tag{4}$$

where S^* is the similarity matrix for the ideal pattern. All elements of the matrix that have no value as well as the diagonal elements are not taken into account when computing the score. Additionally, to improve the accuracy of the metric, we exclude individuals who do not have at least four off diagonal elements with finite values, two during the evening and two during the day. Although our work lacks ground truth data to evaluate the metric, we observe expected and intuitive behavior in Fig 1. For the average week similarity matrix on the left, we observe a much stronger similarity to the ideal commute matrix than we do for the average weekend similarity matrix in the weekend.

In contrast to the often used heuristics for employment, the employment score provides a richer and continuous proxy for employment. It does, however, fail to take into account the locations people traverse during the actual commute as well as important third places other than home or work locations. Additionally, it falsely classifies individuals as unemployed when they work very close to home or have a more mobile job for which they constantly move around the city. Finally, it falsely classifies individuals as employed when they are not employed but spend their time in, for example, the same park during office hours.

4 Results

4.1 Social integration

In studying local integration of refugees in Turkey, we focus on three regions of interest (see Fig 2). The first region is Istanbul, Turkey's largest city and host of 30% of all refugees in our dataset. The second region is Southeastern Anatolia, an official geographical region that we analyze due to its proximity to Syria and its large percentage of refugees. To ensure that the analysis is not influenced by behavior in refugee camps, records registered at cell phone towers in the vicinity of refugee camps are excluded from the data using a list of refugee camps compiled by the U.S Department of State, Humanitarian Information Unit [7]. The third region comprises provinces that have the highest number of tourists per capita according to the official 2017 statistics released by the Turkish Ministry of Culture and Tourism [8].

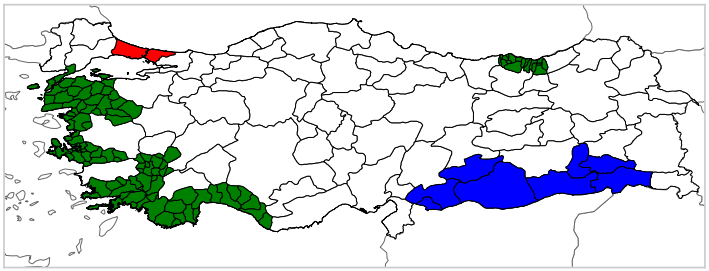


Fig. 2: Three regions of interest. The region of Istanbul is colored red, tourist provinces green and provinces bordering Syria blue.

Aggregated over all individual refugees throughout the year in the fine-grained dataset, Fig 3 shows the difference in social and spatial integration in the three regions of interest. In terms of social integration, measured by the number of calls to locals as a percentage of total calls, all three regions have a similar distribution. They are not similar, however, when comparing the spatial

[7] https://data.humdata.org/dataset/syria-refugee-sites

[8] http://www.kultur.gov.tr/EN,153018/number-of-arriving-departing-visitors-foreigners-and-ci-.html

integration distributions. For the encounter index, a proxy for the probability of encountering a local, we measure similar behavior for Istanbul and tourist areas, but observe a bi-modal distribution in the Southeast. We hypothesize that this is because refugees either move to more urban areas with many locals or they are living isolated in remote areas where they mostly encounter other refugees. Also when measuring spatial integration through the Gini index with respect to their 50 nearest neighbors, we observe clear differences in the distributions. In tourist areas, refugees are spatially most segregated, while the largest integration is observed in Istanbul. We hypothesize that the main reason for refugees to travel to tourist areas is for (seasonal) employment in the services sector where refugees live only temporarily together grouped with mostly other refugees. In Istanbul, however, refugees live more permanently and, over time, blend in more with non-refugees. Aggregated integration measures for all provinces in Turkey and all districts in Istanbul can be found in Appendix A.

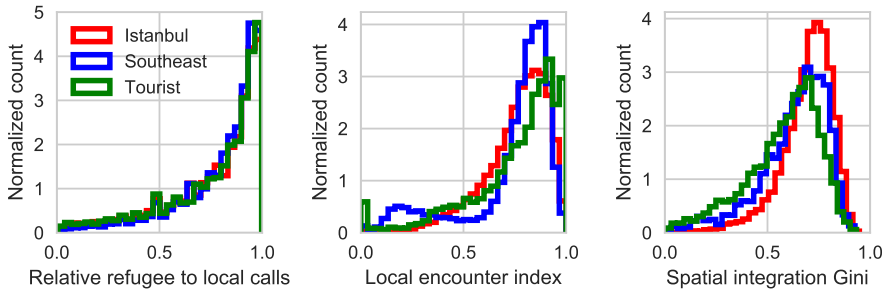


Fig. 3: Refugee social and spatial integration for the three regions of interest specified in Fig 2. In the left plot we see relative refugee to local calls, in the middle the encounter index and on the right the Gini index spatial integration.

From the literature, one would expect a high correlation between spatial and social integration. If a refugee has a high probability of encountering locals, they are also more likely to communicate with locals. In Fig 4, we observe the expected behavior for refugees in Istanbul. There is a strong positive correlation between the encounter index decile and the mean of the relative number of calls to locals. Surprisingly, however, we do not observe the same behavior in the Southeast and the tourist areas. In tourist areas this could be because most refugees come for jobs in the tourism industry and have similar jobs, backgrounds and social network structures independent of whether they spend time in areas with a large number of refugees.

The third dimension of local integration, *economic integration*, is measured using employment score, a measure for the regularity of the refugee's commute

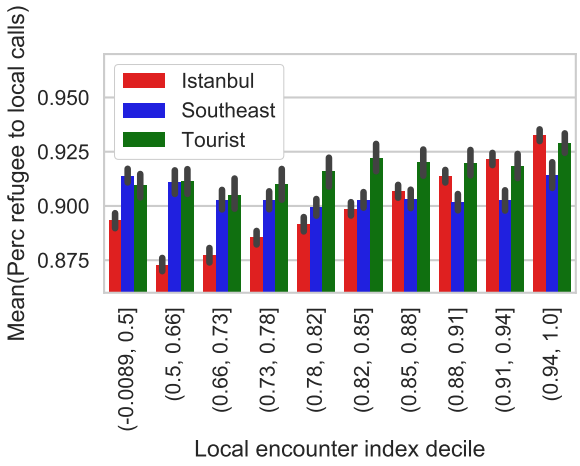


Fig. 4: Relation between social integration and spatial integration. The refugees in each region of interest are binned in deciles by the encounter index, while we compute the mean of the percentage of refugee to local calls for each bin.

patterns as introduced in Section 3. The employment score therefore measures the likelihood of being employed. From Fig 5, we observe a very similar distribution for all three regions of interest. Nonetheless, interesting differences between regions emerge when comparing the relation between employment score and the other dimensions of local integration, social and spatial integration, in Fig 6. To reveal these differences, we bin the employment score by decile and compute the per-bin mean of the social and spatial integration measures. When comparing employment to social integration in the left figure, we observe generally a positive trend for Istanbul but, similar to Fig 4, not for the southeast and the tourist provinces, meaning that either employment has only little effect on social interaction with locals or that bridging capital plays no crucial role when looking for a job. For spatial integration, we measure a clear positive correlation between employment score and the encounter index for the southeastern provinces, while for Istanbul and the tourist provinces the correlation even seems slightly negative.

4.2 Effects of integration on voting behavior

As a test case to show the effectiveness of these integration measures, we analyze the effects of local integration on voting. There are four national parties represented in the Turkey's Parliament. The leading party has the most explicit policy towards refugees, emphasizing Turkey's duty to support refugees.

Previous literature indicates that an influx of refugees influences election outcomes. In Italy and Germany, migration inflow resulted in additional votes for parties with a conservative migration agenda [7, 27], whereas in Austria the

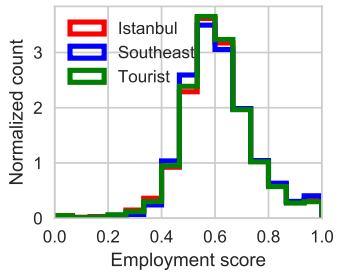


Fig. 5: Distribution of the employment score as defined in Section 3 for the three regions of interest.

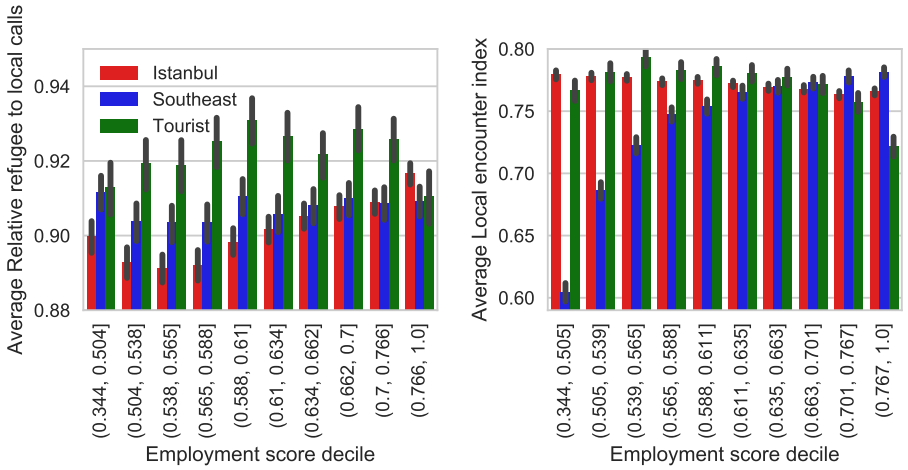


Fig. 6: Relation between spatial, social and economic integration for the three regions of interest. The employment score is binned in deciles while we compute the mean of the fraction of calls to locals and the local encounter index for the left and right figures, respectively.

results are mixed. One study shows that immigration strengthens support for a far-right party while another study argues that refugee inflows weaken the same far-right party [32, 19]. In Turkey, there is one previous study that shows a small but insignificant impact of refugee influx on election outcomes [5].

Whereas previous studies focus on measuring voting behavior against refugee influx, we extend them by measuring it against our three dimensions of integration: spatial, social, and economic. Specifically, we investigate outcomes within Istanbul, measuring results at the neighborhood (*mahalle*) level. For each mahalle we define the refugee percentage as the percentage of home locations that belong to refugees computed with the fine-grained dataset. For the social integration and employment score, we simply compute the averages for all the refugees with a home location in that neighborhood. For the Gini spatial integration, we use the cell tower traffic data instead of the fine-grained data, because it includes all data in 2017 instead of a single sample and is therefore more accurate. The dependent variable in the regression in Table 1 is the change in the percentage of votes for the leading party between the general election either in June 2015 or November 2015 and the latest election in June 2018. Ideally, we would use the differences in integration variables between 2015 and 2018 as independent variables, but since we only have data for 2017 we use the average over that year.

Table 1: OLS estimates of the impact of integration and the percentage of refugees on votes for the leading party. Significance levels are indicated by *< .1, **< .05 , ***< .01.

Variable	Change in %votes for the leading party	
	May 2018 - June 2015	May 2018 - Nov 2015
Refugee Percentage	0.12***	0.05***
	(0.02)	(0.02)
Social integration refugee to Local Calls	0.10**	0.01
	(0.05)	(0.04)
Spatial integration Gini	0.03**	-0.02*
	(0.01)	(0.01)
Median Refugee Employment Score	-0.02***	-0.01***
	(0.01)	(0.00)
Const.	-0.05**	0.00
	(0.02)	(0.02)
R-squared	0.147	0.159
Adj. R-squared	0.133	0.144
N	294	294

First, results indicate a highly significant positive relationship between the percentage of refugees in a neighborhood and the change in votes for the leading

party in Table 1. This could be explained by refugees moving to more conservative areas where an increasing amount of people vote for the socially conservative party that is currently leading the country. Second, we observe the positive correlation of social integration with an increased number of votes for the leading party, which is significant for the June election. Intuitively, the positive effect makes sense: better integration between refugees and locals leads to a more positive attitude towards refugees and thus to a higher number of votes for the leading party. Causally, however, the effect could also be understood differently that neighborhoods that are more positive towards refugees foster better integration. A similar effect and intuitive explanation can be sought for the Gini coefficient measuring spatial integration. High integration leads to a more positive attitude towards refugees, although now the positive effect becomes negative when the difference with the November election is used as the dependent variable. Finally, we measure employment and observe that increases in refugee employment lead to a decrease in support for the pro-refugee party. This could be due to locals feeling more threatened in the job market in these areas, as indicated by previous literature for Turkey and other countries. [19, 10, 6].

5 Conclusions

The unprecedented number of refugees that have entered Turkey since the start of the Syrian Civil War has provided a unique opportunity to study integration of refugees and the effect of refugee integration on the host society. Our study extends the existing literature in two ways.

First, when measuring social, spatial and economic integration, previous work has almost exclusively relied on survey data and interviews. Although both allow one to answer detailed and in-depth questions, they have limited scale and duration of the measurement and inevitably introduce biases [29]. Our study introduces a set of measures that can be applied to study three dimensions of integration directly and at scale through mobile phone metadata. We show how these methods allow us to uncover differences both in the measure of integration along each dimension as well as the interaction between dimensions. In Turkey, the distributions of Gini coefficients for spatial integration show us that, on average, refugees in Istanbul live in more integrated neighborhoods than they do in Southeastern Anatolia and that there is a much stronger correlation between spatial and social integration and between employment and social integration there. Meanwhile, in Southeastern Anatolia, there is a much higher positive correlation between refugee likelihood of working and refugee likelihood of encountering locals, which could speak as to how refugees are finding jobs.

Second, we show that integration of refugees impacts political outcomes and that our set of measures can be used to explain differences in election results even though the time between elections does not overlap our window of measurement.

There is a substantial body of literature devoted to the impact of refugees and immigrants on their host societies, predominantly focused around the impact on economic opportunities for locals. Here we show not only that the presence of refugees and voting behavior are interrelated but also that the extent to which these refugees are integrated along the three dimensions of integration is related to political opinions. While an increase in social integration is positively correlated with an increased number of votes for the pro-refugee party, the opposite is true for economic integration. The intuition is that through increased social integration leads to a more favorable situation for both locals and refugees, other studies have shown that refugee employment could have a detrimental effect on the employment opportunities for locals, especially in the informal sector [6].

6 Recommendations and Future Work

Policies and interventions, facilitating settlement and full participation in the host society, should address issues that relate to local integration by promoting social, spatial and economic integration. Traditional methods like census and employment statistics are typically not collected for refugee populations and fail to accurately capture some of the important dynamics of local integration. Here we report that such measures of integration can have a significant impact on the host society. Therefore, we recommend using these mobile phone based methods to both inform new interventions and to measure the effectiveness and possible side effects of current interventions.

To further extend this work, a longitudinal fine-grained dataset that tracks a consistent large group of users with cell tower precision over a much longer period of time than two weeks could enable researchers to address some of the fundamental questions around integration. One such question regards understanding the causal structure of integration and how it could be used to allow for more optimal allocation of resources in fostering integration. Do refugees build bridging capital through their job or, instead, does their job help them to build bridging capital? How does the causal structure differ in other regions of Turkey and beyond Turkey? Answering such questions could pave the way for important policy decisions and interventions. For instance, if one type of integration (e.g. social) causes other types of integration to follow naturally, then it would be effective to focus on that type of integration.

Finally, further studying the impact of integration on the host society could lead to new insights on how to create a more mutually beneficial situation for both refugees and locals. One shortcoming of this work is the lack of outcome data during the same year that the CDRs describe. Datasets of, for example, employment and spending behavior of Turkish locals during 2017, could help us better understand how strong the effect of integration is on different indicators and how these effects differ over time and across the country.

On the whole, this work marks one of the first systematic attempts at employing fine grained mobility to understand refugee integration at scale. The obtained results, and more importantly, the defined computational methods can be applied to multiple diverse refugee populations across the globe, and help improve the conditions for both the refugee and host population.

7 Acknowledgments

The authors thank MIT and the MIT Media Lab for their support, and Shada Alsalamah for her contributions to the proposal for this project.

References

1. UNHCR operations portal - refugee situations. `https://data2.unhcr.org/en/situations/syria`, accessed: 2018-08-30
2. Article 34, convention relating to the status of refugees (1951), 189 UNTS 150 (entered into force 22 April 1954).
3. Syria war: Thousands return home in south-west. BBC (2018)
4. Almaatouq, A., Prieto-Castrillo, F., Pentland, A.: Mobile communication signatures of unemployment. In: International conference on social informatics. pp. 407–418. Springer (2016)
5. Altindag, O., Kaushal, N.: Do refugees impact voting behavior in the host country? evidence from syrian refugee inflows in turkey (2017)
6. Balkan, B., Tumen, S.: Immigration and prices: quasi-experimental evidence from syrian refugees in turkey. Journal of Population Economics **29**(3), 657–686 (2016)
7. Barone, G., D'Ignazio, A., de Blasio, G., Naticchioni, P.: Mr. rossi, mr. hu and politics. the role of immigration in shaping natives' voting behavior. Journal of Public Economics **136**, 1–13 (2016)
8. Bennett, P.R.: The relationship between neighborhood racial concentration and verbal ability: An investigation using the institutional resources model. Social science research **40**(4), 1124–1141 (2011)
9. Borjas, G.: Immigration and the american worker. Center for Immigration Studies, Washington, DC (2013)
10. Ceritoglu, E., Yunculer, H.B.G., Torun, H., Tumen, S.: The impact of syrian refugees on natives labor market outcomes in turkey: evidence from a quasi-experimental design. IZA Journal of Labor Policy **6**(1), 5 (2017)
11. Cheung, S.Y., Phillimore, J.: Social networks, social capital and refugee integration. Report for Nuffield Foundation: London (2013)
12. Cheung, S.Y., Phillimore, J.: Refugees, social capital, and labour market integration in the uk. Sociology **48**(3), 518–536 (2014)
13. Collard, R.: Syrian refugees return from lebanon to an uncertain future. Financial Times (2018)
14. Crisp, J.: The local integration and local settlement of refugees: a conceptual and historical analysis. UNHCR, Evaluation and Policy Analysis Unit (2004)
15. Dinçer, O.B., Federici, V., Ferris, E., Karaca, S., Kirişci, K., Çarmıklı, E.Ö.: Turkey and Syrian refugees: The limits of hospitality. International Strategic Research Organization (USAK) (2013)

16. Dustmann, C., Glitz, A., Frattini, T.: The labour market impact of immigration. Oxford Review of Economic Policy **24**(3), 477–494 (2008)
17. Erdougan, M.: Thinking outside the camp: Syrian refugees in istanbul. The Migration Information Source (2017)
18. Granovetter, M.S.: The strength of weak ties. In: Social networks, pp. 347–367. Elsevier (1977)
19. Halla, M., Wagner, A.F., Zweimüller, J.: Immigration and voting for the far right. Journal of the European Economic Association **15**(6), 1341–1385 (2017)
20. Hanley, J., Al Mhamied, A., Cleveland, J., Hajjar, O., Hassan, G., Ives, N., Khyar, R., Hynie, M.: The social networks, social support and social capital of syrian refugees privately sponsored to settle in montreal: Indications for employment and housing during their early experiences of integration. Canadian Ethnic Studies **50**(2), 123–148 (2018)
21. İçduygu, A., Diker, E.: Labor market integration of syrian refugees in turkey: From refugees to settlers. Journal of Migration Studies **3**(1), 5–2 (2017)
22. Korac, M.: Integration and how we facilitate it: A comparative study of the settlement experiences of refugees in italy and the netherlands. Sociology **37**(1), 51–68 (2003)
23. Lamanna, F., Lenormand, M., Salas-Olmedo, M.H., Romanillos, G., Gonçalves, B., Ramasco, J.J.: Immigrant community integration in world cities. PloS one **13**(3), e0191612 (2018)
24. Martin, I., Arcarons, A., Aumüller, J., Bevelander, P., Emilsson, H., Kalantaryan, S., MacIver, A., Mara, I., Scalettaris, G., Venturini, A., et al.: From refugees to workers: mapping labour market integration support measures for asylum-seekers and refugees in eu member states. volume ii: Literature review and country case studies. Tech. rep. (2016)
25. Massey, D.S., Denton, N.A.: The dimensions of residential segregation. Social forces **67**(2), 281–315 (1988)
26. Oka, M., Wong, D.W.: Spatializing segregation measures: An approach to better depict social relationships. Cityscape **17**(1), 97–114 (2015)
27. Otto, A.H., Steinhardt, M.F.: Immigration and election outcomesevidence from city districts in hamburg. Regional Science and Urban Economics **45**, 67–79 (2014)
28. Özden, S.: Syrian refugees in turkey. Tech. rep. (2013)
29. Pentland, A.: Honest signals: how they shape our world. MIT press (2010)
30. Salah, A.A., Pentland, A., Lepri, B., Letouzé, E., Vinck, P., de Montjoye, Y.A., Dong, X., Dağdelen, Ö.: Data for refugees: The d4r challenge on mobility of syrian refugees in turkey. arXiv preprint arXiv:1807.00523 (2018)
31. South, S.J., Crowder, K., Chavez, E.: Migration and spatial assimilation among us latinos: Classical versus segmented trajectories. Demography **42**(3), 497–521 (2005)
32. Steinmayr, A.: Exposure to refugees and voting for the far-right:(unexpected) results from austria (2016)
33. Sundsøy, P., Bjelland, J., Reme, B.A., Jahani, E., Wetter, E., Bengtsson, L.: Estimating individual employment status using mobile phone network data. arXiv preprint arXiv:1612.03870 (2016)
34. Toole, J.L., Lin, Y.R., Muehlegger, E., Shoag, D., González, M.C., Lazer, D.: Tracking employment shocks using mobile phone data. Journal of The Royal Society Interface **12**(107), 20150185 (2015)
35. Vinikoor, L.C., Kaufman, J.S., MacLehose, R.F., Laraia, B.A.: Effects of racial density and income incongruity on pregnancy outcomes in less segregated communities. Social science & medicine **66**(2), 255–259 (2008)

A Local integration mapped

For completeness, we visualized social, spatial, and economic integration across provinces in Turkey in Fig 7 and across districts in Istanbul in Fig 9.

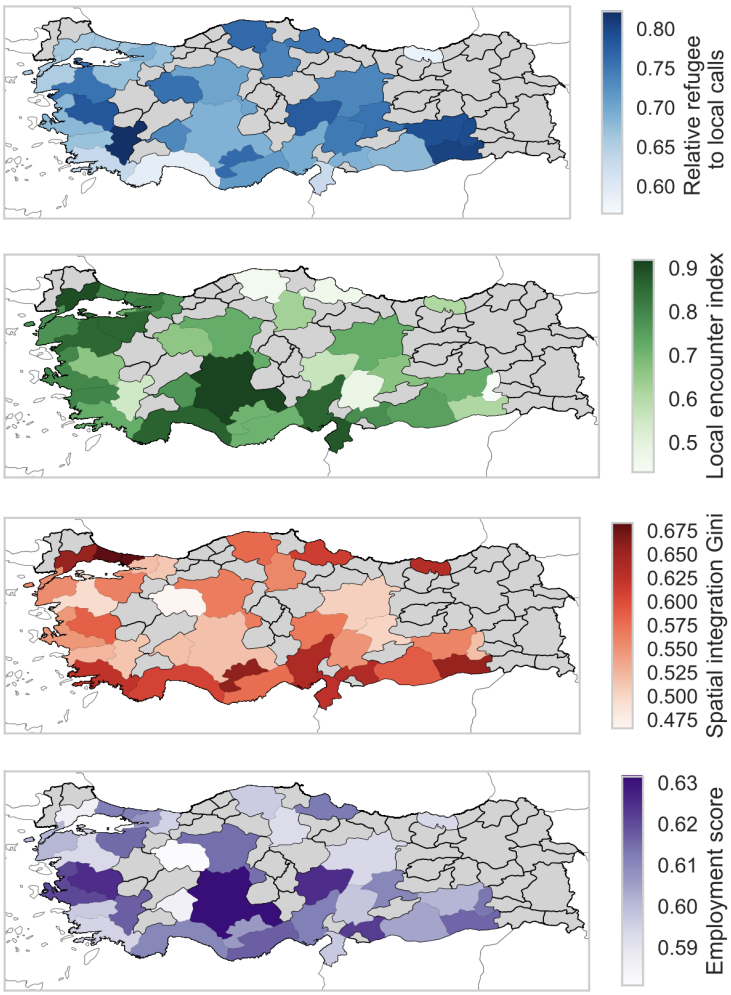


Fig. 7: Average of social, spatial and economic integration measures for each province in Turkey. To reduce noisy values, provinces with less than 500 unique user IDs aggregated over the whole year are greyed out.

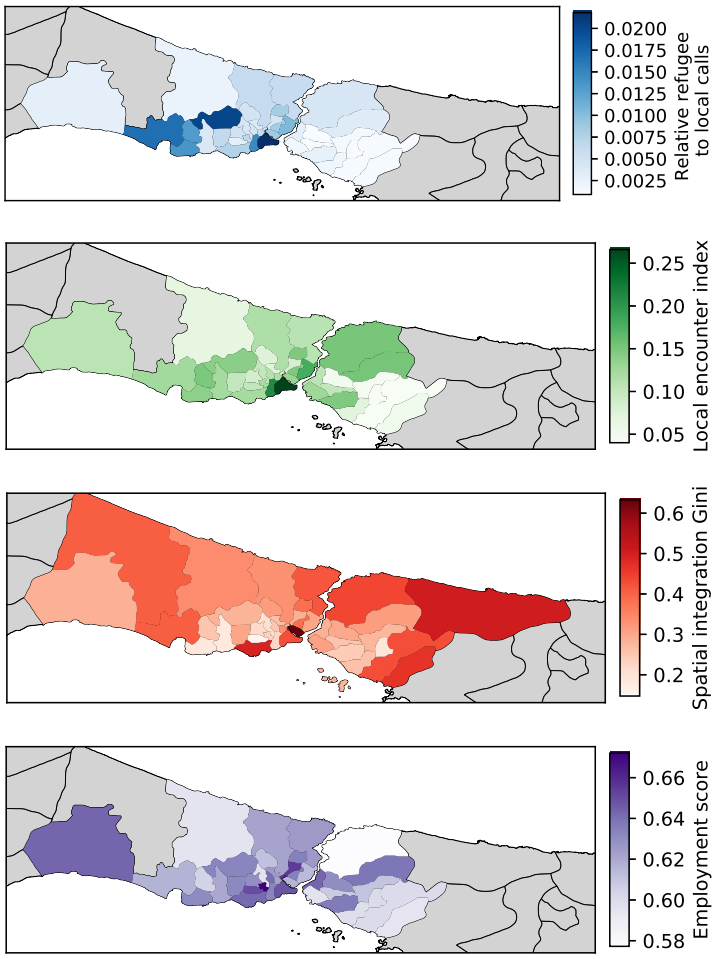


Fig. 9: Average of social, spatial and economic integration measures for each district in Istanbul. To reduce noisy values, districts with less than 500 unique user IDs aggregated over the whole year are greyed out.

Quantified Understanding of Syrian Refugee Integration in Turkey

Wangsu Hu[1,2], Ran He[1], Jin Cao[1], Lisa Zhang[1], Huseyin Uzunalioglu[1], Ahmet Akyamac[1], and Chitra Phadke[1]

[1] Nokia Bell Labs
{firstname.lastname}@nokia-bell-labs.com
[2] Rutgers University
wh251@scarletmail.rutgers.edu

Abstract. Turkey hosts over 3.5 million Syrian refugees. How they integrate into local communities significantly impacts the stability of the host country. In this project, we use mobile users' call-detail records (CDR) and point-of-interest (POI) data to infer users' mobility and activity patterns in order to investigate the level of integration. Using this data, we compare the spatial patterns of refugees against those of citizens. We observe a few patterns that set refugees apart, e.g. smaller travel distances, fewer high-expense activities and separate home locations from the locals. We also establish a metric based on a citizen-refugee classifier to quantify the degree of integration. We are able to rank 11 densely populated cities, and notice that the level of integration varies from city to city. For example, Gaziantep serves as an example of a well-integrated city, whereas Sanliurfa appears to be poorly integrated.

Keywords: Social integration, Syrian refugees, Turkish citizens, CDR, POI, spatial pattern, home-based, non-home-based, integration metric.

1 Introduction

According to United Nations estimates, the Syrian refugee population in Turkey was over 3.5 million as of April 2018 [2]. A vast majority of these refugees live outside the camps and are spread across numerous Turkish cities. Approximately 6-7% live in 21 camps close to the border with Syria. The degree of integration of refugees in Turkey has a huge social and economic impact. The purpose of our study is to quantify the level of integration of Syrian refugees in Turkey.

In this study, we utilize mobile users' call records to investigate how refugees in Turkey are integrated into society. We hope our findings would shed light on the factors that affect integration and drive policies that would encourage better integration. Our investigation focuses on two directions.

Spatial Patterns. We would like to derive movement patterns and frequented locations from call records. We would then be able to compare refugees' spatial patterns against citizens' and detect significant differences, if they exist. Specifically, we consider the following questions:

- Where do refugees reside? Do they live in isolation, or do they mix with the locals?
- Where do refugees typically visit during the day? Where do they shop, work, go to school and have social activities? Do they have similar or different activity patterns compared to citizens?
- How do spatial differences change over time?

Degree of Integration. We would like to develop a metric for measuring the degree of integration and apply it across different cities. Specifically, we consider the following questions:

- Are there regions where the refugees are better integrated into the local community? Are there regions where segregation is more obvious?

To address the questions above, we use three datasets. $D1$ contains Call Detail Records (CDR) for phone calls and text messages over 12 months in 1997; $D2$ contains point-of-interest (POI) information from FourSquare [1]; $D3$ contains city population density and Turk Telekom customer distribution. The first and third datasets are made available by Data for Refugee Challenge (D4R) [9] and the second set is crawled via the FourSquare API. We give a more detailed description of the data in Section 2. We also focus on eleven cities that have large refugee populations. In alphabetical order, these are Adana, Ankara, Antalya, Bursa, Gaziantep, Hatay, Istanbul, Izmir, Konya, Mersin, and Sanliurfa.

Main Findings. In Sections 3, 4 and 5 we focus on spatial pattern analysis. We compute stay points for mobile users to infer potential home-based and non-home-based activities. We make several observations. First, refugees tend to have trips shorter in distance and lower radius of gyration. Second, in comparing home-based and non-home-based activities, we discover that refugees and citizens differ more in the former. Especially in Sanliurfa, refugees tend to have distinct home-based stay points from citizens. However, this separation improves over time. In contrast, the non-home-based stay points and activities of the two groups are less distinct. Using FourSquare POI data, we are able to enumerate 19 major life-style activities including dining, shopping and health care. While refugees and citizens have similar types of activities, it is noticeable that refugees tend to avoid high-expense activities such as fashion and automobile shopping.

In Sections 6, we quantify the level of integration of refugees and citizens. We create a classifier to determine whether mobile users are refugees or citizens based on features derived from CDRs, lifestyle activities and trip lengths and frequencies. The inverse of the accuracy of this classifier can indicate the level of integration. For example, in a well-integrated scenario, it would be more difficult to differentiate refugees from citizens, and the classification accuracy would be low. We refer to this metric as the *inverse classification score*. Using the scores across the eleven cities of interest, we notice that the integration levels differ from city to city. Mersin, Gaziantep Ankara and Istanbul have lower classification accuracy, which translate to better integration. On the other hand, Konya, Sanliurfa and Antalya have higher classification accuracy, which translate to

poorer integration. We also observe that the integration level varies over the twelve months for which the CDR data are available. While we could not find evidence from CDR, we suspect the lower level of integration during the summer months could be attributed to seasonal farm work, a common employment by refugees [8, 6].

We note that *D1* contains CDRs for active communication, i.e. a phone call or text message. This type of data may not capture as much information in comparison to passive records. In addition, the CDR data do not have full recipient information but only retain the "refugee" or "citizen" flag, which would have shed light on whether refugees communicated among themselves or were well connected to the citizens. We believe that further or different insight of integration could be derived with more telecom information.

2 Datasets and Preprocessing

We use three datasets to assess refugee integration in Turkey. In this section, we give a detailed description to each of these datasets and describe our preprocessing steps for the downstream analysis.

2.1 Three Datasets: *D1, D2* and *D3*

The first dataset *D1* is provided by Turk Telekom, via the Data for Refugee (D4R) Challenge [9]. This dataset is based on anonymized mobile Call-Detail Records (CDRs) of 1,211,839 phone calls and SMS messages of 992,457 Turk Telekom customers during the twelve months of 2017. Each includes a "refugee" flag if the caller/sender is likely to belong to a refugee customer, and a "citizen" flag if the caller/sender is likely to be a citizen customer. Each record in this dataset represents a single connection to an antenna and contains the following fields: timestamp, anonymized ID of the user, and the antenna ID the mobile device is connected to. According to [9], the data provider has further anonymized this data, i.e., for each of bi-week periods, an independent set of individuals are randomly sampled, and only calls of these individuals are included.

The second dataset *D2* contains 3,055,216 point-of-interest (POI) records across Turkey, obtained via public Foursquare API [1]. Each POI record includes its latitude and longitude and a business category. FourSquare organizes the categories in a comprehensive multiple-level hierarchy [1]. We use this dataset to correlate with a user's visited locations and to understand of the type of activities that the user may conduct (e.g. lifestyle activities such as eating out and shopping).

The last dataset *D3* provides city-scale population density and Turk Telekom (TT) customer distribution information. It tags each user as refugee or citizen and the user's registered city. At the end of March 2017, there were 75,724,413 mobile customers in Turkey across all operators (94.9% penetration rate). According to the data from the first three months of 2017, the mobile market share of Turk Telekom, from which the D4R challenge data is collected, was 24.7%

[9]. The Cartographic representations of Turkey for *D3* was also provided and visualized in Figure 1 according to the density of population and TT customers in every city where black lines represent the corresponding first-level administrative boundaries. Specifically, we select 11 cities of interest, because the numbers of "citizen" tagged customers in these cities are more than 40,000, while those of other cities are all smaller than 10 according to *D3*, implying no meaningful information can be extracted from data for other cities. In alphabetical order, these 11 cities are

> Adana, Ankara, Antalya, Bursa, Gaziantep, Hatay, Istanbul, Izmir, Konya, Mersin, and Sanliurfa.

Detailed summarizing statistics of population for them, as provided by *D3*, are included in Table 1.

Table 1. Population based statistics for 11 cities of interest for 2016.

City	City population	"Refugee"-tagged TT customers	"Citizen"-tagged TT customers
Adana	2,201,670	2,819	40,415
Ankara	5,346,518	5,581	40,443
Antalya	2,328,555	2,880	40,367
Bursa	2,901,396	3,479	40,359
Gaziantep	1,974,244	14,898	80,655
Hatay	1,555,165	7,024	40,394
Istanbul	14,804,116	84,176	363,334
Izmir	4,223,545	10,425	40,501
Konya	2,161,303	4,718	40,388
Mersin	1,773,852	10,036	40,244
Sanliurfa	1,940,627	9,701	40,321

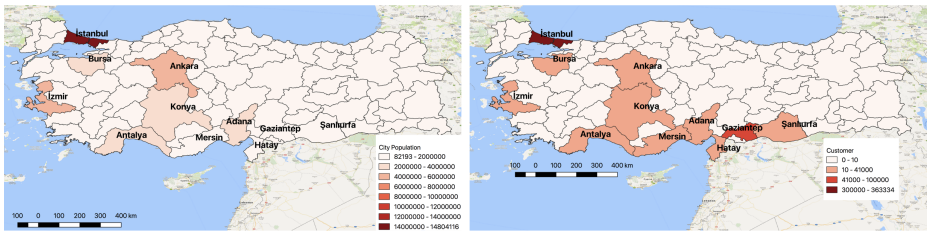


Fig. 1. Cartographic representations of Turkey.(A) Population Distribution. (B) TT Customers Distribution. We shows the names of top 11 cities that contains most TT customers.

This list includes Istanbul, the most populated city with most TT customers, and three major Turkey-Syria border cities, Gaziantep, Mersin and Sanliurfa, as marked in Figure 1. In this integration study, we focus on comparing the behavior of refugees and citizens in these 11 cities and contrast such behavior among these cities.

2.2 Extracting Locations of Stay Points

A stay point is a location where a user stays for a certain duration to conduct activities. Each row of the dataset *D1* represents a stay point. Example of stay points include home, office and places where the user conducts life-style activities such as eating out and shopping.

Though each stay point is mapped to a GSM cell tower, where its location (latitude/longitude) is known, using the location of cell tower as location of a stay point is inaccurate. This is because the CDR data *D1* uses the GSM cell tower as the spatial resolution, even if a user is immobile, his/her phone calls or text messages may be assigned to different co-located cell towers that are very close to each other (e.g., 1 meter). To remove such ping-pong effect on determining the locations of stay points of individual users, we apply DBSCAN [4] to group antennas, and then calculate the centroid of the discovered antenna group to represent an individual user's location. To maintain location consistency, we also assign each POI to its spatially closest antenna group within 5 kilometers.

DBSCAN is a representative of density-based clustering algorithms. It receives two parameters: $MinPts$, minimum number of points in the neighborhood, and Eps, maximum distance between neighboring points. The algorithm starts from the first antenna point and checks if there are any points in the point's Eps neighborhood. If the number of neighboring points is not less than $MinPts$, all previously not assigned neighboring points are added to the new cluster. Then, the cluster is expanded to all other unassigned points which can be reached from the neighboring points with respect to Eps. Here we set the $minPts = 1$ to keep all antennas and set $Eps = 200$ meters. Therefore, co-located antennas will be collapsed into the same antenna group. As a result, the original 93,451 antennas are grouped into 41,212 antenna groups. Then the location of each stay point is actually the centroid of the the antenna group where the antenna corresponding to the stay point belongs to.

2.3 Filtering Out Inactive Users

Although used widely for human behavior studies, mobile phone data such as CDRs provide only a proxy for human activities, as a record is created only when the phone is in use. Not all subscribers use their phones frequently, and this limits the information that can be gained from mobile phone data. Therefore, we use medians of two groups of users as thresholds to filter out those relatively inactive users regarding their record numbers and active days within the bi-weekly periods. The reason we choose median as threshold is simply because many users have too few activities to provide meaningful insights of their

behaviors. Even with a threshold of median, within each bi-weekly period, the thresholds of average daily activities are 2.5 and 3 times per day for refugees and citizens respectively and active days are 4 days for refugee and 5 days for citizen.

2.4 Roadmap for Downstream Analysis

In the following, we will compare the spatial location patterns between citizens and refugees from three perspectives: i) movements, i.e., trips between spatially different locations (Section 3), ii) spatial distributions of stay points (Section 4), where a stay point is defined as a location where a user stay for a certain duration to conduct certain activities, iii) lifestyle categories of stay-points for conducting leisure activities such as eating and shopping (Section 5). For each of these three different perspectives, we point out where the behavior of citizens and refugees differ and compare such behavior differences between different cities. Finally, we create a metric to quantify the level of integration in each of the 11 cities (Section 6).

3 Movement Analysis

Each record in the CDR data (*D1*) represents a cell phone usage by a user with the serving cell tower location at that time. As described in Section 2.2, we associate each record with a specific location. If a user makes a movement, then the consecutive locations will differ. Specifically, we define a trip as the trace between locations of two different consecutively visited stay points and denote its geodesic distance by d to approximate the travel distance.

We apply two indicators to explore the movement patterns of a user: trip lengths and radius of gyration. First, we estimate the probability density $P(d)$ of the individual travel distances d over a period of two weeks. Furthermore, we assess the radius of gyration as another important metric for mobility patterns. The radius of gyration r_g for each caller is the characteristic distance traveled by each caller when observed up to time t, and is computed as follows:

$$r_g^2 = \frac{1}{N}\sum_{k}^{N} ||\boldsymbol{r}_k - \boldsymbol{r}_{mean}||^2, \tag{1}$$

where $\boldsymbol{r}_k, k = 1, \cdots, N$ is the position of kth stay point for an individual and $\boldsymbol{r}_{mean} = \frac{1}{N}\sum_{k=1}^{N} \boldsymbol{r}_k$ is the mean position of all these stay points up to time t.

Observation 1 *Refugees have fewer long-distance trips and lower radius of gyration compared to citizens.*

We plot the trip length distribution of the individual travel distances over a period of two weeks, for refugee (left plot, solid blue line) and citizen (right

plot, solid orange line), as shown in Figure 2. In general, the distributions were qualitatively similar to each other at the country level. However, for refugees, the mean of all the trips is 6.13 and the standard deviation is 30.1; while those for citizens are 6.39 and 35.7. The unit is km. This comparison reveals that refugees have fewer long-distance trips than citizens. Meanwhile, based on the sample distribution of customers tagged as refugees and their registered locations in *D3*, we selected the top 11 cities that contain most customers to investigate the regional differences. Different cities are identified by different scatter marker colors. We observed the refugee user group exhibited greater diversity than similarly defined regions in citizen user group. This would indicate that the likelihood that refugee migrate and commute with respect to distance is much more dependent on what part of the country they are in. Meanwhile, the distributions of radius of gyration show that the refugee users have relatively limited activity area compared to citizen users in general, which can be verified by comparing two estimated densities, where refugees' is lower than that for citizens on the right side.

4 Spatial Distributions of Stay-points

To analyze the pattern of stay points visited by refugees and citizens, we first classify the stay points into one of the following three types based on their relationship with home or employment related activities: *home-based*, *work-based*, *lifestyle*. A location p is a *home-based* for mobile user u if the following two conditions hold:

- p appears in CDRs on at least 70% of the days;
- u appears at p during night hours, defined from $7pm$ to $7am$, in at least 50% of these CDRs.

A location p is a *work-based* if the following three conditions hold:

- p appears in CDRS on at least 60% of the days;
- u appears at location p more often on weekdays than weekends;
- u appears at p more often during day than at night: at least 80% CDRs take place during the day, $7am$ to $7pm$, and at most 30% take place at night.

If p is neither a home base or a work base, we say p is a *lifestyle* location. Furthermore, we also combine *work-based* and *lifestyle* locations and refer them as *non-home-based* locations.

We comment here why we use hard criteria such as the above to partition the stay points into categories instead of applying statistical inference techniques (such as in our previous work [5]). This is due to the infrequent location sampling inherent in CDR records: a user's locations can only be observed when the phone is in use to make calls or send text messages. Therefore even if the user stays at home every night, we would not be able to observe his home cell every night. Such coarse location samples make the location categorization difficult using statistical learning methods.

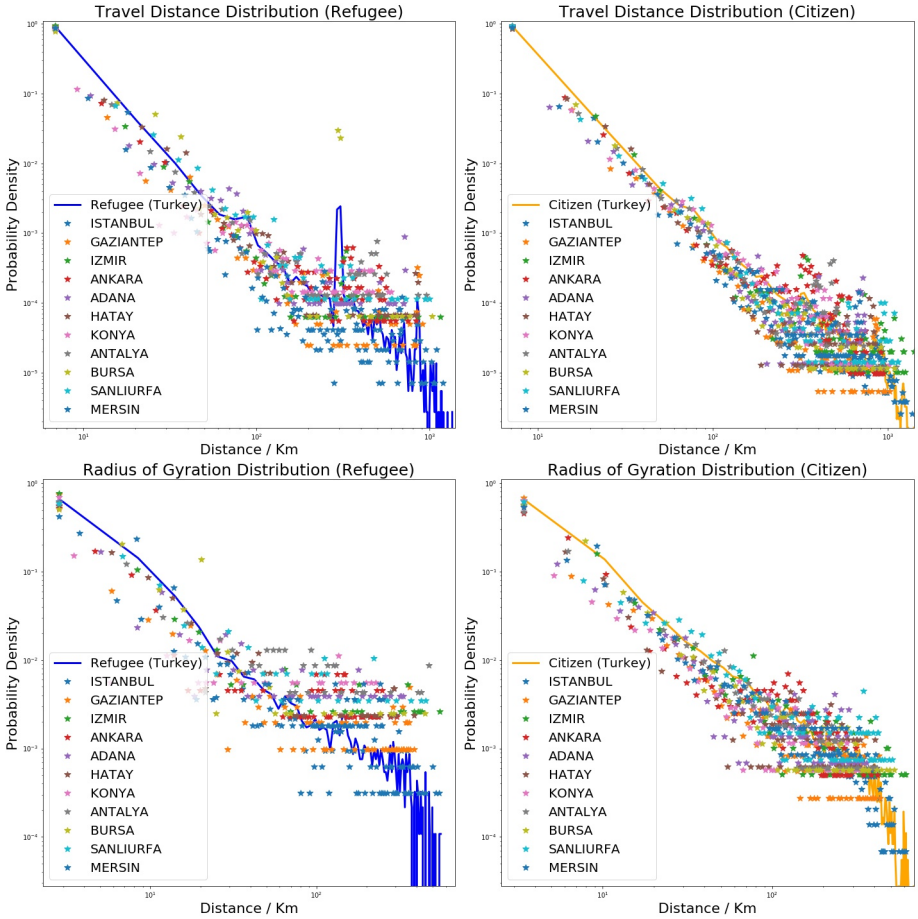


Fig. 2. Mobility Patterns Comparison for Refugee and Citizen

4.1 Spatial Location Distribution Heatmaps

We first visualize the spatial distribution of refugees and citizens' stay points via heat map. In order to do this, we first divide the area of a city into small grids of 1km by 1km and then calculate the population percentages in those grids within the two user groups. At last, we plot those grids on the map with different color scales representing different levels of population percentages. Specifically, to make the grids more visible, we use the square root of the percentages instead of raw percentages so the high densities areas do not dominate the plot. However, due to limited number of customers in each city and sparse locations of stay points (centroid of cell tower groups), often, the colored grids are sparse on the map (where a colored value indicates the presence of visits by users).

Take Sanliurfa as an example. Figure 3 illustrates the spatial distribution of non-home-based activities for refugees and citizens respectively. The blue shades in the left panel reflect the square roots of the following values. In order to calculate population percentage in each 1km by 1km grid, we divide the population of the refugees who have stay points of non-home-based activities in that grid by the overall such refugees in the entire city. The red shades in the right panel reflect similar calculation for citizens. Comparisons of the two panels indicates insignificant differences between refugees and citizens, although citizens have activities in slightly more grids. Furthermore, these non-home-based activities are spread out in the entire map (in the city and towns).

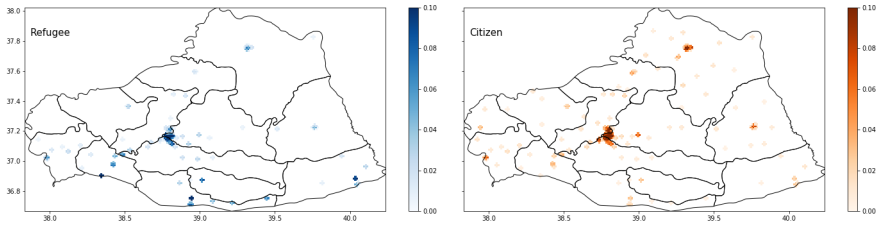


Fig. 3. Heat map of non-home-based activities in Sanliurfa in late May. Left panel is for refugees, while right panel is for citizens.

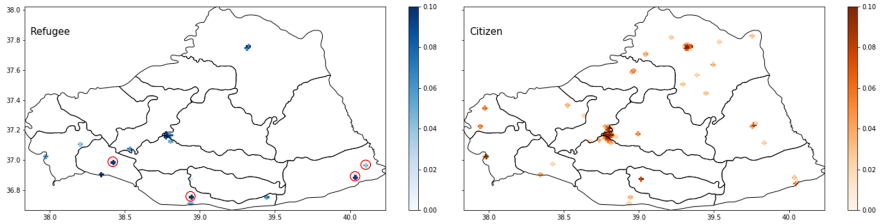


Fig. 4. Heat map of home-based activities in Sanliurfa in late May. Left panel is for refugees while right panel is for citizens. Red circles in the left panel indicate the identified refugee camps (see also Figure 5).

Figure 4 shows the spatial distribution of home-based activities in Sanliurfa for refugees and citizens respectively. We observe here significant differences between the two spatial distributions. To be specific, we see several dark points at the bottom of the refugees' heat map, as circled in red, while those grid points are much less visible on the citizens' heat map. Comparing with the reference refugee camp map from United Nations Refugee Agency (Figure 5)[2], these dark points actually represent the locations of refugees' camps in Sanliurfa. This is

not surprising because refugees go back to camps at night while they go to cities or towns for work during daytime. Another interesting observation is that in contrast to Figure 3, the home-based activities have a much more focused spatial distribution then non-home-based activities, especially for Sanliurfa refugees.

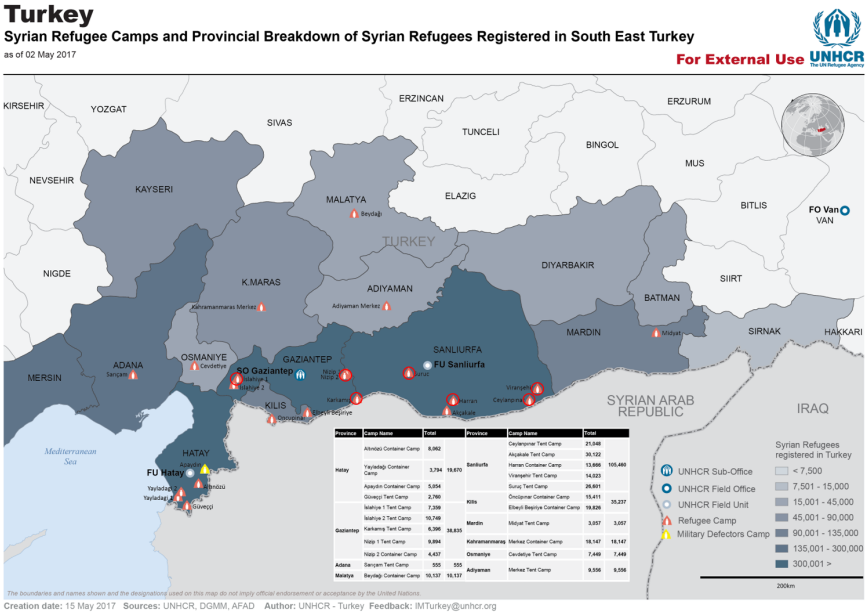


Fig. 5. Refugee camp map from United Nations Refugee Agency. Red circles indicate the identified refugee camps corresponding to Figure 4 and Figure 6.

However, one interesting finding is that the case is different for Gaziantep, another border city just like Sanliurfa, where there are also several refugee camps. Figure 6 shows the comparison of spatial distribution of home-based activities between refugees and citizens. Though we are still able to find several dark point in refugees' heat map that represent refugee camp, the majority of dark points are around center area, consistent with citizens' distribution. This reveals that refugees are more likely to reside in city center area, the same as citizens, indicating a better integration. This observation will also be verified by our developed quantitative measure of the difference. The SPAEF score for Sanliurfa of Figure 4 is -0.54 while that for Gaziantep of Figure 6 is 0.06. See next section for more details. Such distinction between Gaziantep and Sanliurfa was also pointed out in [8], and was linked to the economical and cultural differences between these two cities.

In conclusion, we have the following observation via the stay point spatial distribution analysis.

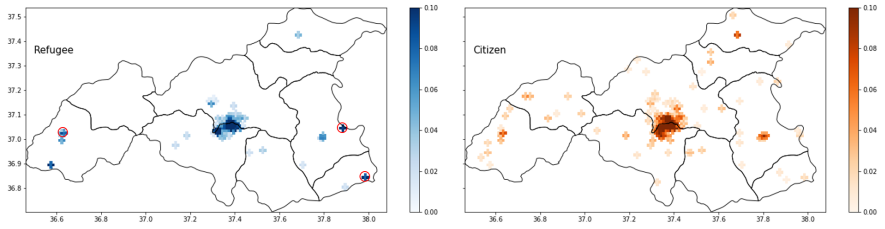


Fig. 6. Heat map of home-based activities for refugees (left) and citizens (right) in Gaziantep in late May. Red circles in the left panel indicate the identified refugee camps (see also Figure 5).

Observation 2 *Between refugees and citizens, their spatial locations for non-home-based activities are more similar than home-based activities, indicating more segregation of primary residence between the two groups.*

Observation 3 *For border city Sanliurfa, home-based activities of refugees have a much higher concentration around refugee camps, comparing to another border city Gaziantep.*

4.2 SPAEF Scores for Quantifying Spatial Distributional Differences

Spatially distributed models, which represent various components of the geographic system, are commonly applied in policy-making, management and research. To assess the degree of refugee integration, here we adopted a spatial performance metric, referred as SPAtial EFficiency (SPAEF), originally proposed in [7]. To be specific, this metric focuses on the paired vectors of the grid-based stay points for refugees and citizens, with values representing the population densities. These two vectors are visualized in heat maps such as Figure 4 for example. Recall we use square roots of percentages for clarity of visualization in the heat maps. Here the two vectors are from the raw percentages. In order to compare two vectors and to ensure bias insensitivity, the values of two vectors are normalized to the range of 0 to 1 when computing SPAEF. And SPAEF is defined as

$$SPAEF = 1 - \sqrt{(\alpha - 1)^2 + (\beta - 1)^2 + (\gamma - 1)^2}. \tag{2}$$

Denote two vectors of spatial distributions for refugees and citizens are R and C. The parameter α is the Pearson correlation $\rho(R, C)$ between the refugee and citizen stay-points spatial distribution. β is the ratio of two coefficient of variation $\frac{\sigma_C}{\mu_C} / \frac{\sigma_R}{\mu_R}$ capturing the comparison of spatial variability. γ is the percentage of the area of histogram intersection for the histograms of the two vectors mentioned above, while two histograms containing the same number of bins. More specifi-

cally, $\gamma = \frac{\sum_{j=1}^{n} min(K_j, L_j)}{\sum_{j=1}^{n} K_j}$, where K is the histogram of R and L is the histogram of C and n is fixed.

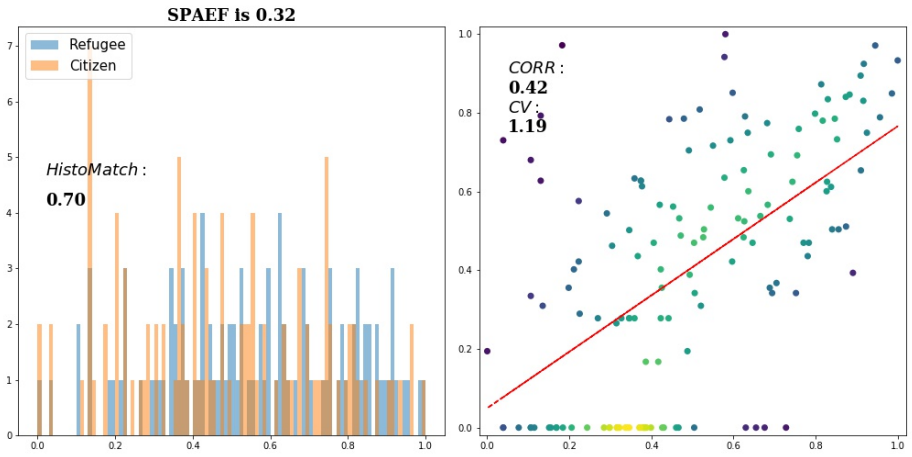


Fig. 7. SPAEF details for non-home-based activities comparison in Sanliurfa, in late May. Left: the histogram intersect after normalization of the grid-based non-home-based activities for refugee (blue bin) and citizen (orange bin). Right: The density scatter plot of the paired vectors of the grid-based non-home-based activities for refugee (y-axis) and citizens (x-axis) using Parula colormap. The yellow points indicate the highest density.

Obviously from (2), the value of SPAEF ranges from $-\infty$ to 1, where larger value indicate two spatial distributions are more similar to or consistent with each other. For example, when SPAEF is 1, all α, β and γ have to be 1, which means two spatial distributions have to be exactly the same; while on the other hand, when two spatial distributions are different, their correlation, covariance ratio and histogram match will not be 1, thus the corresponding SPAEF is smaller than 1 and the more different, the more dispersed from 1. To get a better understanding of how this metric is computed and the intuition behind it, we provide the SPAEF scores for two sets of spatial distribution comparison as illustrated in Figure 3 and Figure 4, the comparison of non-home-based and home-based activities in Sanliurfa. Specifically, Figure 7 represents for Figure 3 and Figure 8 is for Figure 4. The SPAEF score for non-home-based activities is 0.32, which is larger than that for home-based activities, i.e., -0.39. This matches our observation that non-home-based activities integration is better than home-based one via comparing two heat maps.

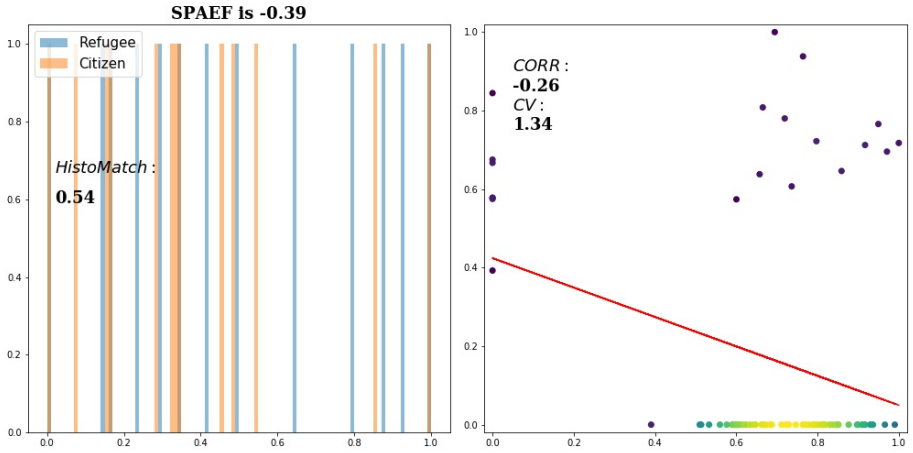


Fig. 8. SPAEF details for home-based activities comparison in Sanliurfa, in late May. Left: the histogram intersect after normalization of the grid-based home-based activities for refugee (blue bin) and citizen (orange bin); Right: The density scatter plot of the paired vectors of the grid-based home-based activities for refugee (y-axis) and citizens (x-axis) using Parula colormap. The yellow points indicate the highest density.

To gain more insights into time trends and location differences, for each of the 11 cities in the study, we calculate the SPAEF value for each of the bi-weekly period (for each of the dataset in *D1*) and for *home-based* and *non-home-based* activities respectively. Figure 9 and Figure 10 plots the SPAEF values for

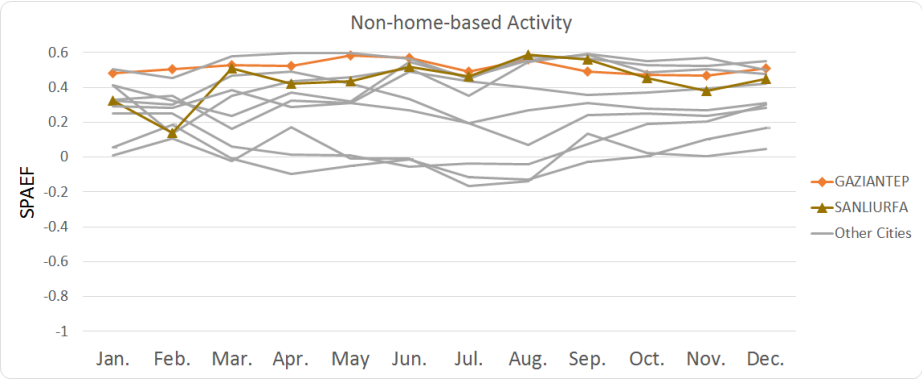


Fig. 9. Comparison of non-home-based activities across 11 cities.

non-home-based and home-based activities and for each of the 11 cities over time, respectively. Not surprisingly, Sanliurfa has overall the lowest scores for

home-based activities, which is consistent with Figure 4 which shows significant differences in home-based activities among its refugees and citizens.

In conclusion, we have the following observations.

Observation 4 *For most cities, the difference in spatial locations visited by refugees and citizens remain throughout the year.*

Observation 5 *Sanliurfa has probably the worst segregation of residential locations between refugees and citizens, although the situation may get improved throughout the year of 2017, as shown by the uptrend in the plot.*

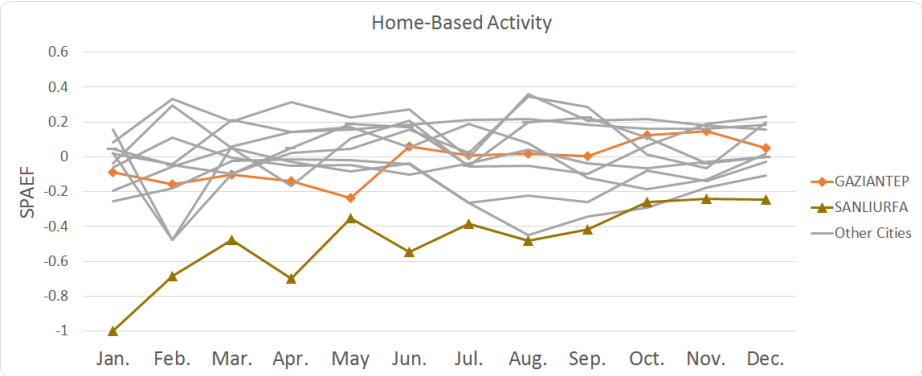


Fig. 10. Comparison of home-based activities across 11 cities.

5 Lifestyle Activity Categories of Stay-Points

Besides comparing the spatial distribution of stay-points between refugees and citizens, we are also interested in the type of lifestyle categorical of these stay-points. The intuition is the type of places a user visit captures his/her personal characteristics. For example, a housewife may be more likely to visit grocery stores or picking children from schools, while foodies are more willing to go to restaurants. This also applies to refugees and citizens. For example, because of the nature of their status and different levels of wealth, citizens may be more likely to visit luxury shops or car servicing places or gas stations.

However, as the smallest location granularity of the CDR data (*D1*)is cell-tower, due to this limited location precision, we are more interested in the nature of the activities than the exact location or the exact point-of-interest (POI) that an individual visits. To compensate the coarseness of location information from cell-based carrier data, we take advantage of third-party POI data source which already classifies POIs into categories. We describe 19 POI categories in Table 2,

Table 2. Lifestyle activity category examples

Activity Categories	Examples
Purchasing and Servicing Cars	Car dealers and leasing, maintenance and repair
Banking	Banks, ATM
Personal Errands	Legal, spa, laundry
Other Personal Errands	Insurance, pet care, costly personal errands
School Activities	Schools
Trips to Public Buildings	Government departments and agencies, post offices
Doctor Appointments	Doctor offices
Other Healthcare	Elder care, vitamin stores
Visiting Parks	Parks, historic sites
Grocery Shopping	Supermarkets, grocery stores
Fashion Shopping	Clothing stores, jewelery stores, department stores
Other Shopping	Winery, Bookstore
Eating	Restaurants, deli stores
Other Social Activities	Bar
Sports and Recreation	Stadiums and arenas
Buying Gas	Gas stations
Commuting	Train stations, bus stations
Other Transportation	Taxi and car services, parking
Travel Related	Lodging, travel agents and tour operators

which are at the right level of granularity, not too fine to make correct inference impossible and not too coarse that the inference is uninformative.

We first extract a POI database *D2* for entire Turkey via Foursquare[1] and assign the corresponding activity category to each POI. Then we assign each POI to its closest cell tower group based on distance. The probability of any activity in the cell group is thus proportional to the number of corresponding activity category in the region of the visited cell tower group.

We associate the location (cell tower group) of a stay point to its potential lifestyle activity as follows. For each row of the raw CDR record (*D1*), denote the cell tower group of the associated stay-point as A. Then we find all POIs from POI database that belong to the region of A. Assume there are n such POIs and $(n_1, n_2, ..., n_{19})$ is the number of POIs in the 19 categories where $n_1 + \cdots n_{19} = n$. Then for this stay-point, the probability of associated activity belonging to category i is $P_i = \frac{n_i}{n}$.

For each 11 cities of interest, we first remove all *home-based* and *work-based* stay-points, and then compute the overall probability distribution $\bar{P}_i$ of the 19 lifestyle categories, averaged among stay points of all users of the city, for refugees and citizens respectively. We reach the following observation by comparing these categorical distributions between refugees and citizens.

Observation 6 *Refugees and citizens have a similar categorical distribution of lifestyle activities. However, refugees engage in fewer high-expense activities.*

Fig. 11. Comparison of activity patterns for refugees and citizens in Sanliurfa.

We illustrate our observation using the result for Sanliurfa, where a radar plot of the probability distribution of the 19 lifestyle activities are shown in Figure 11. Broadly speaking, the two distributions between refugees and citizens are similar. However, if we look closely, we can discern some finer patterns. citizens are more likely to go to regions where there are more gas stations (buying gas), more car dealerships (purchasing and servicing cars), costly shopping places (fashion shopping), compared with refugees, basically because citizens are more wealthy. On the other hand, refugees have higher probabilities of doing low-cost

or basic need personal errands (personal errands) and commuting via public transportations (commuting). Other cities show a similar pattern as Sanliurfa.

6 Quantifying Integration by Inverse Classification Score

Previously, we use CDR records to compare the spatial locations visited by refugees and citizens, from three different perspectives: 1) movement analysis, 2) spatial distributions of stay-points, 3) lifestyle categorical distribution of stay-points. These analysis are conducted by comparing the entire population characteristic of refugees and citizens. Obviously, if the refugees and citizens are well integrated, these population based characteristics should not differ significantly.

In this section, we will study integration from an individual user's perspective, collectively based on all his/her location behavior features. Intuitively, if a refugee is well integrated, then it will be difficult to discern his/her location behavior from the citizens. In other words, if we build a classifier to determine whether a mobile user is a refugee or citizen based on his/her location behavior, then the higher the classification error (or the lower the accuracy), the higher the level of integration. We refer this metric as *inverse classification score* as a way to quantitatively measure the degree of integration between refugees and citizens. We describe our approach in the following.

6.1 Data, Features and Classifier

For each mobile user, we derive the following features for classification purposes:

- *Basic statistics (3 features).* These features summarize the basic information for each user, such as the number of calls or messages and the number of active days.
- *Activity-based features (21 features).* What type of activity an individual does reveals what type of person a user is. These features include indicators of whether home-based activity or work-based activity (as defined in Section 4) can be detected and other 19 features representing the average probabilities of doing 19 life-style activities such as eating or buying gas etc.
- *Distance-based features (8 features).* These features capture how far a user travels, which can help characterize the type of a user. For example, minimum, median and maximum distance of trips and radius of gyration are considered. See Section 3 for definitions.
- *City indicator features (12 features).* To further compare the difference among cities, we have 12 city indicators that are just dummy variables of 11 cities of interest, as described in Table 1, and another dummy variable for all other cities.

We use the CDR records *D1* for classification, after removing inactive users. To be specific, there are $1,075,673$ users across the entire year, while $174,116$ are refugees and $901,557$ are citizens.

We choose the popular gradient boosting tree as our classification algorithm. Specifically, we rely on the xgboost [3] package in R. It is efficient and highly flexible. It is also an ensemble method which improves the performance by building multiple classification trees.

The classification result is measured by the 5 fold cross-validation result, i.e., the mean accuracy of the model on validation subsamples. Since the original data is severely unbalanced, we down sample citizens to create a balanced dataset with same numbers of citizens and refugees for training.

6.2 Classification Performance and Feature Importance

To understand how different types of spatial location behavior features described in Section 6.1 are important for differentiating between refugees and citizens, we build models that only includes part of these feature sets. Table 3 shows the classification scores for different models using different sets of feature set. The accuracy is 74.67% using all features and 72.58% using all features excluding city indicators. More specifically, the False Negative Rate for the model using all features is 30.80% while False Positive Rate is 19.82%. Or in other words, refugees are more likely to be identified as citizens compared to the other way around, which makes sense.

Comparing the first three models in Table 3, we see that distance-based features alone achieve the highest accuracy. Models using only the three basic features is a close second, also indicating their importance. Though activity-based model perform the worst, it still achieves an accuracy of 63.31%, indicating these features can also help distinguish refugees. Therefore, for almost all perspectives of users' location behavior, overall speaking, refugees are distinguishable from citizens.

To understand which specific features (as opposed to the type of features) are important to differentiate between a refugee and a citizen, we compute the feature importance scores from the xgboost package for the full model (last row in the table), and the top 15 features are shown in Figure 12. Clearly, basic features such as number of activities are the most important factor, while distance based features are also very significant. These observations are consistent with our observations earlier.

Table 3. Classification accuracies using different sets of features

Model	Number of features	Classification accuracy
Basic features	3	65.30%
Activity-based	21	63.31%
Distance-based	8	66.01%
All features excluding city indicators	32	72.58%
All features	44	74.67%

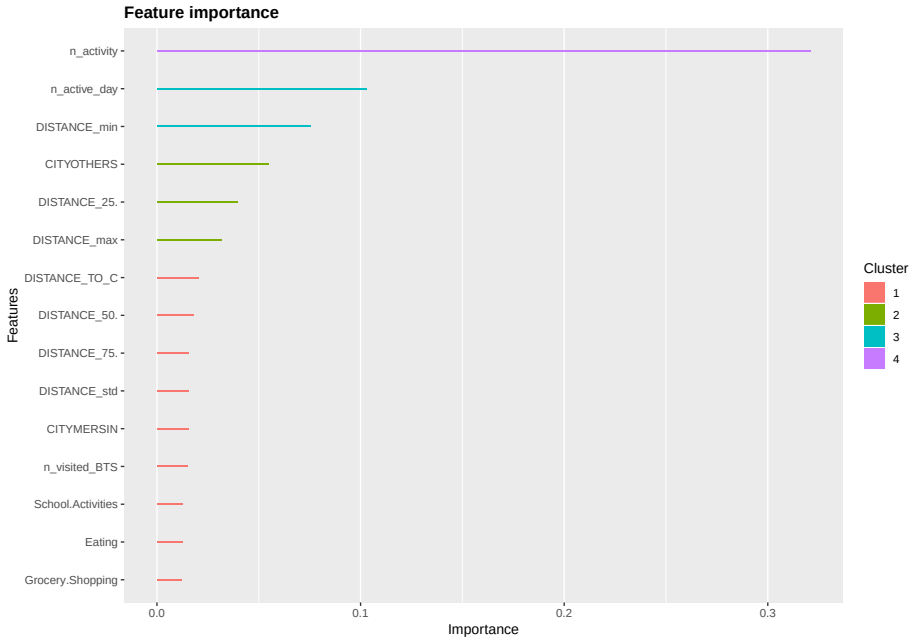


Fig. 12. Top 15 features with highest importance scores from the gradient boosting tree classifier for the full model. Different clusters show grouping of the features in importance, where cluster 1 is the most important, and cluster 4 is the least important among the top 15 features.

Observation 7 *Refugees are most distinguishable from citizens in terms of basic features describing their phone usage (less for refugees and more for citizens). They are also distinguishable in terms of their movement patterns, and to a lesser degree their stay-point activities patterns.*

6.3 Inverse Classification Score

In the scenario of perfect integration, i.e., a refugee is indistinguishable from a citizen, for the balanced dataset, we should have an classification accuracy 50%. On the other hand, if the population of refugees and citizens are completely separable, then the classification accuracy should be close to 100%. This motivates us to devise the following integration metric which we shall refer as *inverse classification score*:

$$\text{integration score} = \text{inverse classification score} = \frac{1}{\text{classification accuracy}} - 1. \tag{3}$$

In the above, the inverse classification score will be 1 when refugees and citizens are completely integrated, and 0 when refugees and citizens are completely disintegrated.

For each of the 11 cites, we obtain its classification accuracy and use (3) to obtain a score to measure its integration. The results are shown in Table 4. Consistent with previous activity-based result, Sanliurfa and Konya are among the cities where integration are the worst, while Mersin and Gaziantep are among the cities where integration are much better.

Table 4. Classification accuracies and integration scores for different cities

City	Classification Accuracy	Integration Score	Integration Rank
Mersin	60.31%	0.658	1
Gaziantep	75.27%	0.329	2
Ankara	77.53%	0.289	3
Istanbul	77.65%	0.287	4
Izmir	80.02%	0.250	5
Bursa	81.10%	0.233	6
Hatay	82.35%	0.215	7
Adana	84.18%	0.188	8
Antalya	85.97%	0.161	9
Sanliurfa	86.15%	0.161	10
Konya	89.34%	0.120	11

Compared with city population in Table 1, top integrated cities are all those with large populations. This tells us that refugees may be better integrated in big cities due to the reason that there are more job opportunities in big cities. One exception is Gaziantep, which is a small city and a border city, where refugees are also integrated well, especially compared with another border city Sanliurfa. This interesting phenomenon needs further study, and it may tell us the reason of this exception. Other cities may learn from it and provide better help to the refugees for a better integration. In summary, our conclusions are as follows.

Observation 8 *Different cities have different degrees of integration. In general, more populated cities tends to be more integrated. Among the 11 cites in the study, Mersin shows the highest degree of integration between refugees and citizens while Konya shows the least.*

7 Conclusion

In this study we use CDR and POI data to infer mobile users' spatial patterns and to investigate how well Syrian refugees are integrated with Turkish citizens. We discover a few behavioral differences that set refugees a side, e.g. shorter trips, fewer high-expense activities and separate residences. We also propose a metric to quantify the level of integration. We observe that the integration level differs from city to city. We see Gaziantep as an example of a well-integrated city and Sanliurfa as an example of the opposite.

Our observations are inferred from data alone. Although we looked for other sources such as news items and refugee studies to corroborate our findings, we did not find many leads. We also believe that additional telecom data such as passive records can provide additional insights to uncover refugee mobility and activity patterns.

References

1. Foursquare. https://developer.foursquare.com/docs/resources/categories
2. UNHCR Syria regional refugee response -Turkey. https://data2.unhcr.org/en/situations/syria/location/113
3. Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. pp. 785–794. ACM (2016)
4. Ester, M., Kriegel, H.P., Sander, J., Xu, X., et al.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: Kdd. vol. 96, pp. 226–231 (1996)
5. Jin Cao, Sining Chen, S.K.N.K., Zhang, L.: Extracting mobile user behavioral similarity via cell-level location trace. IEEE Infocom GI Workshop 2017 (2017)
6. Kaygısız, I.: Suriyeli mültecilerin türkiye işgücü piyasasına etkileri. Friedrich-Ebert-Stiftung Türkei - Dünyadan (August 2017)
7. Koch, J., Demirel, M.C., Stisen, S.: The spatial efficiency metric (spaef): multiple-component evaluation of spatial patterns for optimization of hydrological models. Geoscientific Model Development **11**(5), 1873–1886 (2018)
8. Lordoğlu, K., Aslan, M.: En fazla suriyeli göçmen alan beş kentin emek piyasalarında değişimi: 2011-2014. Çalışma ve Toplum Dergisi (49), 789–808 (2016)
9. Salah, A.A., Pentland, A., Lepri, B., Letouze, E., Vinck, P., de Montjoye, Y.A., Dong, X., Dagdelen, O.: Data for Refugees: The D4R Challenge on Mobility of Syrian Refugees in Turkey. ArXiv e-prints (Jul 2018)

Refugees in undeclared employment - A case study in Turkey*

Fabian Bruckschen[1], Till Koebe[1], Melina Ludolph[1,2], Maria Francesca Marino[3], and Timo Schmid[1]

[1] Freie Universität Berlin, 14195 Berlin, Germany
[2] Humboldt Universität Berlin, 10099 Berlin, Germany
[3] Università degli Studi di Firenze, 50121 Firenze, Italy

Abstract. Exploitation of vulnerable groups such as refugees for cheap labour is a notorious phenomenon in Turkey. Up to 2017, only 1.3% of the around 3 mn Syrian refugees registered in Turkey have been granted a work permit, leaving the overwhelming majority dependent on undeclared employment with all its negative implications: high-risk jobs, pay below minimum wage, lack of access to social security. Mobile phone metadata allow for a detailed view on commuting routines and migration, possibly unearthing employment situations which are not captured otherwise. This study proposes a methodological framework for identifying potentially undeclared employment among refugees in Turkey within the current situation. To do so, it includes an early proof-of-concept based on a Difference-in-Differences approach by analyzing seasonal migration and commuting patterns in two specific cases: during the late-summer hazelnut harvest in the province of Ordu and at the construction site of the Istanbul Grand Airport. The study finds clear indication for work-related migration and commuting patterns among refugees hinting at undeclared employment. The proposed framework therefore provides an analytical instrument to make targeted interventions such as controls more effective by detecting small areas where undeclared work likely takes place.

Keywords: Unemployment · Agriculture · Construction · Migration · Refugees · Mobile phone metadata.

1 Introduction

The production of timely and precise statistics is critical for effective decision-making as well as policy implementations by both governments and non-governmental institutions. Nevertheless, many socio-economic measures are based on lagged and imprecise information. Data on employment, for instance, is typically collected by means of surveys which come with a number of statistical challenges. Moreover, conducting surveys can be costly. Mobile phone metadata has been at the center of recent efforts to find alternative methods to measure socio-economic indicators such as literacy [7], poverty [5] and employment.

* Contribution to the Data for Refugees (D4R) Challenge

To this end, some studies identify certain behavioural indicators based on Call Detail Records (CDRs) that seem to be linked to the employment status of the phone user. Toole et al. use CDRs from an undisclosed European country to identify individuals affected by a mass layoff due to the closure of a large manufacturing plant and to analyze their behavioural change [9]. They find that laid off workers significantly reduce their social interactions (measured by, e.g., the number of outgoing and incoming calls, unique contacts, calls to other individuals in the same town) and mobility (measured by, e.g., the number of unique locations visited, the average distance from the most visited tower). By using these insights and supplementing CDRs from another European country with quarterly unemployment rates from the EU Labor Force Survey, the authors provide evidence that these features can be used to improve forecasts of unemployment rates at the regional level [9]. Similarily, Almaatouq et al. augment records from unemployment benefit programs in Riyadh, Saudi Arabia, with anonymised mobile phone metadata [1]. They find that indicators measuring activity, social behaviour, and mobility patterns which they generate from the mobile phone data are significantly correlated with unemployment rates at district level. While most studies focus on unemployment rates, Sundsøy et al. apply a machine learning model to predict 18 different professions as well as unemployment on an individual basis in a South-Asian developing country. They do so using a very broad number of indicators reflecting financial status, social behaviour, and mobility patterns generated from mobile phone data [8]. They train their model with data from two household surveys and find, for instance, that the most important predictor for being a clerk is a low mobility radius whereas students can be identified by a large number of text messages and a high level of internet usage. However, the link between mobility and socio-economic status which is related to employment can be affected by regional differences. Xu et al. augment urban socio-economic datasets for Boston and Singapore with mobility indicators extracted from mobile phone datasets and find that, for instance, the opportunities of employment alter the link between status and mobility significantly [10].

Another body of literature focuses on the inference of employment based on the mobile phone users location. In a study on measuring economic activity in China, the researchers use geo-positioning data to identify the number of employed workers in certain areas of interest (e.g., commercial area, industrial park) [3]. They construct an unemployment index which can be tracked over time. To verify their approach they analyze two announced mass layoffs and two developments which were associated with sharp increases in employment. De Nadai et al. use insights about the locations and the magnitude of mobile internet activity in several Italian cities to find a positive correlation with the concentration of office workers [2].

However, to the best of our knowledge there has not been made any efforts to use mobile phone metadata to measure undeclared employment. By its nature, undeclared employment is especially difficult to measure with traditional approaches like surveys. As suggested by previous studies, mobile phone metadata

allow for a fine-grained view on commuting routines, possibly unearthing employment situations which cannot captured otherwise. Since the problem of undeclared employment spans across industries, this study proposes a unified methodological framework based on mobile phone metadata. We test the methodology in an early proof-of-concept on two distinct cases by applying a Difference-in-Differences (DiD) approach as a treatment effects model: the hazelnut harvest in Ordu and the construction site of the Istanbul Grand Airport. This choice is not random. First, a quarter of the worlds hazelnut supply comes from one Turkish province alone: Ordu. This particular fact makes it a topic of global impact. Also, hazelnut harvest is seasonal, labour-intensive and it only takes place in a small part of Turkey during a narrow time window. This helps to isolate work as the prime reason for temporary migration. Second, the Istanbul Grand Airport construction site is - with its around 30,000 directly employed workers - of paramount importance both for the construction sector and, once completed, for the Turkish economy as a whole. Furthermore, the choice of these two cases is also intended to cover two distinct characteristics of labour market-related patterns: seasonal, short-term migration as expected during hazelnut harvest and commuting resulting from mid-to long-term migration as in the context of a multi-year construction project.

Consequently, the study provides a unified methodological framework for identifying potentially undeclared employment among refugees in Turkey within the current situation. To do so, the study builds on one main assumption: commuting patterns among refugees most likely link to undeclared employment, since only 1.3% of them are allowed to work. Note, however, that Syrian refugees did not need a permit to work in seasonal agriculture in 2017 [4]. Hence, our findings regarding temporary migration do not imply undeclared work during the hazelnut harvest in Ordu. Nevertheless, our approach might proof to be effective in detecting illegal employment based on seasonal migration once this exemption is repealed. Moreover, the methodology is applicable to other sectors with seasonal working patterns. The study uses mobile phone metadata from 2017 provided through the Data for Refugee (D4R) challenge, i.e. antenna traffic generated from sampled CDRs.

2 Data

The study largely builds on data provided by the D4R challenge. It includes a sample of CDRs from a large Turkish mobile network operator covering the whole year of 2017. The CDRs are preprocessed into three different types: antenna traffic (SET1), fine-grained movement patterns (SET2) and coarse-grained movement patterns (SET3). Also, a tag is introduced that allows to disaggregate SET1 and SET2 by refugee status (refugee vs. non-refugee). Apart from CDRs, contextual data is provided as well, including the antenna locations and details on the composition of the CDR sample. For details on the D4R data, especially its limitations, please see Salah et al. (2018) [6]. We focus solely on SET1 and

the antenna locations as antenna traffic is less prone to privacy issues and thus improves the implementability of the approach.

Further, we use publicly-available information on the administrative boundaries of Turkey from the Humanitarian Data Exchange[4], geographic shapes of refugee camps in Turkey and of the Istanbul Grand Airport extracted from Google Earth based on location information from the Humanitarian Data Exchange and Google Maps, respectively.

3 Methodology

This study proposes a unified methodological framework to identify employment structures via seasonal migration and via commuting patterns in the context of the current situation of Syrian refugees in Turkey. SET1 is used to investigate effects of refugee-specific seasonal migration and commuting patterns using the number of SMS, the number of calls and the volume of calls (network activity). These variables function as proxies, since the actual variable of interest - the number of users - is not available in the data provided through the D4R challenge. The study uses Voronoi tessellation in order to identify relevant antennas and their respective approximated geographical coverage areas. The methodology consists of two parts dealing with specific characteristics of work-related movements: seasonal migration and commuting. Even though very similar in the approach, we present them separately.

3.1 Seasonal migration

1. Identify harvest events by season start and end, agricultural produce and geographic location using publicly available sources
2. Start iteration over harvest events
3. Divide SET1 into three intervals: before, during and after season
4. Compute rate of change Δnet_act between the median values of weekly network activity by interval and harvest location (harvest location vs. non-harvest location) only for the refugee-tagged CDRs, namely:

$$\Delta net\_act = \frac{\bar{net\_act}_{after}}{\bar{net\_act}_{before}} - 1 \tag{1}$$

5. End iteration over harvest events

Under the assumption that a) network activity is a good proxy for the number of network users and b) mobility is the main driver for network user fluctuation, the procedure gives descriptive indication whether there is harvest-related migration of refugees for a specific harvest event. In this study, we concentrate on the hazelnut harvest in Ordu during August/September. Further assuming that c) refugees are unlikely to go on summer vacations in Turkey, the procedure gives indication whether there is seasonal migration of refugees towards employment during the hazelnut harvest in Ordu.

[4] Retrieved August 15, 2018: https://data.humdata.org/dataset/turkey-administrative-boundaries-levels-0-1-2

3.2 Commuting

1. Identify workplaces by sector and geographic location using publicly available sources
2. Reduce SET1 to workdays (Mon-Fri)
3. Start iteration over workplaces
4. Divide it into two groups: Work (7-20h) and Home (20-7h)
5. Compute rate of change Δnet_act between median values of network activity by group and workplace (workplace vs. non-workplace) only for the refugee-tagged CDRs following equation 1.
6. End iteration over workplaces

Under the assumptions a) and b) Δnet_act gives indication whether there is work-related commuting of refugees to a specific workplace. In this study, we concentrate on the construction site of the Istanbul Grand Airport. Following assumptions c) and adding that d) the majority of work performed by refugees is undeclared (only 1.3 % have a work permit), the procedure gives indication whether there is undeclared employment of refugees on the construction site of the Istanbul Grand Airport.

Both procedures, especially the commuting algorithm, can be used in an exploratory way. For example, the magnitude of the Work-to-home activity ratio by refugee-tag and antenna gives indication of work-related commuting of refugees on the antenna-level. This can help to prioritize targeted interventions against employers offering undeclared work.

Further, both procedures can be combined. For example, in order to test the assumption c) - the majority of refugees do not go on summer vacations - for Ordu, commuting patterns in Ordu during harvest can be analyzed, either exploratory or, if the geographic locations of hazelnut farms are available (e.g. identified using satellite imagery), in a predefined setting.

3.3 Proof-of-Concept

While descriptive statistics may be of interest especially for implementation, this study also provides an early proof-of-concept of the proposed methodological framework. Therefore, it uses a DiD approach in order to isolate the commuting/migration effect and to compare it to the descriptive findings. By applying DiD we account for the selection bias arising from the differences in observable and unobservable characteristics of the antennas like, for instance, its geographic location. DiD builds on two main assumptions: First, a control group is available, i.e. a set of antennas located such that the refugee-specific network activity is not affected by the treatment. Second, the refugee-specific network activity processed by antennas of both groups show a common trend prior to the treatment (i.e. start of the harvest season or beginning of the workday).

A downside of the DiD approach is that these assumptions cannot be tested formally. Nevertheless, we can make the argument in both cases - the harvest in Ordu and a workday at the airport - that there are indeed certain locations in

which refugees are likely not affected by these treatments. Theoretically, SET2, the fine-granular movement profiles, could be used to check the hypothesis. However, since only a small fraction of the customer base is sampled in SET2, it does not provide relevant information. Using the location-, refugee- and time-specific network activity provided in SET1 we can check the common trend assumption graphically.

Consequently, the DiD setup looks as follows:

$$net\_act_{i,t} = \beta_0 + \beta_1 * treat_i + \beta_2 * time_t + \beta_3 * treat_i * time_t + \epsilon_{i,t} \quad (2)$$

where $net\_act_{i,t}$ is a measure of absolute weekly or hourly refugee-tagged network activity at antenna i in week or hour t, $treat_i$ is a dummy variable indicating whether antenna i belongs to the treatment group, $time_t$ is a Boolean variable being 1 if the month or week belongs to the treatment period and $\epsilon_{i,t}$ denotes the error term.

The DiD estimator is given by β_3 which represents the average change over time in the outcome variable for the treatment group compared to the average change over time for the control group.

4 Results

4.1 Hazelnut harvest in Ordu

Using descriptive statistics only, we find clear indication of harvest-related seasonal migration to Ordu. Table 1 compares the growth rate of the average weekly number of refugee-tagged calls between the periods before, during, and after the harvest season for Ordu and the rest of Turkey. Differences in these linear relationships indicate the Ordu-specific growth in network activity compared to the rest of Turkey. Around the beginning of the harvest, we observe a growth in network activity that is 77.9%-points higher in Ordu than in the rest of Turkey. Around the end of the harvest, network activity in Ordu shrinks by 39.8%. In contrast, it simultaneously increases by 14.9% in the rest of Turkey, leading to a 54.7%-points difference in growth rates around the end of the harvest.

Table 1. Growth rate of average weekly number of refugee-tagged calls before, during and after harvest

Location	**before → during**	**during → after**
Ordu	241.3%	-39.8%
Rest of Turkey	163.4%	14.9%
Difference	**77.9%**	**-54.7%**

Figure 1 shows that refugee-specific network activity increases both in Ordu and in the rest of the country during the harvest. However, in the former it returns to pre-harvest activity while the rest of the country shows continuously high network activity.

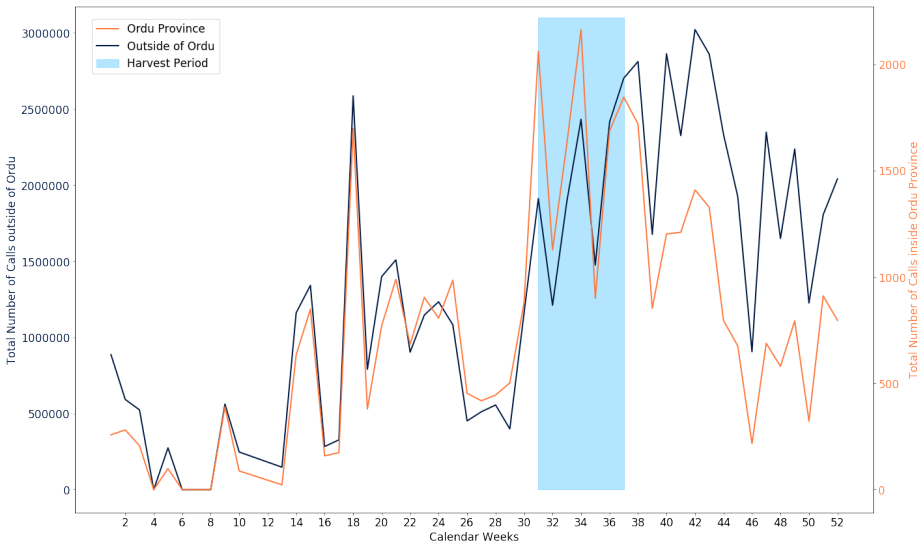


Fig. 1. Weekly number of refugee-tagged calls in Ordu compared to the rest of Turkey

Fig. 2. Hazelnut harvest: Treatment group (blue) and control group (orange). Map of Turkey divided into coverage areas of antennas approximated using Voronoi tessellation.

In order to validate these findings using the DiD approach, we define the beginning of the harvest season (calendar week 31) as our first treatment. Next, we choose our control group based on the assumption that refugees in areas where we can observe a high work-to-home-ratio (the ratio between the network activity during work hours and the rest of the day as defined in the methodology section) are likely to pursue some kind of working activity and are, therefore, less prone towards migrating to Ordu for the hazelnut harvest. Figure 2 illustrates the selected treatment and control group. We then select all antennas that are comparable to our treatment group with respect to the magnitude of the overall network activity.

Next, we check for the common trend assumption. Data from SET1 supports the assumption as the number of calls show similar growth over the weeks prior to the treatment (until week 30). Note that the fluctuations, most notably those in calendar week 14 and 18, are observable in the entire SET1, suggesting that they are due to some unknown factors affecting the behaviour of all Syrian refugees in the sample alike.

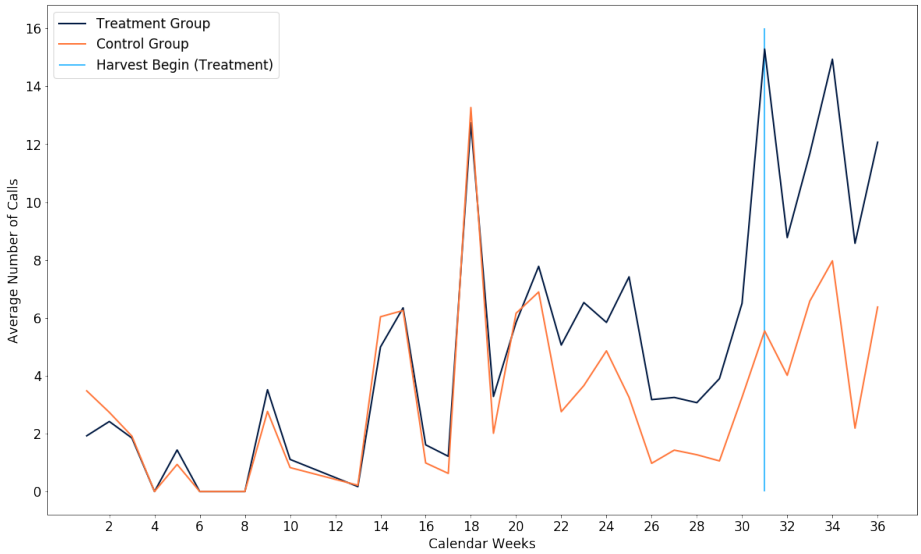


Fig. 3. Treatment I: Weekly number of calls by treatment and control group

Figure 3 indicates that the number of calls made by refugees increased in Ordu with the beginning of the harvest in calendar week 31 compared to areas likely not affected by the treatment.[5] Table 2 gives the treatment effects estimated following formula 2.

[5] Note: The figure shows the weekly total number of calls averaged over all antennas in the control and treatment group. The trends for other network activity measures,

Table 2. Treatment effects of the beginning of harvest on network activity by refugees

Network activity measure	Control group		Treated group		**Absolute TE**	**Relative TE**
	CW 1-30	CW 31-36	CW 1-30	CW 31-36		
No. of SMS	0.46	0.53	0.36	0.70	0.26	59.8%
No. of calls	2.78	5.44	3.75	11.89	5.47	85.3%
Call volume	339	599	327	1327	739	125.6%
No. of interactions	3.24	5.97	4.11	12.59	5.74	83.7%

As expected, the treatment effects for all network activity measures are positive, indicating that the beginning of the hazelnut harvest season in Ordu results indeed in an increase of calls and SMS as well as in the total duration of calls made by refugees. Based on the assumption that refugees living in Ordu prior to the harvest do not significantly increase their communication activities during the summer weeks in question, this spike in activity is likely caused by refugees who migrated to Ordu. An absolute treatment effect of 5.47 for the total number of calls suggests, for instance, that the average weekly number of calls placed by refugees in Ordu increased by 5.47 during the harvest season. However, the absolute numbers have to be interpreted with caution as our sample only represents a small fraction of the total refugees in Turkey. In relative terms this implies that the average weekly number of calls increased by approximately 85% during the harvest relative to the number of calls that would have been placed during calendar week 31 to 36 in Ordu should there not have been the season of hazelnut harvest. Again assuming that refugees living in Ordu prior to the harvest have no reason to increase their network activity during the summer weeks more substantially than refugees living in the control area, this seem to suggest that the number of refugees in Ordu might have increased by roughly 85% due to the hazelnut harvest. The relative treatment effect on network activity measured by the number of SMS or by the total call volume is approximately 60% and 126%, respectively. Consequently, the actual growth of refugee population in Ordu due to the harvest likely lies somewhere within this range. Unfortunately, there are no official figures available on the monthly number of refugees by province to ground-truth these findings.

Interestingly, Figure 3 indicates that there is a slight increase in calls made by refugees in Ordu (blue line) starting in week 21 already. This hints towards an anticipation effect, the so-called Ashenfelters Dip. Refugees seem to have migrated to Ordu even before the harvest started in anticipation of the employment opportunity. Hence, the estimated treatment effect likely underestimates the real effect on network activity due to the hazelnut harvest in Ordu.

To further substantiate our initial findings, we check whether the observed migration during the harvest weeks is only temporary. To this end, we define the end of the harvest (calendar week 37) as our second treatment. We use the same control group as for the previous case. Again, data from SET1 supports

like number of SMS, call volume, or total number of interactions, i.e. SMS and calls, look similar.

the common trend assumption. Figure 4 shows that the number of calls follow a similar trend during the harvest weeks (week 31-36), although with levels of network activity in Ordu (blue line) substantially higher than network activity in the control areas (control line).

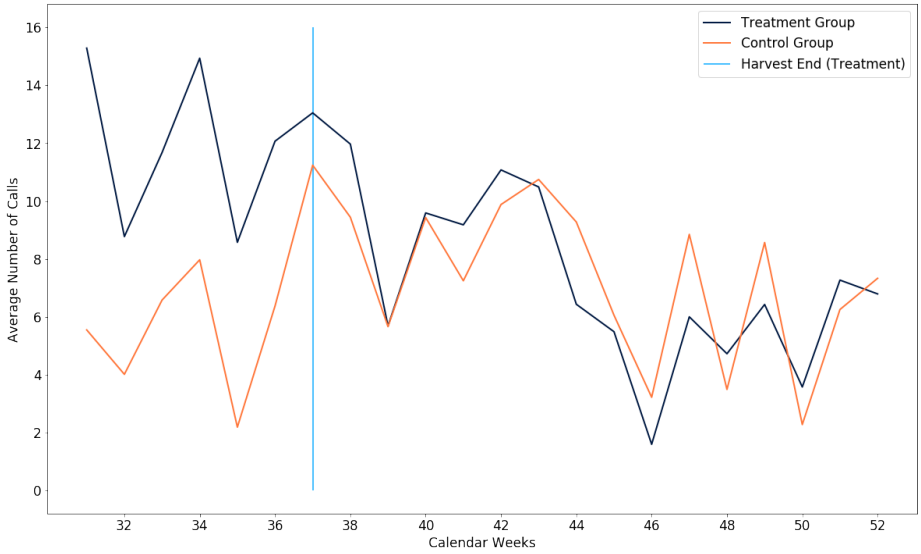


Fig. 4. Treatment II: Weekly number of calls by treatment and control group

The figure also suggests that the level of network activity in Ordu drops to the level of activity in the control areas in week 37, that is at the end of the harvest. These findings are supported by the treatment effects, estimated following formula 2:

Table 3. Treatment effects of the end of harvest on network activity by refugees

Network activity measure	Control group		Treated group		**Absolute TE**	**Relative TE**
	CW 1-30	CW 31-36	CW 1-30	CW 31-36		
No. of SMS	0.53	1.55	0.70	0.59	-1.13	-65.7%
No. of calls	5.44	7.44	11.89	7.53	-6.35	-45.8%
Call volume	599	871	1327	783	-815	-51.0%
No. of interactions	5.97	8.99	12.59	8.12	-7.48	-48.0%

As expected, all difference-in-differences estimators are negative, indicating that the spike in network activity by refugees in Ordu is limited to the summer weeks. These findings further substantiate our hypothesis that refugees migrated to Ordu temporarily during hazelnut harvest. Interestingly, the relative treat-

ment effects due to the end of the harvest are of smaller magnitude than those for the beginning of the harvest (with an exception for the activity measured by number of SMS). While the average number of weekly calls in Ordu increased by approximately 85% due to the start of the harvest it decreased by only 46% due to the end of the harvest. This suggests that part of the refugees that migrated to Ordu to work during the harvest might stay in the region for a longer time. Nevertheless, a substantial part of the increase in network activity is likely to be produced by refugees that migrated to Ordu temporarily.

Table 4 compares the network activity growth from the descriptive statistics with the treatment effect from the DiD analysis. The results provide an early proof-of-concept that the methodological framework laid out above - using descriptive statistics - only give valid indication for harvest-related seasonal migration in Ordu.

Table 4. Hazelnut harvest: Comparing descriptive statistics and treatment effects

	%-change in no. of calls	
	before → during	during → after
Descriptives	77.9%	-54.7%
Treatment effect	85.3%	-45.8%

4.2 Construction site of the Istanbul Grand Airport

Using descriptive statistics only, we find indication of work-related commuting to Istanbul Grand Airport. Table 5 compares the growth rate of the average number of refugee-tagged calls between the hours defined as working hours and home hours for the Istanbul Grand Airport construction site and the rest of Turkey. The difference indicates the airport-specific growth in network activity compared to the rest of Turkey. Around the beginning of work, we observe a growth in network activity that is 966.5%-points higher at the Istanbul Grand Airport construction site than in the rest of Turkey.

Table 5. Growth rate of average number of refugee-tagged calls before, during and after work

Location	**home → work**
Istanbul Grand Airport	1275%
Rest of Turkey	308.5%
Difference	**966.5%**

Similarly to the analysis for Ordu, we first need to identify the treatment for which we choose the beginning of the workday (7:00 am). Consequently, the period prior to the treatment spans the hours from 20:00 pm to 7:00 am. Then,

we select the control group antennas that can be expected to be unaffected by the treatment, i.e., a workday. We argue that refugees living in camps typically do not follow a day-to-day working routine from 7:00 am to 20:00 pm which would affect their network activity. Thus, we choose our control group antennas from all antennas located in the range of refugee camps and select those which are similar to our treatment group (a set of antennas in the range of the airport) with respect to summary statistics.

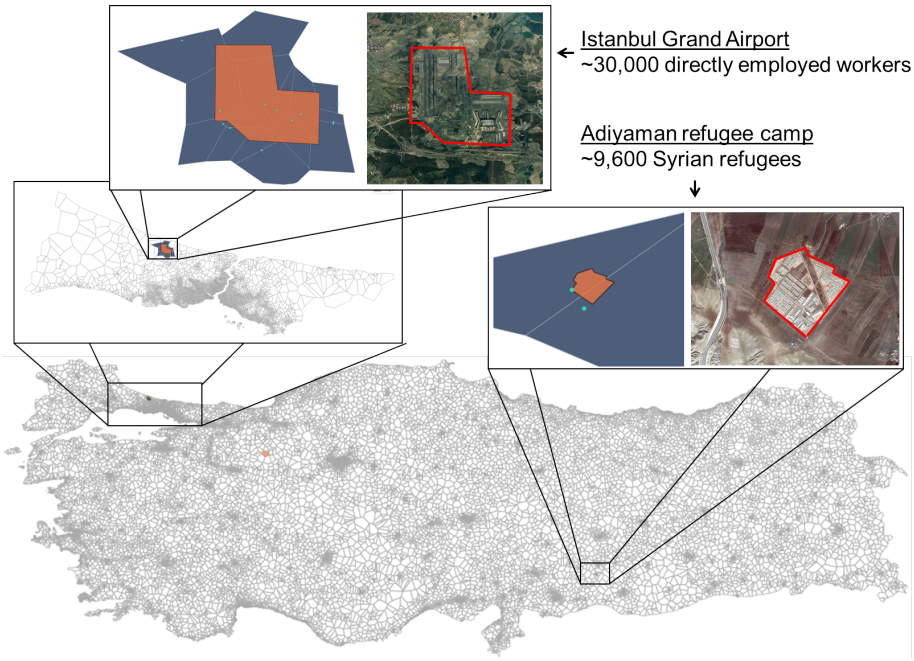


Fig. 5. Istanbul Grand Airport construction site: Treatment group (blue) and control group (orange)

As illustrated by Figure 6, the common trend assumption for the time prior to the beginning of the workday, i.e., for the period from 20:00 to 0:00 and from 0:00 to 07:00, seems to hold.[6]

[6] Note: The figure shows the annual total number of calls in a certain hour averaged over all antennas in the control and treatment group. The trends for other network activity measures, like number of SMS, call volume, or total number of contacts, i.e. SMS and calls, look similar.

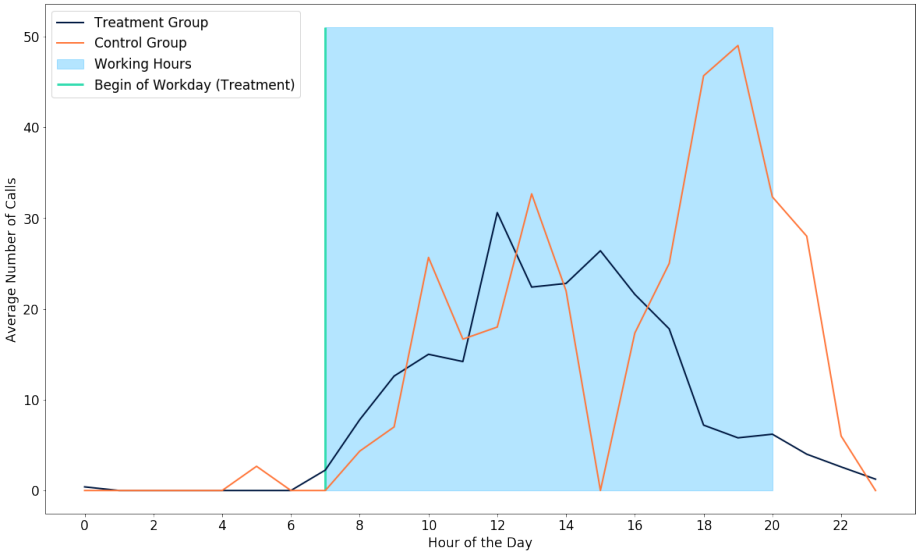


Fig. 6. Treatment III: Hourly number of calls by treatment and control group

We can then calculate the commuting effect following formula 2 and find following treatment effects:

Table 6. Treatment effects of beginning of the workday on network activity by refugees

Network activity measure	Control group		Treated group		**Absolute TE**	**Relative TE**
	CW 1-30	CW 31-36	CW 1-30	CW 31-36		
No. of SMS	0.00	1.05	0.41	1.67	0.21	14.2%
No. of calls	7.39	20.79	1.39	16.09	1.31	8.8 %
Call volume	844	2,284	282	2042	319	18.5%
No. of interactions	7.39	21.84	1.80	17.76	1.51	9.3%

As expected, the results show a positive treatment effect for all four measures of network activity. The treatment effect for calls, for example, is 1.31, suggesting that the average number of calls placed by refugees during the day would decrease by 1.31 in the area of the airport if there would not be a construction site. This interpretation is based on the assumption that there is no other place in the Istanbul Airport construction site area where refugees could find work. Note however, that the actual size of the effect is to be treated with caution as our data set only includes refugee-tagged customers equivalent to 6 % of all registered Syrian refugees (cf Salah et al. (2018)[6]). Nonetheless, the sharper increase of network activity during the day in the construction site area compared to the more moderate increase in the area of refugee camps suggests that there might indeed be a commuting pattern towards the construction site during the day,

hinting towards undeclared labour by refugees at the construction site of the Istanbul Grand Airport. The relative treatment effect for the number of total calls is almost 9%. Hence, the number of refugees that commute to the airport to work at the construction site during the day amounts to approximately 9% of the number of refugees living in the area of the Istanbul Grand Airport. The relative treatment effects for the alternative measures of network activity suggest that this number could likely even be higher. This interpretation, however, builds on the assumption that the fact that a person works does not affect it's network activity. This links to the quality of network activity as a proxy for user counts. The assumption may not hold true as phone use may depend on the profession (e.g. manager vs. construction worker). Consequently, the commuting effects are likely underestimated as construction workers may use their phones less at work than their unemployed alter egos.

Table 7 compares the network activity growth from the descriptive statistics with the treatment effect from the DiD analysis. Even though the results show the same direction, the strength of the effects vastly varies. Here again, this is most likely due to the small sample size as it only includes refugee-tagged customers equivalent to 6 % of all registered Syrian refugees (cf Salah et al. (2018)[6]). Hence, at the construction site a slight change in the total network activity has large effects on the growth rates between home and working hours. While this also holds true for the control group of the DiD analysis, it is not the case for the rest of Turkey as it is used for the descriptives. In order to evaluate the early proof-of-concept for the case of the Istanbul Grand Airport, a larger sample of refugee-tagged CDRs is necessary.

Table 7. Istanbul Grand Airport construction site: Comparing descriptive statistics and treatment effects

	%-change in no. of calls home → work
Descriptives	966.5%
Treatment effect	8.8%

4.3 Exploration

Exploratory analysis may help to detect areas that have a high probability of hosting undeclared employment opportunities for refugees. With a view on impact, this might help to mitigate undeclared work with all its negative consequences by informing effective intervention planning.

Figure 7 shows areas of similar network activity as the Istanbul Grand Airport construction site - low activity during the night, high activity during the day. We identify 60 locations in Turkey with these characteristics and similar magnitude. Comparing the results with information from Google Maps provides a mixed picture of the exploratory approach. While we are able to identify areas

with a high probability of refugee employment such as the industrial park near Çerkezköy (see Figure 7), the approach also produces a high rate of potentially false positives. For example, one of the places identified is Kapalıçarşı, the Grand Bazaar in Istanbul. Here, it is not possible to say whether the activity is caused by refugee-tagged visitors or refugee-tagged workers at the market. Other false positives include shopping malls and hospitals. This shows a fundamental weakness of the exploratory approach: SET1 does not provide information on the duration and regularity of stay derived from individual-level information that could help to differentiate between visitors and workers. Individual-level mobility data as in SET2 can be used for this, however, the sample size of SET2 does not allow for such fine-granular evaluation.

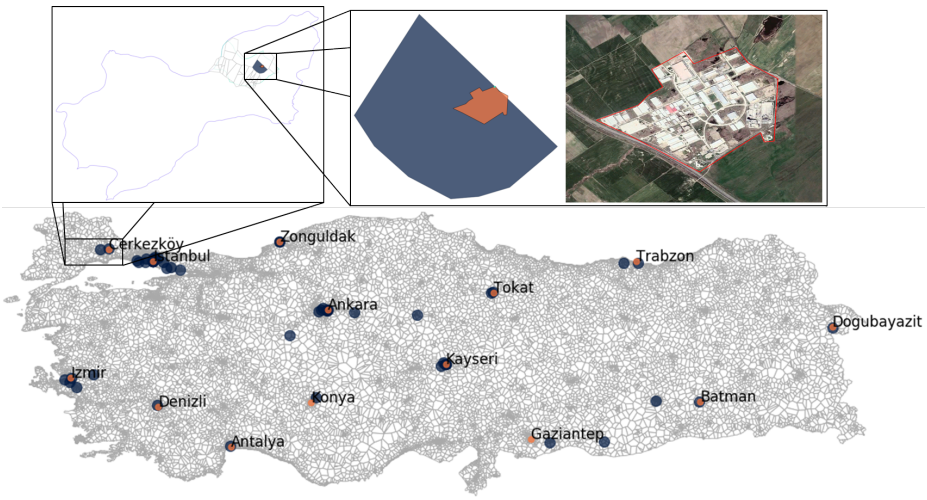


Fig. 7. Antennas (blue) with similar network activity as the Istanbul Grand Airport area. High probability of refugee employment: industrial park near Çerkezköy.

5 Limitations

In general, socio-economic statistics derived from mobile phone metadata carry certain shortcomings. In contrast to statistical data, the data generating process behind mobile phone metadata is usually beyond the control of the analyzing entity and, thus, might not adhere to relevant statistical concepts such as representativeness. The mismatch between the unit of analysis (the individual) and the unit of observation (the device) adds to this as one individual can have multiple phones and vice versa. Further, there are some additional limitations specific to this case study concerning the data, the methodology and the proof-of-concept.

5.1 Data

The dataset at hand cannot be considered as a representative sample for all Syrian refugees in Turkey: the share of refugee-tagged customers in the customer base p_{ref} in SET1 does not match with the actual shares of registered Syrian refugees P_{ref} in the respective provinces (see Table 8 below). This makes it impossible to use CDR information for the estimation of refugee shares by province or other geographical levels. Also other indicators that could be used for representative sampling, such as gender or rural/urban, do not seem to be considered (cf Salah et al. (2018)[6]). Consequently, generalizations of findings based on this sample may not be adequate.

Table 8. Composition of CDR sample

City	N	P_{ref}	n	p_{ref}
Istanbul	15,283,671	3.1%	447,510	18.8%
Gaziantep	2,303,914	14.3%	95,553	15.6%
Izmir	4,332,434	2.5%	50,926	20.5%
Mersin	1,920,783	7.6%	50,280	20.0%
Sanliurfa	2,361,159	17.8%	50,022	19.4%
Hatay	1,939,189	19.8%	47,418	14.8%
Ankara	5,419,716	1.4%	46,024	12.1%
Konya	2,234,748	3.3%	45,106	10.5%
Bursa	3,008,289	3.6%	43,838	7.9%
Antalya	2,328,956	0.0%	43,247	6.7%
Adana	2,352,465	6.4%	43,234	6.5%
Rest of Turkey	39,299,246	1.8%	25,537	99.7%
Others	–	–	3,762	99.8%
Total	**82,784,570**	**3.6%**	**992,457**	**18.6%**

5.2 Methodology

- Refugee-tag: The refugee-tag is a handy way to run analyses disaggregated by refugee status. However, the tag is dependent on business model decisions of the mobile network operator and thus may prove unreliable with the view on implementation.
- Assumption: The main assumption justifying the link between work-related commuting and undeclared employment is the fact that only 1.3% of the Syrian refugees have been granted a work permit up to 2017. This might change in the future, which would negatively affect the power of the approach.
- Descriptive statistics: SET1 shows unexpected strong growth of interactions over the year. While the reason for this may be found in the CDR extraction, it influences the descriptive statistics on migration by overstating migration growth numbers over time.

– Net migration: This study uses indicators related to network activity as a proxy for the number of users in the network. Thus, changes in those variables are not path-dependent and therefore may only inform on net migration. However, net migration is a result of complex in- and out-flows, which eventually understate the true migration through cancelling out.

5.3 Proof-of-Concept

While a DiD approach is used to disentangle effects showcased in the descriptive statistics, the study does not use additional variables to control for other potential influences on commuting/migration patterns such as climate- or geography-related aspects.

6 Conclusion

This study has laid out a framework for identifying potentially undeclared employment among refugees in Turkey. Further, it has provided an early proof-of-concept based on a Difference-in-Differences approach using two case studies: the hazelnut harvest in Ordu in late-summer and the construction site of the Istanbul Grand Airport. We have found clear indication for work-related commuting and seasonal migration patterns among refugees based on which undeclared employment situations can be detected. By informing effective intervention planning, fine-grained information about undeclared employment situations may help to fight undeclared work with all its negative implications: high-risk jobs, pay below minimum wage and lack of access to social security. Limitations of the study such as CDR sampling and applied assumptions have also been discussed. Finally, the study solely uses antenna traffic and other publicly available sources in order to mitigate potential privacy issues and facilitate uptake.

References

1. Almaatouq, A., Prieto-Castrillo, F., Pentland, A.: Mobile communication signatures of unemployment. In: International conference on social informatics, pp. 407–418. Springer, (2016)
2. De Nadai, M., Staiano, J., Larcher, R., Sebe, N., Quercia, D., Lepri, B.: The death and life of great Italian cities: a mobile phone data perspective. In: Proceedings of the 25th international conference on world wide web, pp. 413–423. International World Wide Web Conferences Steering Committee, (2016)
3. Dong, L., Chen, S., Cheng, Y., Wu, Z., Li, C., Wu, H.: Measuring economic activity in China with mobile big data. EPJ Data Science **6**(1), (2017)
4. Fair Labor Association: Integration of Syrian refugees under temporary protection into the Turkish labor market: Challenges and Opportunities. Roundtable - Summary and Outcomes Report. December 2, 2016.
5. Pokhriyal, N., Jacques, D. C.: Combining disparate data sources for improved poverty prediction and mapping. In: Proceedings of the National Academy of Sciences **114**(46), pp. 9783–9792, (2017)

6. Salah, A.A., Pentland, A., Lepri, B., Letouz, E., Vinck, P., de Montjoye, Y.A., Dong, X. and Dadelen, .: Data for Refugees: The D4R Challenge on Mobility of Syrian Refugees in Turkey. arXiv preprint arXiv:1807.00523, (2018)
7. Schmid, T., Bruckschen, F., Salvati, N., Zbiranski, T.: Constructing sociodemographic indicators for national statistical institutes by using mobile phone data: estimating literacy rates in Senegal. In: Journal of the Royal Statistical Society: Series A (Statistics in Society) **180**(4), pp. 1163–1190. Wiley Online Library, (2017)
8. Sundsøy, P., Bjelland, J., Reme, B.-A., Jahani, E., Wetter, E., Bengtsson, L.: Towards Real-Time Prediction of Unemployment and Profession. In: International Conference on Social Informatics, pp. 14–23. Springer, (2017)
9. Toole, J. L., Lin, Y., Muehlegger, E., Shoag, D., Gonzalez, M. C., Lazer, D.: Tracking employment shocks using mobile phone data. Journal of The Royal Society Interface **12**(107), (2015)
10. Xu, Y., Belyi, A., Bojic, I., Ratti, C.: Human mobility and socioeconomic status: Analysis of Singapore and Boston. Computers, Environment and Urban Systems, (2018)

Mobile Data for Mobility: Travel and Communication Patterns of Syrian Refugees

Ervin Sezgin (ORCID ID: 0000-0002-0924-3346)
Istanbul Technical University, Istanbul, Turkey (sezginerv@itu.edu.tr)

Eda Beyazıt (ORCID ID:0000-0002-5526-501X)
Istanbul Technical University, Istanbul, Turkey (beyazite@itu.edu.tr)

Kerem Arslanlı (ORCID ID:0000-0002-6480-5727)
Istanbul Technical University, Istanbul, Turkey (arslanli@itu.edu.tr)

Mehmet Gencer (ORCID ID:0000-0003-1717-8668)
Izmir University of Economics, Izmir, Turkey (mehmetgencer@yahoo.com)

1. Abstract

This research analyses the mobility of Syrian refugees within Turkey. It uses the Coarse Grained Mobility data (Dataset 3) provided by Turk Telekom D4R contest to understand the geographical patterns of refugee mobility and to associate it with the social context.
The research focuses on a sample of unique callers and traces their mobility trajectory throughout 2017. It analyses the number of provinces they visited, and the number and frequency of their visits. Then it links refugee mobility to socio-economic data to understand the external factors influencing the mobility patterns.
Research results suggest that only a limited percentage of Syrian refugees could be considered as mobile (receiving calls in more than one province) and most of them moving only between two and three provinces. Social events, such as Eids increase their mobility, mostly to provinces close to each other.
It is suggested to decision makers to consider transportation policies that increase the mobility of refugees as a part of the integration policies. Particularly, strengthening the cooperation between provinces having high levels of physical interaction may contribute in integration policies. This will connect refugees to their social networks and resources, increase their resilience and strengthen social integration.
The research develops a flexible method that changes the resolution of the analysis from the individual to the whole and combines quantitative analysis of large data with qualitative methods. Although the analysis in this study is limited; an extended analysis with the methodology can point to the geographical distribution of mobilities and social networks of refugees and how these can be mobilized for increasing not only refugees' but also general prosperity.

2. Introduction

Asylum and refugee travel, as Urry [1] describes, is one of the twelve major forms of travel practice in the contemporary world. After a highly risky, complex and expensive escape to places sometimes with

inadequate hospitality, refugees face unequal access to foreign spaces compared to for example business and professional travellers who change their locations at great comfort.

There are around 70 million people in the world forcibly displaced from their home countries, 28.5 million of which are refugees and asylum seekers [2]. Although this equals four percent of the world's population, the escalating numbers demonstrate the degree of an ever-growing crisis, spilling over the border countries of the conflict zones. Considering the climate-oriented refugee crisis at the door with 140 million respective refugees by 2050 [3], inevitable mobility, or even worse the immobility, will become the utmost problem of the world.

Turkey is the top-refugee hosting country with 3.5 million registered Syrian refugees [4] and more thought to be unregistered due to various reasons [5,6]. Although Turkey has not yet granted millions of Syrians who fled their countries with refugee status[1], a situation leading to informal working conditions, exploitation of labour, diminished access to health services and alike, their mobility within the country is not restricted to camps. Unlike the beginning of the Syrian civil war in April 2011 when hundreds of civilians rushed to the Turkish border, followed by thousands, who were settled in refugee camps by large while smaller numbers rented apartments in nearby towns, currently the majority live in cities around Turkey. While Hatay, Gaziantep, Sanliurfa ve Istanbul host more than 300.000 refugees, Izmir, Bursa, Konya, Mersin and Adana are housing more than a hundred thousand Syrians [4] (Figure 2.1).

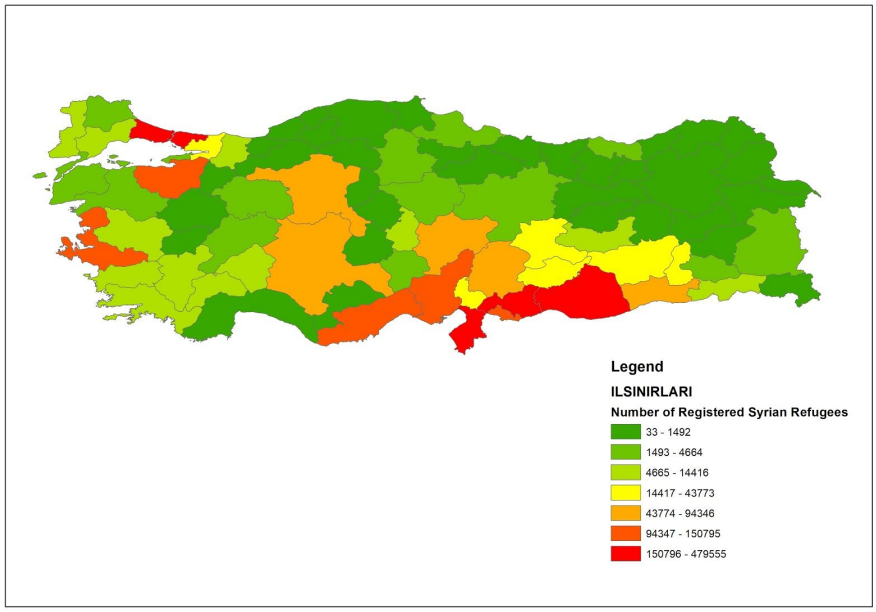


Figure 2.1 Number of Registered Syrian Refugees

Access to job market is an important determinant of residential location for those who leave the camps for better living conditions. While some refugees work in farms or factories as seasonal workers around South-Eastern cities, those who are economically more secure rent apartments in cities like Istanbul where they work in trade and tourism [6]. Some with stronger social networks set out their arrival strategies by

[1] note: temporary protection - may lead to permanent settlement

staying at friends' and relatives' houses across the border first, and then moving towards larger cities (ibid.) As such, Syrians refugees are quite mobile within Turkey.

Mobility emerges as one of the key factors to understand and improve the wellbeing and the level of social integration of refugees. Refugees need to be mobile, whether to reach to employment opportunities, find shelter or connect with their relatives. Moreover, refugee mobility, whether it is one-time or frequent, points to the spatial aspect of their survival strategies. Questions such as how far they need to commute everyday; where do they go to celebrate feasts; how far they can reach in case of need, not only indicate the 'new life' of refugees in their host country, but also to the possible policy areas and action frameworks to improve mobility, strengthen social networks, hence the level of integration.

Integration of Syrian refugees into Turkish society is a crucial step towards combating with refugee crisis around the world as exceeding number of people flee from their home countries every year. As host to the largest number of refugees, Turkey may set an example to other countries with its integration policies fit both for newcomers and local populations.

3. Method

This research aims to investigate the degree and scope of these mobilities. To be specific, it explores the content of these mobilities by asking why, when and where Syrian refugees move within Turkey. It explores whether mobillities have a pattern and if so, what type of temporal, spatial and political elements determine these patterns. It does this by employing a range of quantitative analyses within a spatio-temporal framework in order to answer the 'when and where' questions based on the aggregate data. Yet, understanding the dynamic nature of mobilities of refugees necessitates the use of disaggregate data. Thanks to available data allowing to trace the callers based on their locations, the dynamic state of the refugees is revealed. Finally, in order to answer the 'why' question this paper relies largely basic statistical data on the socio-economic status of origins and destinations of mobility, as well as official reports, media documents and previous studies regarding the mobility of Syrian refugees in Turkey.

Using mobile data in order to understand mobility has been a common practice in academia within the last decade [7, 8]. Mobile phones help gathering important data especially in places where data collection may be troublesome and/or complicated. GPS data acquired from mobile phones help identifying the mobility patterns of individuals at high spatial and temporal resolutions which is not always possible to collect through travel diaries, the conventional data collection method of transport studies. In this sense, the data provided by Turk Telekom D4R contest widens the scope of mobility research in Turkey, especially in terms of a highly timely issue such as the mobility of Syrian refugees.

Turk Telekom provided mobile communication data of 50.000 Syrian refugees and 50.000 Turkish residents. Three datasets were distributed as a part of the contest, providing detailed but anonym information regarding the mobile communication of the sample [9]. This study uses the Dataset 3 ("Coarse Grained Mobility") in order to understand the refugee mobility. This dataset gives the rough location (districts and provinces) and the exact time of the phone calls, disaggregated to unique caller IDs.

Dataset 3 [9] is processed in the following way: First we define Caller_IDs as our unit of analysis, the movements of which could be traced through various calls made from different locations. We isolated the refugee Caller_IDs (Caller_ID number starting with "1"[2]) for this purpose.

Second, we broke down total mobility into two sets of factors. These include time (the months in which mobility increases or decreases) and geography (calls made from each city, the number of calls per Caller_ID from each city, and count of number of cities calls are made in a month and year). This analysis gives a general overview of refugee mobility in Turkey over 2017.

Third, we analyse a 1% sample[3] of Caller_IDs that made calls from at least two different cities during 2017. To process large amount of data, only incoming calls from Dataset 3 were extracted, since at this stage we were not interested in the volume of calls, but only the location of Caller_IDs.

The analysis includes the number of provinces visited by Caller_IDs, how many times they are visited and the frequency of visits. This analysis indicates unique trips as well as patterns of commuting.

Finally, to explain refugee mobility, the identified patterns and unique trips were matched with external factors, including the socio-economic development level of provinces; particular dates and events (such as eids).

4. Analysis

Dataset 3 contains a total of 105.423 unique Caller_IDs, 69.270 of which belong to refugees. Approximately 90% of all refugees did not made or received calls in a different city resulting in a highly immobile pattern.

Looking at the temporal patterns of total refugee mobility (Figure 4.1), received calls peak in September which is the month that corresponds with the Eid (31 August-4 September 2017). Taking out September as an outlier with more than 9.000.000 calls, total number of monthly incoming calls range between 50.000-100.000 from January to March, and increase to 200.000 - 300.000 range in the remaining part of the year. The largest ratio of calls originate from Istanbul, Gaziantep, Izmir, Ankara, İcel, Bursa, Hatay, Antalya, Konya, Adana, Sanliurfa and Kilis; the cities, where the majority of Syrian refugees reside [4].

[2] http://d4r.turktelekom.com.tr/presentation/data

[3] It is recognized that 1% sample (100 Caller_IDs) is not sufficient to draw satisfactory conclusions. However the randomly selected sample provides a spectrum wide enough to apply the methodology and demonstrate the potential of the study. The sample couldn't be increased due technical reasons.

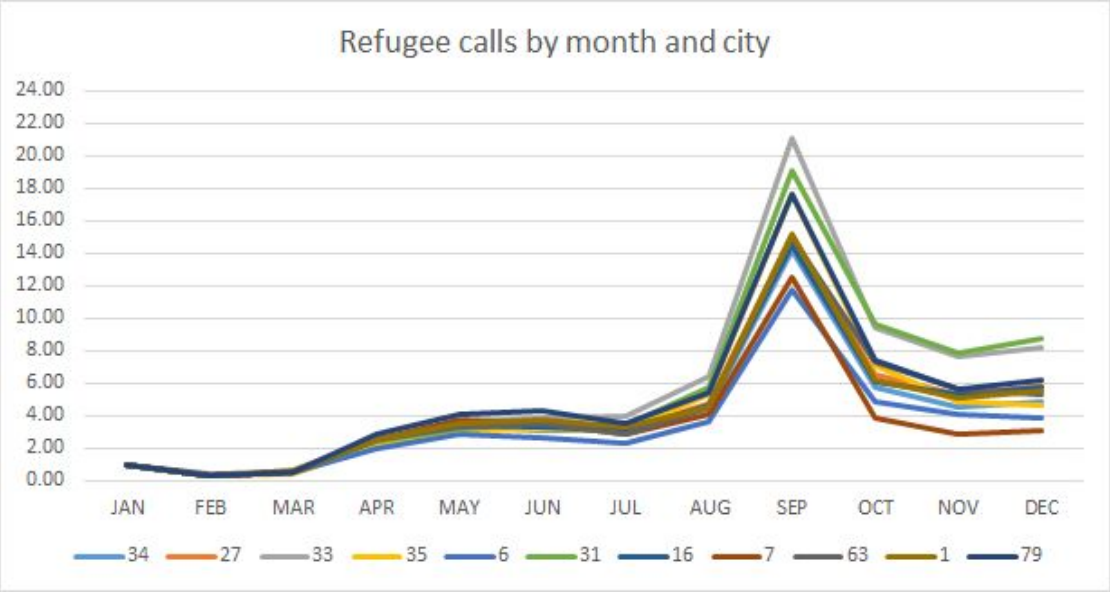


Figure 4.1 Refugee Calls by Month and City

By using the Coarse Grained Mobility information given in Dataset 3, we can identify the mobility levels of individuals in order to obtain a mobility pattern. By doing that we could develop an understanding of the content of their mobility after including socio-spatial and economic structure of these locations. However, this is a highly complicated task. An example of an individual's yearly mobility can be found in Table 4.1 (below).

2800427393	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41
January																																									
February																																									
March																																									
April																																									
May			1		4									20				3	5									7						125			1			5	
June	30	10	17			4					10			2		13										2	26			1		6	100							17	
July	2		4		1	13			4	12	4			4		4		2	4	11	10		2			11		2				10	38	21		2	2			46	
August	22		14		4	2	30			15	1			6		12		3	6	6				1	5	10	8	10	2			22	13	30						31	
September	147		40		8	13	60			31	2			15		29		6	12	12	9			2	10	25	19	20	4		1	57	44	113						76	
October	5		8		7		15				1			6				6	2		8			2	9	7	14		3			5	23			3				12	
November	18		19		3	9	3		2	2	2			5		3	15	1	6	2				3	9	1		7	6			3	10	27		3		9		15	
December	48	1	36		11	57			3		8			10				4	6	9			13	3	22	32		7	1			12	132	2		3				31	
Total	272	11	139	0	38	98	108	0	9	60	28	0	0	68	0	61	15	25	41	40	27	0	15	11	55	88	67	53	16	0	2	0	115	485	193	0	12	2	9	0	233
	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	total
January	193																																								193
February	79																																								79
March	57																																								57
April	342																																								342
May	132										5		4	30				186		12																	2			1	543
June	118	1			3					7			6					66				6					3											6	1	455
July	68		1	8					1	5	2		2	10				18		20		6		2		8			1						3				2	366
August	103			7						4	5		2	6			2	19	2	19	1		21				1	6								1		1		453
September	316			50	1	11				11	10		7	12			4	38	4	38	2	1	52				1	2		12						2		14	3	1346
October	232					11				1			2	2			1	6	1									2										1		395
November	100			43							5		1	20				17	1	11	1		6			1		3		3									1	396
December	112	5	14	47	9					10	1		12	12			10	12	2			4	13		1		2	1		7							1		2	5	729
Total	1852	6	15	155	13	22	0	0	1	38	28	0	36	92	0	0	17	362	10	100	4	13	92	0	3	1	14	9	0	29	0	0	0	0	0	3	6	0	24	13	5354

Table 4.1 A highly mobile non-refugee's annual mobility pattern showing the number of calls received in each city

In order to manage such a complicated task in a limited time with limited software, a random selection of refugees has been made. A hundred refugees (1% of all refugees with trips between 2 or more cities) have been randomly assigned for an in-depth investigation of their mobility patterns.

Spatial structure of refugee distribution is analysed by Morans' Index of spatial autocorrelation measure. where N is the number of spatial units indexed by i and j; × is the variable of interest; × is the mean of ×; wij is a matrix of spatial weights with zeroes on the diagonal (i.e.,w=ii); and W is the sum of all wij.

$$I = \frac{N}{W} \frac{\sum_i \sum_j w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2}$$

Figures 4.2 and 4.3 provide the statistic details of the data sample.

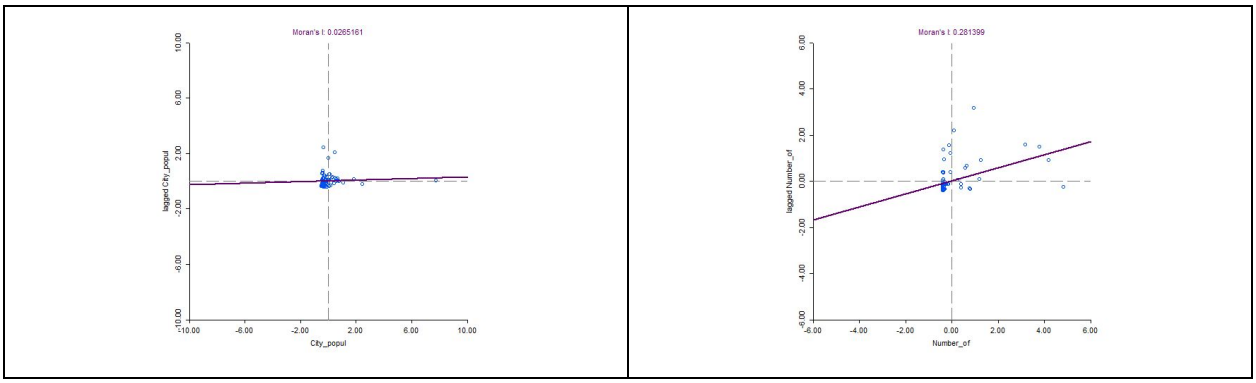


Figure 4.2. Morans' Spatial Autocorrelation of City Population vs Refugee Distribution

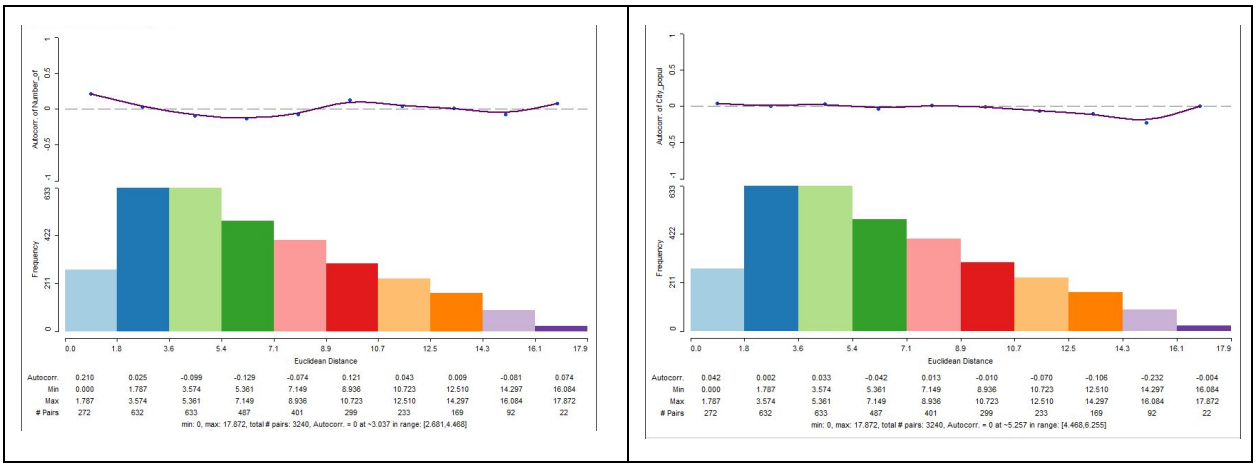


Figure 4.3. Spatial Autocorrelation Distribution of Neighbourhood City by Distance

Positive spatial correlation has been detected in the data layout. In order to correct the correlation Spatial Error Model has been employed on the 2 stage regression analysis. Lambda has been found as positive with significant level of 0.001. and the model robust White test probability over 0.09 shows the clearance of spatial autocorrelation in the residual terms (Table 4.2).

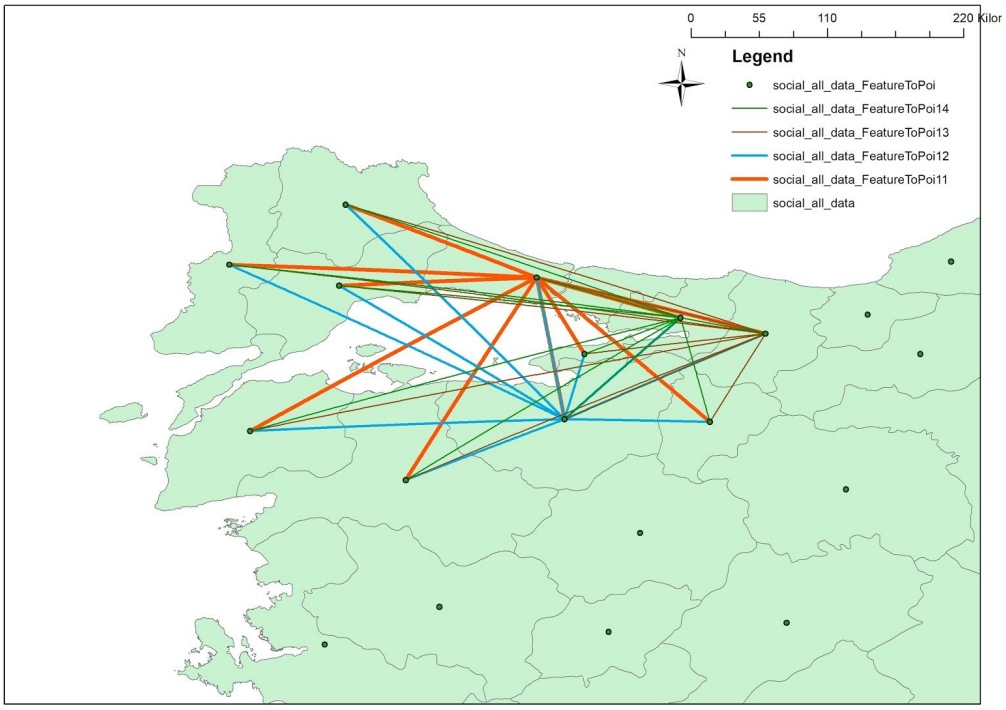
Regression Diagnostics Test on Normality of Errors			
Test	Df	Value	Prob
Jarque-Bera 2	2	117.1433	0.00000***

Diagnostics for Heteroskedasticity Random Coefficients			
Test	Df	Value	Prob
Breusch-Pagan test	10	115.2305	0.00000***
Koenker-Bassett test	10	31.2067	0.00054***

Specification Robust Test			
Test	Df	Value	Prob
White	65	80.2325	0.09657

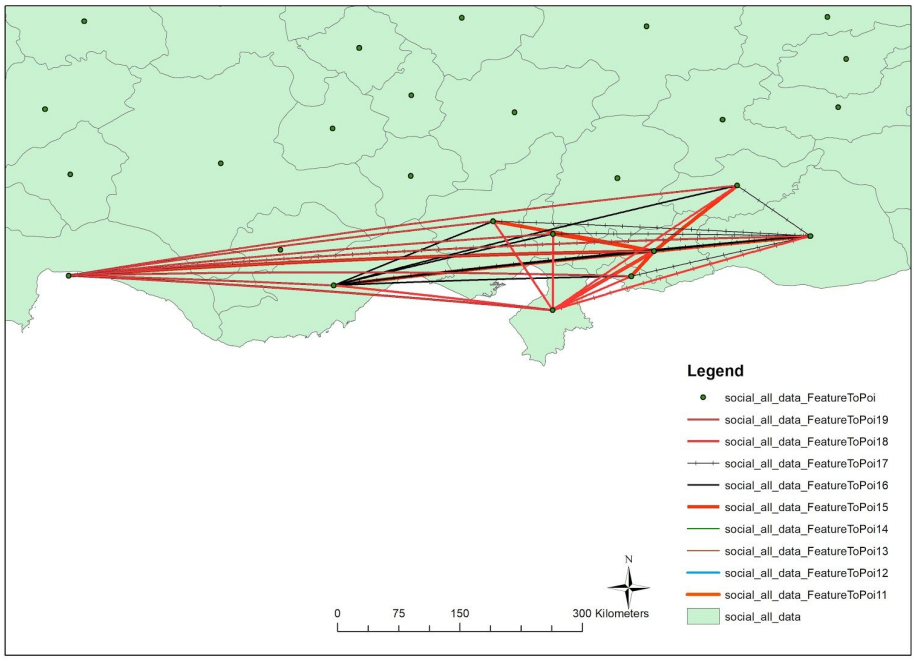
Table 4.2: Spatial Autocorrelation Tests for Model Residuals

63% of the sample has commuted to only one city during 2017. Among the commutes 31 (49%) either originated from Istanbul, or were destined to Istanbul. Here İstanbul - Bursa trips with 9 repeats are the most observed ones. They are followed by Istanbul - Sakarya (4 times) and Istanbul - Ankara (3 times). İstanbul - Kocaeli is observed 2 times, however when three city trips are included Istanbul-Kocaeli-Yalova is observed 3 more times. In general, commuting patterns suggest that, the range of the two city commutes is very limited, overwhelmingly to/from the cities in the Marmara region (Map 4.1). Even commuting to other two major cities (Ankara and Izmir) and to other refugee centres (Gaziantep, Şanlıurfa, Hatay) are not observed. There is no significant commuting between the cities in the Marmara region, either. As such Istanbul remains the centre of Marmara region.



Map 4.1 Commutes in Marmara Region (the map is produced from the whole Data Set 3)

The analysis of the two-city commutes proceeds with the other major refugee destinations, Gaziantep, Şanlıurfa and Hatay. Here Gaziantep is the centre of commuting, again limited to its geographical region. Commuting between Gaziantep and Kilis (5 times) is followed by Şanlıurfa (3) and Hatay (1) and Istanbul (1). Contrary to Marmara region, in Southeastern Anatolia, commuting between other cities is observed, too. Hatay-Kilis, Hatay-Osmaniye are the other two city commuting partners. However trips that include more than two cities increase within the region commuting significantly (Map 4.2).



Map 4.2 Commutes in Southern Turkey (the map is produced from the whole Data Set 3)

The analysis of travels that include more than two cities requires a more detailed analysis, because of the risk of including calls from cities that are passed by while traveling to another place. Therefore, the analysis of these trips is limited to following observations: Istanbul is included in 21 of 37 (57%) more than two city trips. This is followed by Ankara (12), Adana (8) and Gaziantep (6). As a general observation, these trips' range is not exclusively limited to provinces within the same region, and reach to different parts of the country.

To increase the precision, we now concentrate to the commutes in the Marmara region that include Istanbul and Bursa (the two provinces in our sample, between which, the refugees commute the most). Table 4.2 shows the 12 Refugee callers that commuted between Bırsa and Istanbul in our sample. Each sub table shows how many calls a unique caller made from that particular city (34 for Istanbul and 16 for Bursa), in that particular month of 2017. Especially the refugees that commute only between these two cities traveled almost exclusively between May and October, the summer months. Also most of the Callers made fewer calls (between 1-10) from Bursa comparing to their calls from Istanbul. These combined, may lead us to suggest that commutes between these two cities are holiday trips or relative visits and there are no business relationships, which require frequent calls from both of these cities.

	Row Labels	3	6	14	16	23	26	33	34	41	42	44	50	64	68	81	Grand Total
JAN	1.....5								23								23
FEB	1.....5								4								4
MAR	1.....5								5								5
APR	1.....5								45								45
MAY	1.....5								125	7							132
JUN	1.....5							17	128								145
JUL	1.....5				7				95							1	103
AUG	1.....5		1		19				161		9					1	191
SEP	1.....5		2	1	38	32		20	397		22	4	1	6	2	2	527
OCT	1.....5	1					1		11		81						94
NOV	1.....5										55						55
DEC	1.....5								25		108						133

	Row Labels	1	16	33	34	35	42	51	68	Grand Total
FEB	1.....0			4						4
MAR	1.....0		1							1
APR	1.....0		22		6	6				34
MAY	1.....0		38		6					44
JUN	1.....0		72		22					94
JUL	1.....0		15							15
AUG	1.....0	4	3	119			1	1	3	131
SEP	1.....0	9	15	282			2	2	6	316
OCT	1.....0		1							1
NOV	1.....0		18							18
DEC	1.....0		24							24

	Row Labels	16	34	Grand Total
AUG	1.....8		4	4
SEP	1.....8	18	53	71
OCT	1.....8		44	44
NOV	1.....8		30	30
DEC	1.....8		172	172

	Row Labels	16	34	Grand Total
JAN	1.....9	1	25	26
FEB	1.....9		27	27
MAR	1.....9		37	37
APR	1.....9		86	86
MAY	1.....9		75	75
JUN	1.....9		99	99
JUL	1.....9		46	46
AUG	1.....9		69	69
SEP	1.....9		247	247
OCT	1.....9		76	76
NOV	1.....9		86	86
DEC	1.....9		85	85

	Row Labels	16	34	77	Grand Total
JAN	1.....15	14			14
FEB	1.....15	6			6
MAR	1.....15	9			9
APR	1.....15	70			70
MAY	1.....15	20			20
JUN	1.....15	60		3	63
JUL	1.....15	30			30
AUG	1.....15	24			24
SEP	1.....15	101			101
OCT	1.....15	56			56
NOV	1.....15	66	1		67
DEC	1.....15	59			59

	Row Labels	16	34	Grand Total
MAY	1.....18	1	14	15

	Row Labels	16	34	Grand Total
JUL	1.....2	1	3	4

	Row Labels	16	34	Grand Total
JUN	1.....4	2	21	23

	Row Labels	16	34	Grand Total
JUN	1.....0	4		4
JUL	1.....0	9	6	15

	Row Labels	6	16	34	54	Grand Total
APR	1.....82		1	3		4
MAY	1.....82			5		5
JUL	1.....82	2		6	5	13
AUG	1.....82			14		14
SEP	1.....82			28		28

	Row Labels	16	34	Grand Total
oct	1.....58	1	16	17

	Row Labels	16	34	Grand total
SEP	1.....75	7	2	9

Table 4.2 Commutes between Bursa and Istanbul

Now, the resolution of the analysis will be increased to get a clue on the Istanbul-Bursa commutes. We will concentrate on the calls of Ms 1…….8's in February, 2017 and her Bursa-Istanbul-Bursa trip that extends over four days in this month (Table 4.3). This time we include the districts she moves, too. Off course, such analyses are not generalizable, but they allow us to escape of what may be called the data trap (i.e. to treat the human beings as numbers) and relate generalized data analysis to real life trajectories. Accordingly, Ms 1…….8's lives in Osmangazi, Bursa, from where she made most of her calls in this month. She occasionally goes to Nilüfer, Bursa (perhaps she has a part-time work there, or a friend to visit). On 2nd of February she catched the 4 O'clock ferry from Monday to Istanbul. She arrived at Fatih (where the pier is) and stayed at this district overnight. On the following two days she stayed at Fatih and payed long visits during the day of 3rd February to Zeytinburnu, Istanbul. On the fourth, she is back to Osmangazi, where she will continue with her routine. Given that 3-4 February is weekend, we can suggest that this was a relative visit and Ms 1…….8's social networks are not limited to only one district in

Istanbul, but two. Similar analyses that focus on working-day routine trips most probably will point to the trade relationships and economic networks of refugees.

CALLER_ID 1.....8							
TIMESTAMP	CITY_ID	ID (prefecture)		TIMESTAMP	CITY_ID	ID (prefecture)	
02-02-2017 13:16	16	70	OSMANGAZİ	04-02-2017 14:18	34	386	
02-02-2017 13:31	16	70		04-02-2017 14:22	34	386	
02-02-2017 14:01	16	70		04-02-2017 15:13	34	386	
02-02-2017 15:29	16	826	NİLÜFER	04-02-2017 16:12	77	425	ARMUTLU
02-02-2017 15:33	16	70	OSMANGAZİ	04-02-2017 16:27	77	425	
02-02-2017 15:51	16	67	MUDANYA	04-02-2017 17:37	16	826	NİLÜFER
02-02-2017 16:04	16	67		04-02-2017 17:43	16	70	OSMANGAZİ
02-02-2017 16:35	16	67		04-02-2017 17:46	16	70	
02-02-2017 17:05	77	425	ARMUTLU	04-02-2017 18:28	16	70	
02-02-2017 17:13	77	425		04-02-2017 18:29	16	70	
02-02-2017 17:46	34	644	AVCILAR	04-02-2017 19:19	16	70	
02-02-2017 18:36	34	386	FATIH	04-02-2017 19:27	16	70	
02-02-2017 18:55	34	249	ZEYTINBURNU	04-02-2017 20:29	16	70	
02-02-2017 21:04	34	249		04-02-2017 20:42	16	70	
02-02-2017 21:23	34	386	FATIH	04-02-2017 20:51	16	70	
02-02-2017 21:32	34	386		04-02-2017 21:04	16	70	
02-02-2017 21:43	34	386		04-02-2017 22:09	16	70	
02-02-2017 21:53	34	386		04-02-2017 22:19	16	70	
02-02-2017 21:57	34	386		27-02-2017 11:02	16	70	
02-02-2017 22:40	34	386		27-02-2017 11:11	16	70	
03-02-2017 00:36	34	386		27-02-2017 11:19	16	70	
03-02-2017 14:15	34	386		27-02-2017 11:42	16	70	
03-02-2017 14:40	34	386		27-02-2017 14:29	16	70	
03-02-2017 14:48	34	386		27-02-2017 14:34	16	70	
03-02-2017 14:52	34	386		27-02-2017 14:51	16	70	
03-02-2017 16:17	34	386		27-02-2017 16:09	16	826	NİLÜFER
03-02-2017 16:50	34	249	ZEYTINBURNU	27-02-2017 16:27	16	826	
03-02-2017 17:02	34	249		27-02-2017 17:22	16	70	OSMANGAZİ
03-02-2017 17:06	34	249		27-02-2017 17:45	16	70	
03-02-2017 18:29	34	249		27-02-2017 18:22	16	70	
03-02-2017 19:08	34	249		27-02-2017 20:44	16	70	
03-02-2017 19:36	34	249		27-02-2017 20:46	16	70	
03-02-2017 20:23	34	386	FATIH	27-02-2017 20:58	16	70	
03-02-2017 22:24	34	386		27-02-2017 21:01	16	70	
04-02-2017 13:54	34	386		27-02-2017 21:13	16	70	
04-02-2017 14:12	34	386		27-02-2017 22:31	16	70	

Table 4.3 February, 2017 calls and commute of Caller_ID 1……...8

5. Evaluation of analysis results

Mobility is defined by, and usually confined within, a range of factors including employment, labour, gender, age, intergenerational factors, ethnicity, race, physical and mental capabilities, life aspirations and opportunities to name a few [10]. It is often shaped by individual or household needs for education, health, social well-being and else. In order to move beyond the immobile structures of refugee camps and to meet their needs as well as to continue their daily lives, refugees started moving to cities in Turkey in the early days of the Syrian conflict. However, as our analysis demonstrates they have become inhabitants of these cities and voluntarily submitted themselves to the immobility of their new settlements. Whether this is out of choice or necessity is unknown to our big-data driven research. Yet, we can conclude that

cities as escape mechanisms may encapsulate the escapees. Nevertheless, this is also likely to point out to the fact that refugees' needs can be met in these new harbours.

In addition to immobility levels of refugees our analysis demonstrate the locations where mobilities are performed. The majority of annual mobilities generate between large urban areas and their adjacent cities alongside the cities bordering Syria. Yet, a low number of refugees maintain high mobility levels throughout the year.

6. Conclusion and policy recommendations

6.1 The limiting factors and drawbacks of the research

First, despite being limited to only one mobile operator, the data is precious for understanding the mobility patterns of refugees in Turkey. However time and data constraints limited the present analysis, which may be improved in the following ways:

- The limited time for processing this amount of data, did not allow us to perform various complementary analyses. These include a detailed analysis of contextual factors with multiple tools, including media and discourse analysis; the geographical and sociological factors affecting refugee mobility and a comparative study of refugee and non-refugee mobility.
- Comparative analysis of call/ sms volumes and physical im/mobility can more precisely identify the missing physical links in refugees' networks
- The sample represents the total number of refugee callers, yet they can be further grouped to establish a typology of refugee mobility.
- A more fine grained analysis could be done to include within-city mobility.
- The application of transport modelling methods and programs (such as Visum and TransCad) to D4R data can provide an improved understanding of refugees' travel behaviours.

6.2. The value and potential of the research and the method

The advantage of the applied method is its flexibility in changing the resolution of the analysis. It combines multiple levels, stretching from individual (Caller_ID) to the whole refugee (sample) population included in the data set. This method allows us to combine the quantitative analysis of large amounts of data with qualitative research tools; as well as to have a mixture of computer and human brain data processing capacity. In this way generalized interpretation of large data clusters is avoided.

The method also proves that mobile communication data is not only about mobile communication, but also indicates physical mobility. Physical mobility is difficult and expensive to measure, and requires extensive surveys. Although mobile communication data cannot be a substitute (it does not include those who don't have a phone), it can be used effectively as a supplementary tool to understand the patterns of mobility.

This method can be extended to multiple years to identify the social networks of refugees and to assess not only the relevant policies, but also the contribution of refugees to the host society by mobilization of their networks and resources.

6.3 Policy recommendations

This research primarily targets the integration theme of the competition. Analysis suggests that only a small amount of refugees travel across provinces. Limited mobility reduces the level of communication between refugees, access to their networks, hence their access to jobs.

Syrian refugees are required to reside in the

Increasing the mobility of refugees and benefitting from their networks for social integration and economic development are the main policy recommendations of the present research.

The research indicates that refugees, commuting with only one city do not reach beyond the neighbouring provinces. Such commuting is especially high in the regions bordering Syria, and in the provinces in geographical proximity to Istanbul. However, even these relatively shorter trips, which are likely to indicate economic interactions as much as social relationships, are possible for only a small part of refugees. To increase inter-city mobility through public transport allowance will contribute significantly to refugees' accessibility to their social networks and resources.

On the other hand, a limited number of refugees travel comparatively more, the most mobile reaching to eleven different provinces. These persons travel across Turkey, not limiting themselves to their neighbouring provinces. Apart from their reasons for travel, they can be considered as agents linking different parts of their communities. As such, they are not only links between refugee groups, but also indicators of geographical distribution of refugee networks. As a result of forced migration, it is expected that various Syrian groups, be they neighbours, relatives or business partners are teared apart. Promoting their interaction, hence the continuity of previously established social networks in the Turkish geography can contribute to the socio-economic development of Turkey, too.

7. References:

1. Urry, J. (2007) Mobilities. Polity Press., p.263
2. UNHCR, 2018 (http://www.unhcr.org/figures-at-a-glance.html)
3. The World Bank, 2018 (https://www.worldbank.org/en/news/press-release/2018/03/19/ climate-change-could-force-over-140-million-to-migrate-within-countries-by-2050-world-bank-report)
4. UNHCR, 2018 (https://data2.unhcr.org/en/situations/syria)
5. http://t24.com.tr/haber/suriye-dostluk-dernegi-turkiyede-1-milyonu-kayit-disi-4-milyon-suriyeli-var,349342
6. Ozden, S. (2013) Syrian Refugees in Turkey. Migration Policy Centre http://cadmus.eui.eu/bitstream/handle/1814/29455/MPC-RR-2013%2005.pdf?sequence=1&isAllowed=y
7. Porter G, Hampshire K, Abane A, Munthali A, Robson E, Mashiri M, Tanle A. 2012. Youth,mobility and mobile phones in Africa: findings from a three -country study. Journal of information Technology for Development 18(2): 145–162
8. Porter, G. 2015 Mobile phones, mobility practices and transport organisation in sub-Saharan Africa. *Mobility in History*6, 81-88
9. Salah, A.A., Pentland, A., Lepri, B., Letouzé, E., Vinck, P., de Montjoye, Y.A., Dong, X. and Dağdelen, Ö., 2018. Data for Refugees: The D4R Challenge on Mobility of Syrian Refugees in Turkey. arXiv preprint arXiv:1807.00523.
10. Ohnmacht,T., Maksim, H., Bergman, M.M. (2009) Mobilities and Inequality. First edition. Ashgate Publishing

Segregation and Sentiment: Estimating Refugee Segregation and Its Effects Using Digital Trace Data

Neal Marquez[1,2], Kiran Garimella[3], Ott Toomet[1], Ingmar G. Weber[4], and Emilio Zagheni[1,2]

[1]University Of Washington
[2]Max Planck Institute for Demographic Research
[3]Ecole polytechnique fédérale de Lausanne
[4]Qatar Computing Research Institute

September 2018

Abstract

In light of the ongoing events of the Syrian Civil War, many governments have shifted the focus of their hospitality efforts from providing temporary shelter to sustaining this new long-term population. In Turkey, a heightened focus has been placed on the encouragement of integration of Syrian refugees into Turkish culture, through the dismantling of Syrian refugee-only schools in Turkey and attempts to grant refugees permanent citizenship, among other strategies. Most of the existing literature on the integration and assimilation of Syrian refugees in Turkey has taken the form of surveys assessing the degree to which Syrian refugees feel they are part of Turkish culture and the way Turkish natives view the refugee population. Our analysis leverages call detail record data, made available by the Data 4 Refugees Challenge, to assess how communication and segregation vary between Turkish natives and Syrian refugees over time and space. In addition, we test how communication and segregation vary with measures

of hostility from Turkish natives using data from the social media platform Twitter. We find that measures of segregation vary significantly over time and space. We also find that measures of inter group communication positively correlate with measures of public sentiment towards refugees. Attempts to address the concerns of Turkish natives to minimize the traction of online hate movements may help to improve the integration process.

Keywords: segregation, Syrian refugees, Turkey, CDR, social media, social integration

1 Introduction

In the spring of 2011, at the beginning of the Syrian Civil War, Syrians began to find themselves displaced by the armed conflicts between the Syrian Arab Republic and numerous other forces who sought to challenge the authority of the government in the wake of perceived injustices committed by the regime led by Bashar al-Assad [7]. During this time, Turkey had an open door policy with Syria and assured that those migrating in would be able to stay until Syria was once again safe for return [9]. By later that year, it was apparent that extensive measures would need to be taken to accommodate the growing number of refugees. During the first years of the Syrian conflict it was unclear how long the crisis would last and refugees would be seeking asylum, in Turkey and other locations. Initial measures addressed short-term issues by setting up temporary schools, camps, and health care facilities [9]. By 2015, however, it became clear that the conflict was not to conclude in the near future and the flow of refugees into Turkey continued, reaching over 2.5 million Syrians in Turkey by the end of the year. [1]

The strategy of the Turkish government shifted from short- to long-term plans, as policies were developed to ease the transition of Syrians into Turkish life. A new, worldwide visibility of the plight of Syrian refugees allowed Turkey to coax greater action from the international community to share a portion of the economic and resource burden created by housing refugees. Though other European countries have stepped up their contributions to the crisis by way of accepting more refugees and offering Turkey financial compensation [24], Turkey has by far the largest Syrian refugee population

[1] `https://data2.unhcr.org/en/situations/syria/location/113`

to date, more than 3.5 million as of August 2018[2], and continues to struggle with integrating its new population. The difficulty of integrating refugees into Turkish culture is a battle that has two fronts, as the government not only looks to facilitate a smooth transition for refugees but also to ease the concerns of Turkish natives, who fear the extended stay of Syrian refugees may come at the expense of their desired lifestyle [13].

Segregation and social isolation can exacerbate the differences between these two groups by limiting the amount of cultural overlap they experience. To date almost no measures of segregation of Syrian refugees and Turkish natives are available. The rapid increase in the number of refugees in Turkey in the past few years has made it difficult for traditional methods of data collection to capture this phenomenon.

This analysis leverages call detail record (CDR) data, made available by the Data 4 Refugees Challenge, to assess how communication and segregation differ between Turkish natives and Syrian refugees over time and space. Using CDR data, we create metrics of geographic activity space and residential dissimilarity as measures of segregation. We also calculate spatial-temporal measures of the probability of refugees contacting Turkish citizens through phone calls and texts, as a measure of group isolation. Finally, we examine how communication between the two groups is altered by differing levels of segregation as well as changes of expressed opinions from Turkish citizens toward Syrian refugees.

2 Background

2.1 Segregation

Segregation has long been seen as a mechanism that isolates individuals from accessing greater opportunities if their isolated enclave is poor in group resources [19]. In addition, greater isolation of communities has been linked to increased xenophobic attitudes toward minority migrant groups in the the global south [16]. Previous policy research has advocated for working towards greater cohesion between groups in the form of public education campaigns as a way of combating negative opinions towards these minority groups [8]. The extent to which segregation between native populations and refugees is an issue in Turkey is not yet well understood.

[2] `https://data2.unhcr.org/en/situations/syria/location/113`

To date no studies have systematically or comprehensively quantified the level to which segregation exists between Syrian refugees and Turkish natives. We use the word segregation here simply to mean the separation of two or more groups of people, in our case Syrian refugees and natives within Turkey. Most studies that address segregation do so within the context of economic approaches. For example, in a recent publication by Balkan et al. (2018), the authors found that increases in the refugee population led to increased rent costs in higher end properties, which is seen as evidence for increased value of housing that is geographically segregated from refugee populations [5]. Additionally, İçduygu et al. (2017) found that integration efforts made by Syrian refugees to participate in the legal labor force were thwarted by difficulties to obtain visas, thus limiting chances to integrate socially and culturally [25].

2.2 Turkish Attitudes Toward Refugees

How segregation ties in with attitudes from Turkish citizens is at the moment unclear. Because the level of segregation between the two populations is not well known in Turkey, it is difficult to say something about the effect that it has on Turkish citizens' opinions of refugees, if any at all. We do know, however, that sentiment towards refugees has been negatively trending. While early studies showed a more neutral stance on the Syrian refugee population, recent studies show strong negative attitudes [9]. In the Syrian Barometer Study 2017, Erdoğan found that over 80% of Turkish survey respondents claimed that the Syrian and Turkish culture do not overlap at all [10]. In addition, several studies have found that some populations who have experienced large Syrian refugee intake have taken to social media platforms to voice their dissatisfaction with the presence and government handling of Syrian refugees [20, 4].

Social media offers a way to study how populations react to events without the time or expense requirements of conducting a poll. While Twitter is known to not have a representative population of users for most if not all countries, previous studies have found that text analysis in the form of sentiment extraction can provide reliable predictions for events such as changes in the stock market [6]. Additionally, researchers have been able to track changes in attitudes toward minority groups in response to policy announcements [12].

A content analysis of tweets – posts from the social media platform Twit-

ter – about Syrian refugees across Europe found that when users attack refugees they often do so by attacking the character of male refugees, labeling them either as cowards or terrorists [20]. A more recent report within Turkey found that several anti-Syrian hashtags had gained traction in 2017, undermining efforts to foster greater cohesion between refugees and citizens [14]. The events co-occured with a threefold increase in intergroup violence between 2016 and 2017, lending evidence that events on Twitter may in fact well represent attitudes of the greater population despite Twitter only having a 15% penetration rate in Turkey[1].

Our analysis tests how segregation, both geographic and social, varies over space and time between Syrian refugees and Turkish natives using CDR data. Using this information, we will be able to make better informed decisions regarding the way that refugees have integrated into the Turkish population differentially within the country. We can do this by examining both residential and activity space dissimilarity as measures of geographic segregation. Furthermore we can test social isolation by assessing the kind of persons that refugees call, either fellow refugees or Turkish citizens, as a measure of social isolation. Lastly, using Twitter data that contain subjects related to refugees, we will examine how variation in the sentiment of tweets alters with changes in refugee-citizen segregation over space and time.

3 Data

The analysis utilizes call detail records from the Turkish mobile network carrier Turk Telecom (TT), a member of the group TTG, as part of the Data for Refugees in Turkey (D4R) challenge [21]. The goal of the challenge is to give researchers access to privately-owned data from TTG that has user details removed for anonymity, such as names and telephone numbers. The time stamp and the location of the call are available in the data set. Each call also has a randomized ID assigned to it which indicates a unique user and whether TTG has the individual recorded as a refugee or not. This classification does not perfectly identify refugees and should be seen as an imperfect measure [21]. The specific data set that we utilize in our analysis tracks users for two weeks at a time with an undisclosed portion of their calls and text messages both in and out provided. The CDRs provide time stamp data, to the hour, and the cell phone tower that was pinged for the particular record. The records consist of 212,364,027 unique records from 5,006,222 and

1,082,603 unique non-refugee and refugee users, respectively. The call records span 26 two-week segments from January 1, 2017 to December 31, 2017 with the number of calls and users being not necessarily equally distributed across time (Figure 1).

Individuals were oversampled for areas that had relatively high refugee populations, such as border provinces and the major metropolitan areas of Istanbul and Izimir [11]. Each record in the data is given a tower ID which can be linked via a database with towers and their corresponding latitude and longitude. Any tower with a location outside of Turkey's administrative bounds was removed from the dataset. To verify that we can adequately capture mobility of individuals to an adequate level, we analyzed the degree to which district level (administrative level 2) population size correlated with the number of cell phone towers in an area. In a log-log linear model the tower count explained 81.67 of the variation in the 2014 population, taken from the 2014 Turkey national census, at the district level. The areas that had the most discrepancy between the number of cellular towers and the population count can be seen in the Figure 2. Refugee status of the other individual participating in the phone call is also provided in the dataset.

To estimate changing attitudes over time and space in Turkey, we pulled Twitter data from 2011 to 2017 from the Twitter Stream that matched several topics related to Syrian refugees (see Appendix). The Twitter Stream is an ongoing project from the Internet Archive Team that consistently collects a 1% stream of all Twitter data produced.[3] While the Twitter API only allows users to collect data that has recently been created, this archive allows us to search trends that overlap with our CDR record dataset. Tweets were only considered for our dataset if they were from 2017. We further restricted our analysis to include only tweets from Turkish language users, users that specified their location to be within Turkey, or tweets that could be geolocated within Turkey. Individual tweets could be geolocated either by providing the exact coordinates of the location of the tweet, i.e. "Tweet with a location" option, or by designating a "place" from a pre-specified list provided by Twitter which contains geographic coordinates. If these coordinates fall within the administrative boundaries of Turkey, the tweets are kept. Users could be identified as being from Turkey based on their user-specific location string. To geotag this string, we use the Open Street Maps API and select the location coordinates with the highest match to determine

[3] `https://archive.org/details/twitterstream`

if the user is located within Turkey. This filtering process left us with 65,778 tweets for our analysis.

Several other variables were collected for modeling purposes. Population data at the province level was taken from the 2014 Turkish census. Land use data was collected from CORINE Land Cover surveys 2006-2012 to calculate the percent human created land coverage, a proxy measure for urban space [3]. These data were then population weighted using population rasters created by satellite imagery from the gridded population of the world v4 [2].

4 Methods

To calculate residential and activity space dissimilarity for a district, we created subunits within each district by way of Voronoi tessellation from the cell phone towers within the district. Voronoi tessellation creates areal units which define a two dimensional space that is the least distance from a particular point, in our case a cell phone tower [15]. If many towers exist in a district, then the areas that are created are relatively granular, given that the towers are evenly spaced. Using Voronoi cells as subdivisions of districts, we calculate an activity space dissimilarity index for each district. While traditional residential dissimilarity indexes measure differences from the perspective that individuals are situated in a single location, activity space dissimilarity measures the probability of remaining isolated from another group or 1 - "potential to encounter" as defined in Wong et al. [23]. Activity space dissimilarity scores were calculated for each district for each week of the analysis using the formula in Equation 2 where i is a Voronoi cell, j is an individual, p_{ij} is the percentage of time individual j is in Voronoi cell i, A is the refugee population size, and B is the non-refugee population size. In addition, we calculate residential dissimilarity by taking the modal call location of an individual between the hours of 9 p.m. and 6 a.m. and calculate a traditional dissimilarity index, using the modal location as the place of residence, with the formula in Equation 1. For residential dissimilarity, we only calculated one score per district rather than weekly scores because the values did not change significantly over time, which is to be expected as residential segregation is slow to change.

$$\text{Residential Dissimilarity Score} = \frac{1}{2}\sum_{i}^{N}\left|\frac{a_i}{A} - \frac{b_i}{B}\right| \tag{1}$$

$$\text{Activity Space Dissimilarity Score} = \frac{1}{2}\sum_{i}^{N}\left|\frac{\sum_{j}^{A} p_{ij}}{A} - \frac{\sum_{j}^{B} p_{ij}}{B}\right| \quad (2)$$

To test whether the dissimilarity values were different than expected for a district given the number of Voronoi cells and number of refugee and non-refugee calls, we randomized the caller type for each record 1000 times and re-calculated dissimilarity scores from the simulated distribution, often referred to as a permutation test. Z-scores were then calculated for the district's observed dissimilarity score against the simulated values. Uncertainty for our measures of dissimilarity were calculated by bootstrapping, where individuals were sampled with replacement for each unit of analysis, district for residential dissimilarity and district-week for activity space dissimilarity.

For each district, we also compiled a connectivity score of refugees to non-refugees as a measure of intercommunication between the two groups. The percentage of calls going from refugees to non-refugees was calculated for each district. We excluded records from non-refugees to refugees because of the small sample size they represented in the data, less than .1%.

Tweets were analyzed using a Turkish translated version of the AFINN, a common sentiment analysis tool with words valence rated on a scale from -5 to 5. Each tweet is rated by the sum of individual word scores. Though this process only allows us to attribute sentiment on a word by word basis, it has been extensively tested [18] and is more easily translated into other languages than other sentiment tools. To match Twitter sentiment with CDRs we aggregated sentiment by week and calculated the average weekly sentiment from the Turkish tweets with Syrian related content (Figure 3).

To test the relationship between Twitter sentiment and intergroup connectivity, we run a series of logistic regressions where each outgoing call made by a refugee is the response variable. The outcome is 0 if the call/text was made to a fellow refugee or 1 if made to a non-refugee, with a total of 10,235,988 records. Call records were connected with covariates by their district of call location (for population size, urban area coverage), the biweekly time period that they occurred (for Twitter sentiment), or the combination of the two (for activity space dissimilarity index). We tested a number of covariate combinations to test the robustness of the relationships between covariates and the outcome. To account for the bias in the data from repeated calls from a single user, we ran a mixed effects model with a random

intercept on individual. Equation 3 shows the structure of the model where i represents an individual, j represents a particular call that was made, $\boldsymbol{\beta}$ is a vector of beta coefficients, X_{ij} is a vector of coefficients for individual i call j for the particular time and location that the call took place, and ζ_i is the individual level random effect. We did not adjust for spatial autocorrelation as our outcome of interest did not show evidence for it.

$$\begin{aligned} y_{ij} &\sim \text{Binomial}(\hat{p}_{ij}) \\ \hat{p}_{ij} &= \text{logit}(\boldsymbol{\beta} \bullet X_{ij} + \zeta_i) \\ \zeta_i &\sim \mathcal{N}(0, \sigma) \end{aligned} \tag{3}$$

We also tested the ability to predict the sentiment (both positive or negative as well as score) of a tweet as a function of the above mentioned covariates linked by location of the tweet at the province level. Geocoded tweets left us with a considerably smaller sample size from the original dataset, as only 53,793 tweets were from 2017 forward and could be reliably geocoded to a specific province within Turkey. All model covariates were included at the province level and were time invariant.

5 Results

Our analysis of spatial overlap found a significant difference between the observed values of activity space dissimilarity and their expected values. Of the 970 districts in our analysis, around 75% had observed values that were more than 4 standard deviations away from their simulated permutation distribution. Of the major metropolitan areas, Ankara had the highest average observed values of dissimilarity, while Istanbul had the lowest, though district level variance was twice as high in Ankara (Figure 4).

Urban land coverage of a district was inversely correlated with dissimilarity although the effect was non-significant when running a simple linear model. Using bootstrapped estimates of the uncertainty of our calculations for activity space dissimilarity, we found that there were significant differences over time at both the district and province level. We also found that residential dissimilarity was strongly correlated with activity space dissimilarity with a correlation coefficient of 79.96 at the district level and 83.83 at the province level. In line with previous literature, we found that activity space dissimilarity was more often less than residential dissimilarity [22].

Twitter sentiment was also found to change significantly over time but not over locations. Because our province level analysis required that users tweets be geocoded at least to the provincial level, our sample size was dramatically reduced when examining geographic differences in tweets (Figure 5). Analysis of changes over time found that sentiment of tweets were lower in the months of June through September than in the other months (Figure 3). This pattern is noteworthy in that it also appears in 2016, again with lower sentiment scores in the months between June and September. The content of the tweets was examined and the most negatively rated words for June through September drastically differed from other months, and were consistent with they way previous research found Syrians to be negatively characterized (Figure 6).

Analysis of tweet sentiment using our collection of covariates was not statistically significant. While the covariates were largely in the expected direction (higher dissimilarity and urban areas led to lower predicted sentiment), our restricted sample size and noisy signal limit our ability to detect small differences in sentiment across provinces in Turkey. An increased sample size would allow us to detect differences despite a noisy signal and analyze effects at a district level where we expect measures of activity space dissimilarity to be more informative than at the provincial level.

Models for predicting calls and texts from refugees to non-refugees showed a significant positive relationship between sentiment and connectivity. As weekly Twitter sentiment scores increased, i.e. more positive text occurred in tweets about refugees, we observe higher probabilities of refugees contacting non-refugees. To evaluate the robustness of the relationship and remove potential confounding effects, we constructed a number of models with additional covariates. The effect was consistent across all models, and robust to the inclusion of other variables as seen in panel 3 of Figure 7. The probability between cross group connections was larger in urban areas than non-urban, and higher when dissimilarity was higher. However, this pattern is sensitive to the definition of urban area. The full specification of all models which follow the structure of Equation 3, can be found in the Appendix along with an extended definition of each covariate and which covariates were included in each model.

6 Discussion

We find activity space differences between major metropolitan areas by analyzing the movements of refugees and Turkish citizens through CDR data. Meaningful differences of activity space dissimilarity exist both within and between provinces. Furthermore, the differences that we observe between locations appear to be consistent over time (Figure 8). Previous research has shown that heightened segregation between groups can lead to an inability of marginalized groups to access opportunities [19] and is connected to higher rates of xenophobia, especially when related to immigrants [16]. Decreasing segregation between groups should be seen as a goal in and of itself, especially in population-dense areas where contact with other groups is more easily attainable because of spatial proximity.

Though the effect size that we find for the relationship between measures of segregation and Twitter sentiment is small, its presence persists across all models. Previous research concerning changes in the way that social media negatively portrays Syrian refugees are few [14, 20], and most often do not make connections between how changes in portrayal co-occur with other events. Our analysis finds that negative changes in sentiment towards refugees – as calculated from sentiment analysis of Twitter posts – are correlated with a decrease in the probability of refugees connecting with non-refugees. Though the majority of calls and texts made by refugees go to non-refugees, it should be noted that the non-refugee group covers a broad range of individuals (Turkish citizens), groups (Turkish entities), and services (such as Arabic speaking call centers with information on social services for refugees). Refugees rely heavily on their phones to navigate their new environment in Turkey [17]. Even small changes in the reduction of connections made by refugees to others could prove to be damaging. Events that deter refugees from connecting with non-refugees, such as changes in online portrayal and attitudes towards Syrian refugees, should be closely monitored.

The inability to detect significant differences in the average tweet score between geographic regions does not give this analysis enough signal to leverage in order to test different geographic sentiments. This does not mean that different regions tweet in a similar matter, but rather that our current resources did not allow us to capture the signal in an adequate way. There are two ways that we can potentially overcome this obstacle in future studies. One way is to use a more sophisticated process to classify tweets as positive or negative via statistical training. By labeling tweets as either positive or

negative via manual coding for a small set of tweets, we would be able to train a statistical model on features extracted from the text. This would allow us to focus our sentiment detection on the language that is specific to the topic of Syrian refugees. Alternatively, by increasing the sample size of our tweets, we would be able to better detect differences in signals over time and space. This could be done by using a proactive data collection strategy with the Twitter API which would allow us to collect a much greater sample than the 1% historical records provide.

7 Conclusion

This analysis is the first to provide comprehensive measures of segregation, both activity space and residential, between Syrian refugees and Turkish natives. We find that there are significant differences between major metropolitan areas within Turkey that are home to a significant share of the refugee population. Given that segregation has been a reported factor in the continuation of xenophobic language toward minority groups we find that it would be of interest to policy makers to continue to measure the level of both activity space and residential segregation in the near future.

Furthermore, we find that there is significant variation over time in attitudes towards refugees in Turkey on the social media platform Twitter. These variations could prove to be helpful as a gauge of changing attitudes toward Syrian refugees in light of particular events. The evidence for a relationship between segregation and changes in attitude towards Syrian refugees is limited; however, the consequences of reducing connections between Syrian refugees and Turkish natives could have dramatic consequences. Better data collection or sentiment detection could enable us to better make connections between geographic and temporal differences in sentiment and should be pursued further.

Figure 1: Number of Texts and Calls present in each BiWeekly dataset broken down by ID type of the TTG user, either Registered Refugee or non-Refugee. Values are shown on a log scale.

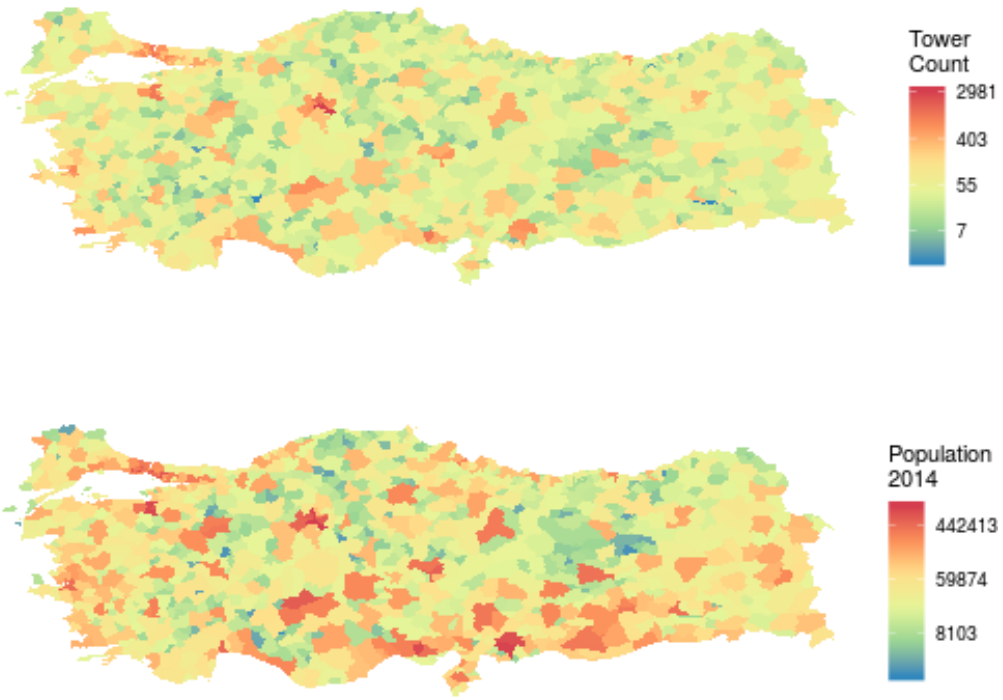


Figure 2: Population counts (lower panel) & cell phone tower users (upper panel) geographic distributions at the district level.

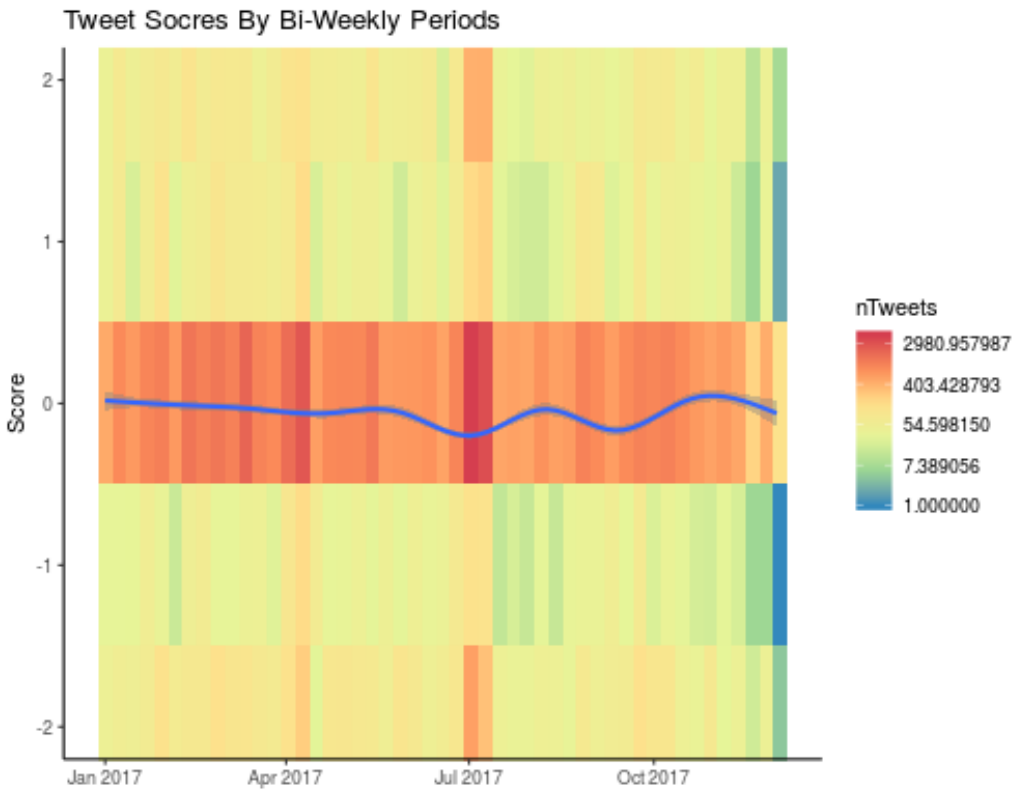


Figure 3: Weekly Sentiment Heat Map with Loess Smoothed Scores. Each rectangular bin is a week(x-axis)-tweet score(y-axis) combination where the hue indicates the number of tweets in a week that had a particular sentiment score.

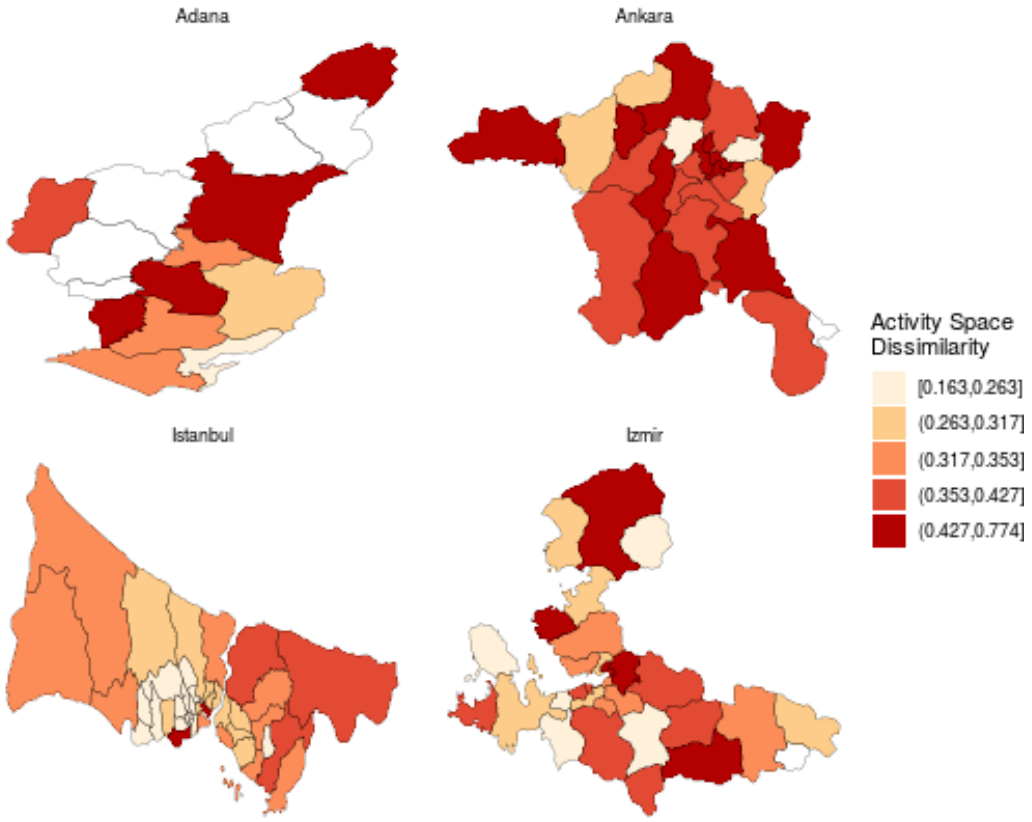


Figure 4: Activity Space Dissimilarity Scores for selected provinces. Results with observed dissimilarity less than 4 standard deviations away from mean are whited out.

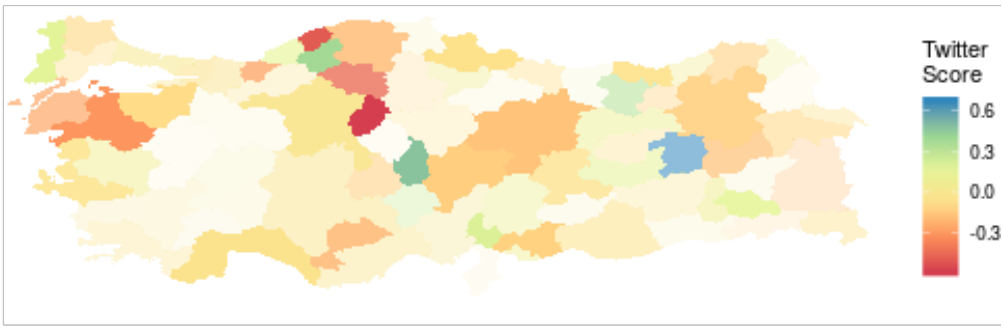


Figure 5: Province Level Average Tweet Scores from sentiment analysis. Opaqueness is adjusted for 0 value z-score. High values indicate more positive (or less negative) sentiments.

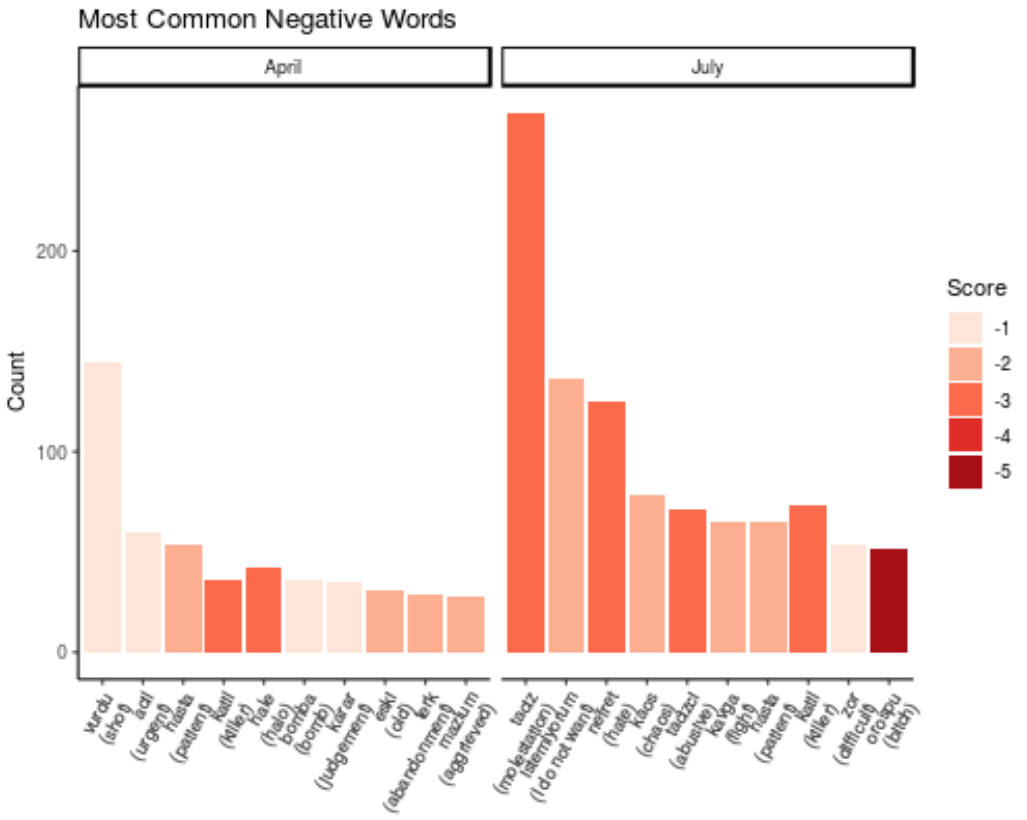


Figure 6: Comparison of most common negative words in our data set of tweets about refugees for selected months.

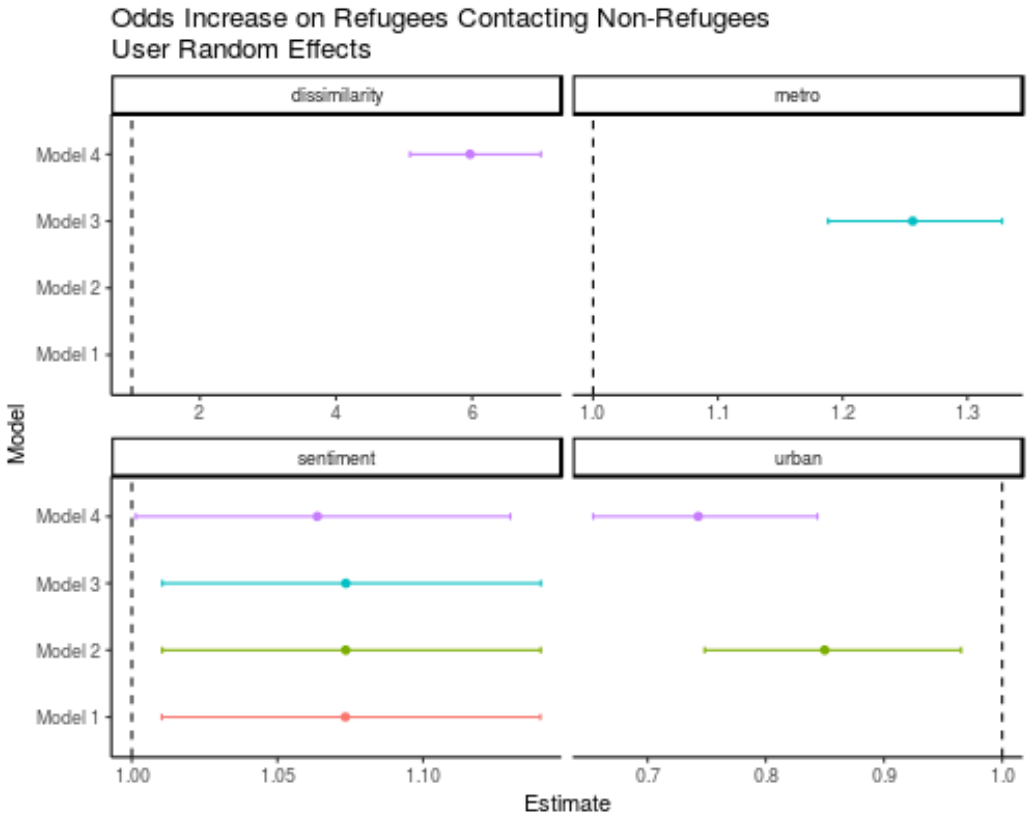


Figure 7: Model Odd Ratios Coefficient Estimates for Select Covariates. Error bars not overlapping with dotted line indicate significant result. Four models are presented in the figure on the y axis and coefficients are placed in separate panels. Full explanation of models and covariates can be found in the Appendix.

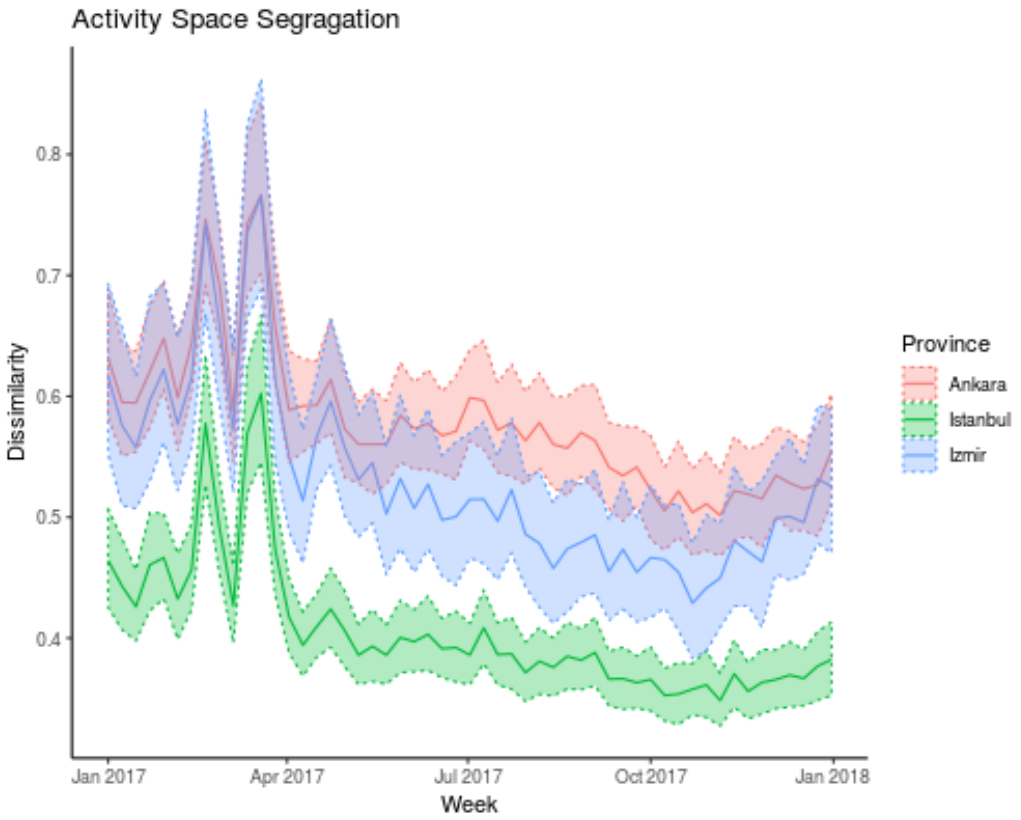


Figure 8: Change in Dissimilarity by Week for Select Provinces. Uncertainty calculated from bootstrapped samples with 95% confidence intervals shown.

References

[1] We Are Social - Digital Report 2018.

[2] Gridded Population of the World, Version 4 (GPWv4): Population Density. Technical report, Center for International Earth Science Information Network - CIESIN - Columbia University, Palisades, NY: NASA Socioeconomic Data and Applications Center (SEDAC), 2016.

[3] Land Cover Change (LCC) 2006-2012, Version 18.5, 2016.

[4] Katty Alhayek. Double Marginalization: The Invisibility of Syrian Refugee Women's Perspectives in Mainstream Online Activism and Global Media. *Feminist Media Studies*, 14(4):696–700, jul 2014.

[5] Binnur Balkan, Elif Ozcan Tok, Huzeyfe Torun, and Semih Tumen. Immigration, Housing Rents, and Residential Segregation: Evidence from Syrian Refugees in Turkey — IZA - Institute of Labor Economics. Technical report, IZA Institute of labor Economics, Bonn, 2018.

[6] Johan Bollen, Huina Mao, and Xiaojun Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8, mar 2011.

[7] T. G. Carpenter. Tangled Web: The Syrian Civil War and Its Implications. *Mediterranean Quarterly*, 24(1):1–11, jan 2013.

[8] Jonathan Crush and Sujata Ramachandran. Xenophobia, International Migration and Human Development — Human Development Reports. Technical report, United Nations Development Programme, New York, 2009.

[9] O. B. Dinçer, V. Federici, E. Ferris, S. Karaca, K. KiriŞci, and ÇarmIKlIE.Ö. Turkey and Syrian refugees: The limits of hospitality. Technical report, Brookings Institute, 2013.

[10] M. Murat Erdoğan. Syrians Barometer 2017. Technical report, İstanbul Bilgi University, Istanbul, 2018.

[11] M. Murat Erdoğan, Burcuhan Şener, Elif Sipahioğlu, Yudum Kavukçuer, and Esin Yılmaz Başçeri. Urban Refugees From Detachment To Harmonization. Technical report, Marmara Municipalities Union's Center for Urban Policies, 2017.

[12] René D. Flores. Do Anti-Immigrant Laws Shape Public Sentiment? A Study of Arizona's SB 1070 Using Twitter Data. *American Journal of Sociology*, 123(2):333–384, sep 2017.

[13] International Crisis Group. Turkey's Refugee Crisis: The Politics of Permanence, 2016.

[14] International Crisis Group. Turkeys Syrian Refugees: Defusing Metropolitan Tensions, 2018.

[15] D. T. Lee and B. J. Schachter. Two algorithms for constructing a Delaunay triangulation. *International Journal of Computer & Information Sciences*, 9(3):219–242, jun 1980.

[16] Jerome S. Legge. *Jews, Turks, and other strangers : the roots of prejudice in modern Germany*. The University of Wisconsin Press, Madison, 2003.

[17] N Narli. Life, Connectivity and Integration of Syrian Refugees in Turkey: Surviving through a Smart Phone. *Questions de Communication*, 33(1):269–286, 2018.

[18] Finn Årup Nielsen. AFINN. Technical report, Informatics and Mathematical Modelling, Technical University of Denmark, Kongens Lyngby, 2011.

[19] Michael Pacione. *Applied geography : principles and practice : an introduction to useful research in physical, environmental and human geography*. Routledge, New York, 1999.

[20] Jill Walker Rettberg and Radhika Gajjala. Terrorists or cowards: negative portrayals of male Syrian refugees in social media. *Feminist Media Studies*, 16(1):178–181, jan 2016.

[21] A.A. Salah, A. Pentland, B. Lepri, E. Letouzé, P. Vinck, Y.A. de Montjoye, X. Dong, and Ö. Dağdelen. Data for Refugees: The D4R Challenge on Mobility of Syrian Refugees in Turkey. *arXiv preprint arXiv:1807.00523*, 2018.

[22] Ott Toomet, Siiri Silm, Erki Saluveer, Rein Ahas, and Tiit Tammaru. Where Do Ethno-Linguistic Groups Meet? How Copresence during Free-Time Is Related to Copresence at Home and at Work. *PLOS ONE*, 10(5):e0126093, may 2015.

[23] David W.S. Wong and Shih Lung Shaw. Measuring segregation: An activity space approach. *Journal of Geographical Systems*, 13(2):127–145, jun 2011.

[24] Ahmet İçduygu. Syrian Refugees in Turkey: The Long Road Ahead — migrationpolicy.org. Technical report, Migration Policy Institute, Washington, 2015.

[25] Ahmet İçduygu and Eleni Diker. Labor Market Integration of Syrian Refugees in Turkey: From Refugees to Settlers. *Göç Araştırmaları Dergisi*, 3(1), 2017.

Appendix

Twitter Collection Keywords

Tweets were collected from the Twitter Archives for the period between January 1st, 2017 and December 31st 2017. Any tweets that contained the following words which pertain to Syrian refugees were included in our analysis.

Suriye	mültecileri	Suriye Makedonya
Suriyeli	göç dalgası	şişme bot göçmen
suriyeli	Suriye Yunanistan	sahil güvenlik göçmen
mülteci	Suriye Macaristan	düzensiz göçmen
mülteciler	Yunanistan'a göç	göçmen iadesi
mültecilere	Yunanistan göçmen	ÜlkemdeSuriyeliİstemiyorum

Covariate Abbreviations

Covariate	Description
sentiment	Weekly sentiment score derived from Tweets about Syrian refugees in Turkey.
lrPop	Natural log population of district derived from 2014 census.
urban	The percentage of man made land coverage from CORINE Land Coverage Database.
metroTRUE	Dummy variable where True indicates a district is in one of top 5 urban provinces.
borderTrue	Dummy variable indicating whether a district is in a province that borders Syria.
diss	Activity space dissimilarity at the district level calculated from a single week of data.

Model Specifications

Model 1

$$\hat{p}_{ij} = \text{logit}(\beta_0 + \beta_1 \text{ sentiment} + \zeta_i)$$

Model 2

$$\hat{p}_{ij} = \text{logit}(\beta_0 + \beta_1 \text{ sentiment} + \beta_2 \text{ lrpop} + \beta_3 \text{ urban} + \zeta_i)$$

Model 3

$$\hat{p}_{ij} = \text{logit}(\beta_0 + \beta_1 \text{ sentiment} + \beta_2 \text{ lrpop} + \beta_3 \text{ metroTRUE} + \beta_4 \text{ borderTRUE} + \zeta_i)$$

Model 4

$$\hat{p}_{ij} = \text{logit}(\beta_0 + \beta_1 \text{ sentiment} + \beta_2 \text{ lrpop} + \beta_3 \text{ urban} + \beta_4 \text{ diss} + \zeta_i)$$

Model Results Table

Model	Covariate	Estimate	Std. Error	Pr(>|z|)
Model 1	sentiment	0.07	0.03	< .05 *
Model 2	sentiment	0.07	0.03	< .05 *
Model 2	lrPop	-0.02	0.01	0.11
Model 2	urban	-0.16	0.06	< .05 *
Model 3	sentiment	0.07	0.03	< .05 *
Model 3	lrPop	-0.11	0.01	< .05 *
Model 3	metroTRUE	0.23	0.03	< .05 *
Model 3	borderTRUE	-0.01	0.02	0.64
Model 4	sentiment	0.06	0.03	< .05 *
Model 4	lrPop	0.05	0.01	< .05 *
Model 4	diss	1.79	0.08	< .05 *
Model 4	urban	-0.30	0.06	< .05 *

Integration of Syrian refugees: insights from D4R, media events and housing market data*

Simone Bertoli[1][0000−0002−6512−0834], Paolo Cintia[2], Fosca Giannotti[3][0000−0003−3099−3835], Etienne Madinier[4], Caglar Ozden[5][0000−0003−3424−185X], Michael Packard[6][0000−0003−3127−6221], Dino Pedreschi[2][0000−0003−4801−3225], Hillel Rapoport[4], Alina Sîrbu[2][0000−0002−3947−7143], and Biagio Speciale[4]

[1] Université Clermont Auvergne, CNRS, IRD, CERDI, Clermont-Ferrand, France
[2] Department of Computer Science, University of Pisa, Pisa, Italy
[3] ISTI-CNR, Pisa, Italy
[4] Paris School of Economics, Paris, France
[5] The World Bank, Washington DC, United States
[6] Georgetown University, Washington DC, United States

Abstract. We explore various means of quantifying integration using two of the D4R Challenge datasets. We propose various integration indices and discuss their output. We combine the data from the D4R Challenge with data from the GDELT Project and with data on transactions on the housing market in Turkey. We also describe research directions to be undertaken should an extended access to the data be provided.

Keywords: Social integration · Communication patterns · Segregation · GDELT · Housing market

1 Overview

Responding to a sudden arrival of large number of refugees is a daunting task for many host societies and governments. After addressing the immediate humanitarian needs of millions of people fleeing from civil war and violence, destination countries turn their attention to medium and long-term issues since the refugees are unable or unwilling to return home in most cases, in fear of their safety and well-being. At that point, assimilation and integration become the key concern, in order to reduce the burdens on the host society and the refugees. Identifying the extent and determinants of refugee integration will help policymakers mitigate the negative impacts (perceived or real) of refugees and further facilitate more refugee settlement.

The integration of minority groups, whether native- or foreign-born, has been a focus of academics across a variety of fields. Measures of integration will vary

* The findings in this paper do not necessarily represent the views of the World Bank's Board of Executive Directors or the governments they represent. Any errors or omissions are the authors' responsibility.

based on the dimension of interest and the data available. For example, in economics, wage convergence or occupational placement, obtained via labor force surveys, are the most commonly used measures (see, for instance, [2]). In sociology or political science, commonly used indices are based on social interaction, language acquisition, residential integration or cultural convergence (see [6]). Again, individual or household level surveys are the most commonly used data collection methods. These types of sources, despite their value, have a major shortcoming: They do not provide high frequency data in terms of a time or space dimension due to the cost and complexity of conducting such surveys.

One type of data that addresses this shortcoming is the use of social big data that is generated by phone records, social media, print media or daily economic transactions. This paper aims to contribute to our knowledge in this direction by combining D4R (see [7] for details) datasets with other big data sources to assess the economic, social and physical integration of Syrian refugees in Turkey. In addition to constructing various geographic integration and communication indices based on D4R, we merge the D4R-based data with real-estate market data and media data to explore their interaction.

While we acknowledge the need for further investigation, we identify several interesting patterns in the data regarding refugee integration. First, we observe heterogeneous segregation across provinces, this heterogeneity appears to be correlated with the size of the refugee population. More specifically, areas with higher refugee shares of the population are, on average, more integrated. Potentially indicating that refugees are settling in areas in which they are more accepted. Interestingly, though, spatial integration is not correlated with more inter-group phone calls. Second, segregation appears to be declining over time, this is to be expected as refugees expand their social networks and become more intertwined in the local economy. Third, segregation tends to be lower during the day than at night, indicating that refugees tend to work more closely to native Turks than where they live. Finally, there are clear linkages with events and residential markets but requires further analysis.

There are numerous policy implication of the observations and the results presented in the paper. Naturally, further analysis, more detailed data and a certain level of policy experimentation are needed to design appropriate policy instruments that can be employed. These policies, especially those that enable faster and smoother economic and social integration of the refugees, will benefit both the refugees and the host communities. The first policy measure is on labor market access. Even though the data do not reveal any direct information in this regard, formal access to labor markets is shown to be a critical policy measure in many different contexts. ([8]). Our analysis in the paper shows that refugees "daytime" integration is higher relative to the "nighttime" which indirectly indicates labor market integration (proxied by the former) is higher than social or residential integration. Legal access to labor markets will reduce "resentment" among hosts who would otherwise view refugees either as stealing their jobs by working under the table or simply free-riding by receiving welfare checks. Furthermore, refugees can enter many higher skilled occupations that

require formal employment, instead of being informally employed in low-skilled occupations. Similarly, in order to improve residential integration and to prevent refugees from living in isolated urban slums, it is necessary to impose laws that punish discrimination against refugees by landlords as well by real estate agents. Another option is to encourage refugees or subsidize their rents in areas where there is more housing but relatively low level of refugee presence. This would also provide a boost to the real estate markets in these areas.

The rest of the report is structured as follows: Section 2 presents some basic measures of communication built from the D4R datasets; Section 3 describes how we have extracted dyadic call propensities from Dataset 2; Sections 4 and 5 present two time-varying and spatially dis-aggregated measures, the EI index and the dissimilarity index respectively; Section 6 combines the D4R data with geo-localized events related to refugees extracted from the GDELT database; Section 7 explores data on the evolution of the housing market in Turkey using price and sales data from local real estate markets; finally, Section 8 presents the concluding remarks, and it describes the scope for future research.

2 Basic measures of communication

Datasets within D4R have different strengths and, as a result, are better suited for different purposes. For example, Dataset 1 is a more complete universe of observations while Dataset 2 has more detailed information for a smaller sample of users (see [7]). Dataset 2 is the only dataset in the D4R collection which contains point-to-point communication where both the caller and the callee have a *refugee* (R), or *non-refugee*[7] (N) label. Because dataset 2 is the only one to identify inter-group calls, much of the trends derived regarding call patterns are done using these data. Dataset 2 contains a series of 26 two-week long panels (which we call "waves"), following roughly between five to eighteen thousand refugees and fifty to sixty-five thousand non-refugees in a given wave. One observation in this dataset is a *communication* between (1) a sampled individual and (2) another individual that may or may not be part of the sample. A communication is either a phone call or a text message, that may be either outgoing, i.e., initiated by the sampled user, or incoming, i.e., received by the sampled user. For each communication, we know in which province the sampled user was, based on the antenna location. Out-of-sample individuals are labeled as *unknown*; for the purpose of our analysis, we drop communications involving unknown users.

Figure 1 plots the number of outgoing and incoming calls for R and N users in Dataset 2, showing that the number of calls involving N users is relatively stable over time, with an increase between Wave 21 and Wave 24, while the number of calls involving R users follow a steady increase up to Wave 21. This increase reflects a steady growth (again, up to Wave 21) in the number of R users included in the various 2-week samples, rather than in increase in call activity.

[7] At times, we refer to this group also as *natives*.

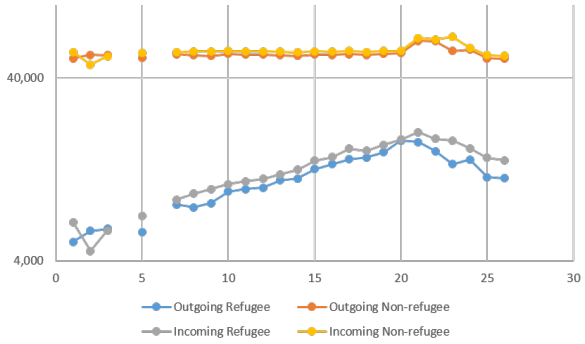


Fig. 1. Number of outgoing and incoming calls for R and N users in Dataset 2; each point corresponds to a different 2-week sample of users; no data are available for Wave 4 and Wave 6; the $y-$axis is in logarithmic scale.

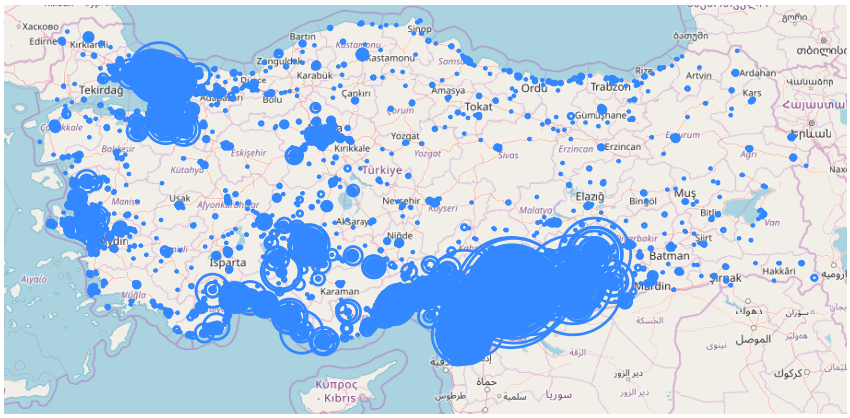


Fig. 2. Geographical distribution of voice calls in Dataset 2. Each circle corresponds to an antenna, and the radius of the circle is proportional to the number of calls made and received by that antenna.

For a high-level overview of Dataset 2, we show in Figure 2 the geographical distribution of calls. We observe a higher concentration of calls in large urban areas and also in the region close to the Syrian border. This is due to the fact that both R and N users were sampled based on the distribution of refugees from official records. If we only consider calls involving R users (see the map in Figure 3), the number of calls decreases but the spatial distribution appears to be similar, with higher concentrations of calls in the same areas as before, and especially around the Syrian border. At the province level, refugee call volume is highly correlated with the official population numbers, though there are a

few outliers. For example, Antalya, which has relatively few refugees according to the official numbers (ranked 66^{th} overall) has the eighth highest call refugee call volume. On the contrary, the provinces of Sirnak and Edirne (ranked 21^{st} and 27^{th} in official numbers) rank only 66^{th} and 57^{th} in refugee call volume, respectively.

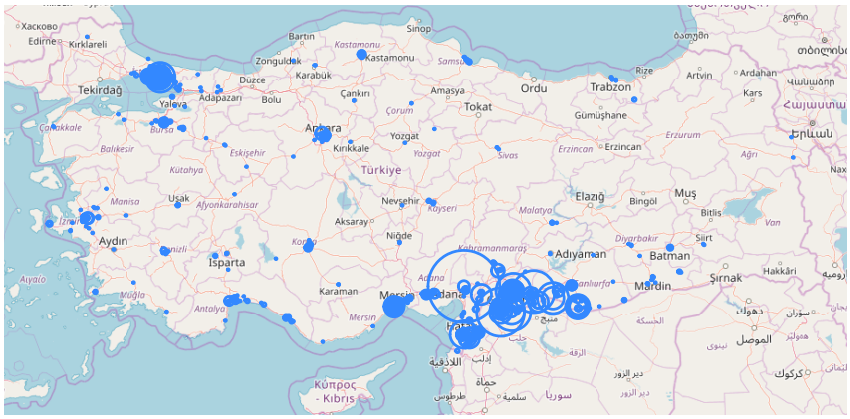


Fig. 3. Geographical distribution of voice calls in Dataset 2, including only calls that involve refugee users. Each circle corresponds to an antenna, and the radius of the circle is proportional to the number of calls made and received by that antenna.

To compare the call activity of refugees and non-refugees, we provide Figure 4, which shows histograms of individual call volumes for two waves, namely Wave 20 (September 25 to October 8) and 23 (November 6 to November 19), which correspond to a high and a low distance between the distributions of the calls for R and N users measured by the Kolmogorov-Smirnov statistic (not reported). A first observation is that, for both groups of users, most of the population is involved in a low number of calls, with a few individuals displaying very large numbers, i.e., heavy-tailed distributions. When comparing R to N, we see that the number of calls made and received are smaller for R, with a smaller fraction of users involved in a very large number of calls. This is difference is larger for incoming calls than outgoing calls. That is, while refugees and non-refugees make similar numbers of outgoing calls (it is still greater for non-refugees), refugees receive much fewer calls than do non-refugees.

Figure 5 presents the smoothed daily number of calls between each of the possible four pairs of caller and callee from Dataset 2, separately for calls recorded in the outgoing and in the incoming portions of the dataset. Daily calls are normalised so that yearly averages become 1 for each of the four possibilities. Figure 5 shows an increase in the number of calls where the caller is a R user. Over time, the share of calls going to refugees, from both other refugees as well

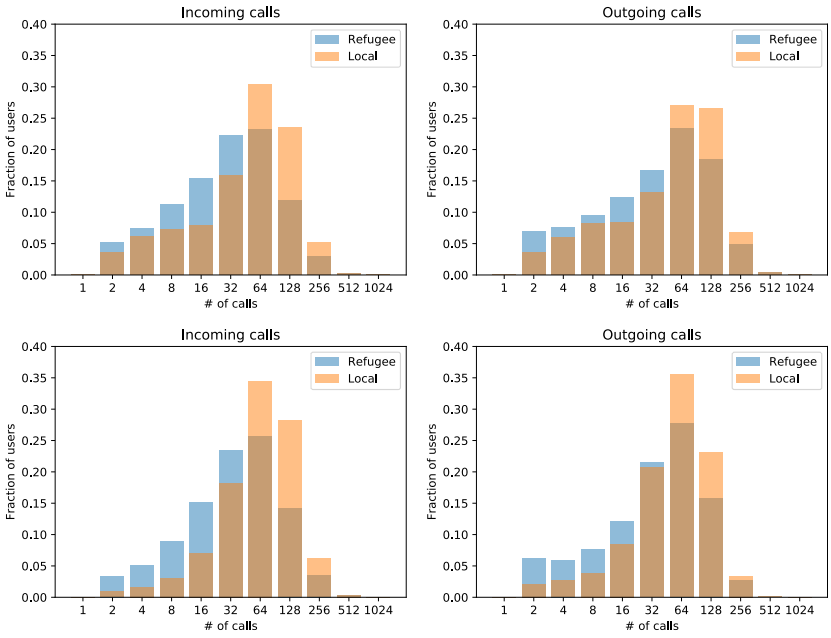


Fig. 4. Activity patterns for Wave 20 (top) and 23 (bottom). The plots show normalised histograms of the number of calls per user for R and N groups, separated into outgoing and incoming call; please note the logarithmic bin size.

as non-refugees, is also increasing. It is unclear whether this trend reflects integration of refugees over time, an increase in the overall refugee population, or simply an increase in the refugee population of Turk Telekom users. Finally, for both outgoing and incoming calls, there is a sharp (and persistent) drop in the number of calls around mid-2017. We are not clear about the cause of this sudden change, but the timing is consistent with changes in other measures we present in subsequent sections.

Figure 6 reports the actual and smoothed number of outgoing calls for R and N users on a daily basis from Dataset 1. The timeline of the call density of R users is in line with the one emerging from Dataset 2 (see Figure 5). The consistency of the two datasets is important, as we combine the information coming from Dataset 1 and 2 in Section 4.

3 Call propensities

Dataset 2 provides information on the status (R, N or U for unknown) for both the caller and the callee. The Dataset is partitioned into outgoing or incoming call depending on whether the user included in the sample is the caller or the

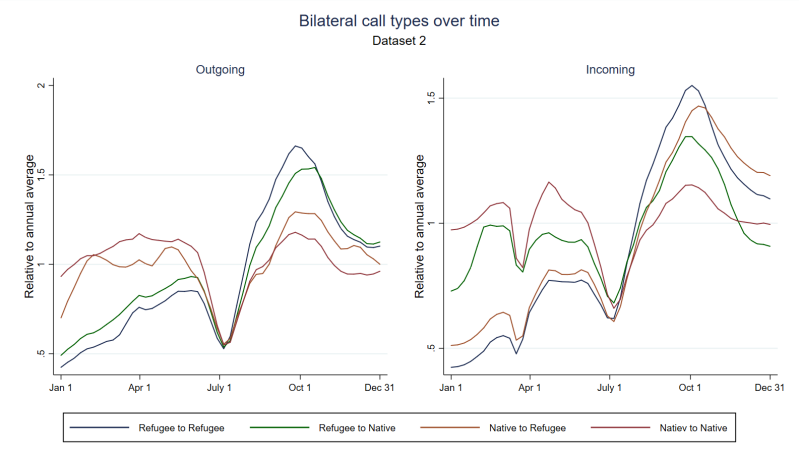


Fig. 5. Smoothed normalized daily number of outgoing (left panel) and incoming (right panel) calls between a caller $g \in \{R, N\}$ and a callee $h \in \{R, N\}$ over 2017; the average number of each type of dyadic calls over the year is normalized to 1; there are 82 days for which the Dataset 2 contains no data

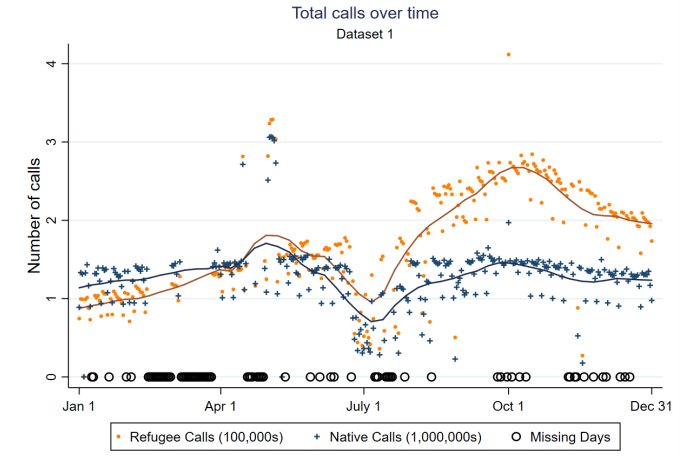


Fig. 6. Actual and the smoothed daily number of outgoing calls for R and N users over 2017; calls for R and N users are reported using a different scale; there are 82 days for which Dataset 1 contains no data.

callee. These two portions of Dataset 2 allow us to estimate the propensity of each type of call (R-to-R, R-to-N, N-to-R and N-to-N) for an average user

in a given wave, these propensities are estimated separately for both incoming and outgoing calls.[8] This analysis is done at the province level, providing 8 propensities measures for each province-wave pair. A simple analysis of these propensities lead to a few basic facts. First, a majority of calls are made to non-refugees, this is true of both non-refugee and refugee users. Second, non-refugees make and receive calls at higher rates than refugees, making around 60 outgoing calls over a two-week period as opposed to refugees who make around 40, as we also saw in Figure 4. Also, the probability that a refugee calls another refugee is directly related to the number of refugees in a given area; Figure 3 plots the share of R-to-R calls over all the calls made by R users for each province using the data from Wave 22 (October 23 to November 5), provinces where a larger number of refugees are located tend to have higher rates of refugee-to-refugee calls (see Figure 3). These propensities will also be combined with the data on the antenna traffic (separately by type of user) from Dataset 1 to obtain a time-varying estimate of the number of R and N users at various degree of spatial resolution.

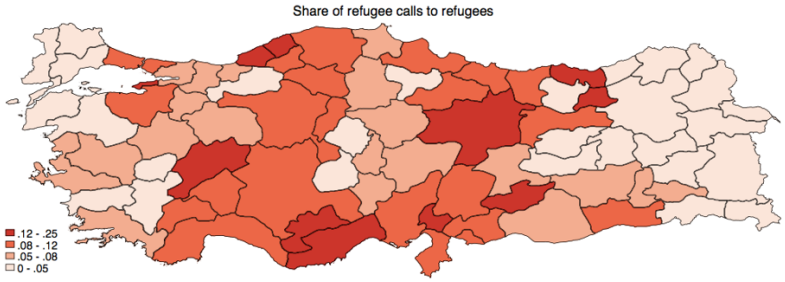


Fig. 7. Share of R-to-R calls among all calls made by R users with data from Dataset 2 (Wave 22).

4 The EI index

We rely on the EI index, introduced by [5] and originally proposed to measure homophily to analyze the frequency of calls across the R and N groups. We compute the index for every month (more precisely, 13 four-week periods). For a given province i in month $t = 1, ..., 13$, and for $g, h \in \{R, N\}$, we define $m_{it}(g, h)$ as the number of communications where the sampled user was located in province i and belongs to group g, while the other end of the communication

[8] The propensity is defined as the average number of calls a type of user performs towards another type of user.

belongs to group h (his or her location is unknown). We can also define $m_t(g,h)$, equivalently defined at the national level.

The EI index is defined as the ratio between the difference of between-groups (or external) and the within-groups (or internal) calls over the total number of calls:

$$EI_{it} = \frac{m_{it}(R,N) + m_{it}(N,R) - [m_{it}(R,R) + m_{it}(N,N)]}{m_{it}(R,N) + m_{it}(N,R) + m_{it}(R,R) + m_{it}(N,N)} \quad (1)$$

We clearly have that $EI_{it} \in (-1,1)$, with low values indicating few connections between groups, while high values indicate many connections between groups, i.e., better integration.

In Figure 8, we plot for each t the distribution of EI_{it}.[9] Overall, the distributions are strongly centered around negative values, indicating few between-groups communications. Still, we must remain cautious in interpreting the absolute values of this index. $EI_{it} = 0$ is equivalent to having the same amount of communications between groups and within groups, but given the strong imbalance between group-sizes, we can only expect negative values. However, the distribution seems to evolve toward more integration, or at least to have a larger dispersion overtime.

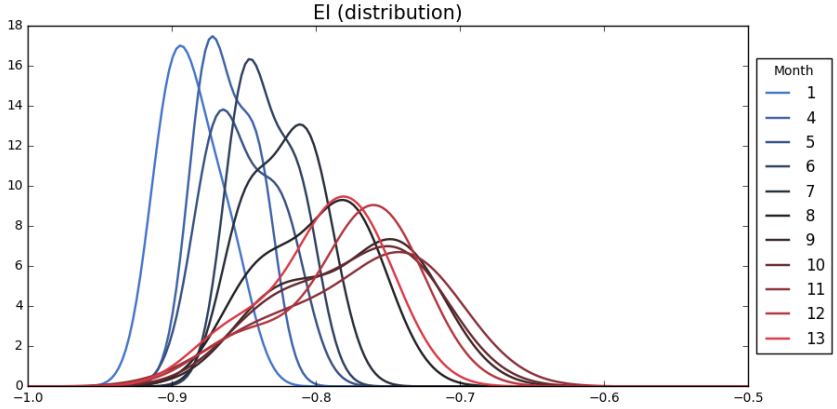


Fig. 8. Histograms showing the distribution of the EI in the 82 provinces. Each line corresponds to one 4 week time period.

In Figure 9, we plot the evolution of the EI index overtime for the five largest cities and for the index computed at the national level. The same general trend appears: integration seems to improve overtime. However, while the time-window is too narrow to make any definitive statement, all series suggest that this increase eventually stabilizes if not reverses at the end of the year.

[9] Data are mostly missing for "month" 2 and 3 so they are dropped.

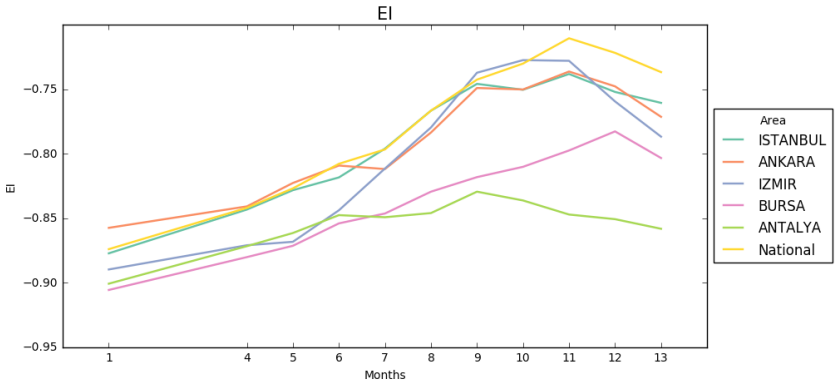


Fig. 9. Time evolution of the EI for the 5 provinces with the largest sample sizes.

It is important to mention, however, that part of this increase in integration could be an artifact of the change in the ratio of R and N users over time. An increase in the (absolute and relative) number of calls involving R users (see Figure 5) entails a reduction in the EI index, as the number of within-group calls for N users remains unchanged, while the three other types of calls increase.[10,11]

5 Dissimilarity Indices

The most commonly used metric of segregation is the index of dissimilarity (D) originally introduced by [3] and [4]. The basic formula for the index is given by:[12]

$$D_i = \frac{1}{2} \sum_j \left| \frac{r_{ij}}{r} - \frac{n_{ij}}{n} \right| \tag{2}$$

where r_{ij} is the population of group R in the $j-$th area of the province i. and r is the total population of the group in the province (n_{ij} and n are similarly defined). In our context, provinces (such as Istanbul or Mardin) are the regions and each area j is the catchment area of each cell-tower within a province. The dissimilarity index for a province i, D_i, is a measure of the evenness of the distributions of the two groups across the area of that province. We can interpret

[10] This happens even if we allow the number of within-group calls for R users, i.e., $m_{it}(R, R)$, to be a quadratic function of the number of R users in the sample.

[11] We have computed an Herfindahl-Hirschmann Index of the concentration across Turkish provinces of R users in the sample for each two-week sample in Dataset 2; higher (lower) number of R users in the sample are associated with a lower (higher) value of the Herfindahl-Hirschmann Index, revealing a weaker (stronger) concentration.

[12] Eq. (2) omits the time subscript, but the index D_i is actually time-varying.

the index as the percentage of a group's population that would have to move to obtain the same percentage of that group within the overall province. The index D_i ranges from 0.0 (complete integration) to 1.0 (complete segregation). Notice that the index D_i is unaffected by a time-varying size of the sample of R users, provided that the spatial distribution of the samples is uncorrelated with their size. Under random assignment, D_i will still be greater than zero as population sizes will vary slightly due to random variation. To test the extent of this factor, we calculate the segregation in call volume for the roughly 1.3 million non-refugee callers that appear in all waves of dataset 2 by randomly assigning the users to two equally sized groups. Under this scenario, the 'random' level of dissimilarity averages .11 across all provinces, ranging from .16 in the most segregated province and .03 in the least segregated. As is shown further down, this level 'random' segregation is significantly lower than the amount observed between refugees and non-refugees.

Even though Dataset 2 has call volume of the individual towers from which calls are originating, we choose to use Dataset 1 to measure segregation. Dataset 1 has a significantly larger sample size and there simply are not enough (or even any) observations for a majority of the towers on a given day in Dataset 2. This sampling bias would distort our analysis considerably. The downside of Dataset 1 is that we only know the number of calls from a given tower, not the number of people. In order to link the two databases, we take the provincial level propensities that we calculated from Dataset 2 (R-to-R, R-to-N, N-to-R and N-to-N) for each of the 26 waves. Then, we divide the number of calls originating from each tower in each time period for each group (R or N) by these propensities to estimate the number of refugees (R) and non-refugee (N) populations for each tower area for each time slot, i.e., r_{ij} and n_{ij} in Eq. (2) above. Populations are calculated by dividing the total call volume over a two week period by the province level propensities calculated in Section 3 (using dataset 2). The main assumption of this approach is that call propensities are constant across antennae within a province.

One of our key innovations is that we calculate the dissimilarity index for working hours and non-working hours separately[13]. This distinction allows us to compare and comment on residential segregation and employment segregation between the refugee and non-refugee populations in each province. Figure 10 below plots the Dissimilarity Index D_i for each of the 26 waves of 2017 for both the working and non-working hours. The plot is a weighted average of all provinces in the country. There are two immediate observations. The first is that there is a certain degree of dissimilarity (or segregation) between the refugees and non-refugees but it is declining over time. The temporary jump during the waves of 11-16 corresponds to the time frame where the number of phone calls (in both datasets) decline significantly (see Figures 5 and 6 above). We suspect this is due to the biased sampling issues that need to be explored further. The second observation is that the dissimilarity index for the working (day) hours

[13] Working hours are defined as 8 am to 5 pm, all other hours are assigned as non-working.

is always below that of the non-working (evening and nighttime) hours. This pattern indicates that refugees are more integrated in terms of their work and employment relative to residences.

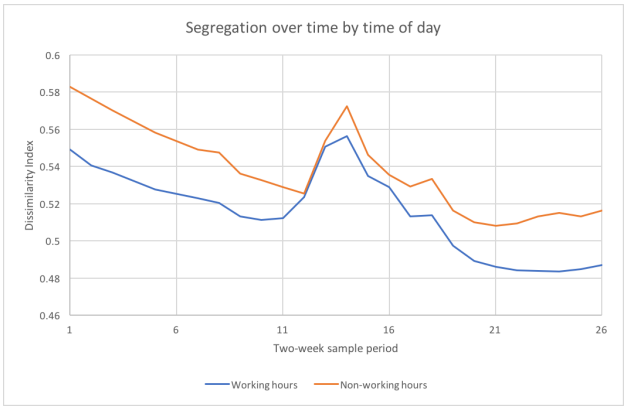


Fig. 10. Evolution of the segregation index D_i; the figure reports the country-level evolution of the segregation index D_i defined in Eq. (2), separately for day (8 am to 5 pm) and night time hours.

Fig. 11. Evolution of the segregation index D_i for Istanbul, Mediterranean provinces and Southeast Anatolia.

As mentioned above, the dissimilarity index is calculated at the province level over time. Figure 11 presents the index for three different geographic areas: (i) Istanbul, home to over half a million refugees, (ii) Mediterranean coast (provinces of Adana, Antalya, Burdur, Hatay, Isparta, Mersin, Osmaniye) and (iii) Southeast provinces, mostly along the border (Adiyaman, Batman, Diyarbakir, Gaziantep, Kahramanmaras, Kilis, Mardin, Siirt, Sanliurfa, Sirnak) using calls from all hours of the day. The declining segregation index over time for all regions are observed in this figure as well. Another striking observation is that the Southeast region, where the refugees make up the largest share of the total population, has significantly lower degree of segregation than the other regions. This could be because, due to their proximity to the Syrian border, non-refugees in those areas already have a greater familiarity with the Syrian population. Or perhaps the refugee populations along the borders represent the earliest to arrive and thus have had more time to integrate into their host communities.

Figures 12 and 13 show dissimilarity indices for the 40 provinces with the highest share of refugees (according to official numbers) for the first and last time periods of 2017. The maps are color coded so that darker shades show the higher levels of segregation. Figure 12 is for the first wave of the year (Jan 1-15, 2017), while Figure 13 is for the last wave (Dec 18-31, 2017). We can see that the southeast provinces are lighter in color than the rest of the country. Furthermore, the overall map for the last wave is much lighter in color, indicating all provinces became more integrated over time.

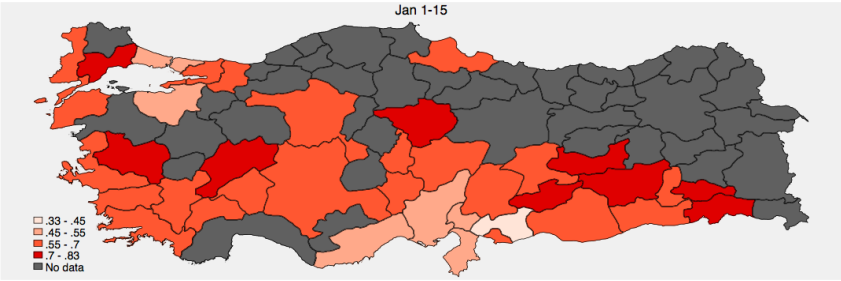


Fig. 12. Province-level measure of the segregation index D_i defined in Eq. (2) computed with the data from Wave 1.

The dissimilarity index D_i in Eq. (2) is the leading index among a large set of indices that have been constructed and analyzed over the last four decades of active research on residential segregation of different communities in many different countries, cities and regions. There are numerous indices that measure other dimensions of communal interaction and integration. Among these are the isolation index (measuring the extent to which minority members are exposed

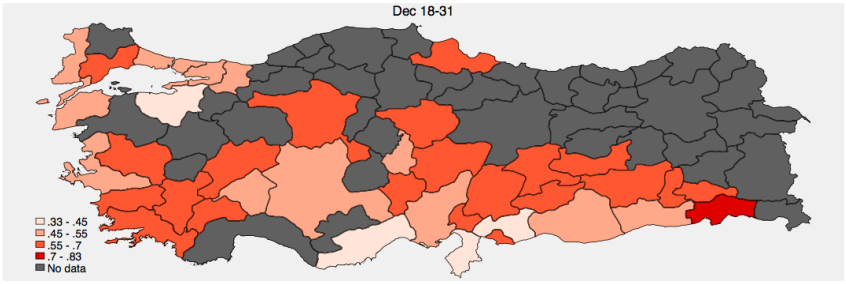


Fig. 13. Province-level measure of the segregation index D_i defined in Eq. (2) computed with the data from Wave 26.

only to one another, see [1]), concentration index (measuring the relative amount of physical space occupied by a minority group in the metropolitan area, see [6]), centralization index (measuring the degree to which a group is spatially located near the center of an urban area, see [6]), or clustering index (measuring the extent to which areal units inhabited by minority members adjoin one another, or cluster, in geography, see [6]). We have performed preliminary analysis of these indices with each one providing important insights on the geographic and social distribution of Syrian refugees within Turkey. They are not included in this report due to space constraints but the same trends across space and time are evident across all indices.

We would also like to identify which province characteristics explain high or low segregation of refugees. For example, preliminary results (see Table 1) indicate that high-refugee provinces (as a share of their total population), as well as larger provinces, experience lower levels of segregation as compared to other low population or low-refugee provinces. Interestingly, when we control for refugee share and overall population, there is no relationship between the amount of cross-group calling and segregation levels.

6 Global Database on Events, Language and Tone (GDELT)

The arrival of Syrian refugees in Turkey is a dramatic social and cultural event with important political ramifications, for the Syrian refugees, Turkey as well as the rest of the world. The geographic and time dimension of the phone call data can be exploited to measure the linkages between concentration of refugees, their social interaction and political events. For this purpose, we integrate the D4R data with another unique database - Global Database of Events, Language, and Tone (GDELT), which we use to measure the extent of refugee-related events across both time and space. GDELT collects news media articles from around the globe in over 100 languages, going back to 1979. Each media observation

Table 1. Simple regression of the segregation index D_i

	D_i
Share of outgoing calls made by R to N	0.015
	(0.012)
Share of outgoing calls made by N to R	-0.014
	(0.012)
Ln(population)	-0.037**
	(0.014)
Refugee share of population	-0.252***
	(0.082)
Observations	40
R^2	.369

Notes: Table displays results from a linear regression on a single cross-section of the 40 largest refugee provinces using data from the first two-week period in January, 2017. ***, ** and * denote significance at the 1, 5 and 10 percent level respectively; standard errors between parentheses; call shares transformed to have a mean equal to 0, and a standard deviation equal to 1.
Source: Authors' elaboration on Datasets 1 and 2, refugee and overall population data comes from Turkish Ministry of Interior.

is classified into an event data, a form of data common in political science to study political history in a systematic way. Events are classified by location, a set of actors (e.g., governments, NGOs, refugees, private companies, etc.), a set of actions (e.g., announcements, diplomatic meetings, accidents, etc.) as well as other information that attempts to predict the tone and impact of an event.

For our purposes, we queried all events from January 2016 to June 2018 that were located in Turkey and included refugees as at least one of the actors. This query yielded 119,000 events over the 2.5-year period, although only 22,431 events occur in 2017 (the year which overlaps with the D4R phone data). Of those events in 2017, 9,498 include a specific province in which the event occurred. Other events are either national in nature without specific assignment to a province or the GDELT text-processing algorithm was simply unable to assign a location.

Using the events data, we constructed a daily panel of events across 81 Turkish provinces. Observations include the number of daily events as well as the average tone of events (tone is calculated from a textual analysis of the media article and is done by GDELT.) We also include a weighted measure of events and tones in which the weight of each event is calculated as the square root of the number of news articles that mention an event. Figure 14 presents the distribution of the events that were extracted from GDELT for the whole country. Spatially, events are most prevalent in Istanbul, Ankara (as the political center of Turkey), the southeast region of Turkey which borders Syria, as well as west-

ern regions along the Aegean coast (coinciding with common departure points of refugees attempting to enter Europe). There is also significant variation over time; a substantial portion of events occur in the first three months of 2017 and there are important surges in June and September (see Figure 15).

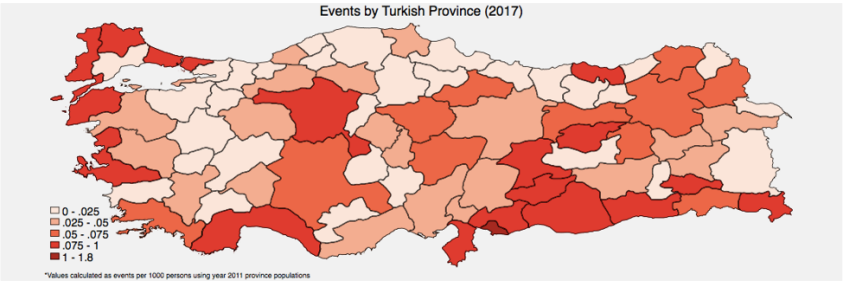


Fig. 14. Events by Turkish province in 2017

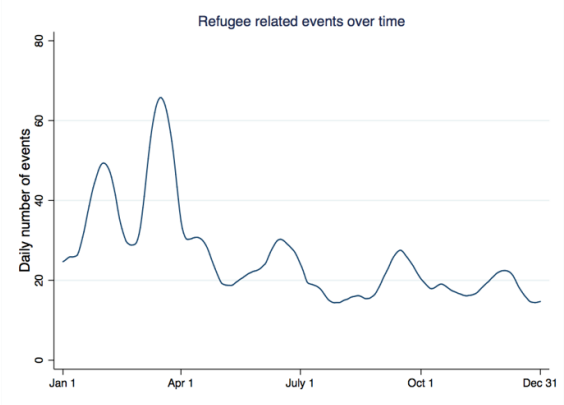


Fig. 15. Daily number of events from GDELT

The critical feature of the GDELT dataset is that it has both the time and space dimension and can be matched with the D4R datasets. There are several potential paths we can follow, linking the two datasets. For example, Figure 16 below plots (natural log of) number of calls to the (natural log) number of events in GDELT where each dot represents (binned) province-day level of observations.

The plot shows that, controlling for province-level effects, refugee-related events are correlated with increased call volume. Further analysis (not shown) implies that this increase is driven by native, rather than refugee, call volume. These results are also robust to removing Ankara from the analysis (because of it's political importance, Ankara is an area associated with roughtly 40% of refugee related events). We can go further and link the call propensities (R-to-R, R-to-N, N-to-R or N-to-N) or the dissimilarity/segregation indices with the GDELT indices we constructed. In addition to the number of events in GDELT database, another valuable measure is the emotional tone of the events. This feature is especially informative on a topic such as refugees and their social and economic integration in the host community. The whole issue is highly charged in terms of politics and emotions and this dimension is one of the key issues we intend to explore further.

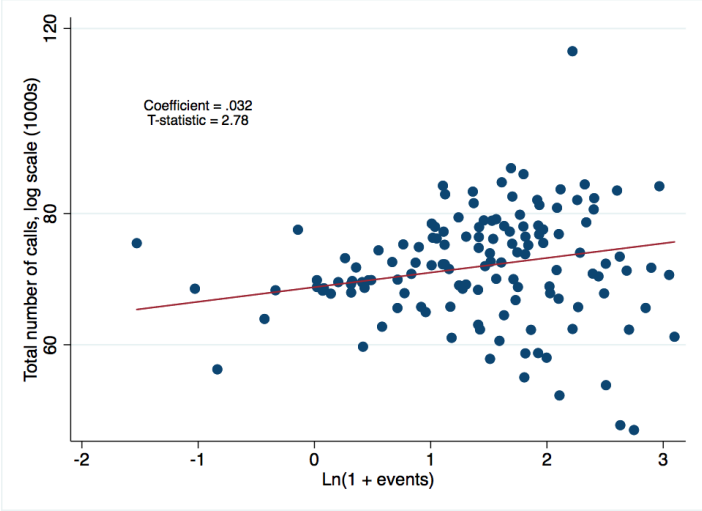


Fig. 16. Events from GDELT and calls from Dataset 1; the figure shows a binned scatterplot of the log number of events the against the log number of phone calls at the province-day level and the corresponding linear best fit line, only including observations that experience more than 10,000 calls per day. Plot controls for province-level fixed-effects

7 Housing data

Economic and cultural assimilation of refugees depends critically on where they live and work. Section 5 showed the existence of segregation between the refuges

and the non-refugees, with considerable variation across provinces. Furthermore, we saw that segregation was declining over time across all provinces in the period covered by the D4R datasets.

In order to further explore the determinants of these integration/segregation patterns, we turn to data on Turkish real estate markets.[14] The data includes monthly indices for both rental and sales prices for close to 1,000 distinct real estate markets across the country. Some of these markets are at the provincial level (for smaller provinces) and others are at the neighborhood level for big cities like Istanbul. For the time being, we aggregated their real estate sales and rental price data to the provincial level but the data would allow us to conduct quite disaggregated analysis taking advantage of the geographic distribution derived from the D4R dataset. For 62 of 81 provinces, our indices begin in 2012 or earlier (before the largest inflows of Syrians began), while the remaining 19 indices do not begin until 2015. In addition to price data, there is also data on residential sales volume, again, at the provincial level. The sales data include the monthly number of sales disaggregated by primary and secondary sales, which represent new construction and resale of existing houses, respectively. These indices begin in 2013 for all provinces and are based on government registration records.

A cursory look at the data indicates a distinct break in trend between high and low refugee areas beginning in 2014 among both prices and volume. Figure 17 presents the rental price indices for three regions of the country–Istanbul, Mediterranean coast and the Southeast Anatolia along the border. The surprising observation is that prices in the Southeast, the region with the largest relative number of refugee inflows, have trended below the other regions since 2014. The price difference between Istanbul and the Southeast increased by more than 50 percentage points between 2014 and 2018, even though they were following a nearly identical trend prior to 2014. Given the sharp increase in demand due to the refugees, we would expect the opposite trend and is not consistent with a sharp housing demand shock.

There are a few forces that can explain this rapid and surprising price divergence between low and high-refugee markets. We believe this phenomenon is explained by rapid supply response and changing composition of housing quality. Figure 18 shows the sales volume of primary housing markets respectively (sales of new construction) where sales in Southeast Anatolia increased drastically compared to other regions. If the Southeast Anatolia region had followed the same path as the comparable Mediterranean region, it would have experienced 16.9 thousand fewer primary market sales. This rapid increase is indicative of a sharp positive supply response of the construction sector. Similarly, Figure 19 shows the sales in the secondary market (of existing homes) where we again see rapid increase in sales. When we look at the prices in the secondary market, we again see a decline, implying increased sales of lower quality homes.

[14] The data come from REIDIN Data and Analytics, a leading provider of real estate data and information for emerging markets, under a confidentiality agreement.

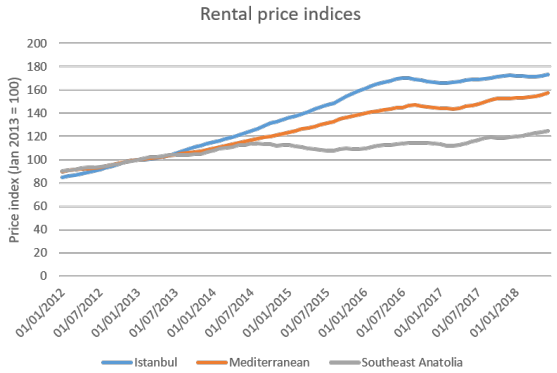


Fig. 17. Evolution of rental price indices for Istanbul, Mediterranean provinces and Southeast Anatolia

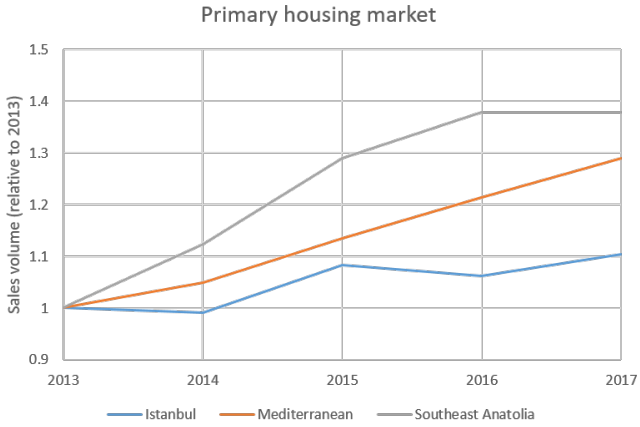


Fig. 18. Primary housing market for Istanbul, Mediterranean provinces and Southeast Anatolia.

Our next step is to link the segregation indices with rental/sales price data to identify the causal links between real estate markets, integration and social interaction of refugees.

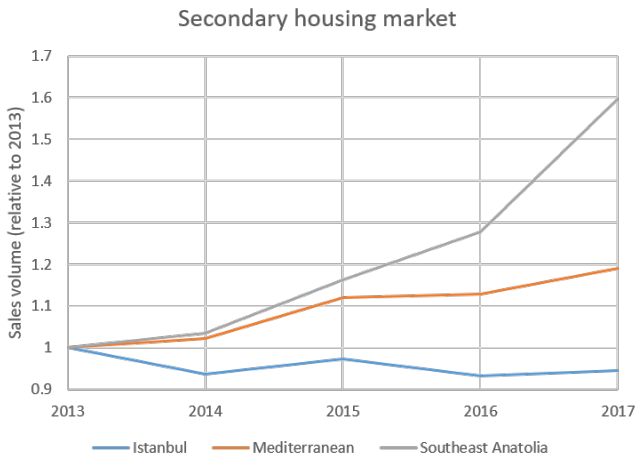


Fig. 19. Secondary housing market for Istanbul, Mediterranean provinces and Southeast Anatolia.

8 Conclusion

The analysis presented in the previous sections reveals that Syrian refugees in Turkey have become more integrated (in terms of communication) and less spatially segregated over the period covered by the D4R Challenge, albeit the various measures of integration (notably, the EI and dissimilarity indices) exhibit a certain degree of spatial variation across provinces. In terms of specific results, we find that the communication between refugees and non-refugees increased over time as indicated by the propensities to call each other. Similarly, spatial segregation of refugees as measured by the dissimilarity index has declined, especially in provinces where refugees make up a higher share of the population. Finally, spatial segregation during the day is lower than at nighttime, implying labor market segregation is lower than residential one. All of these measures indicate improved integration of the refugees into the society.

We performed two additional analyses using GDELT database on events and Reidin database on real estate prices. Both of these analyses were more exploratory in nature, highlighting the possible research avenues while providing preliminary results. GDLET data show there is positive correlation between events and call volume while the housing data reveal that real estate prices did not increase as much as expected, possibly due to increase construction.

The value of D4R dataset for academic research and policy evaluation can be significantly increased by extending the amount of information included in the D4R datasets. For example, a more detailed description of the data collection and sampling procedures would be useful, and possibly by including a larger

sample of the non-refugee population. Since the results depend highly on the way natives and refugees were selected to be included in the in the D4R sample, any bias in the sampling procedure will influence results. Furthermore, it would be useful to be able to extract all the calls initiated by R/N users in given province since this is the only dataset that has information on point to point (R to N) communication. We are hopeful that the path paved by this initial D4R dataset will stay open and data from later years will also be made available to explore critical economics, social and cultural integration issues of refugees. The lessons learned will not only be useful for the Syrians in Turkey but for millions of other refugees all over the world.

References

1. Bell, W.: A probability model for the measurement of ecological segregation. Social Forces **32**(4), 357–364 (1954)
2. Chiswick, B.R.: The effect of americanization on the earnings of foreign-born men. Journal of Political Economy **86**(5), 897–921 (1978)
3. Duncan, O.D., Duncan, B.: A methodological analysis of segregation indexes. American Sociological Review **20**(2), 210–217 (1955)
4. Duncan, O.D., Duncan, B.: Residential distribution and occupational stratification. American Journal of Sociology **60**(5), 493–503 (1955)
5. Krackhardt, D., Stern, R.N.: Informal networks and organizational crises: An experimental simulation. Social Psychology Quarterly **51**(2), 123–140 (1988)
6. Massey, D.S., Denton, N.A.: The dimensions of residential segregation*. Social Forces **67**(2), 281–315 (1988)
7. Salah, A., Pentland, A., Lepri, B., Letouzé, E., Vinck, P., de Montjoye, Y., Dong, X., Dağdelen, O.: Data for refugees: The d4r challenge on mobility of syrian refugees in turkey. arxiv preprint arxiv:1807.00523. (2018)
8. World Bank: Moving for Prosperity: Global Migration and Labor Markets (Policy Research Reports). World Bank Publications (2018), http://www.worldbank.org/en/research/publication/moving-for-prosperity

Mobile phone records for exploring spatio-temporal refugee mobility: Links with the Syrian war and socio-economic variations in Turkey

Fatima K. Abu Salem[1], Al-Abbas Khalil[1], Mohammed Al Khatib Al Khalidi[1], Sara Awad, Yasmine Hamdar, Hani Sami, Joachim Diederich[1], Wassim El Hajj[1], and Shady El Bassuouni

American University of Beirut, P. O. Box 11-0236, Beirut, Lebanon
fatima.abusalem@aub.edu.lb

Abstract. This report attempts to explore mobile phone records in order to gain insight into Syrian refugee mobility patterns in Turkey due to large scale events surrounding the Syrian war, and socio-economic variations across Turkish provinces. Our indicators for mobility revolve around the volume of phone calls, Shanon's entropy, and the radius of gyration. The complexity and the high dimensional data for uncovering refugee mobility patterns makes it compelling to consider machine learning predictive modeling techniques. To that end, our work is focused on feature engineering of the data in order to prepare it for for predictive modeling using rule-based regression and model trees, artificial neural networks, thus exploring the relationship between mobility and large scale events. It also paves the way for our data to be handled as a multivariate time series prediction problem, where time and large scale events are the prominent input. We also dwell on exploratory analysis of the mobility data derived in order to spot any spatio-temporal trends. Our preliminary results show significant variance of mobility measures for refugees as opposed to citizen callers in cities with low socio-economical development indices, peaking at junctures of large scale events in the Syrian war, and in places that are within close proximity to those events. We hypothesis that refugees feel more welcome and probably less intimidated in cities with lower socio-economic status, and that the effects of the Syrian war are visible in time and space on the disposition of the refugee population in the host country.

Keywords: Social Integration · Mobility Patterns · Large Scale Events · Socio-economic development indices · Mobile Phone Records.

1 Introduction

This report attempts to explore mobile phone records in order to gain insight into Syrian refugee mobility patterns in Turkey due to large scale events surrounding

the Syrian war, and socio-economic variations across Turkish provinces. Our indicators for mobility revolve around the volume of phone calls, Shanon's entropy, and the radius of gyration. Our guiding research question is: can the aggregated mobile phone data recordsprovided by Turk Telekom shed light on whether Syrian refugees been opening up to their surroundings with time? or whether they have been experiencing venturing away from their surroundings with time? Are those indicators of mobility spatially and temporally associated with large scale events of the Syrian war in 2017? as well as with the socio-economic variations across Turkish provinces? The complexity and the high dimensional data for uncovering refugee mobility patterns makes it compelling to consider machine learning predictive modeling techniques. To that end, our work is focused on feature engineering of the data in order to prepare it for for predictive modeling using rule-based regression and model trees, artificial neural networks, thus exploring the relationship between mobility and large scale events. This is meant to pave the way for predictive modeling using rule-based regression and model trees as well as artificial neural networks (see [2], for example). In the event that certain strong associations may be uncovered, one can harness the power of predictive modeling in order to empower policy makers and disaster relief operations managers to be able to predict mobility patterns on time, once a large scale event has been detected. Rule extraction helps render the predictive modeling interpretable to a non-expert. Particularly, SVMs and ANNs, two of the most powerful machine learning algorithms, remain opaque to the public so long as the process by which they they they arrive at their classification or prediction, or the overall knowledge embodied in them, are not made transluscent to the user. For the results of our work to translate into actionable intelligence that can be undertaken by policy makers, rule extraction techniques on SVMs and ANNs can help a layperson validate their under all possible input conditions.

Following our feature engineering, we dwell on preliminary explorations of the engineered data in order to spot any spatio-temporal trends. Our initial observations reveal significant variance of mobility measures for refugees as opposed to citizen callers in cities with low socio-economical development indices, peaking at junctures of large scale events in the Syrian war, and in places that are within close proximity to those events. We hypothesis that refugees feel more welcome and probably less intimidated in cities with lower socio-economic status, and that the effects of the Syrian war are visible in time and space on the disposition of the refugee population in the host country.

Despite certain noticeable trends distinguishing refugee from citizen mobility measures across rural versus some major cities, as well as obvious associations with certain large scale events, the task remains daunting because of the high dimensionality of the data. This makes a compelling case for automated machine learning techniques to be employed thereafter. Our work also paves the way for our data to be readily tackled as a multivariate time series prediction problem (for example, using deep learning), where time, large scale events, and socio-economic development indices, comprise our input.

Our manuscript is organised as follows. In Sec. 2 we describe the feature engineering process we have undertaken, in which we create input features using large scale events surrounding the Syrian conflict in 2017 (Sec. 2.1) and socio-economic development indices for Turkish provinces (Sec. 2.1), as well as output features using the Turk Telekom data and measures of mobility (Sec. 2.2). In Sec. 3, we explore the mobility metrics in certain chosen cities in Turkey, and visualise the spikes in large scale events alongside visualisations of those metrics. In Sec. 4 we conclude with immediate consequences of this work and venues for improvement that can be immediately undertaken following the submission of this report, particularly with regards to rule extraction methods.

2 Feature Engineering

In this section we are guided by the intuition that refugee mobility patterns, which in turn can shed light on refugee social integration or openness, may be associated in time and in space, with large scale events happening around the Syrian conflict, as well as socio-economic variations across Turkish provinces. Despite that we do not yet have a clear understanding as to when the intial time coordinate should be, we now focus exclusively on events in 2017, taking place in Syria and globally. Our understanding of mobility patterns goes by two aspects of individual mobility: the volume of mobility, indicating how large the typical distance traveled by an individual is, and the diversity of mobility, describing how the trips of an individual are distributed over the locations they visited. Those individual metrics are then aggregated at the population level. We describe all of that in the sections below.

2.1 Engineering Input Data: Large Scale Events and Socio-economic indices

Large Scale Events Our scope for large scale events covers many potential sources. Those can be major political statements made by leading policy makers around the Syrian refugee crisis, to shifting political and military alliances happening in the Middle East, to the status of financial aid dedicated for refugees, and finally, to peaks in violations happening in Syria. To this end, we choose to map the information in our dataset against data collected in real time from the Syrian Violation Documentation Center (VDC)[1].

The Violation Documentation Center (VDC) is a non-profit, non-governmental organisation registered in Switzerland that tracks and documents human rights violations from the Syrian war[2]. The VDC accepts funding solely from independent sources. Since its onset in 2011, the VDC data records, in real time, war-related deaths as well as missing and detained people. As stipulated on its website, the VDC adheres to international standards for the documentation of

[1] `https://vdc-sy.net/en/`
[2] https://vdc-sy.net/en/

its data. The VDC relies on reports from investigators and a ground network of internationally trained field reporters, who attempt to cover every governorate in Syria. Reporters collect data in three steps. First, initial information on one or more victims is gathered, from immediate and local sources (for example, hospitals, morgues, accounts of relatives/friends, ..etc). Second, supporting information such as videos or photographs are sought. With this, the account gets confirmed and a record gets established. The VDC remains the only human rights group documenting deaths in the Syrian conflict over the entire duration of the conflict, and making the distinction between civilian or combatant status. The VDC has been a source of valuable information for a wealth of notable public health publications on the human cost of the war in Syria (see [4, 6, 7], for a few examples).

Data from the VDC is available in both Arabic and English, despite that we had to deal with inconsistencies occurring between the two databases. For example, certain violations where reported only in the Arabic version, and vice-versa. Additionally, for some of the names reported in the Arabic version, various different spellings were used in the English version. We aggreated data from the VDC and generated plots that showed frequencies of attacks based on the actor/perpetrator (e.g. government forces, rebel forces, ISIS, ...etc), frequencies of civilan casualties, frequences of certain types of attacks (e.g. chemical attacks, air bombardments, streets shootings, ...etc). From these plots, we identified peaks in those events in the year 2017, and appended them to the remainder of large scale events.

Below are randomly chosen examples of the type of large scale events our dataset consists of:

- VDC Data: ISIS. In Homs. Field Execution. 270 killed.
- Cash Assistance: USA announces an increase in its financial aid to the Syrian crisis. Aid goes to refugees inside Syria and refugees in other hosting regions. More than 566 million dollars in additional assistance.
- Political Statements: Israel cancels a plan to accept 100 Syrian child refugees due to opposition within its government.
- Inaugration: Donald Trump 2017 presidential inauguration.
- Alliances: The Trump administration approved a plan to arm Kurdish forces in Syria, despite opposition from Turkey. The move comes with the aim of helping Kurdish forces to capture Raqqa from the Islamic State of Iraq and the Levant.
- Changing realities on the ground: The Islamic State seizes the town of Al-Qaryatain in the province of Homs in a surprise attack against government forces.

The Large Scale Events dataset contains the following features:

1. Date: The date at which the event occurred.
2. Event: A brief description of the event.
3. Type: Type of event, e.g., attacks, elections, major public statement etc.
4. Alliances: Main actors involved in the event.

5. Against: Opposing actors to alliances that are in turn involved in the event.
6. DateFrom: The date at which the event started (mostly, the same as the "Date" column.
7. DateTo: The date at which the event ended. If the event is still ongoing then this field is left empty.
8. Location: Where the event took place.
9. Longitude: longitude for the location of the event.
10. Latitude: The latitude for the location of the event.
11. Relevant Links: Links to online resources related to the event.

The events were ranked in an ad-hoc manner taking into account common expertise as to the severity of a certain event. Fig. 1 shows the ranking we have adopted on a scale from 1 to 10.

Event Ranking			
Type	Actor	Rank Description	Rank/10
Territorial/Military gains, Attacks	Turkey	Armed Conflict - High	10
	Turkish Army		
	Syrian Army		
	Syria		
	Iraq	Armed Conflict - Medium	10
	ISIS		
	Other values	Armed Conflict - Low	8
Shifting Alliances, Arms Sales	Turkey	Political Changes/Decisions - High	6
	ISIS		7
Aid Status, Financial Aid	Civilians	Aid Status - High	8
	Other values	Aid Status - Low	4
Major Public Statements, Elections	Turkey	Public Announcement - High	7
	Turkish Army		
	Syrian Army		
	Syria		
	USA	Public Announcement - Medium	5
	Russia		
	Other values	Public Announcement - Low	2
Calendar Events	All values	Event/Occasion - Medium	4
Global incidents e.g. terrorist attacks outside the MEA	All values	Accident - Medium	3

Fig. 1. Event Ranking

Socio-economic indices and refugee distribution One of the guiding metrics we rely on is the Socio-Economic Development Index (SEDI) which studies the quality of living of a society based on several economical metrics. We adopt the published rankings of SEDI ranging from 1 to 5 for Turkish provinces as produced by [8]. The rankings are shown in Fig. 2 also adapted from [**?**].

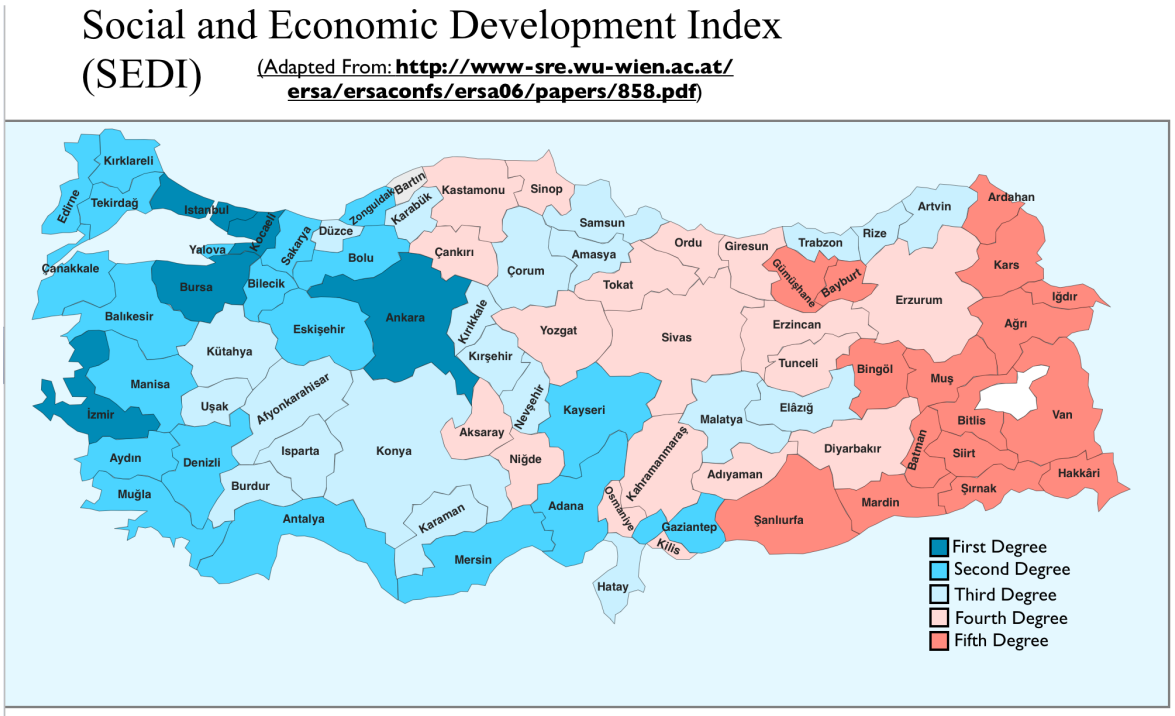


Fig. 2. SEDI Ranking

Finally, to frame our assessment of all mobility metrics in the proper context, we obtain the percentage of refugees per city from *ReliefWeb*[3].

2.2 Engineering Ouput Data: Measures of Mobility

This section makes use of dataset2 provided by Turk Telekom. This dataset provides the cell tower identifiers used by a group of randomly chosen active users to make phone calls and send texts. The data was timestamped and different groups of users were observed for a period of two weeks. At the end of the two-week period, fresh samples of active users were drawn at random. Each sample contains 3% of the refugee base plus equal amount of non-refugee users. To protect privacy, new random identifiers are chosen in every time period. This dataset contains bi-weekly *text* and *voice* traffic data. We do not tackle individual records of this dataset but rather aggregate at the population level distinguishing between callers with a refugee status versus citizen.

[3] `https://reliefweb.int/report/turkey/sultanbeyli-istanbul-turkey-case-study-refugees-towns`

Mobility Diversity: Entropy In line with [11], we measure mobility diversity of a group of people in any particular city by using Shanon's entropy formula (see Formula (1) below). Say we are interested in the mobility diversity of the group of callers who have refugee status. Those individuals can place calls/texts to callees with the following status (1) other refugees, (2) non-refugees, or (3) people of an unknown status. From dataset2 of Turk Telekom, one can obtain the total number of calls/texts made by all refugees to each of the three groups of callees, for each two weeks in 2017. From that, one obtains the probability p_i of a call/text made from the part of a refugee to an individual belonging to each of the three aforementioned groups (where $i = 1, 2, 3$). Entropy can now be calculated as

$$\mathcal{E} = -\sum_{i=1}^{3} p_i \log p_i \tag{1}$$

We have proceeded by calculating weekly entropies for when the callers have been refugees as well as for when the callers have been non-refugees, for each city in the dataset. Entropy captures the measure of disorder in a set. Entropy can be seen as an indicator of diversity and a population's increasing ability to open up to its surroundings. As entropy increases, the more random are the contacts made by individuals in a particular group, and the calls/texts are distributed across many connections. In the same way, as entropy decreases, the less random are the contacts made, and the calls/texts are distribued across a few preferred connections.

Volume of Mobility: Radius of gyration The volume of mobility is a measure that captures how large the typical distance traveled by an individual is. The radius of gyration is a measure of mobility volume that depends, among many factors, on the center of mass of a given individual, defined as the weighted mean point of the phone sites visited by this individual. For example, if an individual visits points $p_1, \ldots, p_j$ for $k_1, \ldots, k_j$ times respectively, the center of mass would be the weighted mean of the points with the weights being given by the k_j's.

The radius of gyration of an individual captures the characteristic distance traveled by an individual [5, 9, 10]. It characterises the spatial spread of the phone sites visited by an individual from their center of mass, and is defined as follows:

$$r(u) = \sqrt{\frac{1}{N} \sum_{i \in S} n_i \|\mathbf{r}_i - \mathbf{r}_c\|^2} \tag{2}$$

where T is the set of phone site visited by the individual, n_i is the individuals calls/texts made from phone site i, $N = \sum_{i \in S} n_i$, $\mathbf{r}_i$ is the vector of coordinates of phone site i, and $\mathbf{r}_c$ is the center of mass of the individual, respectively.

An increasing radius of gyration over time can be used to infer that an individual (or group of individuals) is able to contact people further away from their center of mass, which can shed light onto the level of venturing that they are able to do away from their given place of residence.

We aggregate the radii of gyration for all groups of individual callers in each given city as follows. At the start of week t, if the first outgoing call that an individual makes is from city c, then it is assumed that this individual resides in city c for week t. We then calculate the average radii for all individuals residing in each city during some week t. These groups of callers can be refugees or non-refugees, and the average radii for each such group are calculated separately.

2.3 Final Dataframe

Our final dataframe representing input and output variables consists of the following attributes:

- Input:
- Date: Day the event occurred.
- Military: True/False. States whether the event is a military movement or not.
- Political: True/False. States whether the event is political or not.
- Global: True/False. If false, it means the event concerns Turkey and/or Syria. If true, its an event that might have happened in other countries, but affect Turkey/Syria.
- Location: Nation/City/Null. The place where the event happened. Specific to either nation or city. If Null, it means the event does not have a specific location (for example, religious or other holidays).
- Bombing/Attack: True/False. States if this event is a bombing or attack.
- Injured: Integer. Number of individuals injured during this event.
- Killed: Integer. Number of individuals killed during this event.
- Cash Assistance: Integer. Amount of money given for assistance in US dollars.
- Aggressive: True/False/Neutral. Nature of the event. If true, it is aggressive. If false, it is peaceful. If neutral, it is neither.
- Description: Text description of the event.
- Distance: Distance between the location of the event and the Turkish city. Calculated in meters and by using the longitude and latitude of each location.
- Week: Integer. What week during 2017 did this event occur. It is related to the date of the event. So if "Date of the day" belongs to week 5, then this event would have occurred in week 5.
- City: Name of the Turkish city.
- Ouput:
  1. Radius: Average of Radius of gyration of non-refugees of the Turkish city.
  2. RadiusVariance: Variance of Radius of gyration of non-refugees of the Turkish city.
  3. RadiusRef: Average of Radius of gyration of refugees of the Turkish city.
  4. RadiusRefVariance: Variance of Radius of gyration of refugees of the Turkish city.

5. Entropy: Entropy of the number of calls (text or voice) made by non-refugees. This entropy is calculated with three probabilities: (1) calls/texts made to refugees (2) calls/texts made to non-refugees (3) calls/texts made to unknown.
6. EntropRef: Entropy of the number of calls (text or voice) made by refugees. This entropy is calculated with three probabilities: (1) calls made to refugees (2) calls made to non-refugees (3) calls made to unknown.

A brief snippet of the dataframe is shown below:

Date	Military	Political	Global	Location	Distance (m)	Bombing/Attack	Injured	Killed	Cash Assistance	Aggressive	Description	Week	City	Radius	RadiusVariance	RadiusRef	RadiusRefVariance	Entropy	EntropyRef
20-01-17	FALSE	TRUE	TRUE	USA	10471490.6	FALSE	0	0	0	Neutral	Donald Trum	3	GAZIANTEP	0.057816	0.085380003	0.06958489	0.160104168	0.201642	0.6933822
20-01-17	FALSE	TRUE	TRUE	USA	9677333.98	FALSE	0	0	0	Neutral	Donald Trum	3	ISTANBUL	0.066755	0.123354869	0.06062493	0.097698103	0.138727	0.6590725
20-01-17	FALSE	TRUE	TRUE	USA	10138566.5	FALSE	0	0	0	Neutral	Donald Trum	3	KONYA	0.058505	0.059920141	0.02929218	0.015346681	0.21571	0.5029802
22-10-17	FALSE	FALSE	FALSE	Homs	973689.164	TRUE	0	270	0	TRUE	The organiza	42	ISTANBUL	0.079105	0.126336423	0.08212824	0.153399604	0.122123	0.6749111
22-10-17	FALSE	FALSE	FALSE	Homs	908036.61	TRUE	0	270	0	TRUE	The organiza	42	BURSA	0.087358	0.108069527	0.08515986	0.123563579	0.134921	0.6432252
22-10-17	FALSE	FALSE	FALSE	Homs	954122.385	TRUE	0	270	0	TRUE	The organiza	42	BALIKESIR	0.119924	0.076455475	0.10891508	0.08120277	0.099105	0.401583

Fig. 3. Sample Input/Ouput

3 Visual Data Exploration

In this section we pursue visual exploratory analysis of the data engineered using dataset1, dataset2, and dataset3 from Turk Telekom. In Sec. 2.3 and Fig. 3) we describe the feature engineering performed using dataset2. Dataset1 focuses on antenna traffic and tracks the total number and duration of calls, as well as the total number of SMS messages, from refugees and citizens on each base station in the reported Turkish cities. Dataset3 is similar to dataset1, except that it omits tracking SMS messages, and reduces the spatial resolution by replacing the base stations by the city-district geo-idenitification. To handle each of these datasets we aggregate the individual records at the sub-population level, distinguishing between between callers with a refugee status versus citizen.

3.1 Volume of calls across SEDI's and geolocation

Kilis has the highest ratio of refugees followed by Hatay and Gazaintep. Kilis is the closest city to the Syrian border where it begun to function as a safe zone and buffer city to refugees. Hatay was part of Syria of 1939 before being reclaimed by Turkey and is also on the Syrian border. Moreover, Gazaintep is only 97 kms north of Aleppo, Syria. Furthermore, from dataset3, we observe tha the highest volume of calls in each bi-week throughout the whole year came from the district of Shahinbey in Gaziantep. Quoting from AL-Jazeera, "One of the districts that live with this migration in the heaviest way in Gaziantep, in which Turkey places most of the Syrian [refugees], is Shahinbey district".

Looking into the relationship between SEDI's and the ratio of refugees, Fig. 4 shows that SEDI and ratio refugees are inversely proportional. We hypothesise that refugees probably feel more welcome and comfortable as well as less intimidated in cities with low SEDI's,

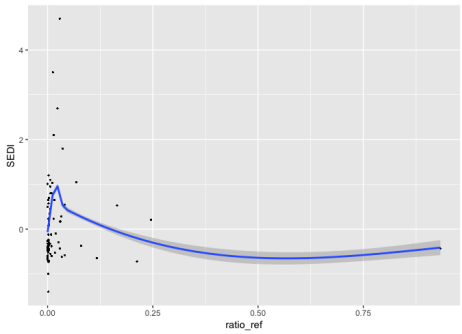


Fig. 4. SEDI vs ratioref

3.2 Spatio-temporal analysis of volume of calls across large scale events

Figure 3.2 tracks volume of calls by refugees for each bi-week of 2017, and shows a massive increase in the number of refugee calls in bi-week 18 (weeks 35-36), designating the end of August and the Beginning of September 2017. During this exact period, the "Qalamoun Offensive" has ended in Syria, followed by the Syrian Defense Forces seizing full control of "Raqqa". Also, in these weeks, the Turkish Prime Minister declares his intentions to "settle the crisis with the involvement of all the key players, including Syrian President Bashar al-Assad".

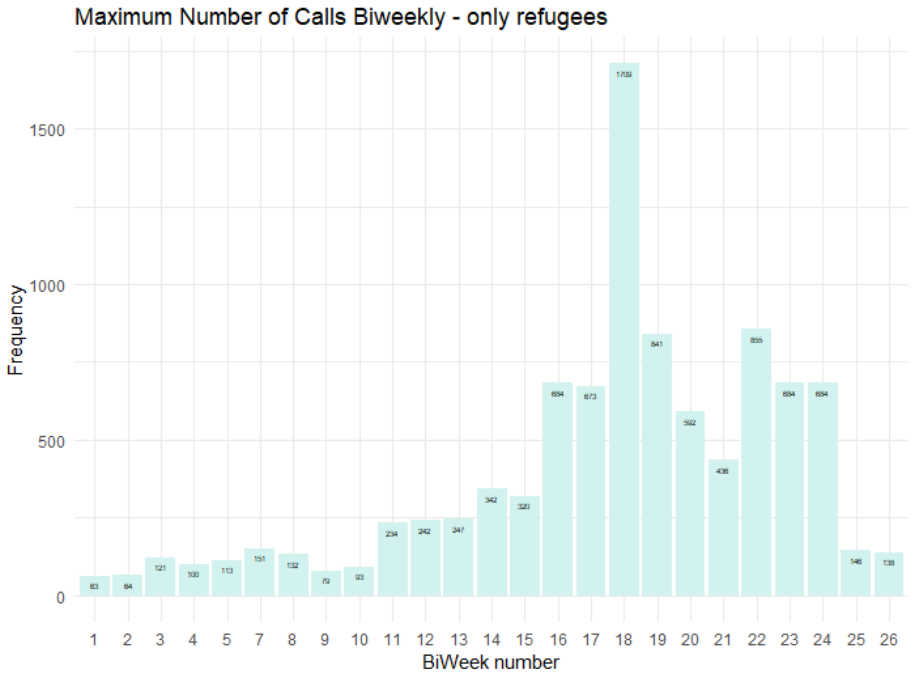


Fig. 5. Maximum Number of Calls Done by Refugees Bi-weekly

During the Qalamoun offensive period, we observe that the district of Fatih in Isanbul shows the highest volume of calls. Particularly, during bi-week 16, the number of outoing calls made by refugees from inside of Fatih is around 30,000 calls. However, the incoming calls from refugees outside of Fatih to inside is around 100,000. During bi-week 17, we observe that the number of outgoing calls made by refugees from inside of Fatih is around 50,000. The number of incoming calls made by refugees from outside of Fatih to inside is a staggering 180,000. We hypothesis that refugees in Fatih keep strong ties with people from Syria-Qalamoun.

Fig. 3.2 shows that a boost in call duration occurred in bi-week (weeks 45-46), signifying the month of November when the Eastern Syria campaign was launched - a large scale military operation to free the city of "Deir ez-Zor". Also, this was a time where the region of Eastern Ghouta was extremely short on food and supplies.

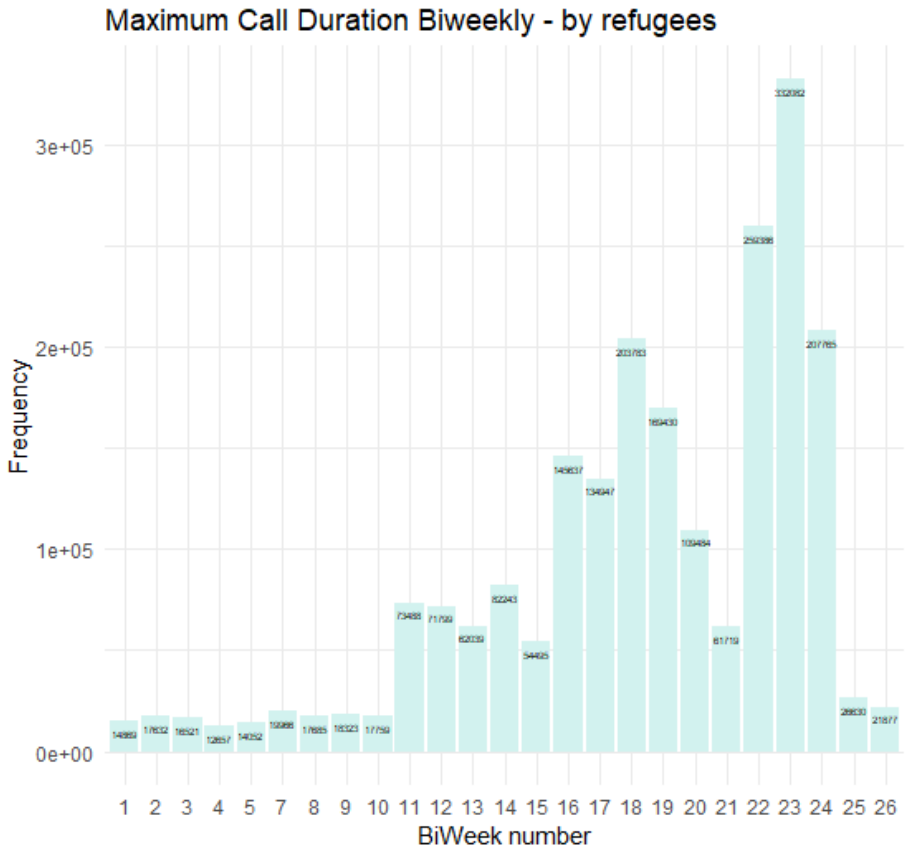


Fig. 6. Maximum Call Duration Done by Refugees Biweekly

Inference 1: Calls made by citizens showed a high increase in a holiday month. However, calls (and their duration) made by refugees showed a massive increase in months coinciding with the named large-scale events related to war in Syria occurred.

Inference 2: The lowest number of calls (and their duration) from refugees came from the first 5 biweekly period. This tells us that a probable influx of refugees into Turkey must have occurred immediately afterwards.

3.3 Mobility Measures across SEDI's

We begin by searching for associations with SEDI's of cities and whether trends are different for the population of refugee callers versus citizen callers. Given the weekly measures from groups with refugee callers, we compute the weekly entropy measures and weekly average radii and variance of radii of gyration, for

each group of cities of a particular SEDI rank $i \in \{1, \ldots, 5\}$. A high variance indicates significant variations in mobility diversity for refugees than other cities. For each of those group of cities, we plot the weekly entropy and radii of gyration averages and variance for groups of refugee callers as well as citizen callers. The aim of this batch of comparisons is to compare how mobility measures differ among refugees versus citizens.

From Fig. 7 and 8: Entropy for refugees is noticeably higher than entropy of citizens across cities of all SEDI's. The entropy for refugees is monotonically increasing until it reaches a peak in the weeks 30-40 of 2017, indicating that the refugees tracked in this dataset managed to diversify their connections significantly across this particular year.

From Fig. 9 and 10, we observe that both reguees and citizens have comparable averages of radii of gyration. However, our most interesting observations come from Fig. 11 and 12. There, we see that generally, the variance of radii of gyration has been higher for refugee callers than citizen callers. Yet, the conspicuously high radii variances for refugees are obesrved in cities with lower SEDI's. We interpret those results to say that, not only do refugees cluster in cities with low SEDI's. Over there, the disparity in openness that refugees are able to achieve in those places is higher, so that despite some refugee callers remain somehow niched within their center of mass, a good number of them are able to venture out with time. This could be possible due to more social or economical integration, where a high variance indicates a healthy dynamism in interactions.

3.4 Mobility Measures across large scale events

In this section we investigate potential associations between refugee mobility measures and large scale events, temporally and spatially, across each given group of cities of a particular SEDI rank $i \in \{1, \ldots, 5\}$. In Figures 13 to 17, and for each group of cities of SEDI rank i, we stack the plots of weekly entropy, average radii of gyration, and variance of radii of gyration, for refugee callers. In the same stacked figures, we plot the average of ranks of large scale events happening in each week of 2017. This captures the temporal component of our analysis. To capture the spatial component, we compute the distance in meters from each large scale event and each of the given cities, and plot the average of such distances from the large scale event and all cities in a given SEDI rank i. Our notable observations are as follows. There exists a clear temporal trend that associates peaks in all tracked mobility measures with the peak in which the large scale events achieved highest rank/impact. This can be seen for all groups of cities in a given SEDI rank to be around weeks 30-40. A large number of remarkable events are tracked from the VDC database during those weeks:

- Iraq launches a U.S.-backed campaign to liberate Mosul from the Islamic State.
- Syrian Desert Campaign.
- Operation to seize Raqqa from ISIL started leading to the battle of Tabqa.
- Syrian government forces retake the Damascus-Palmyra highway from ISIL.

- U.S. begins providing weapons to Syrian Kurds.
- Quneira offensive.
- Qalamoun offensive.
- ISIL withdraws from Aleppo province.
- Trump ends CIA arms support for Anti-Assad Syria rebels.

Spatially speaking, all of the observed peaks take place at a time when the large scale events tracked are at an extremely close average proximity to cities of the given SEDI group. Our interpretation is as follows. In times of insecurities, people begin connecting more within their surroundings and further. They also begin developing a sense that their stay within the host country might be extended which reflects on the diversity and extended stretches of their connections. During those times also, more refugees might be coming into Turkey, which triggers an update in both diversity and displacement metrics. We see this to be completely intuitive as increased group anxiety leads to a desire to maximise information sharing and acquisition. When the number or impact of large scale events decreases, enough information would have been gathered or shared among refugee members, and mobility metrics begin going down.

4 Conclusion

The exploratory analysis presented in this work almost totally paves the way towards predictive modeling in the aim of understanding and being able to predict refugee mobility and social integration and openness as a result of large scale events affecting the Syrian conflict. The dataframes we produced for each city can be readily tackled as a regression/classification problem or as a multivariate time series problem using deep learning, where time and large scale events are input parameters. Rule extraction can be employed to extract the rules behind such associations so that they can be put at the service of policy makers and relief agencies, a task we already began exploring. The following rule extraction methodologies can be explored. Rule extraction from local function networks can be achieved via decompositional algorithms that directly decompile weights to generate rules. For example, Rulex, a tool developed by Andrews and Geva in [1], converts the numeric weights of an artifical network into symbolic *if-then* rules that explain the decisions made by the network. Pedagogical techniques such as classification via decision trees will also be explored. See5/C5 is a system commercialized by Rulequest Research in [12] for analysing data and generating classifiers in the form of decision trees and/or rule sets. The patterns are expressed in the form of a decision tree or a set of *if-then* rules. The Ripper rule learner is a system for inducing classification rules from a set of pre-classified examples and will be used for benchmark purposes. Ripper (Repeated Incremental Pruning to Produce Error Reduction) is an efficient, noise tolerant propositional rule learning algorithm based on the separate and conquer strategy. The basic strategy used by Ripper is to find an initial model and then to iteratively improve that model using an optimisation procedure [3].

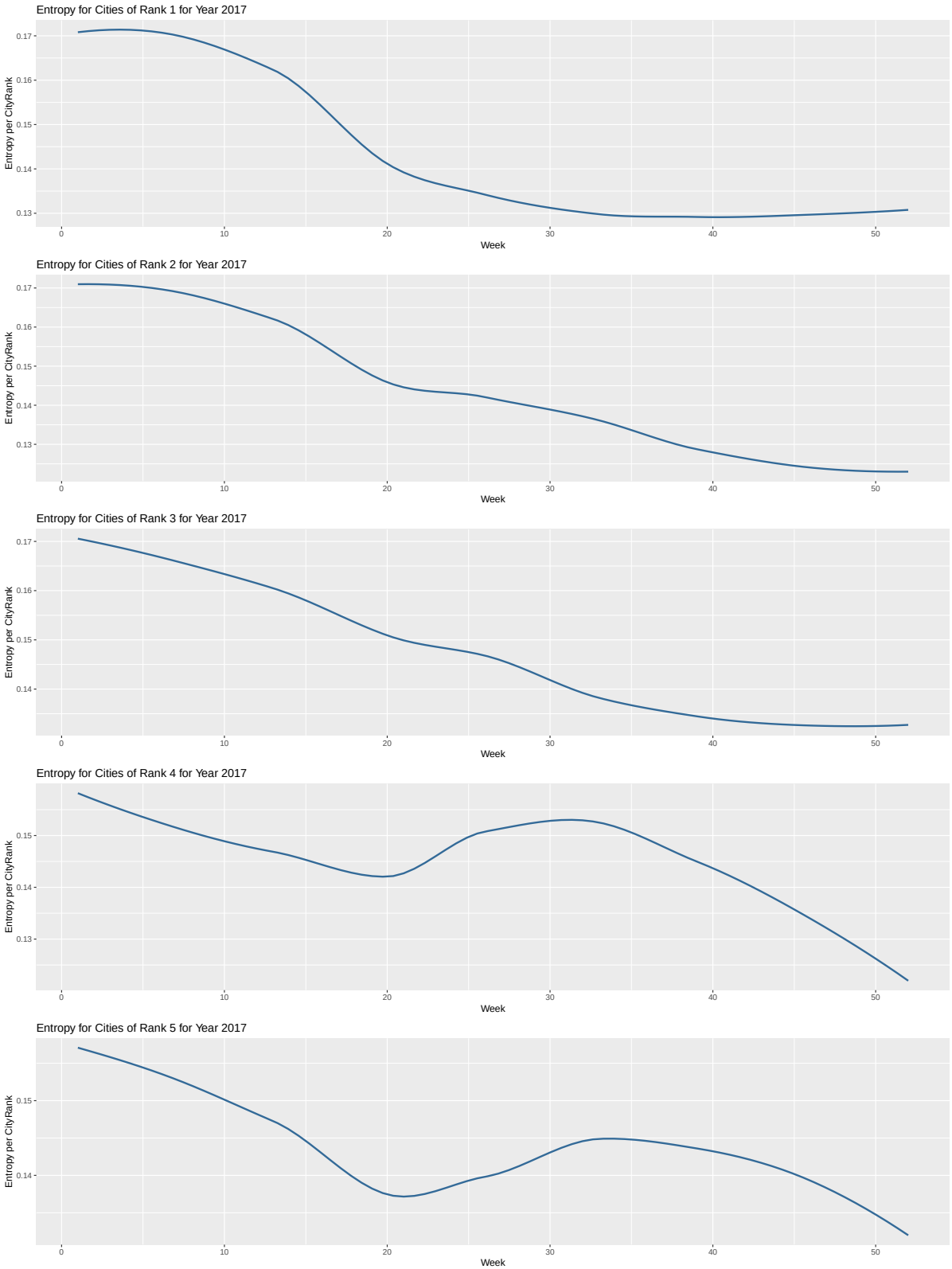


Fig. 7. Weekly Entropy for Citizen callers – All cities

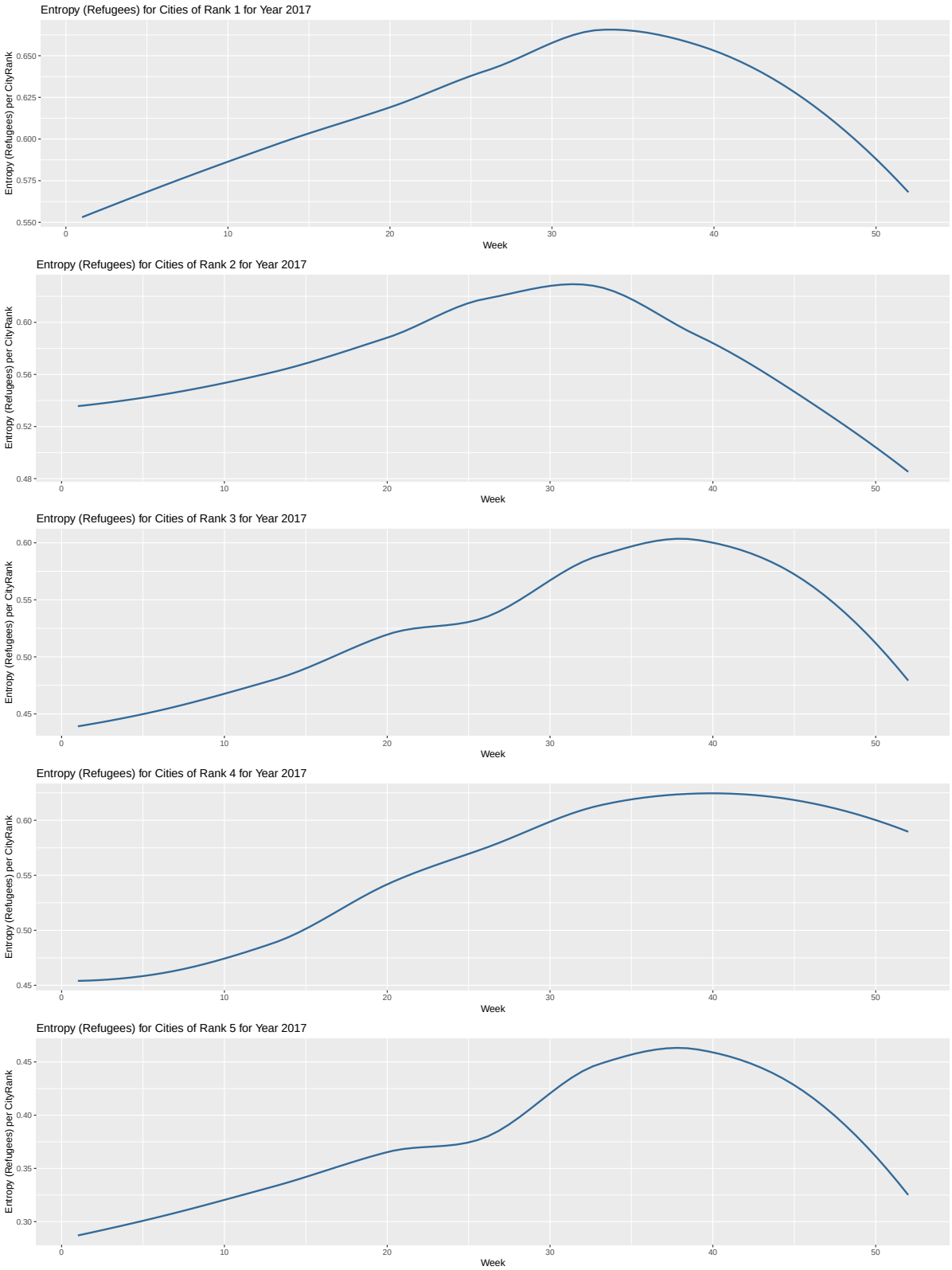


Fig. 8. Weekly Entropy for Refugee callers– All cities

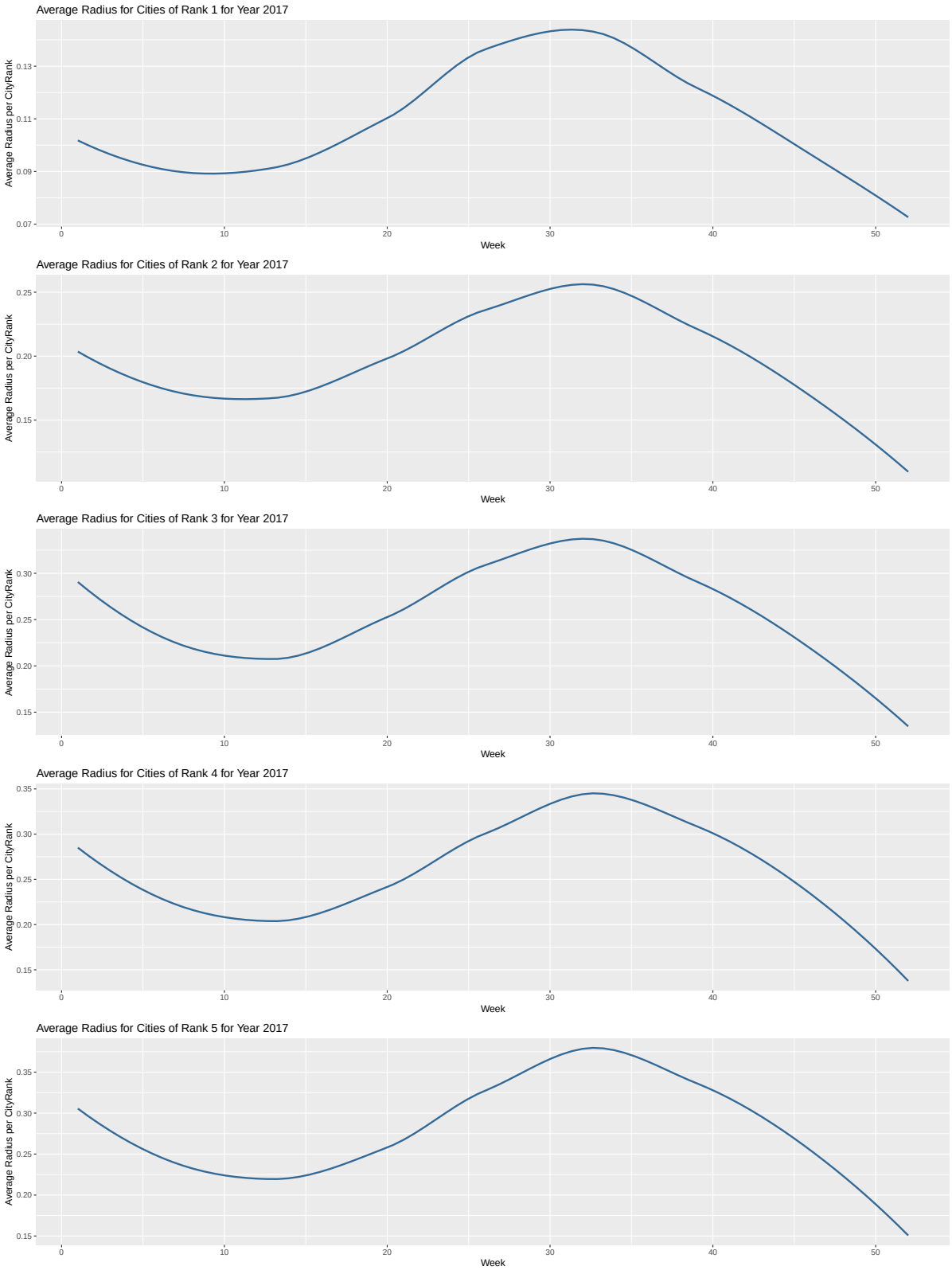


Fig. 9. Weekly Average Radii for Citizen callers – All cities

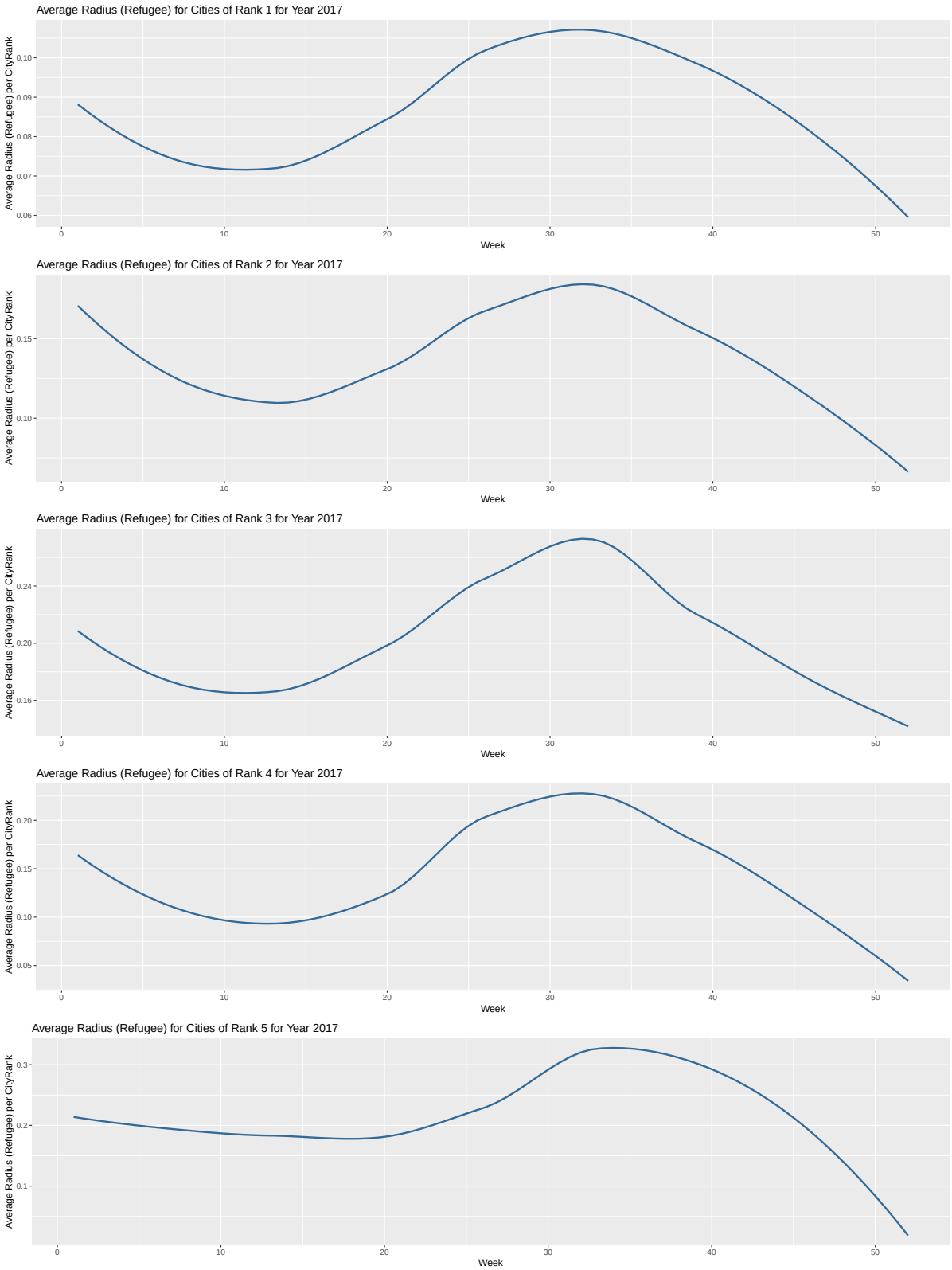


Fig. 10. Weekly Average Radii for Refugee callers – All cities

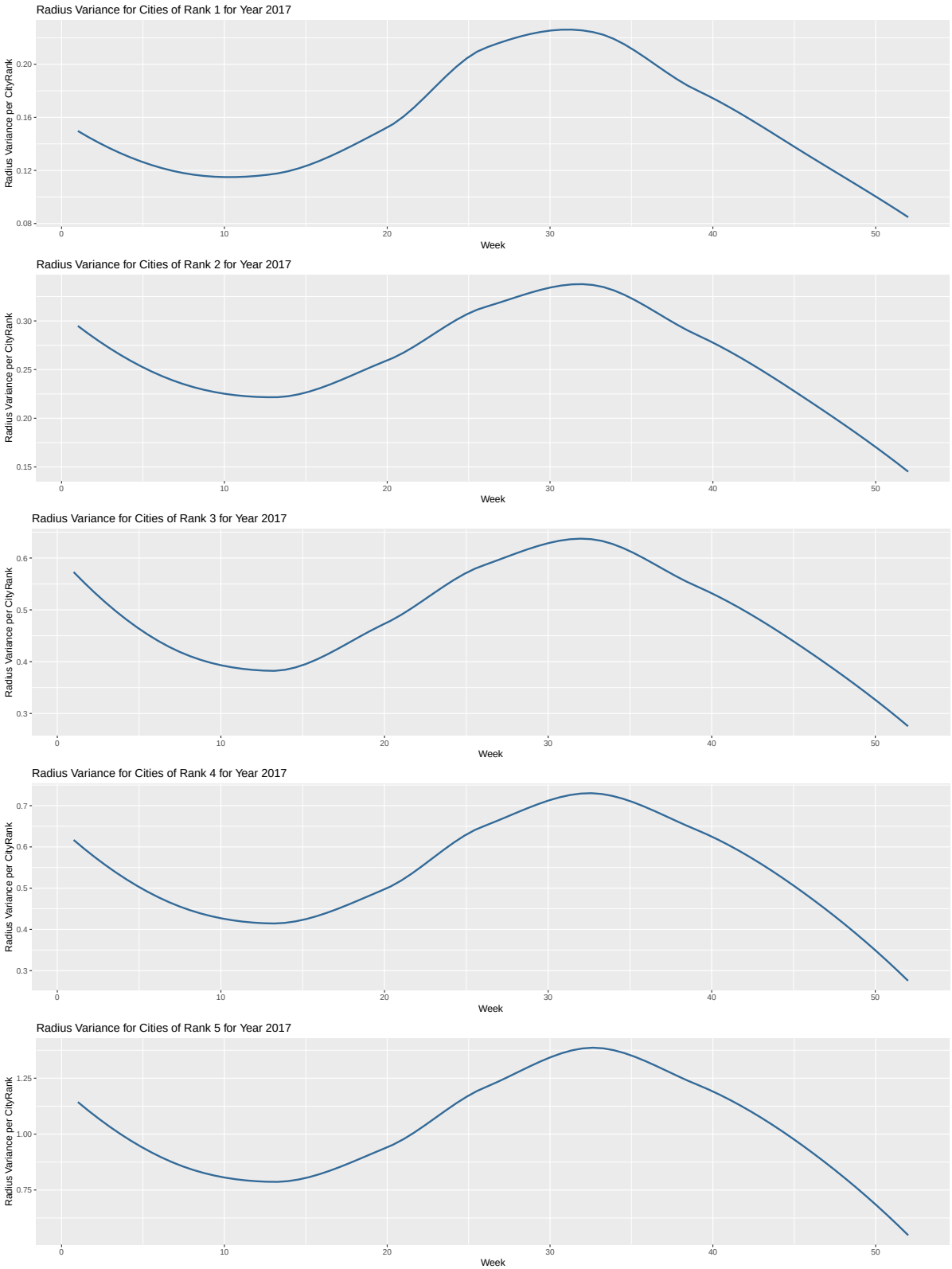


Fig. 11. Weekly Variance Radii for Citizen callers – All cities

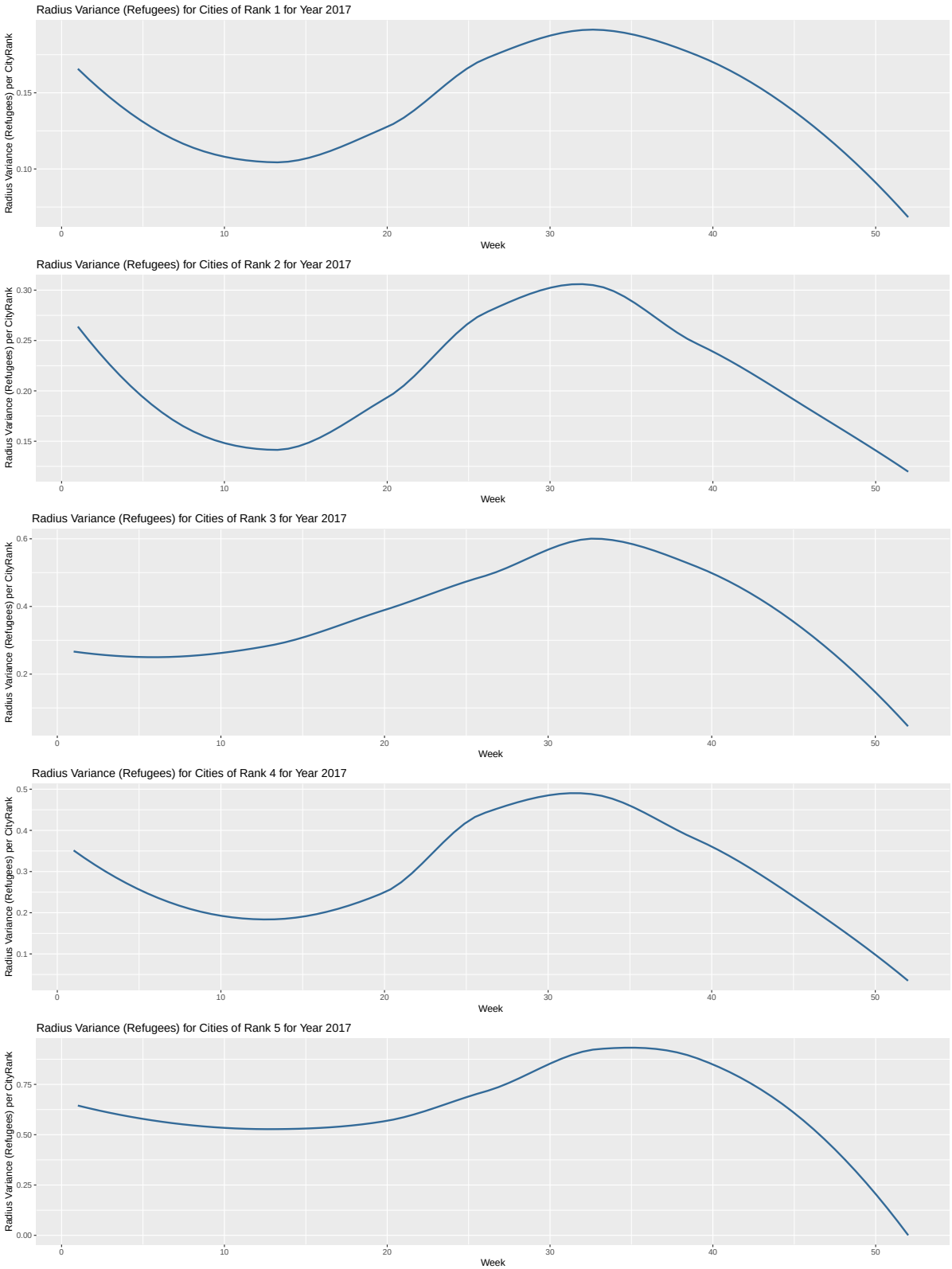


Fig. 12. Weekly Variance Radii for Refugee callers – All cities

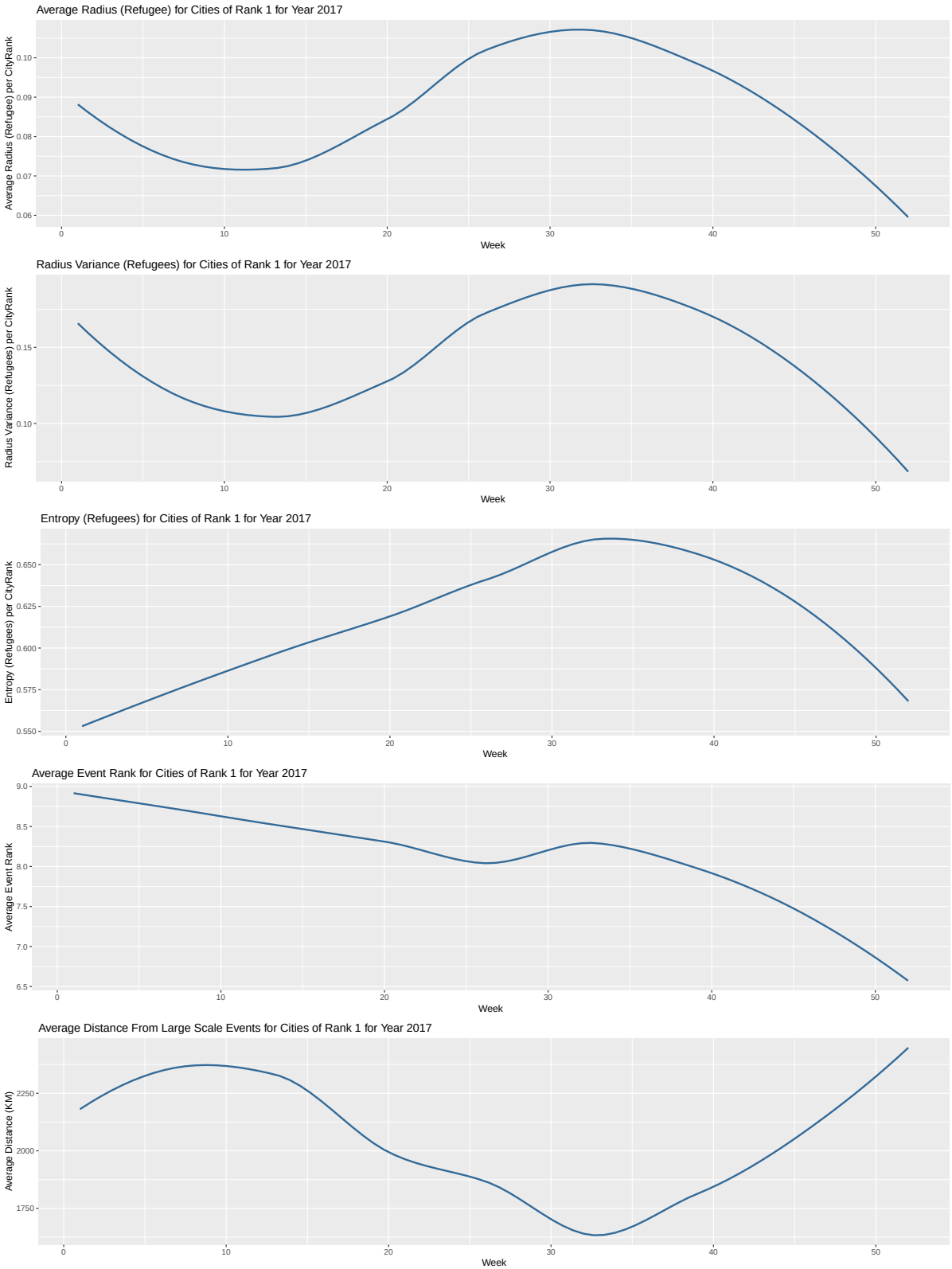


Fig. 13. Weekly Mobility Measures for Refugee callers – Cities with SEDI rank 1

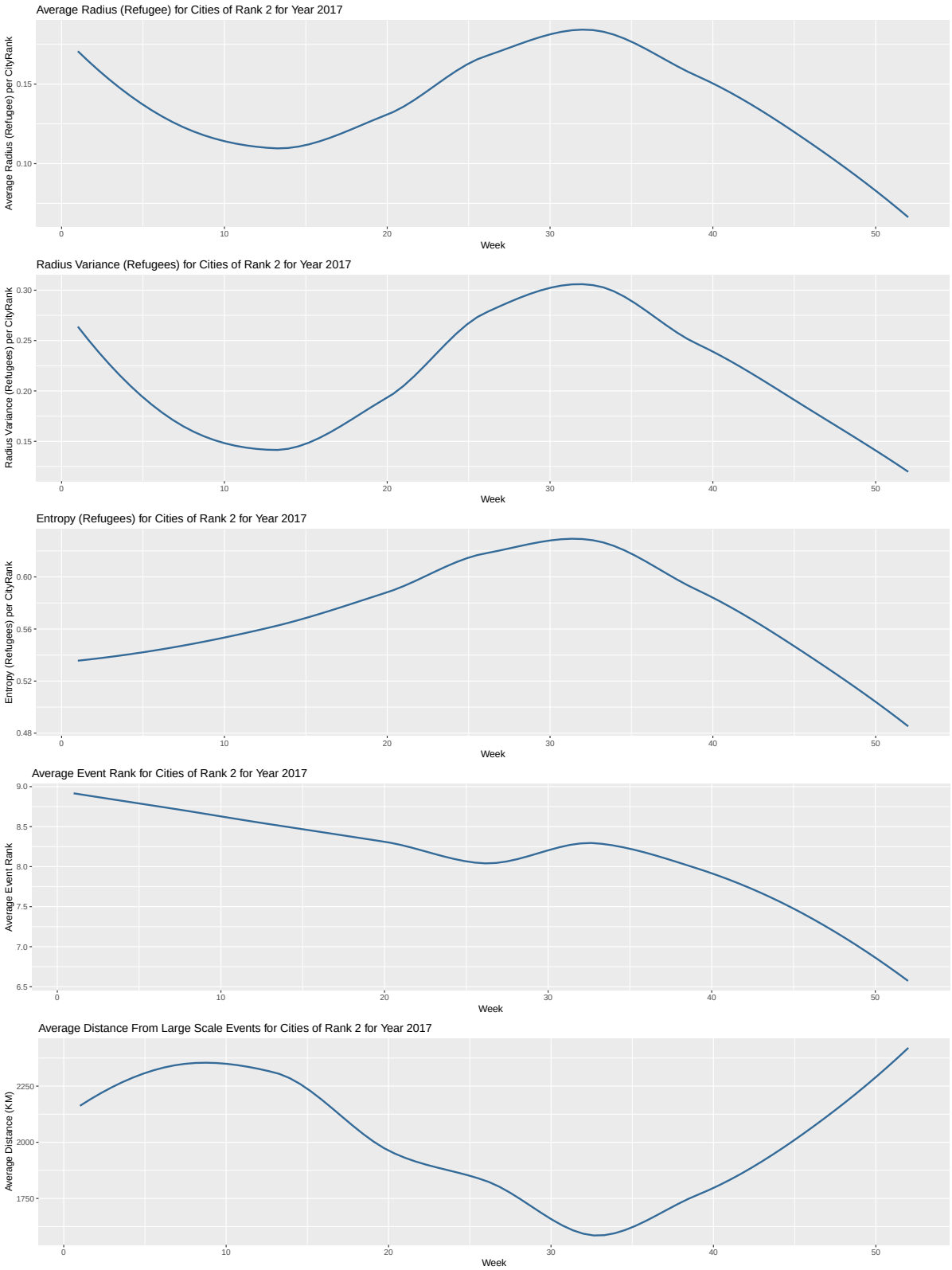


Fig. 14. Weekly Mobility Measures for Refugee callers – Cities with SEDI rank 2

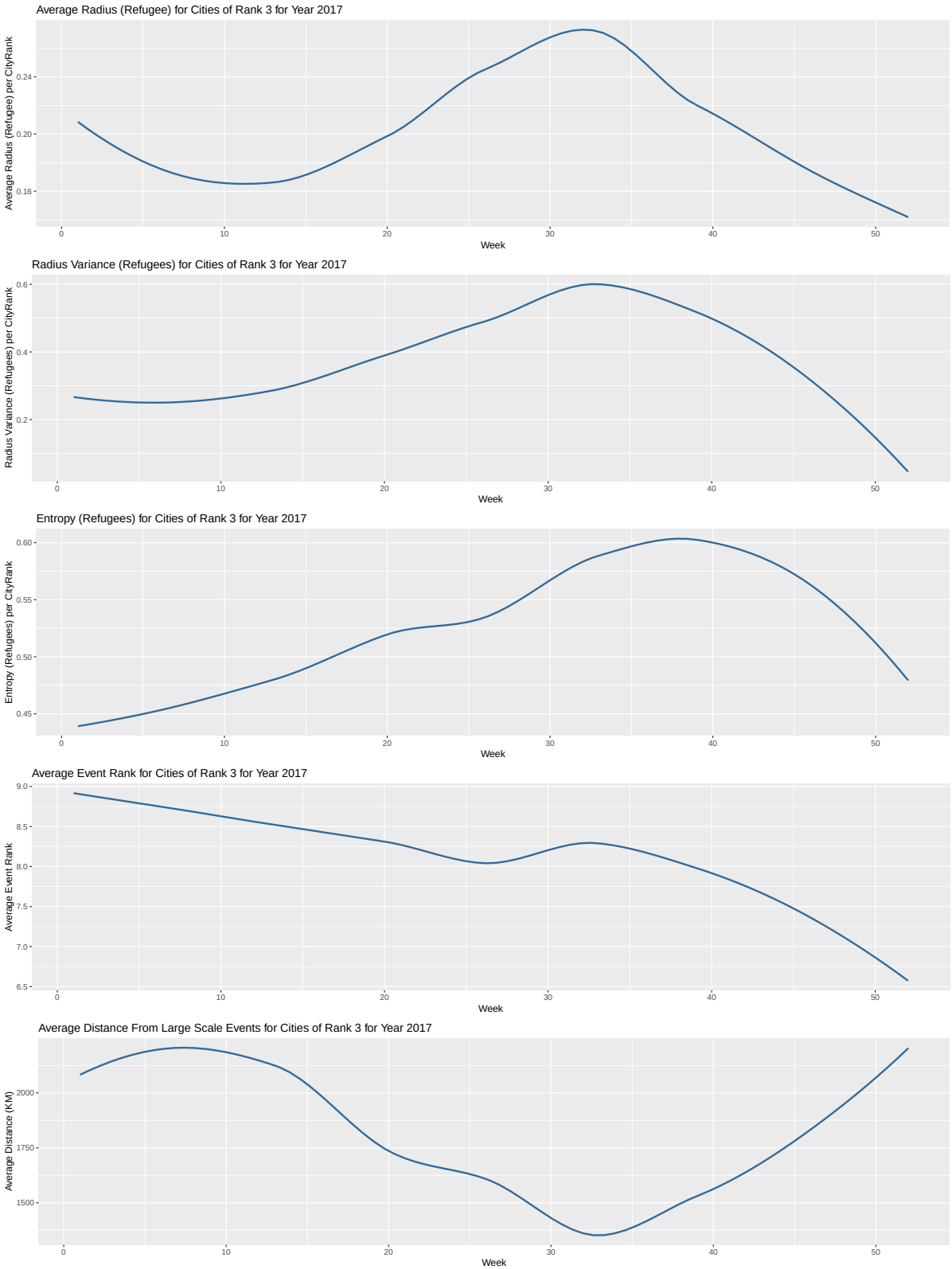


Fig. 15. Weekly Mobility Measures for Refugee callers – Cities with SEDI rank 3

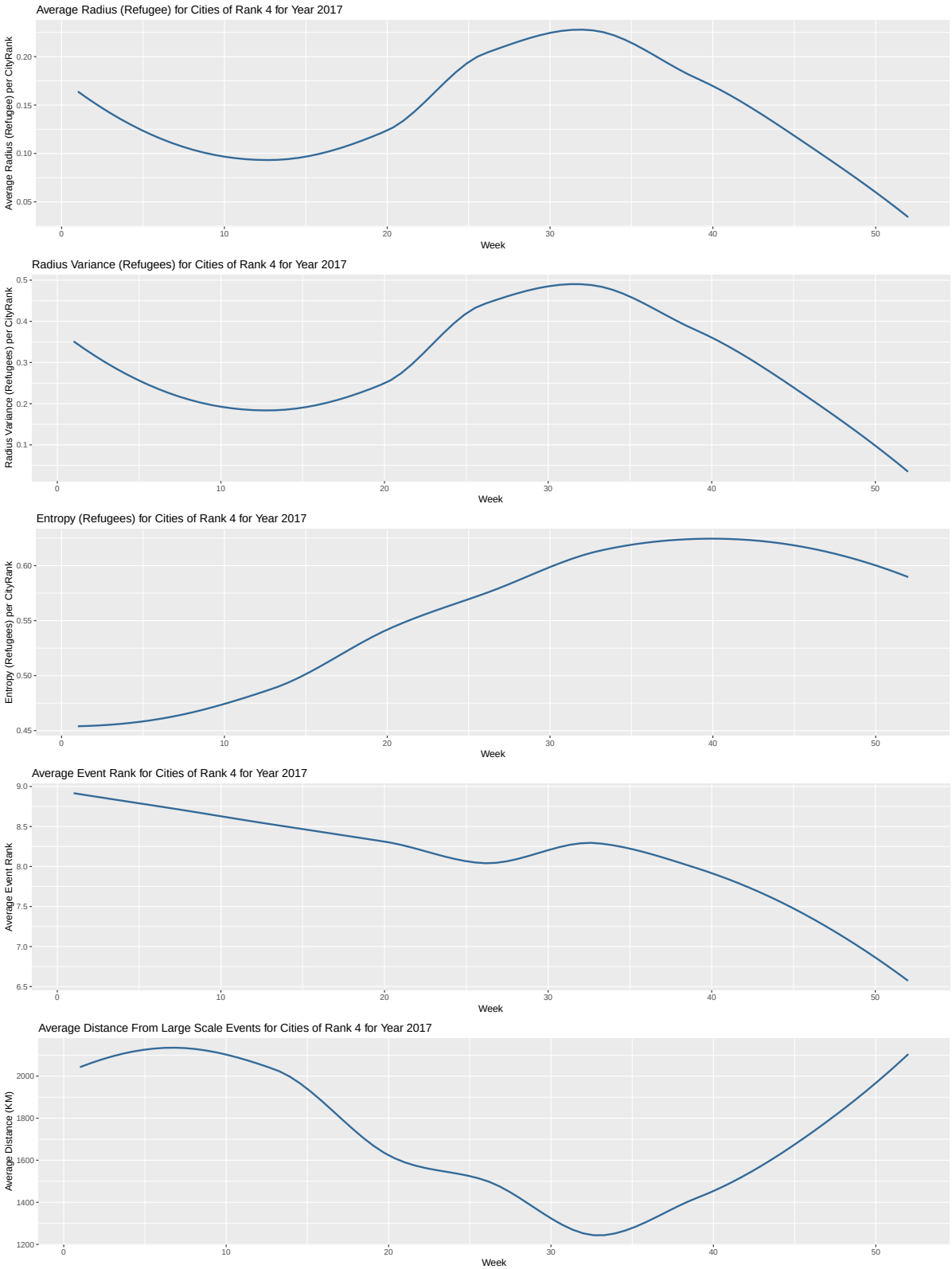


Fig. 16. Weekly Mobility Measures for Refugee callers – Cities with SEDI rank 4

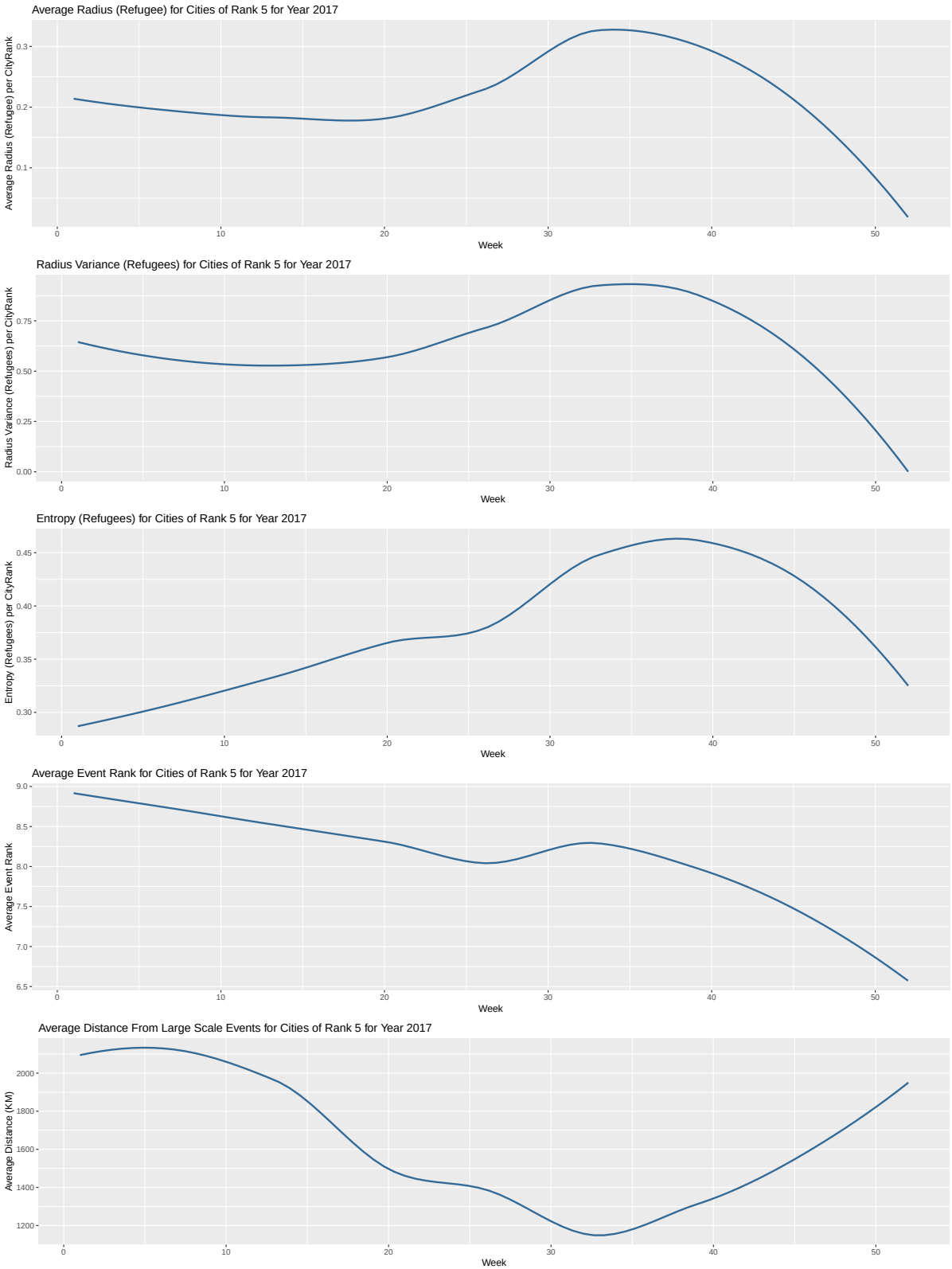


Fig. 17. Weekly Mobility Measures for Citizen callers – Cities with SEDI rank 5

References

1. Andrews, R., Geva, S.: Rules and local function networks. Rules and Networks, R. Andrews, R. & J. Diederich (eds), Queensland University of Technology, Neurocomputing Research Centre.
2. Andrews, R., Geva, S.: Rule extraction from local cluster neural nets. Neurocomputing **47**, 1–4 (2002).
3. Fast effective rule induction (Ripper). Proceedings of 12th International Conference of Machine Learning, Lake Tahoe, California (1995).
4. Fouad, F. M. , Sparrow, A., Tarakji, A., Alameddine, M., El-Jardali, F., Coutts, A. P., Arnaout, N. E., Karroum, L. B.,Jawad, M., Roborgh, S., Abbara, A, Alhalabi, F., AlMasri, I., and Jabbour, S.: Health workers and the weaponisation of health care in Syria: a preliminary inquiry for The LancetAmerican University of Beirut Commission on Syria. The Lancet, **390** 10111, 2516-2526 (2017).
5. González, M. C., Hidalgo, C. A., Barabási, A.-L.: Understanding individual human mobility patterns. Nature **453**(7196), 779-782 (2008).
6. Guha-Sapir, G.: Patterns of civilian and child deaths due to war-related violence in Syria: a comparative analysis from the Violation Documentation Center dataset, 201116. The Lancet Global Health **6**(1) (2018). [Online]. Available: `https://doi.org/10.1016/S2214-109X(17)30469-2`
7. Mowafi, H., Leaning, J.: Documenting deaths in the Syrian war. The Lancet Global Health **6**(1) (2018) [Online]. Available at `https://doi.org/10.1016/S2214-109X(17)30457-6`
8. Özaslan, M., Dincer, D., Özgür, H.: Regional Disparities and Territorial Indicators in Turkey: Socio-Economic Development Index (SEDI). Proceedings of the 46th Congress of the European Regional Science Association (ERSA) (2006).
9. Pappalardo, L., Rinzivillo, S., Qu, Z., Pedreschi, D., Giannotti, F.: Understanding the patterns of car travel. EPJ Special Topics **215**(1), 61–73 (2013).
10. Pappalardo, L., Simini, F., Rinzivillo, S., Pedreschi, D., Giannotti, F., Barabási, A.-L.: Returners and explorers dichotomy in human mobility. Nature Communications **6**(8166) (2015).
11. Song, C., Qu, Z., Blumm, N., Barabási, A.-L.: Limits of predictability in human mobility. Science **327**(5968), 1018–1021 (2010).
12. Rulequest Research: Data Mining Tools See5 and C5.0 (1997). Available at `http://www.rulequest.com/see5-info.html`

An Overview of Group Behavior on Turkey

Humberto T. M-Neto[1], Jussara M. Almeida[2], Artur Ziviani[3], Jaqueline Oliveira[1], Douglas C. Teixeira[2], and Haron C. Fantecele[3]

[1] PUC Minas - Pontifical Catholic University of Minas Gerais
humberto@pucminas.br, jaqueline.oliveira@sga.pucminas.br
[2] UFMG - Federal University of Minas Gerais
jussara@dcc.ufmg.br, douglasdocouto@gmail.com
[3] LNCC - National Laboratory for Scientific Computing
ziviani@lncc.br, haroncf@hotmail.com

Abstract. Comprehending human mobility and behavior can assist governments and companies to provide better services to the population. Although human mobility reveals a high degree of predictability, some external factors and non-routine events can create unusual patterns difficult to predict. After the Syria war started, numerous citizens started looking for better life quality and security in other countries. Comprehending behavior of this group can improve safety, social integration and other aspects of life for them. In this research, we analyze data from the D4R Challenge from Turkey and more detailed analyses from Istanbul City. Our findings identify key regions of activity of refugees in Turkey, and we verify the refugees and non-refugees have similar signature workload imposed on the network. We identify bursts of activity by refugees on the dataset, most specifically in Batman region. In the city of Istanbul, we find that Refugees, as well as non-refugees, generally live close to their corresponding group, and refugees probably work near there homes. This report shows preliminary results on refugees behavior, showing the potential of working with the provided data, together with improving life quality for refugees.

Keywords: Human behavior · Syria Refugees · Human Mobility · Mobility Analysis · Social Integration

1 Introduction

Human behavior patterns uncover numerous facets of the dynamics of a city, a region, or a whole country. The analyses of these patterns have great potential to comprehend and improve diverse sectors such as transportation, land use and health [22, 33]. Gathering data with information of human behavior can be a challenging task, but with the advent of smart devices and information communication technologies (ICT), our personal devices are not limited to store only contacts and pictures, it started to generate data which could be used to understand a single user behavior or large groups of citizens, providing an opportunity to create smart services and improve life quality of the population, as well the infrastructure of a city [3, 31, 25].

A broadly studied topic on human behavior is mobility. Comprehending human movement is essential to understand the structure of a city and its dynamics, contributing to an improved city plan [3], improvement of transportation services and action taking on emergencies [21]. Moreover, human mobility can be used to predict the spread of dynamics, contributing for a better distribution of health resources (e.g. vaccines) [29, 6].

Although human mobility presents high degree of predictability [26, 18], diverse events of distinct nature (e.g. migration, natural disasters, sports events) can significantly alter the pattern of human mobility and the use of city services and mobile services [9, 19]. These diverse events occurs occasionally, and even though some of them are planned in advance (e.g. sport events and concerts), it is difficult to predict the impact on services such as transportation and network infrastructure during these events [30]. Other events last for longer periods (e.g. hurricanes, migration) and analyzing the temporal and spatial aspect of these diverse events is a challenging task.

A non-routine global event started in 2011-12 after the beginning of the Syrian Civil War, where an increasingly number of civilians sought refuge in neighboring countries. This mobility of civilians between countries requires a special attention because essential many of the refugees after leaving conflict zones and moving to other countries face problems of integration, discrimination and unemployment. Comprehending these patterns of migration is essential by the governments to provide essential services such as health, education, safety and social integration. Turkey had received the largest refugee population in the world, mainly because it borders Syria. Turkey already received over 3 million refugees [4].

Analysis of movement and behavior of refugees within the country could assist governments and companies provide better services to them. Aiming to provide ways to analyze the behavior of the population, Türk Telekom in partnership with the Turkish Academic Research and Council (TUBITAK) and Bogăziçi University launched the Data for Refugees Challenge (D4R), providing a database to the scientific community based on anonymised Call Detail Records (CDR) of phone calls and SMS messages of Türk Telekom [24].

Models of mobility are usually built using CDR data, due to its easier acquisition compared to other types of data (i.e. available by carries on partnerships) and generally covers information of a large number of users [1]. The CDR is a record that documents details of phone calls (e.g. antenna source, duration of call, destination antenna) or other phone services such as Short Message Service (SMS). These records are stored by the telco companies, and can be shared to the scientific community based agreements and partnerships. These records are anonymised to preserve users' identity. On Türk Telekom dataset, due to the sensible data of refugees, just a sample of the data was provided.

The general goal of the D4R challenge is to contribute to the welfare of refugee population, gaining insight on key issues (e.g. security, health, social integration) based on the mobile traffic generated during 2017 by this group.

[4] https://data2.unhcr.org/en/situations/syria/location/113

These insights can help governments and international bodies discover vulnerabilities, and propose new services and solutions for refugees in Turkey and in other countries.

CDRs present a challenge related to the representativity of the truth behavior of users, that is, the record is stored only when a service (i.e. phone call, SMS) is made or received, this way the temporal and spatial information as well as the frequency of visits of a user on a given cell can differ from the reality [23, 32]. On Türk Telekom dataset this challenge is even bigger once the sample is selected randomly on different days for the entire Turkey. Nevertheless, CDRs are widely used and proved to be a good way to study human mobility patterns [28, 10, 17, 4].

We focus on the main challenge of social integration of refugees into local communities. This way, we compare the mobile signature of use of the mobile network by refugees and non-refugees to explore regions of the city used by these two groups, and how similar and diverse they are. We then split the time-scale of our signatures to verify use of regions by locals (e.g. residence, work, leisure). Finally, we perform mobility analyses on a national scale to search for moments of burst of migration, or activity on the mobile infrastructure.

Our findings identify key regions of activity of refugees in Turkey, and we verify the refugees and non-refugees have similar signature workload imposed on the network. In the city of Istanbul, we find that Refugees, as well as non-refugees, generally live close to their corresponding group and refugees probably work near there homes. The refugees and non-refugees have the same signature workload imposed on the network. We identify bursts of activity by refugees on the dataset, most specifically in Batman region, but with our investigation, we could not identify the cause of the peaks at the base, we believe that a more profound investigation can clarify the facts. The results presented are preliminary, and it is essential to emphasize the importance of more in-depth analyzes at all bases since we only analyze base 1. It is important to correlate D4R data with other databases like socioeconomic, socio-demographic and global-news.

This work is organized as follows. Section 2 present some related work that motivates the development of our approach. Section 3 present an overview of the D4R Dataset. Section 4 explains the methodology used to conduct the analyses. Section 5 we present the results obtained so far. The conclusions and future work are presented in Section 6.

2 Related Work

Several works on the literature demonstrate the potentials of comprehending human behavior and mobility and its capabilities to enhance personal and collective aspects (e.g. health, public transportation, social interaction, security) [8, 16, 20].

Although human mobility patterns present high degree of predictability [13], diverse events alters these patterns dramatically impacting the use of services such as SMS, calls and data [15]. In other work, [19], we analyze how user be-

havior patterns during large-scale events (e.g. New Year's Eve, musical concerts) present a high correlation to the mobility of the user and the event. Users attending a specific type of event presented the same behavior on future events of the same type. Not only mobility patterns alters during non-routine events, but as well as service usage as demonstrated on Bagrow [2], where the activity of a user during the Boston Marathon Bombing incident changed in comparison of other usual moments.

Human behavior is intrinsically related to the structure of the region, as well as the usage of the mobile network infrastructure. The characterization of the signature produced by the use of the mobile network infrastructure reveals information of the use of a specific region of the city according to the day, and the time of the day. The work of Furno et. al [11] unveiled the strong relation between the mobile traffic activity and the urban fabric consisting of the use of the region. The results could be applied to automated land use detection, and network management. The notion of mobile traffic signature was introduces on Girardin et. al [12], condensing representation of the typical mobile demand dynamics in a given geographical region. The same idea was then applied on other works such as Calabrese et. al [7] which analyze the traffic signature in the city of Morristown identifying differences between the demand in downtown and high school areas, and Becker et. al [5] which analyze the entire conurbation of Boston.

Distinct regions of the city plays different roles on the city services according to its use. Industrial and commercial areas present high mobility and mobile traffic during working-hours, while residential areas present higher activity during late-hours of the day [33]. Comprehending each region can assist governments on a better distribution of resources and public actions, such as placement of public transportation, improving life quality of citizens .The work of Toole et. al[27] present a technique for automatic identification and classification of land use from data obtained by cell phone. Their approach computes the aggregated calling patterns of the network infrastructure, and find cluster comparing the activity between each antenna. This way, the authors find optimum clusters distribution to automatically identify how citizens use different geographic regions within a city.

3 Dataset

Türk Telekom made three datasets available to the D4R participants [24]. The datasets include one year of mobile CDR data, collected between January, 2017 and December, 2017. The main difference of these datasets to ordinary CDRs datasets, it is that D4R contains a flag which indicates if a record of the data belongs to a refugee or not. This flag on the data is given to the Türk Telekom subscriber based on one of the following criteria:

- Subscriber have an ID number of refugee or foreigner in Turkey;
- Subscriber registered with Syrian passport;

– Subscriber use special tariffs reserved for refugees.

Different vulnerable groups (e.g. migrants, asylum seekers, temporary protected foreign) receive an specific ID number when registering in Turkey. None of the records of the data are guaranteed to belongs to a refugee, however, the flags indicates a high probability of belonging. We based our analyses on this flag to separate the analyses conducted on refugees and non-refugee mobile traffic, but we consider the bias in the data related to the flag.

	Refugees	Non-Refugees	Total
Customers	184,949 (18.6%)	807,508 (81.4%)	992,457 (100%)
Subscribers	231,142 (19%)	980,697 (81%)	1,211,839 (100%)

Table 1. Number of Customers and Subscribers on D4R Dataset.

Table 1 present an overview of the number of subscribers and customers on the dataset. Some of the customers had multiple phone lines, this way, each line correspond to a single subscription. As mentioned before, the refugee data contain some noise due to the bias of the data. The dataset contains information of approximately 6.18% of the total subscribers of Türk Telekom, based on the 2017 quarterly report[5].

Three datasets are provided by D4R Challenge. Dataset 1 contains the total number and duration of calls and SMS exchange per cell tower. This dataset contains no information which can be used to identify users. This dataset enables analysis of activity at different areas of the city. We focus our analyses on this dataset since we aim to discover refugees interaction on different regions of the city at different times. The data is structured as follows:

– **Timestamp** : Day/Hour formatted as YYYY-MM-DD HH
– **Outgoing Site ID** : ID of site the call originated from
– **Incoming Site ID** : ID of site receiving the call
– **Total Number of Calls**: Total number of calls between the given antennas during the given timestamp
– **Number of Calls Originated from Refugees**: The number of calls originated from refugees.
– **Total Call Duration**: Total duration of all calls between the given antennas during the given timestamp
– **Total Call Duration Originated from Refugees**: Total duration of calls between these two antennas during the given timestamp originated from refugee IDs.

For the SMS dataset, besides *Timestamp*, *Outgoing Site ID* and *Incoming Site ID*, the dataset contains the following information:

[5] http://www.ttyatirimciiliskileri.com.tr/RaporlarEN/2017/Press%20Release%20Q4'17-%20Final.pdf

- **Number of SMS**: Total number of SMS messages between the given two antennas at the given timestamp
- **Number of SMS Originated from Refugees**: Number of SMS originated from numbers with refugee IDs.

Dataset 2 contains data related to a group of randomly chosen active users. Each group of users is observer for a period of 2 weeks, and then a fresh sample of active users is drawn at random, to protect privacy. Each user is identified by a tag (i.e. refugee, non-refugee, unknown). In addition, the call type of the activity is recorded, that is, if the user is receiving or placing the call. Finally each data is rounded to minutes. If one party uses a different operator, the antenna information is set as missing.

On Dataset 3, trajectories of 50,000 randomly selected refugees and 50,000 randomly selected non-refugees are provided for the entire observation period, but with very coarsely spatial resolution, that is, replacing antenna identifiers with broader area identifies. All personal information is excluded to protect user identity, however, the refugee flag is indicated to each user. Dataset 2 and 3 are not considered for the scope of this work, but we intend to conduct further analyses on these datasets aiming to understand mobility behaviors considering social indicators such as demographics, financial indicators and public transportation placement, analyzing how closely these data are linked, and how changing one data could impact the lives of refugees.

In addition to the three Datasets, a dataset containing cell tower information was provided containing the following information:

- **Latitude**: DMS latitude of the base tower.
- **Longitude** DMS longitude of the base tower.
- **City**: The registered city of the base tower.
- **Population Type** : An unofficial note about the population type around the base tower used by Türk Telekom. It takes values such as Rural, Sub-urban, Industrial, Seasonal areas, Dense Urban, etc.

4 Methodology

Consider the Dataset 1 as D, describing the communication activity of subscribers during a set of days $\mathbf{d} = \{d\}$. For each day $d \in \mathbf{d}$, the mobile demand is stored as the aggregate of the traffic generated within a specific time slot $\mathbf{t} = \{t\}$. The traffic generated at specific day $\mathbf{d}$ and time slot $\mathbf{t}$ is associated to the respective antenna $\mathbf{a} = \{a\}$. This way, $D = v_a(d, t)$ where every element $v_a(d, t)$ describes the total mobile communication activity within each antenna a at time slot t of day d.

4.1 Behavior Analyses

To construct a representative set of mobile traffic signatures to analyze behavior and verify usage of similar regions, we process the dataset through five phases explained below:

Data Cleaning

This step consist in identifying missing values for the dataset. Once the dataset consist of registers of mobile activity on a site (i.e. call or sms sent or received), if no activity is placed, then no record in inserted on the dataset. In order to compare traffic signature between sites, the time series must have the same length. If no activity was found on a given $v_a(d,t)$, we set the value to 0 for refugee and non-refugee activity.

The dataset contains information of total number of calls and SMS, and if the information was originated from refugee. This way, we inferred the aggregated activity of non-refugee ($NREF$) users as the difference between the total of the site a and the activity of refugee (REF), that is:

$$v_a(d,t)_{\text{NREF}} = v_a(d,t)_{\text{TOTAL}} - v_a(d,t)_{\text{REF}}$$

As mentioned previously, the bias on the refugee tag must be taken into account, so the number of non-refugee could vary a bit from the reality.

Finally, we converted all coordinates of the DMS system (Degree, Minutes and Seconds) to Decimal Degrees to facilitate some analyses and the application of tools.

Typical Week Signature

The work of Grauwin et al. [14] propose a Typical Week Signature (TWS). The signature metric adds up voice and text volumes. The signature aggregate the signature considering the day of the week, from Monday to Sunday, i.e., $\delta = \{MON, TUE, WED, THU, FRI, SAT, SUN\}$. Let us denote as $d^\delta \subset d$ the set of days in the dataset D which correspond to the day of the week δ. For instance, d^{SUN} groups all Sundays in the dataset. This way, the generic signature for each site is given as follows:

$$s_a(\delta,t) = \mu(\{v_a(d,t) \mid d \subset \mathbf{d}^\delta\}), \forall a \subset \mathbf{a}$$

The signature is given for each time slots t during day δ. The μ symbol represent the mean of set within parenthesis. This method implies high compression on the data resulting on 7x24 (days x hours) values for each antenna. At this point we have the signature traffic for each of the sites on our dataset and for the two groups of citizens tagged on the dataset, that is, $s_a(\delta,t)_{\text{REF}}$ and $s_a(\delta,t)_{\text{NREF}}$.

Creating Grid

We intend to verify the mobility aspects and land use of Turkey cities considering refugees and non-refugees behavior. Analyzing each site separately could enhance deeper knowledge of behavior patterns at a small scale, however as we aim to analyze the whole city behavior and identify regions and its use, we proposed a method of separating the whole city into unit cells of a grid. For this, we selected a square which comprised the whole desired region, and created unit

cells of approximately 1km^2 area adding 1 to the third significant digit of the latitude and longitude (e.g. $\text{Cell}_{1\ \text{Start}}$ = (lat, lon), $\text{Cell}_{1\ \text{End}}$ = (lat + 0,001, lon + 0,001)). We chose this approach to simplify calculation of the grid cells.

Group Signature by Cell Unit

After creating the grid for a given region, we mapped antennas to unit cells based on its location. We group the traffic signatures of same unit cells to find the mean activity of the region. Let us denote U the set of unit cells of a given city of a selected city or region, and u_a the set of antennas a located at the unit cell u. The signature traffic of the unit cell (i.e. $s_u(\delta, t)$) is given as follows:

$$s_u(\delta, t) = \mu(\{s_a(\delta, t)\}), \forall a \subset u_a$$

After this step, the created grid cell contains distinct traffic signature for each unit cell where occurred an activity of refugee or non-refugee.

Cluster Cell Units After summarizing the mobile traffic activity in each unit area into meaningful profile, we group similar unit areas signatures into a limited set of classes based on the time series distance calculation. Clustering similar unit areas signatures indicate a similar use for a given set of regions. This step can consider the total time series as input, or specific time intervals of the time series such as working-hours or night hours.

4.2 Spatial Mobility Analyses

As seen earlier, diverse events alters mobility patterns of groups and individuals. We perform initial investigation on the mobility of refugees on the dataset. Our goal is to verify the occurrence of burst of refugee movement and the relation to worldwide news, and location with high concentration of refugees.

In order to perform the mobility analyses, after the *Data Cleaning* step detailed previously, we perform the following analyses on the data:

Monthly Traffic Volume

Opposite to the behavior analyses where we intend to find small regions and areas within a city with similar behavior, on the mobility analyses we aim to find refugee movement patterns on a national scale. This way, we apply a higher level of compression compared to the previous analyses, where we aggregate information of a whole month of a specific city, using the city data of the antennas dataset as parameter.

As we intend to verify volume and not behavior, the refugee mobile traffic volume of a city is given by the sum of all mobile traffic generated data during the period. Consider C the set of cities, M the set of months running from

January to December of 2017 and c_{a} the set of antennas within the city c. The traffic volume (i.e. $v_{(\mathrm{c,m})}$) of a city on a given month is defined as:

$$v_{(\mathrm{c,m})} = \frac{\sum(Calls, SMS)_{\mathrm{a}}}{len(c_a)len(m)}, \forall a \subset c_a$$

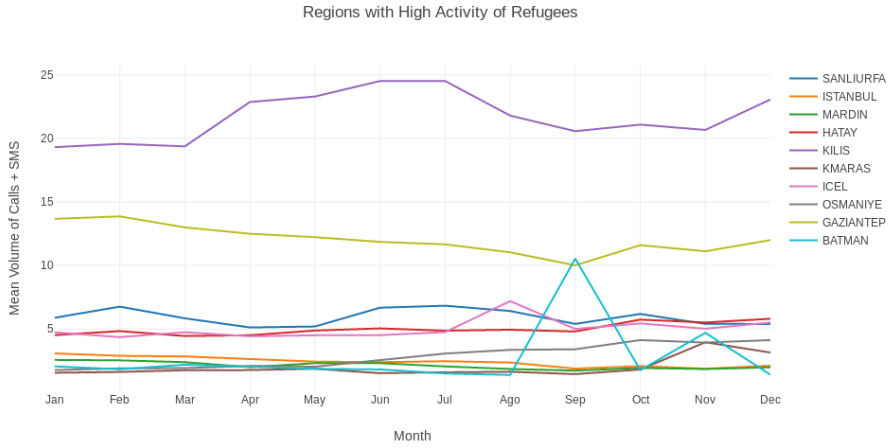
where $len(c_a)$ is the number of antennas at the city c, and $len(m)$ the number of days at the month m. We normalize the monthly data considering the total activity performed by refugees for the whole country on the month.

5 Results

In this section we present initial results obtained after applying the methodology described previously. Further investigation would be conducted on the data to gain more insights about mobility and social behavior of refugees aiming to integrate this group into the society.

Prior to analyzing refugee behavior, we performed the mobility analyses to verify the volume of traffic generated at a national scale, and then conduct behavior analyses on cities which presented higher level of activity.

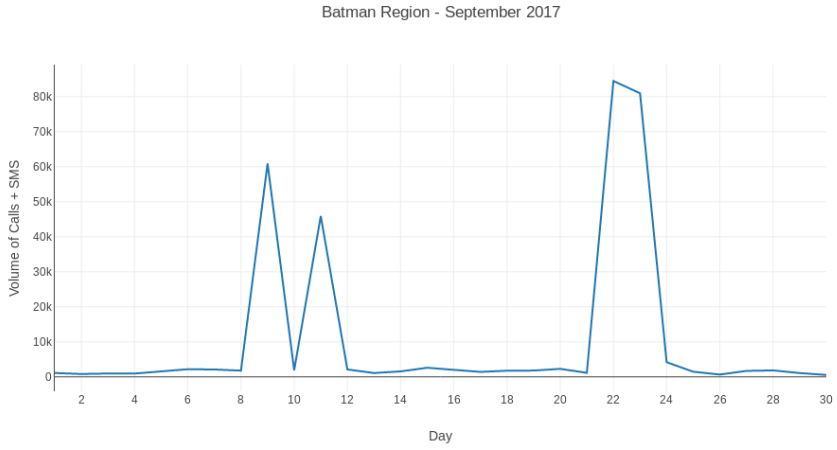
Fig. 1. Regions with Higher activity of Refugees.



This way, we applied the spatial mobility methodology proposed, and obtained the volume of activity for all regions at the dataset, taking as input the region division provided by Türk Telekom. We filtered the 10 regions which presented higher activity during the year to verify possible bursts or anomalies on the data. Figure 1 presents the overview volume of calls aggregated by day.

The regions that presented higher activity of refugees are located next to the Syria border as expected, where the concentration of Refugees is higher compared to other regions. The activity on this regions remain similar along the year, except by Batman region, which presented a peak during September and a smaller peak of activity at November. Verifying the dynamics of the volume during September for Batman region (see Figure 2), we clearly see the distinct set of days where the increase on the activity occurred, i.e., 9, 11, 22 and 23.

Fig. 2. SMS and Calls of Refugees for the Region of Batman on September, 2017.

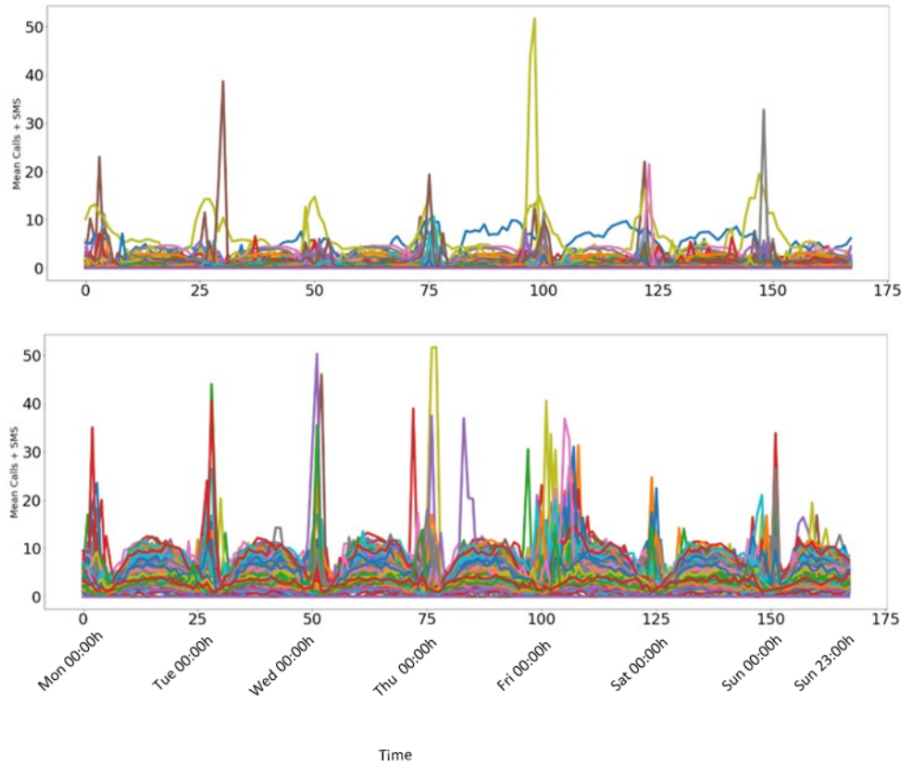


Although the current dataset present some bias considering refugee activity, we expect this peak to be related to some event that occurred in the region, or increase in the number of users marked as refugees collected in this specific period of the dataset. Further analyses will be conducted to understand the pattern shown, and verify possibles motifs of the disturbance.

The behavior analyses consider behavior in different regions of the city. Although regions such as Batman and Kilis presented high presence of Refugees, we focused our analyses on Istanbul for a reason the reason which is the most populated city in Turkey, with a broad territorial area, and with a great population tagged as the refugee. Considering smaller cities such as Batman and Kilis could difficult our region clustering and spatial analyses.

We applied the proposed behavior methodology to verify the signature workload imposed on the network by refugees and non-refugees on the dataset. Comparing all time series generated for Istanbul highlights a same pattern of usage between the two groups, i.e., refugees and non-refugees (see Figure 3). The time series are ordered by aggregate hour, starting from Monday to Sunday.

Fig. 3. Time Series by Grid generated for Istanbul city. Top: Refugees Time Series - Bottom: Non-Refugees Time Series.

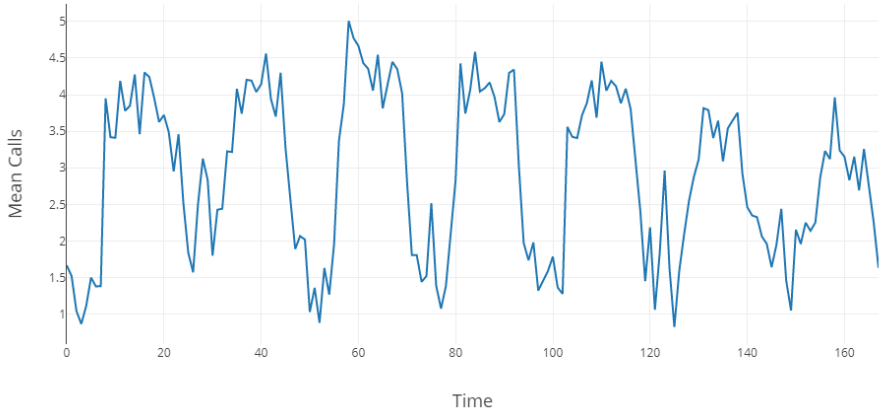


As non-refugees contains more data on the dataset and represent a major part of Istanbul's population, the volume generate by this groups is larger, consequently the workload imposed on the antennas tend to be higher. Despite this, we see that some grids of Istanbul present mean traffic values as high as non-refugees, e.g. the yellow line near time 100 (around Thursday late night hours and Friday early hours).

It is important to emphasize that this aggregation considers the whole dataset ranging from January 2017 and December 2017. The high level of compression generates noise on the data as seen on Figure 4, however, the general shape of the time series is maintained, which we clearly see a pattern of use during the weekdays.

In order to verify regions which present similar patterns, we clustered the time series by hour using hierarchical cluster technique and euclidean distance as metric. Figure 5 shows the result of the clustering.

Fig. 4. Refugee Traffic Volume for a single grid cell of Istanbul.



The dataset presented some bias related to the users selection, that is, users are selected randomly, varying from time to time, this was expected, as reported in the D4R report. When we aggregate all the information, a specific behavior of a user is merged with behaviors of other users, and the clustering gets somehow messy. On refugee data, three single cluster were found: one on the far right region of the city, and three others at far left of the city. A bigger cluster was created at the center of the city. Using the dataset containing fine grained information of users could enhance the clustering analyses.

We split the time series into period of time and aggregated the information according to the mean traffic generated. We set the time intervals as [0-6] hours, [7-12] hours, [13-18] hours and [19-23] hours, respectively as early morning, morning, work hours and night hours. The mean activity of each region is presented on the legend.

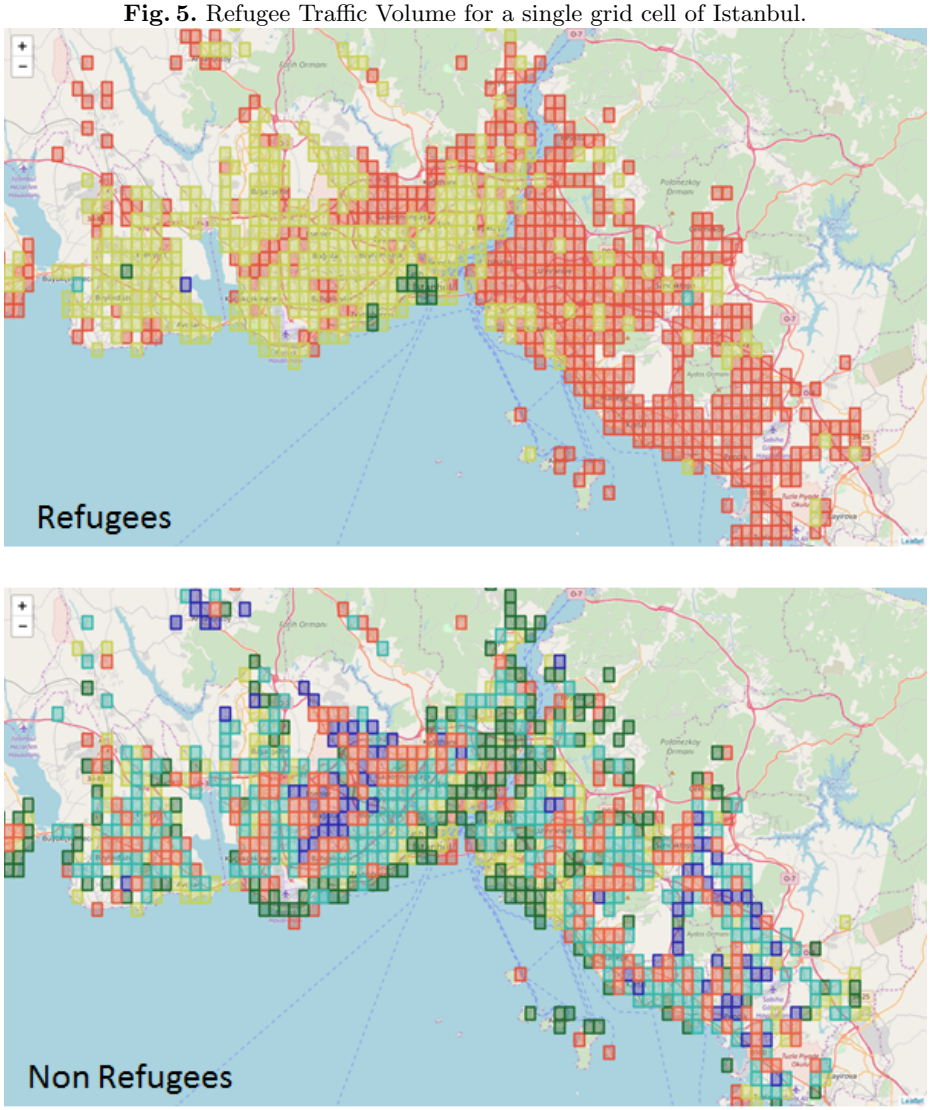
We normalized the information by group, once the non-refugee activity is higher than the refugee. The maps (see Figure 6) shows the increasing activity during the day, mainly on the working hours interval (i.e. 13-18 hours). Comparing closely, we observe that refugee activity are higher on areas different than non-refugees. We believe that these groups are more likely to live together and work on near areas, and integrating these two groups is a challenge for governments and institutions.

6 Discussion

This work presented some overview of the dataset challenge focused on social integration of Refugees and Non-Refugees.

The first analyses tried to identify key regions of activity of refugees and non-refugees, however, bias on the data and reduced information of users avoiding identification created a challenge on inferring the use of specific regions.

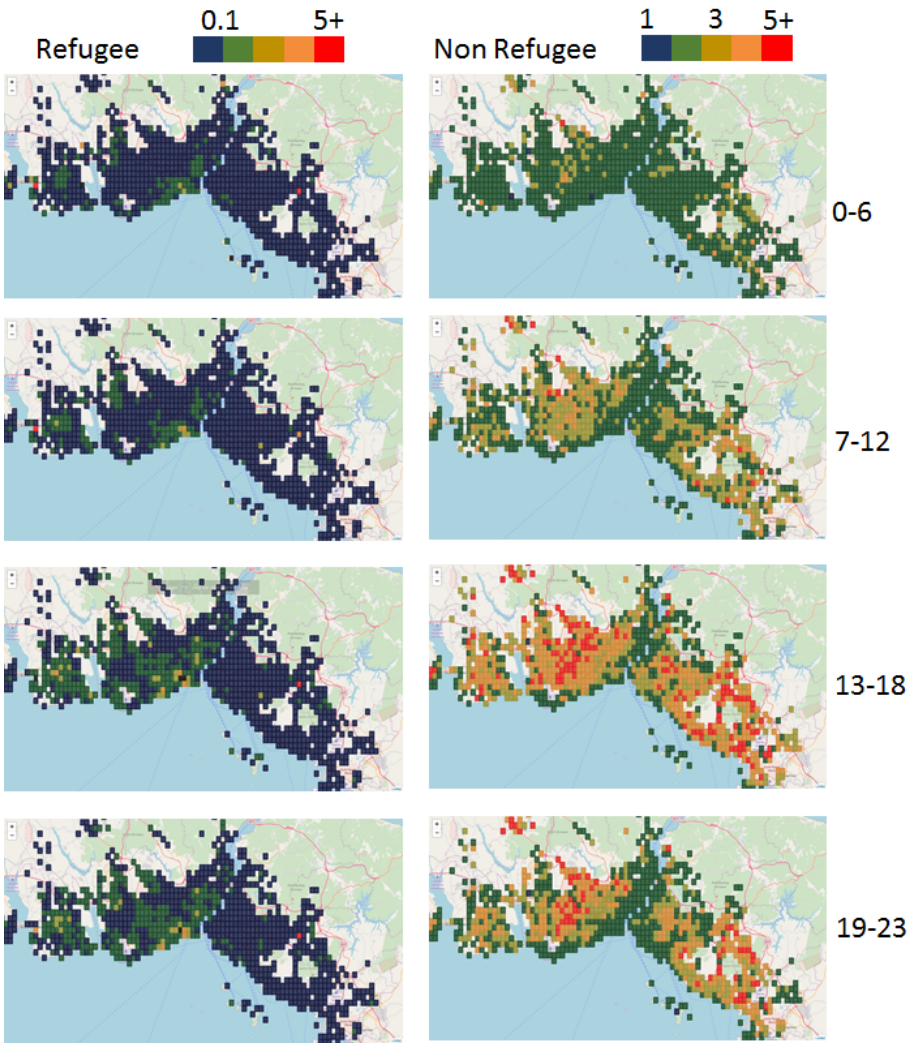
Fig. 5. Refugee Traffic Volume for a single grid cell of Istanbul.



The proposed methodology helped identify key aspects that will be analyzed further:

– Refugees and Non-Refugees usually lives close to the corresponding group. Inserting sociodemographic and socioeconomic data to this information can reveal insights about regions used by refugees.

Fig. 6. Grid Activity by Specific Intervals



– Bursts of activity by refugees were identified on the dataset. Merging this information with global news and outside factors could reveal the reason of those bursts.

– The methodology identified regions of distinct use at different time, however, when clustering the whole dataset the clusters got messy. Alternating the number of clusters and different ways to create traffic signature will be conducted in order to get more confident information.

– Finally, we intend to aggregate social media news, which is already being collected together with other datasets to reveal new insights about social interaction between refugees and non-refugees.

References

1. Asgari, F., Gauthier, V., Becker, M.: A survey on human mobility and its applications. arXiv preprint arXiv:1307.0814 (2013)
2. Bagrow, J.P.: Information spreading during emergencies and anomalous events. In: Complex Spreading Phenomena in Social Systems, pp. 269–286. Springer (2018)
3. Batty, M., Axhausen, K.W., Giannotti, F., Pozdnoukhov, A., Bazzani, A., Wachowicz, M., Ouzounis, G., Portugali, Y.: Smart cities of the future. The European Physical Journal Special Topics **214**(1), 481–518 (2012)
4. Becker, R., Cáceres, R., Hanson, K., Isaacman, S., Loh, J.M., Martonosi, M., Rowland, J., Urbanek, S., Varshavsky, A., Volinsky, C.: Human mobility characterization from cellular network data. Communications of the ACM **56**(1), 74–82 (2013)
5. Becker, R.A., Caceres, R., Hanson, K., Loh, J.M., Urbanek, S., Varshavsky, A., Volinsky, C.: A tale of one city: Using cellular network data for urban planning. IEEE Pervasive Computing **10**(4), 18–26 (2011)
6. Brockmann, D., David, V., Gallardo, A.M.: Human mobility and spatial disease dynamics. Reviews of nonlinear dynamics and complexity **2**, 1–24 (2009)
7. Calabrese, F., Reades, J., Ratti, C.: Eigenplaces: segmenting space through digital signatures. IEEE Pervasive Computing **9**(1), 78–84 (2010)
8. Caminha, C., Furtado, V., Pequeno, T.H., Ponte, C., Melo, H.P., Oliveira, E.A., Andrade Jr, J.S.: Human mobility in large cities as a proxy for crime. PloS one **12**(2), e0171609 (2017)
9. Candia, J., González, M.C., Wang, P., Schoenharl, T., Madey, G., Barabási, A.L.: Uncovering individual and collective human dynamics from mobile phone records. Journal of physics A: mathematical and theoretical **41**(22), 224015 (2008)
10. Dong, Y., Pinelli, F., Gkoufas, Y., Nabi, Z., Calabrese, F., Chawla, N.V.: Inferring unusual crowd events from mobile phone call detail records. In: Joint European conference on machine learning and knowledge discovery in databases. pp. 474–492. Springer (2015)
11. Furno, A., Fiore, M., Stanica, R., Ziemlicki, C., Smoreda, Z.: A tale of ten cities: Characterizing signatures of mobile traffic in urban areas. IEEE Transactions on Mobile Computing **16**(10), 2682–2696 (2017)
12. Girardin, F., Calabrese, F., Fiore, F.D., Ratti, C., Blat, J.: Digital footprinting: Uncovering tourists with user-generated content. Institute of Electrical and Electronics Engineers (2008)
13. Gonzalez, M.C., Hidalgo, C.A., Barabasi, A.L.: Understanding individual human mobility patterns. nature **453**(7196), 779 (2008)
14. Grauwin, S., Sobolevsky, S., Moritz, S., Gódor, I., Ratti, C.: Towards a comparative science of cities: Using mobile traffic records in new york, london, and hong kong. In: Computational approaches for urban environments, pp. 363–387. Springer (2015)
15. Imran, M., Castillo, C., Diaz, F., Vieweg, S.: Processing social media messages in mass emergency: A survey. ACM Computing Surveys (CSUR) **47**(4), 67 (2015)
16. Isaacman, S., Becker, R., Cáceres, R., Martonosi, M., Rowland, J., Varshavsky, A., Willinger, W.: Human mobility modeling at metropolitan scales. In: Proceedings of the 10th international conference on Mobile systems, applications, and services. pp. 239–252. Acm (2012)

17. Jiang, S., Ferreira, J., González, M.C.: Activity-based human mobility patterns inferred from mobile phone data: A case study of singapore. IEEE Transactions on Big Data **3**(2), 208–219 (2017)
18. Liang, X., Zhao, J., Dong, L., Xu, K.: Unraveling the origin of exponential law in intra-urban human mobility. Scientific reports **3**, 2983 (2013)
19. Marques-Neto, H.T., Xavier, F.H., Xavier, W.Z., Malab, C.H.S., Ziviani, A., Silveira, L.M., Almeida, J.M.: Understanding human mobility and workload dynamics due to different large-scale events using mobile phone data. Journal of Network and Systems Management pp. 1–22 (2018)
20. Mitchell, T.M.: Mining our reality. Science **326**(5960), 1644–1645 (2009)
21. Moat, H.S., Preis, T., Olivola, C.Y., Liu, C., Chater, N.: Using big data to predict collective behavior in the real world 1. Behavioral and Brain Sciences **37**(1), 92–93 (2014)
22. Pentland, A.: Reality mining of mobile communications: Toward a new deal on data. The Global Information Technology Report 2008–2009 **1981** (2009)
23. Ranjan, G., Zang, H., Zhang, Z.L., Bolot, J.: Are call detail records biased for sampling human mobility? ACM SIGMOBILE Mobile Computing and Communications Review **16**(3), 33–44 (2012)
24. Salah, A.A., Pentland, A., Lepri, B., Letouzé, E., Vinck, P., de Montjoye, Y.A., Dong, X., Dağdelen, Ö.: Data for refugees: The d4r challenge on mobility of syrian refugees in turkey. arXiv preprint arXiv:1807.00523 (2018)
25. Sherly, J., Somasundareswari, D.: Internet of things based smart transportation systems. International Research Journal of Engineering and Technology **2**(7), 1207–1210 (2015)
26. Song, C., Qu, Z., Blumm, N., Barabási, A.L.: Limits of predictability in human mobility. Science **327**(5968), 1018–1021 (2010)
27. Toole, J.L., Ulm, M., González, M.C., Bauer, D.: Inferring land use from mobile phone activity. In: Proceedings of the ACM SIGKDD international workshop on urban computing. pp. 1–8. ACM (2012)
28. Wang, D., Pedreschi, D., Song, C., Giannotti, F., Barabasi, A.L.: Human mobility, social ties, and link prediction. In: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 1100–1108. Acm (2011)
29. Wesolowski, A., Eagle, N., Tatem, A.J., Smith, D.L., Noor, A.M., Snow, R.W., Buckee, C.O.: Quantifying the impact of human mobility on malaria. Science **338**(6104), 267–270 (2012)
30. Xavier, F.H.Z., Silveira, L.M., Almeida, J.M.d., Ziviani, A., Malab, C.H.S., Marques-Neto, H.T.: Analyzing the workload dynamics of a mobile phone network in large scale events. In: Proceedings of the first workshop on Urban networking. pp. 37–42. ACM (2012)
31. Xu, Y., González, M.C.: Collective benefits in traffic during mega events via the use of information technologies. Journal of The Royal Society Interface **14**(129), 20161041 (2017)
32. Zhao, Z., Shaw, S.L., Xu, Y., Lu, F., Chen, J., Yin, L.: Understanding the bias of call detail records in human mobility research. International Journal of Geographical Information Science **30**(9), 1738–1762 (2016)
33. Zhong, C., Arisona, S.M., Huang, X., Batty, M., Schmitt, G.: Detecting the dynamics of urban structure through spatial network analysis. International Journal of Geographical Information Science **28**(11), 2178–2199 (2014)

New Approaches to the Study of Spatial Mobility and Economic Integration of Refugees in Turkey*

Steven Reece[1], Franck Duvell[2], Carlos Vargas-Silva[2] and Zovanga Kone[2]

[1] Machine Learning Research Group, Oxford University and the Alan Turing Institute, London, UK,
reece@robots.ox.ac.uk,
WWW home page: http://www.robots.ox.ac.uk/~parg
[2] Compas, Oxford University, UK,
{franck.duvell,carlos.vargas-silva,zovanga.kone}@compas.ox.ac.uk,
WWW home page: http://www.compas.ox.ac.uk

Abstract. We utilise mobile telephone data to investigate the economic activities and related spatial mobility of Syrian refugees in Turkey. To do this we develop a model that relates mobile phone usage data to manual and non-manual job classes and use this model to determine the extent, type and spatial dispersion of the refugee labour force in Turkey. Our data analytic approach exploits multiple weak rules of mobile usage by workers in each job class and the model parameters are derived from the data using a novel extension to an unsupervised machine learning classification algorithm. We find that refugee workers are more manual than non-manual, which is opposite to Turkish workers, and we map the dispersion of manual and non-manual workers, both Turkish and refugee, within Turkey. We conclude with a synopsis of further telephone data fields and other data that may improve our analysis and also suggestions for future work.

Keywords: Bayesian machine learning, unsupervised learning, refugee dispersal, economic integration, Syrian refugees in Turkey, unemployment and social integration

1 Introduction

There is scarce data on the spatial dispersal and, in particular, on the economic integration of refugees both in Turkey and around the world. In this project we explore whether and to what extent we can utilise mobile telephone data to explore the economic activities and related spatial mobility and thus economic integration of Syrian refugees in Turkey. We began with the idea of finding out

* Steven Reece was supported through a fellowship at The Alan Turing Institute, sponsored by the Programme on Data-Centric Engineering and funded by Lloyds Register Foundation.

about employment of and trade by Syrian refugees, in particular cross-border trade between Turkey and Syria. To this end we first established some experimental collaboration between social scientists at the Centre on Migration, Policy and Society and a machine learning researcher at the Department of Engineering Science at the University of Oxford. Second, we developed a framework to express weak social science models appropriate for data analysis. Third, this led to the development of a novel machine learning algorithm that learns the social science models from data. These models can then be used to predict social behaviours. Fourth, we determined patterns in the Türk Telecom data and propose economic demographics of refugees within Turkey. And fifth, we identified weaknesses in the telecom data and propose additions to the data to help improve our analysis of refugee economics.

2 Background

In international law the term 'refugees' refers to a broad category of people who are forced to migrate under conditions defined by the UN Convention on the Status of Refugees. Turkish law, the law on Foreigners and International protection, distinguishes between refugees from Europe, conditional refugees from non-European countries who are supposed to be resettled, and subsidiary protection, also understood as temporary protection, which usually applies to Syrian displaced persons. With regards to the spatial dispersal of persons under international protection these are required by the Law on Foreigners and International protection (LFIP), article 71, to reside in designated places, so-called satellite cities [14] [3]. Persons under temporary protection, hence Syrians, if not allocated to a refugee camp are mostly registered in urban areas and, to some extent, are restricted in their movements within the country. This policy is not uncommon and also practiced, for instance, in Germany. However, it has been observed that persons in both categories do not always follow these regulations and de facto reside in places other than those where they are registered or are registered in more than one city. Statistics held by the Directorate General on Migration Management (DGMM) on the dispersal of Syrians by province [2] might thus be distorted by irregular behaviour of refugees. For instance, the DGMM records 563,963 Syrian refugees in the province of Istanbul whereas the International Crisis Group suggests there could be up to 700,000 [5]. This surplus of Syrians in Istanbul partly coincides with some under-recording of Syrians in Turkey as well as with some over-recording of Syrians in other provinces who relocated to Istanbul but did not re-register. These occurrences undermine the accuracy of data on the spatial dispersal of persons under international or temporary protection in Turkey and subsequently also targeted policy interventions.

[3] By the end of 2017, there were 62 designated satellite cities though numbers keep changing

With regards to the economic integration of Syrian refugees, who usually have a temporary protection status [4], they were granted access to permission to work in January 2016 [5]. This is subject to a fee payable by the employer and various other conditions and restrictions [10]. Furthermore, Syrians under temporary protection are exempted from work permits for seasonal agriculture or animal husbandry works. They can individually apply for the work permit exemption to provincial Turkish Employment Agencies [3]. Prior to this, they were not permitted to take on employment. Meanwhile, as of 31 March 2018, 19,925 work permits have been granted to Syrians under temporary protection [3] [6]. Thus, only 0.95% of the Syrian refugee population of 2,097,174 working-age individuals have permission to work. It is assumed that instead most work is irregular and outside the formal sector [5]. However, this is not uncommon as 31-35% of all economic activities in Turkey are claimed to be falling into the shadow economy [7]. Therefore, there is little hard statistical data on the employment of (Syrian) refugees in Turkey.

The other main economic activity, apart from employment, is running a business and here there is some data. As of 31 March 2018, another 13,776 work permits were granted to Syrians who set up their own business [3]. In addition, the number of newly founded businesses that have a Syrian business partner has increased more than 40-fold between 2010 and 2015, from 30 new foundations in 2010 to 1,257 in 2014 [7]. This is mostly due to new businesses in the Southern border provinces and in Istanbul. In addition, Turkstat (TÜIK) data implies that exports from Turkey to Syria dropped significantly from $0.79 bn in 2007 to $0.49bn in 2012 and from that low quadrupled again to 1.36bn in 2017 [12]. This increase is most pronounced in the provinces of Kilis ($0.13bn in 2007 to $0.7bn in 2017), Gaziantep (displaying a rise from $2.5 bn in 2007 to 6.6 bn in 2017), Mardin ($0.3bn in 2007 to $0.9bn in 2017) and Hatay (displaying a rise from $1.2 bn in 2007 to $2.3 bn in 2017) [13]. It can be assumed that a large proportion of this international trade is with neighbouring Syria and that the increase of exports to Syria is largely due to an increase in exports from the border provinces to Syria. We thus also hypothesise that there is a nexus between the increase of Turkish businesses with Syrian partners and the rise of international trade in the southern provinces. However, such macro-level data does not enable any micro-level analysis of the economic actors behind these developments. We believe it is possible to address this knowledge gap by analysing mobile phone data on the micro-level of individual users, notably patterns of usage in order to establish how the macro-level developments play out on the individual level, no-

[4] See article 91 of the Law on Foreigners and International Protection, No. 6458 and article 7 of Regulation on Temporary Protection

[5] See Regulation on Work Permit of Refugees Under Temporary Protection, Official Journal No. 2016/8375, 15 January 2016

[6] Another 20,993 work permits were issued to Syrians with residence permits but that groups is not considered in our study

[7] `http://www.tepav.org.tr/upload/files/1460720443-3.Trade_Relations_with_Syria_after_the_Refugee_Influx.pdf`

tably with regards to Syrian refugees. We do not explore this possibility further in this report.

Of interest to us is the distribution of manual and non-manual labour amongst the refugees. This tells us about opportunities and economic integration as well as social mobility. Table 1 [11] provides a breakdown of labour within the Turkish workforce. Total manual and non-manual workforce are estimated to be $6,678$ thousand and $20,527$ thousand respectively, a ratio of about 1 in 4. As we have already mentioned, a breakdown of the refugee workforce in Turkey is not available in the literature. However, we will demonstrate that we can infer the manual/non-manual dispersion of refugees using data analytics and mobile phone activity records.

Table 1. Labour distribution of Turkish nationals in 2016

	Population in 1,000		
Occupation (ISCO 08)	Total	male	female
All	27,205	18,893	8,312
Non-manual			
Managers	1,402	1,190	212
Professionals	2,786	1,507	1,279
Technicians and associate professionals	1,531	1,128	403
Clerical support workers	1,945	1,100	845
Service and sales workers	5,130	3,524	1,606
Skilled agricultural, forestry and fishery workers	4,044	2,504	1,540
Craft and related trades workers	3,689	3,244	444
Manual			
Plant and machine operators and assemblers	2,519	2,251	268
Elementary occupations	4,159	2,444	1,715

The next three sections describe our methodology: the data; the model and the machine learning data analysis methods.

3 Caller Mobility Data Set

Figures 1 and 2 show the distribution of the number of callers and the total number of calls/SMS messages made during 2017 obtained from the Türk Telecom data set *Dataset 2: Fine Grained Mobility* [6]. This data provides cell tower identifiers used by a group of randomly chosen active users to make phone calls and send texts. The data is timestamped and each particular group of users is observed for a period of two weeks. At the end of the two-week period, a fresh sample of active users is drawn at random. Each sample contains 3% of the refugee base plus an equal amount of non-refugee users.

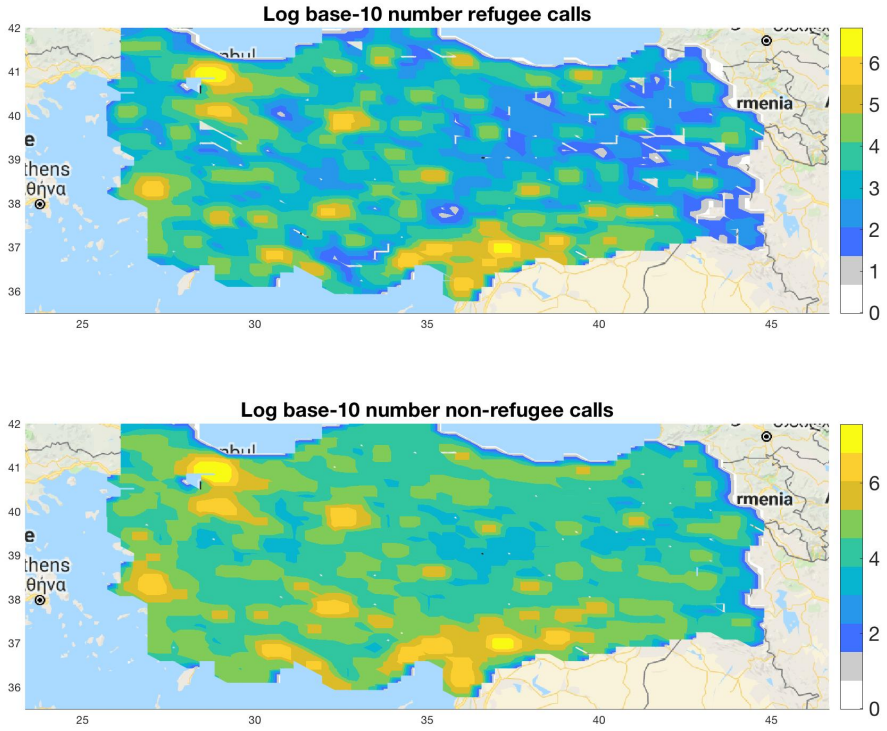


Fig. 1. Distribution of refugee and non-refugee calls and SMS texts.

The term 'refugee' is used as a blanket term in the data set, and includes migrants, asylum seekers, and foreigners who have acquired a temporarily protected foreign individual ID number in Turkey [6]. 45% of the refugee customers were registered in Istanbul. The data set contains samples of the Turkish citizen customers mainly from the cities which have significant registered refugee presence. For further details see [6].

We assume no sample bias of non-manual or manual workers in Dataset 2. This assumption will allow us infer the proportion of manual workers at each location using this data set.

4 Labour/Mobile Phone Usage Model

Our objective is to determine the relationship between mobile phone user behaviour and employment. Unfortunately, there are no existing strong models connecting mobile phone usage to user labour class (i.e. manual or non-manual). However, we are able to propose weak rules of behaviour from intuition and by interviewing several Turkish and Syrian nationals. These rules are informed

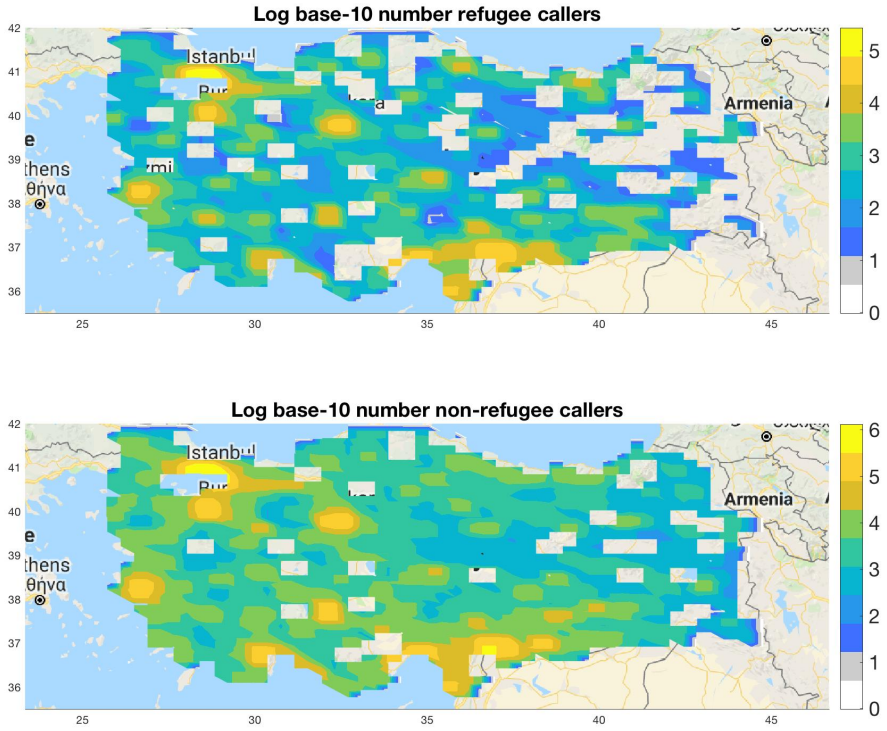


Fig. 2. Distribution of refugee mobile phone users (voice and text).

hunches and certainly not scientifically proven and hence 'weak'. We will determine the efficacy of these rules using a machine learning algorithm that establishes their consistency with Türk Telecom mobile phone data.

The assumption is that manual workers do not use their phone during working hours, 7am to 1pm and 2pm to 8pm, However, non-manual workers may use their phones for business purposes and thus more during working hours. Workers' usage of telephones for private purposes spike around lunch time and after work, hence after 7pm. Calls between 7pm and 11pm are of a private nature. Manual workers may make a call before 7am to determine work availability. We note that non-workers do not use their phone as early in the morning as workers of any kind, and they also use it more evenly across the day. Research suggests that students display specific patterns and rather use their phones in the afternoon and evenings [1]. Furthermore, mobile phones "still accentuate social inequalities insofar as their factual usage patterns are tightly correlated with the various purposes of social actions, as well as with different situations, social relationships and social roles" [4].

We use these insights to build rules associating Dataset 2 to manual and non-manual work and consider both mobile calls and SMS on working days only

(Monday to Friday). Calls/SMS before 0800 or after 1800 are deemed to be out of working hours. Calls/SMS between 0900 and 1700 are assumed to be during working hours. We assume both Turk nationals and refugees exhibit very similar mobile phone usage behaviours to each other.

The following rules characterise the different mobile phone behaviours for two job classes (i.e. manual and non-manual), for both refugees and Turk nationals. A *rule* maps a *condition* onto the job class.

The following conditions provide evidence that worker is a manual worker:

Condition 1 With an empirical probability greater than 0.8, when a worker sends more than 3 SMS messages in a day the worker sends fewer than 3 SMS between 0800 and noon and between 1400 and 1700 on a working day.

Condition 2 Worker makes a call before 0700 during working day.

Condition 3 With an empirical probability greater than 0.2, worker makes one or more calls between noon and 1400 during working day.

Condition 4 Worker makes a call between 0800 and 1700 in rural area and then somewhere else between 1800 and midnight on a working day.

The following conditions provide evidence that the worker is a non-manual worker:

Condition 5 The worker makes more than 3 calls or SMS in total between 0900 and noon or between 1400 and 1700 on working day.

Condition 6 If caller is a refugee and calls a refugee between 0800 and noon and between 1400 and 1700 during a working day.

For each worker who makes a call or sends an SMS we investigate if these conditions apply. For those that do a label, either *manual* or *non-manual*, is assigned to the worker by that condition. If the condition is not satisfied then no corresponding label is assigned to the worker. Consequently, the above rules provide 'censored' data in that, they only provide a job classification when the condition holds true. Thinking in terms of a decision tree, the apex node is the condition and there is one branch for when the condition holds that has a corresponding leaf value (*manual* or *non-manual*). The other branch, when the condition does not hold, has an unknown leaf value. Note, it does not follow that the worker should be assigned the opposite job class when the condition does not hold. We applied these conditions to the mobile phone data from Türk Telecom Dataset 2 on both Turkish and refugee calls and SMS. Table 2 shows the proportion of Turk and Syrian workers for which each condition holds.

The rules developed in this section provide a tentative definition of manual and non-manual classes [8]. In the next section we develop an unsupervised

[8] We note here our initial intention to investigate cross border trade between Syria and Turkey using a similar rule-based approach. However, we were not able to formulate rules appropriate to this task that exploited the D4R challenge data sets.

Table 2. Proportion of refugee and non-refugee workforce assigned either *manual* or *non-manual* by each employment rule.

	Proportion of workforce (%)					
	Cond. 1	Cond. 2	Cond. 3	Cond. 4	Cond. 5	Cond. 6
Refugee	5.3	19.3	15.3	4.0	13.0	4.4
Non-refugee	11.2	23.5	20.4	8.0	0.0	4.9

machine learning algorithm to determine worker employment from their mobile phone behaviour using our employment rules.

5 Data Analysis Method

Our aim is to classify each phone caller or texter as manual, non-manual or non-worker. We use the rules presented in the previous section to inform our algorithm of the definition of manual and non-manual work. However, the rules are weak and so should be subject to some refinement by the mobile phone data. To this end, we use an extension of an unsupervised Bayesian data analysis technique, the Independent Classifier Combination Algorithm (IBCC) [8, 9], to classify mobile phone users from their mobile phone usage. The IBCC learns the relationship between the employment rules, presented above, and the data in Dataset 2 (See Section 3). The IBCC does not require telephone call/SMS data labelled as *manual* or *non-manual* to train it. Instead it uses the assumption that the majority of the rules are accurate.

Our method models both refugee and non-refugee workforces within a single principled Bayesian approach and thus can infer both refugee and non-refugee labour classifications without loss of information. The approach provides both an estimate of each Turk's and refugee's job class, a statement of confidence in the classification as well as the efficacy of the job classification rules. We note that our approach is not biased by the relative number of rules for manual and non-manual work. Furthermore, non-employed phone users are deemed to be those whose behaviours do not conform to manual nor non-manual worker rules.

The IBCC methodology is extended to accommodate censored data arising from our rules (see Section 4) and also includes prior knowledge of manual and non-manual work amongst the Turkish workforce, as per Table 1. A graphical representation of the extended IBCC model is presented in Figure 3. The nomenclature follows that in [8]. The Dirichlet prior distribution parameters, $\nu_{T,0}$, for the probability, κ_T, of manual, non-manual or non-worker of a Turkish national for the whole of Turkey are derived from Table 1 (i.e. 6679, 20528, 44388 respectively). The Dirichlet prior for the refugee community is uninformative. The ground truth employment type, t, for each refugee, indexed i or Turk, indexed j, is drawn from the respective Dirichlet distributions. The employment status, $c_{T,j}^{(k)}$, for Turk j inferred using rule k is drawn from the confusion matrix, $\pi^{(k)}$, for rule k and the true worker status $t_{T,j}$. Likewise for the refugee employment

status. The confusion matrix is assumed drawn from a set of Dirichlet distributions, one for each manual, non-manual or non-worker, with hyper-priors $\alpha_0^{(k)}$. The reader is invited to contact the lead author for further details.

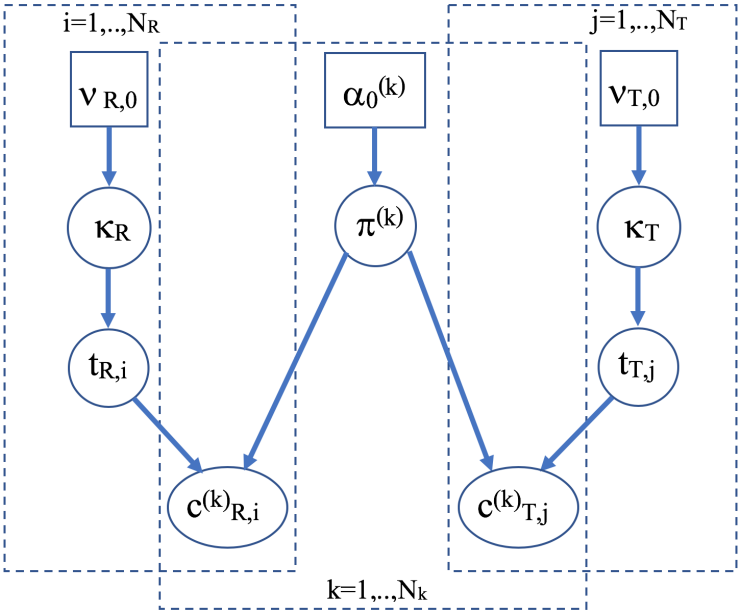


Fig. 3. A graphical model representation of the extended IBCC Bayesian approach to worker job type classification. N_R and N_T are the number of refugees and Turks in the sample data set, respectively. N_k is the number of rules (six in our case).

The approach learns a confusion matrix for each rule presented in Section 4. The confusion matrix models the likelihood of a worker type for each rule condition. In detail, consider a 2×2 confusion matrix: the upper left value represents the probability of a condition holding true given a manual worker (i.e. $p(C \mid M)$); top right, the probability that the condition does not hold true for a manual worker (i.e. $p(\neg C \mid M) = 1 - p(C \mid M)$); the lower row, left entry is the probability that the condition holds for a non-manual worker (i.e. $p(C \mid \neg M)$ and finally, the bottom right entry is the probability that the condition does not hold for a non-manual worker (i.e. $p(\neg C \mid \neg M) = 1 - p(C \mid \neg M)$) [9]. Each rule condition holds most often for one worker type only (manual or non-manual): Condition 1 holds more often than not for manual workers; Condition 6 holds for non-manual workers more often than not. However, we have no prior knowledge about the validity of the rule when the worker is not of that type. It is possible,

[9] The sign $\neg$ indicates 'not' or 'non'. So $p(\neg C \mid \neg M)$ is the probability that the condition C does not hold for non-manual workers.

in the case of Condition 1, that the condition could hold more often than not for non-manual workers too. Thus, we assign the following prior, α_0 in Figure 3, for manual rules:

$$\begin{bmatrix} 1.5 & 1.0 \\ 1.0 & 1.0 \end{bmatrix}$$

For a non-manual rule the rows are permuted:

$$\begin{bmatrix} 1.0 & 1.0 \\ 1.5 & 1.0 \end{bmatrix}$$

as the probability that its condition holds for a non-manual worker is greater than for a manual worker. The low values of α encode our uncertainty in the conditional probability values.

6 Results

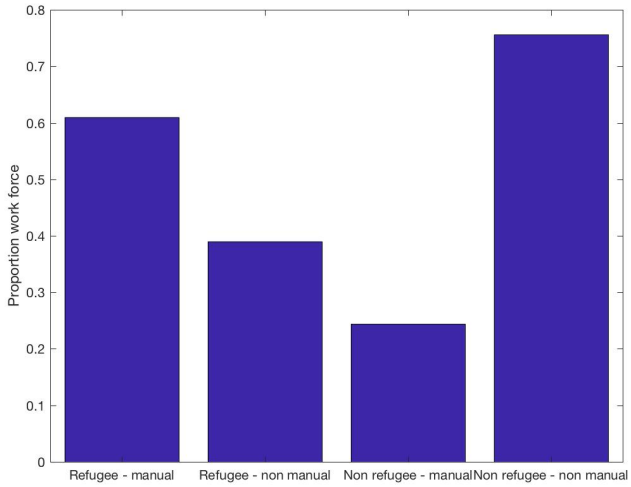


Fig. 4. Manual and non-manual worker distribution of workforce for both refugee and non-refugee communities within the Türk Telecom Dataset 2. The standard errors in the 'refugee - manual' and 'refugee - non-manual' proportions are both 0.01.

Figure 4 shows the distribution of manual and non-manual workers for both refugee and non-refugee communities. Note that the non-refugee posterior distribution is not significantly different to the prior information, as per Table 1, indicating that our model is flexible enough to reconcile the mobile phone data for the Turkish community with our prior knowledge of Turk employment. However, the distribution of refugee workers is noticeably different to that of the

Turks. There are approximately 60% manual workers in the refugee working community. This is consistent with studies which suggest a strong concentration of Syrian workers in the informal sector [10].

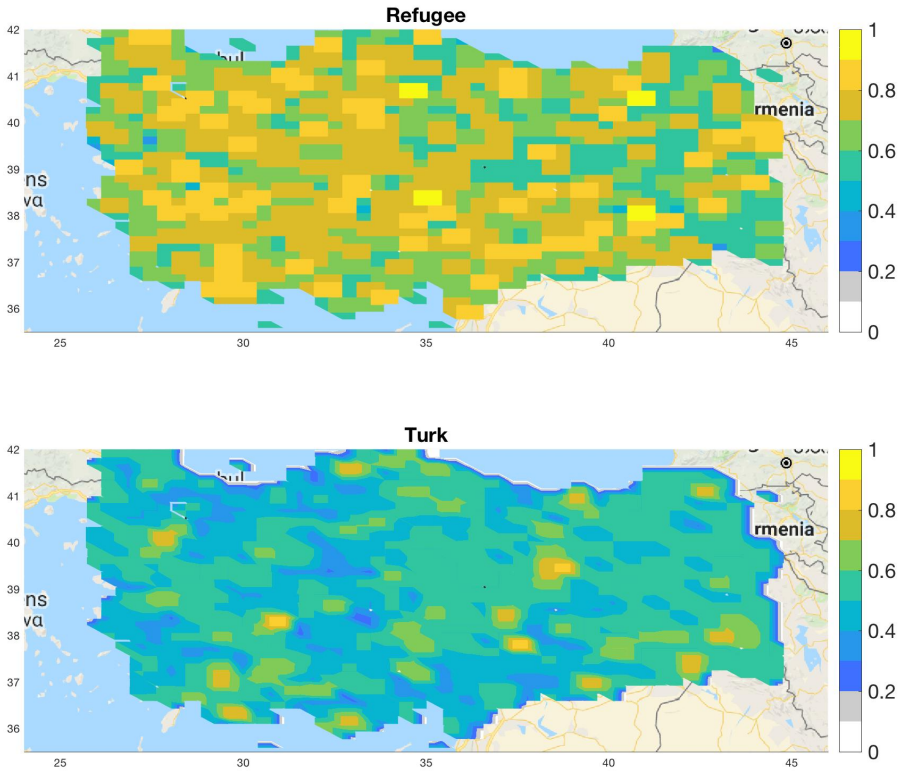


Fig. 5. Distribution of manual workers as a proportion of all workers from the same community.

Figure 5 shows labour dispersion maps, the mean probability that the mobile phone user is a manual worker for refugee and non-refugee communities respectively (again, the probability of non-manual worker is one minus the probability of manual worker). As our statistical methods yield uncertainty in these probabilities Figure 6 shows where our results are statistically significant. In these percentile maps we are 95% confident that the proportion of manual workers is no less than the indicated value in the lower map in each figure and 95% confident that it is no larger than the value in the upper map.

[10] http://documents.worldbank.org/curated/en/505471468194980180/The-impact-of-Syrians-refugees-on-the-Turkish-labor-market

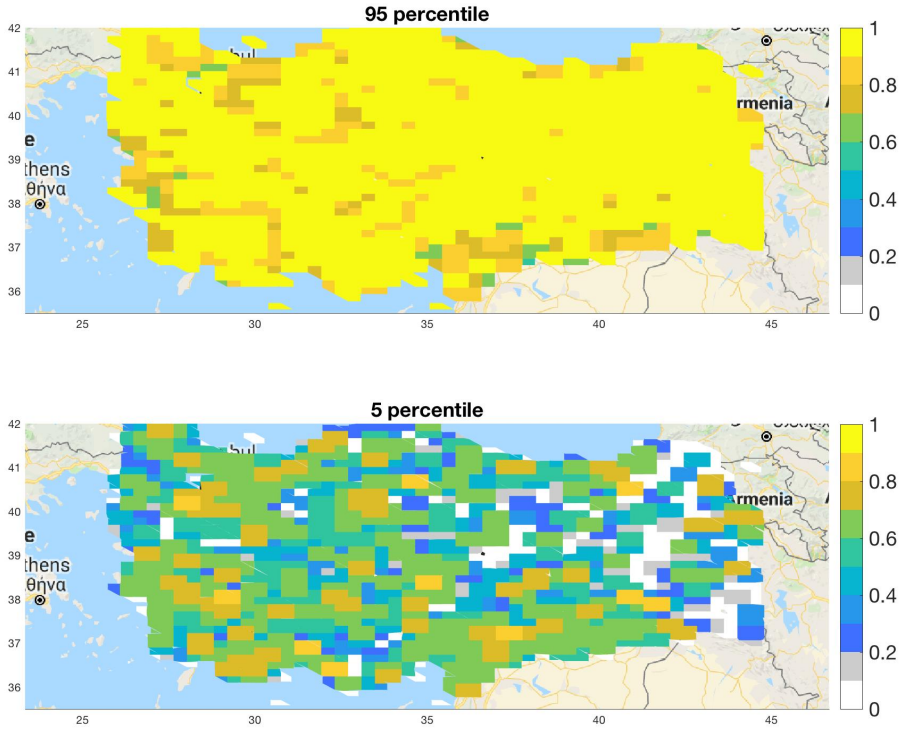


Fig. 6. 5 and 95 percentile confidence of refugee manual workers as a proportion of all refugee workers.

Figure 7 shows the differences in the probability between refugees and Turks of being in manual work. These differences could point to differentiated access to the labour market and to jobs and thus to structural or individual discrimination. Similarities in the probability between refugees and Turks of being in (non-) manual work could point to similar structures and functioning of the local labour market with regards to refugees and Turks.

We also investigated the probability of non-employment by refugees and non-refugees. Figures 8 shows a heatmap of the probability that a draw of an non-employed person from each location in Turkey is a refugee. Figure 9 also shows percentile heatmaps, as per Figure 6, indicating where we are confident in these results. Our findings with regards to probability of refugees not in work suggest that in most places of high concentration of refugees they are also more probably unemployed than Turks. However, there is one exception and that is the region of Bursa. Here refugees seem much less likely to be unemployed than Turks in the province. Bursa is an industrial hub and its surrounding are a centre of agricultural activities, notably growing of grapes and olives. Thereby, it

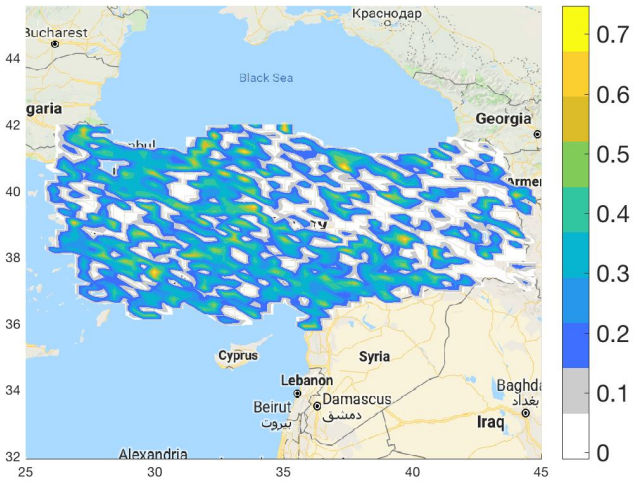


Fig. 7. Difference in probability between refugees and Turks being a manual worker.

offers a favorable opportunity structure for refugees who can find employment in manufacturing and/or agriculture.

7 Conclusions and Future Work

In this study we found that refugee workers are more likely manual than non-manual workers (in contrast to Turkish workers). Refugees don't normally have permission to work and only have access to informal employment. Our results not only provided country-wide statistics of employment but also gave a detailed breakdown of employment characteristics via heatmaps across Turkey. This information is valuable since it would allow GOs and NGOs to refine and target appropriate policy to generate opportunities and economic integration as well as social mobility specific to each area of Turkey.

We appreciate that we are at the very start of big data analytics for social analysis from mobile telephone data. Several immediate follow-on research directions include the integration, where available, of telephone call durations. Our current model uses SMS texts as a proxy for short duration calls. We believe further useful information is embedded in call duration. Furthermore, we were only provided with an approximate position of each caller, i.e. within the radius of an antenna but not the exact location. Unfortunately, for urban environments, where residential, industrial and shopping areas lie close to one another, this is not detailed enough. Further, our analysis would benefit from a clearer distinction between non-manual workers, notably people in employment and people in education. An additional rule to be applied to the analysis could be based on school hours, roughly 9am to 4pm versus working hours, roughly 8am to 6pm, as well as days at school, and, on that basis, search for change of telephone

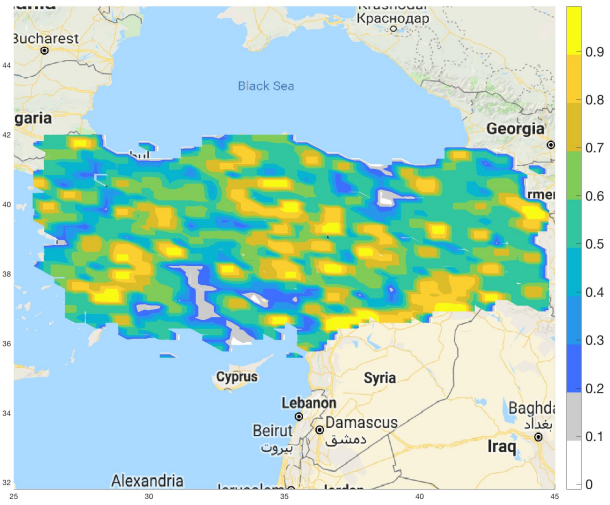


Fig. 8. Probability that an non-employed person is a refugee.

use patterns around these times and days and distinguish, in the data set, between people in work and education. We believe our analysis would benefit from explicit rules of telephone behaviours by the unemployed. Age and gender metadata for each Türk Telecom user would provide an opportunity to dig deeper into unemployment demographics.

We have shown how weak social science rules can be folded into the data analysis and how knowledge of Turk employment demographics can provide a spring board to determining refugee demographics within a Bayesian framework. Our approach would improve by using other employment indicators. For example, a reviewer kindly pointed out to us that the manual workforce should increase in size during harvest season. The Bayesian framework is flexible enough to accommodate this kind of information, complementing our telephone behaviour rules, and should provide a more robust worker classifier. Identifying and deploying further employment indicators will be a key element of our future research.

References

1. Chakraborty, S.: Mobile phone usage patterns amongst university students: A comparative study between india and usa (2006)
2. Directorate General on Migration Management: Temporary protection, distribution of Syrian refugees in the scope of temporary protection by province (2018), `http://www.goc.gov.tr/icerik6/temporary-protection_915_1024_4748_icerik`
3. European Council: Assistance to Syrian refugees in Turkey, document presented to conference 'Supporting the future of Syria and the region', Brussels, 24-24/4/2018

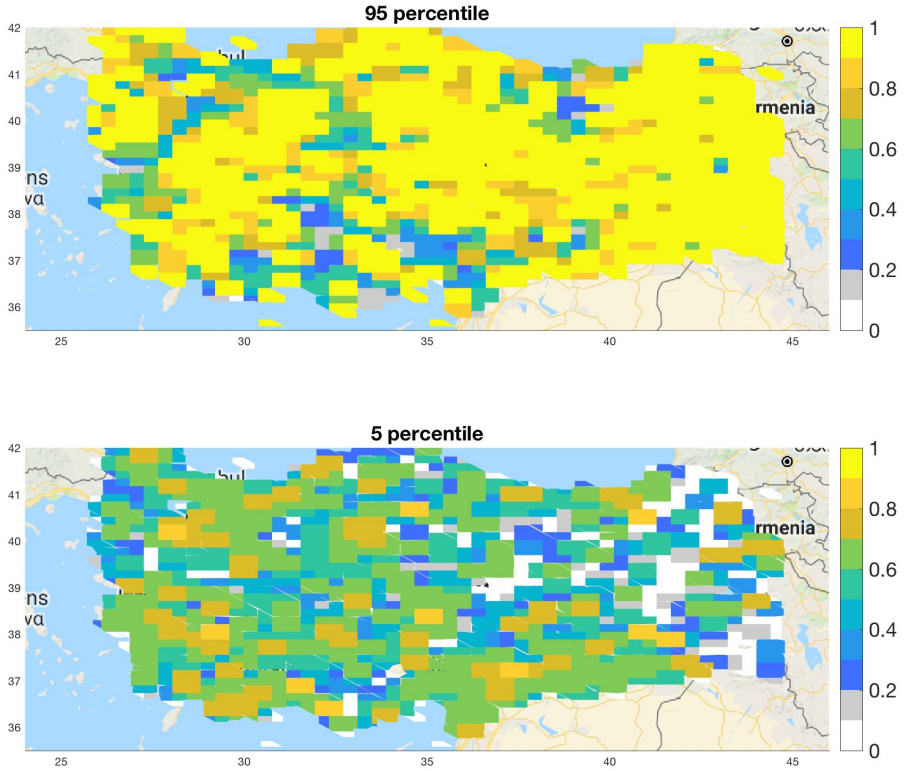


Fig. 9. 5 and 95 percentile confidence for the probability that an non-employed person is a refugee.

(2018), https://www.consilium.europa.eu/media/34146/turkey-partnership-paper.pdf

4. Geser, H.: Towards a sociological theory of the mobile phone. E-Merging media: communication and the media economy of the future pp. 235–260 (2004)
5. International Crisis Group: Turkey's Syrian Refugees: Defusing Metropolitan Tensions (Jan 2018), https://www.crisisgroup.org/europe-central-asia/western-europemediterranean/turkey/248-turkeys-syrian-refugees-defusing-metropolitan-tensions
6. Salah, A.A., Pentland, A., Lepri, B., Letouzé, E., Vinck, P., de Montjoye, Y.A., Dong, X., Dağdelen, O.: Data for refugees: The d4r challenge on mobility of syrian refugees in turkey. arXiv preprint arXiv:1807.00523 (2018)
7. Schneider, F., Savaşan, F.: The Size of Shadow Economies of Turkey and of Her Neighbouring Countries from 1999 to 2005, Working Paper no. 31, Linz: JKU Economics Department (2006)
8. Simpson, E., Roberts, S.J., Smith, A., Lintott, C.: Bayesian combination of multiple, imperfect classifiers. In: NIPS 2011. Oxford (December 2011)

9. Simpson, E., Reece, S., Roberts, S.J.: Bayesian heatmaps: probabilistic classification with multiple unreliable information sources. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 109–125. Springer (2017)
10. Turkish Labour Law (206): Turkey grants work permit for Syrian refugees (Jan 2016), https://turkishlaborlaw.com/news/legal-news/362-turkey-grants-work-permit-for-syrian-refugees
11. Turkish Statistical Institute: İSTATİSTİKLERLE TÜRKİYE, Turkey in Statistics (2016), https://ec.europa.eu/eurostat/documents/7330775/7339623/Turkey+in+Statistics+2016.pdf/0fd9e008-7045-48ba-9839-484bb4761097
12. Turkstat: Exports by country (2018), http://www.turkstat.gov.tr/PreTablo.do?alt_id=1046
13. Turkstat: Exports by province, 2002-2018 (2018), http://www.turkstat.gov.tr/PreIstatistikTablo.do?istab_id=646
14. UNHCR: The practice of satellite cities in Turkey, http://www.unhcr.org/50a607639.pdf

Syrian Refugee Integration in Turkey: Evidence from Call Detail Records *

Tugba Bozcaga[†] Fotini Christia[‡] Elizabeth Harwood[§] Constantinos Daskalakis[¶]
Christos Papadimitriou[‖]

December 26, 2018

Abstract

Over the past seven years, the needs of the three and a half million Syrian refugees have shifted from emergency response to programs focused on their integration. Using D4R call detail records (CDRs), this report focuses on questions derived from the academic literature on refugee integration and explores whether and how local characteristics and service provision affect refugee integration. Unlike existing studies, this study addresses multiple factors in a single analysis, accounting for potential confoundedness between different factors that might otherwise bias the results. Our analysis employs linear regression, regularization techniques, and multiple data sources. We find that social integration is affected by multiple socioeconomic, welfare-related, or spatial factors such as economic activity level, availability of health facilities and charity foundations, network centrality, and the location of the district. In addition, long-term over-time movement of refugees is motivated by the availability of scarce welfare resources such as health clinics, as well as economic activity levels and the availability of religious facilities in the district. Our results show that policy makers concerned with social integration must site projects and services in targeted ways.

1 Introduction

Displaced for nearly seven years, the over three and a half million Syrian refugees currently in Turkey have had their needs shift from emergency response to programs focused on their integration. Using D4R call detail records (CDRs), this report focuses on questions derived from the academic literature on social integration and explores whether and how the local context and service provision affect refugee integration. We test these hypotheses using primarily linear regression analyses. We combine D4R CDRs with a variety of geolocated or district-level administrative data gathered from government organizations.

To date, scholars have largely explained social integration by focusing on a single factor (Dancygier and Laitin 2014). This study addresses multiple factors in a single analysis. The triangulation

*We are grateful to the Republic of Turkey Ministry of Interior, the Red Crescent (Kızılay), and the Disaster and Emergency Management Authority (AFAD) for their data support. We thank Ahmet Utku Akbiyik for his research assistance.

[†]Ph.D. candidate, Political Science, Massachusetts Institute of Technology, Email: bozcaga@mit.edu
[‡]Professor, Political Science, Massachusetts Institute of Technology, Email: cfotini@mit.edu
[§]Graduate Student, IDSS, Massachusetts Institute of Technology, Email: eharwood@mit.edu
[¶]Professor, CSAIL, Massachusetts Institute of Technology, Email: costis@mit.edu
[‖]Professor, Computer Science, Columbia University, Email: cp3007@columbia.edu

of the CDR data with our diverse data types helps us to analyze different potential factors of integration simultaneously, thereby accounting for potential confoundedness between different factors that might otherwise bias the results. We analyze potential factors affecting social integration under three categories: socio-economic, welfare-related, and spatial. In addition, we look at refugee mobility, particularly how the available services and other factors spur or dampen refugee movement across Turkish districts, which may inform Turkish government officials' efforts to manage future refugee flows.

We use multiple data sources for our analyses. Specifically, we have compiled a dataset on development indicators and public services available in all 972 Turkish districts by scraping thousands of public government websites. To our knowledge, this is the most comprehensive database on Turkish local governance to date, with information ranging from basic development indicators such as literacy rates to more complex data on number and locations of health clinics, schools, and charity foundations, allowing us to estimate various types of service provision available to refugees and their impact on integration. We also make use of data on two governmental projects for social assistance directed to refugees, the Ministry of Interior's district-level data on Syrian refugee residence in Turkey, and data on Turkish general elections.

We find that social integration is affected by multiple factors. Economic activity level is negatively correlated with integration. The number of health facilities is positively correlated with integration, but only in districts with very low refugee population, suggesting that it may be easier for refugees in such districts to get the requisite information and healthcare services without burdening the local system. Religious-oriented foundations are associated with higher integration, presumably due to the material aspect of charity they offer. Among spatial factors, both refugee and Turkish callers *living* in camp and coastal districts are less likely to make inter-group calls, while refugees and Turks *visiting* those areas are more likely to do so. In addition, if a refugee lives in a place with high centrality among refugees, her integration level is lower. Similarly, if a Turkish citizen lives in a place with high centrality among Turks, she is less likely to interact with refugees. Finally, long-term over-time movement of refugees seems to be motivated by the availability of rather scarce welfare resources such as health clinics, as well as economic activity levels the availability of religious facilities in the district.

While CDRs have been used to address several phenomena ranging from natural disasters (Tomaszewski 2014), to disease (N. Baldo and P. Closas 2013; Mari et al. 2017; Tompkins and McCreesh 2016; Lima et al. 2015), and poverty (Pokhriyal and Jacques 2017)(D4R:p3), along with the other studies in D4R, this study will be one of the first studies that utilize CDRs in questions related to social integration. In addition to the contribution of this study to the literature on CDRs and social integration, it also has important policy implications, as discussed in the last section of the paper.

The structure of the paper is as follows. We allocate the next section to a discussion of the data. In the section, we explain the methods we used to create certain measures from CDRs and make a detailed analysis of antenna- and individual-level calling behavior. The third section is composed of our main analysis. It investigates the impact of various factors in the existing literature on the social integration of refugees. The fourth section analyzes the over-time movement of refugees. The final section summarizes our policy-relevant findings and policy suggestions.

2 Data

When Syrian refugees started flowing into Turkey in April 2011, the government upheld an "open door" policy. Turkey now hosts over 3.5 million Syrian refugees, notably more than its critical threshold (ORSAM 2015:12). As Turkey has ratified the 1951 Geneva Convention on the legal

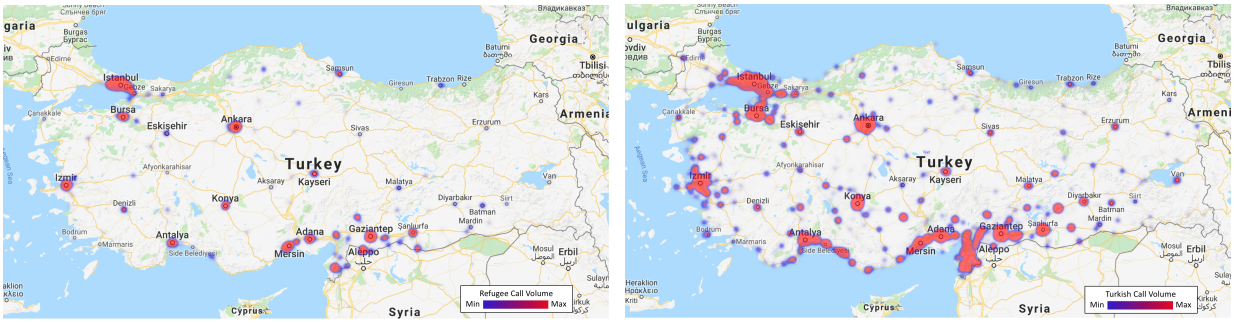


Figure 1: Geographic Spread of Call Volume for Refugees (left) and Turks (right)

status of refugees with geographic limitations, it cannot accept Syrians as refugees. Rather, Syrian refugees are considered "guests" and based on an October 2011 Turkish Interior Ministry decree, registered Syrians are given "temporary protection" that grants them rights to indefinite residence, protection from coerced return to Syria, and some aid for emergency needs (ORSAM 2015:7, 12). Roughly 91% (2,774,018) of Syrian refugees live outside camps as compared to 9% (246,636) that live in camps. Though there are 81 provinces in Turkey, the majority of refugees reside in 20 of them as per Appendix Table 3.

Based on March 2017 data, Turkey has a mobile penetration rate of about 95% across all mobile carriers, with Turk Telekom accounting for roughly 25% of that (BTK 2017; D4R 2018:p.4). We don't have exact penetration rates for the refugee population and though Turk Telekom used the whole refugee customer base to sample the individual-level CDR, there are geographic fluctuations in penetration (D4R 2018:p.5).[1] The CDRs include information on a total of 105,277 antennas distributed across Turkey's 81 provinces as per Appendix Figure 8, which gives a sense of antenna concentration across Turkey's 972 districts.

We start out by examining Dataset 1, the total call volume per antenna for Syrian refugees and Turks. The proportion of Turks to refugees in the dataset is approximately 4.37 to 1 (D4R:p3).As expected, the overwhelming majority of calls are made by Turks. There are a total of 408,520,715 calls in the year, with refugees making 49,455,533 and Turks the remaining balance of 359,065,182. Refugees thus account for 12% of the total call volume, with Turks making 7.2 times the number of calls as compared to Syrians. Turks make an average of 1,268,781 calls per day as compared to 174,754 for Syrians, with a frequency of 0.5 calls per hour for Syrians and 3.7 calls per hour for Turks. Sampling for the refugee customers was from all geographic locations, while Turkish citizens were primarily sampled from the 11 cities with significant refugee populations (D4R:p3-4). In both cases, Istanbul accounted for approximately 45% of the customers (D4R:p4).Thus there is notable geographic variation for these calls, as seen by the heat maps in Figure 1 that indicate call volume for Turks and Syrian refugees across Turkey. Out of a total volume of 164,682,860 SMS, Syrian refugees sent 10%, amounting to 16,749,745, while the remaining 147,933,115 were sent by Turks. Refugees sent an average of 54,382 SMS per day compared to Turks, who sent 480,302, i.e. 8.8 times as many messages.

[1]Note that given the structure of the data released by D4R we cannot tell if the calls are directed to Syrians inside Turkey or in Syria. Over 90% of calls in Dataset 1 originate from a known site in Turkey and are directed to an unknown site or operator noted as 9999 that could be inside Turkey or abroad.

2.1 Identifying Antennas in Camp, Coast, and Border Locations

We create different indicators using CDRs to be used in our regression analyses. We operationalize three dimensions of heterogeneity among refugees that could affect their attitudes and behaviors toward integration: 1) whether they reside inside or outside camps 2) whether they reside in areas on the Aegean coast with accessible points of departure towards the Greek isles, as this could indicate a desire to migrate to Europe and 3) whether they reside in areas on the Syrian border that makes potential movement in and out of Syria and into Turkey easier, which may be affecting attitudes towards long-term settlement and integration in Turkey.

For refugee camps, we rely on data from the Ministry of Interior to identify the size and district location. There is a total of 232,992 refugees in 20 camps across 18 districts. Using Google Maps, we found the exact geolocation of each of the camps and estimated their spatial size.[2] We then assigned antennas to each camp location by distance from the camp and in proportion to the refugee population of the camp, as reported for June 2017. We included only antennas with a nonzero refugee call volume in Dataset 1 and those which had at least 25% total call volume made by refugees. Any overlapping antennas were assigned to the camp larger by population unless the other camp had no other antennas assigned. To see whether the measure calculated by this decision rule is valid, we look at the correlation between the total call volume by camp and camp refugee population, which gives us an estimate of 0.73. See Appendix Table 4 for camp names, district location, geolocation, camp size in sqms and number of refugees and number of antennas assigned to each camp.[3] We also use a more inclusive decision rule where we use all antennas in the districts associated with camp locations as antennas capturing refugee traffic for camp districts. This can be seen as an upper bound of camp-related refugee communication.

For refugees on the Aegean coast, we use UNHCR data for refugee flows from the Turkish coast into the Greek islands and associate them with the specific departure points on the Turkish coast that have the shortest distance to the islands that receive refugee flows. We identify 12 such geolocation points across five Greek islands for which UNHCR collects refugee inflow data (Leros, Samos, Lesvos, Chios, and Kos) and look at the closest cell tower with nonzero refugee call volume to determine antennas associated with these departure points. We also use a more inclusive decision rule where we use all antennas in the districts associated with coastal departure points as antennas capturing refugee traffic for districts that are departure points for migration to Europe, an upper bound for such communication. See Appendix Table 5 for crossing point names, district location, geolocation, and the number of antennas assigned to each departure point.

Lastly, for refugees on the Syrian border, we geolocate UNHCR data on border crossings along the Syrian-Turkish border. As with the coastal departure points, we locate the closest cell tower with nonzero refugee call volume to identify the antennas associated with these border points. We also use a more inclusive decision rule where we use all antennas in the districts associated with Syrian border crossings as the upper bound of such refugee communication traffic. See Appendix Table 6 for border crossing names, district location, geolocation, and the number of antennas assigned to each crossing.[4]

[2] Please see Appendix Figures 9 and 10.

[3] The two refugee camps Nizip 1 and Nizip 2 are in, effectively, the same location, so they are treated as one camp for the purposes of antenna assignment. Together they have 5 antennas associated with them.

[4] We note that two antennas assigned to border crossings closest cell tower rule were also assigned to refugee camps. See Table 7 in the Appendix for a breakdown.

2.2 Identifying Antennas with Potential Concentration of Sectarian and Friday Prayer Activity

We operationalize three measures for religiosity in the form of sectarianism and Friday prayer attendance. We employ distinct Shi'a and Sunni Syrian holidays to estimate potential points of geographic concentration for Sunni and Shi'a Syrian refugees and explore the role, if any, of sectarian identity in refugee integration. To predict Shi'a refugee concentration, we examine which antennas tend to make more calls on the Ashura holiday compared to a regular day; while to predict Sunni refugee concentration, we focus on Mawlid-al-Nabawi. Specifically, we identify antennas where there is significantly more call volume by hour during these *Syrian religious holidays* as compared to the relevant baseline level of calls using a paired t-test. Looking at measures of sectarianism, no camp districts or metropolitan areas show exclusively high Shi'a or Sunni concentration, suggesting a more heterogeneous clustering of Syrians in such districts. As shown by some province and district-level maps depicting antenna locations, there appears to be some neighborhood-level clustering among antennas that show similar types of behavior on religious days.

We also focus on Friday prayer as a way to identify mosques attended by Syrian refugees as potential focal points for socialization and collective action. We narrow in on the 11 am - 12 pm time interval for refugee calls for Fridays, the time when they may be coordinating on where to meet for prayer, and try to identify specific hotspots of refugee concentration at those times. Specifically, using a paired t-test, we examine whether the hourly average for 11am-12pm (time before Friday prayer) is significantly higher than the hourly average for 12 pm - 2 pm (Friday prayer time falls into this time range if we consider all the provinces in Turkey) throughout 52 weeks. In 500 of around 27,000 antennas for which we have complete data, refugee calls are significantly higher between 11am-12pm, at the 10% significance level. While no metropolitan areas show a high concentration of Friday prayer activity among refugees, indicating less homogeneity and collective socialization of refugees in those places, some camp districts show greater Friday prayer activity and potential for collective action.

2.3 A First Check of Antenna-Level CDRs

When we check whether the total volume of refugee calls (Dataset 1) reflects the underlying population, we find that it is positively correlated with refugee rates of concentration in each province. Specifically, we compare how daily refugee call volumes across Turkey –as proxied by antenna traffic data– compares to province-level data on refugee concentration from June 2017, provided by the Ministry of Interior. A 10,000 increase in the refugee population is associated with 630 added calls per day (Appendix Table 10). This correlation reassures our confidence in the representativeness of the data, given that the average refugee population is 35,560 and the average number of calls per day is 2,075 across Turkey's 81 provinces, suggesting an average of approximately 583 calls for every 10,000 refugees.

The correlation continues to hold when we narrow down our focus to the refugee population in camps (Appendix Table 10). It turns out that an increase of 10,000 refugees in camps is associated with an increase of 370 calls, where the average refugee population is 12,262, and the average total number of calls per day is 373 across all camps. Furthermore, the association between the population and the number of calls is statistically significant at the 1% level.[5] This value is around 60% of the increase we identified for the overall refugee population. Assuming that our decision rule to identify camps is a good approximation, the lower increase in the daily number of calls in camps, as compared to that outside camps, might indicate a lower use of phone calls overall in a more confined

[5]Please see the Appendix for detailed results and summary statistics.

space.

Finally, we also look at refugee call volume on the Aegean coast, an area of departure for Syrian refugees who want to move onwards to Europe. The left panel of Figure 2 shows a heat map that focuses on the closest points on the Turkish coast across from the Greek isles with the highest inflow of refugees (as per UNHCR 2017 on the top 5 Greek islands of arrival) (see Appendix Table 5). We look at the correlation between refugee arrivals to the Greek islands and calls associated with the potential departure points described above. To that end, we use UNHCR's data on daily refugee flows from the Turkish coast into the Greek islands, and call data from antennas for each specific departure point, as described above. Employing this panel data with island dummies to control for unobservable characteristics for each point of arrival, we see a positive yet slight association between the number of calls at each of the 20 antennas and refugee flow, calculated by linear regression. For each additional refugee arrival, we see an increase of 0.002 in the number of calls per antenna per day, with a statistical significance at the 1% level [6]. At the average number of arrivals per island per day, 14.6, this implies a 23 percent increase relative to the sample mean, 1.35. (Appendix Table 12).

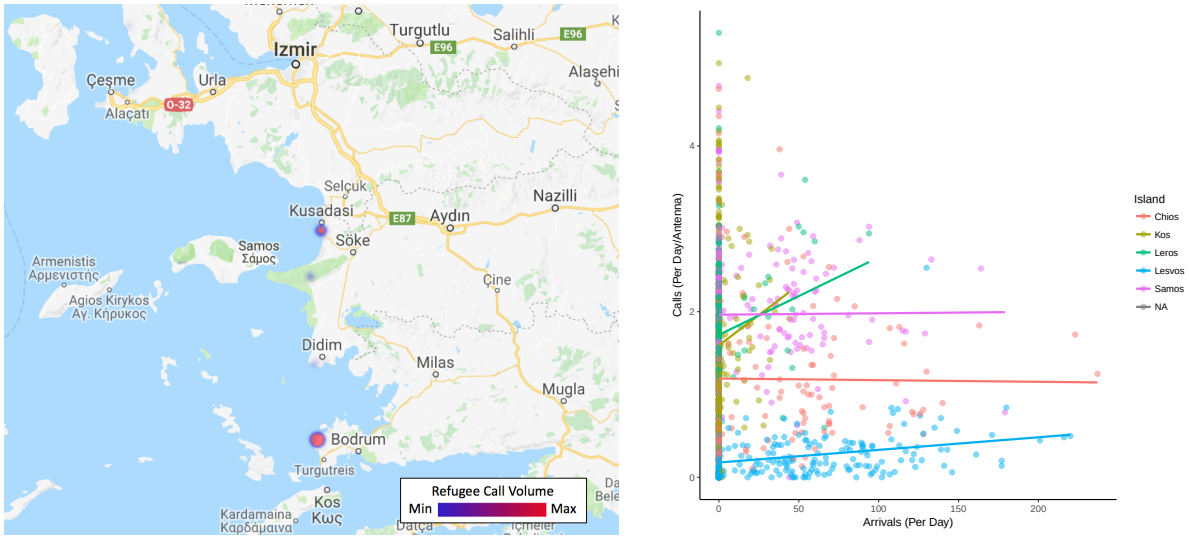


Figure 2: Refugee Call Volume at Focal Coastal Departure Points (left) and Daily Refugee Flows and Call Volume (per Antenna) at Focal Coastal Departure Points (right)

Table 1: Summary Table: Population Correlation

Factor	Effect	Significance Level
10,000 increase in Refugee population	Increase of 630 calls per day	1%
10,000 increase in Refugee population (camps)	Increase of 373 calls per day	1%
14.6 increase in the number of refugee flows	Increase of 0.03 calls per day per antenna	1%

Despite the rather limited SMS usage by refugees (survey data from the Ritsona refugee camp in Greece shows that 64% of Syrian refugees use their mobile phones to make calls, while only 16% of them use them to send text messages with 94% using WhatsApp (Latonero et al. 2018)), the

[6] We conducted a least trimmed squares regression trying different alpha values. The coefficients changed between 0.0015 and 0.002 for the possible range of alpha values, from 1 to 0.5

positive correlation of the overall refugee population as well as those in camps with SMS volume persists, albeit lower in magnitude than call volume. A 10,000 increase in the refugee population is associated with 150 added text messages per day across Turkey's 81 provinces, and 160 added SMS across the 20 camp locations (Appendix Table 11). We find no association between refugee flows to Greek Islands and SMS, perhaps due to the inconvenience of SMS in time constrained situations (Appendix Table 13).

2.4 Calling Behavior in Antenna-Level CDRs

As per several survey reports on Syrian refugees in Turkey (see AFAD 2017, ORSAM 2015, and EDAM 2014) there are differences in attitudes and behaviors among Syrian refugees. We look at heterogeneity in behavioral patterns between citizens versus refugees; refugees living in camps versus outside camps, refugees living in the coastal areas close to the Greek islands versus those that are not, as well as refugees living close to the Syrian border.

While Turkish call volume is only slightly higher in December 2017 than in January 2017, following a seasonal rather than increasing trend, there is a notable increase of refugee calls, starting out with 7% of the total call volume in January 2017 and reaching as high as 16% of the total call volume in October. The volume of text messages sent by Turks is slightly lower in December 2017 than in January 2017 and does not follow a discernible trend. As with call volume, the refugee SMS volume increases considerably over the year, starting out with 5% of the total call volume in January 2017, peaking in September with 19% of the total call volume, and ending at 10% in December.

Meanwhile, the share of Turk Telekom users rose from 18,560,000 to 19,590,000, corresponding to an approximately 1 million (5.5%) increase compared to the previous year. Another 1 million increase is seen in the number of users of alternate operators (BTK, 2017). The increasing overtime trend in refugee call volume could thus be due to increased Syrian refugee inflows into Turkey;[7] and/or an overall increase in mobile phone usage, along with an increase of Syrian refugee subscribers into the Turk Telekom network; and/or an improvement of refugee circumstances that allows them to spend more money on calls.

We also focus on variation in calling patterns among Turks and refugees over weekdays, weekends and holidays to see whether their different calling behavior suggests different socialization patterns that leave a footprint and are indicative of a differential likelihood for integration. First, we explore Syrian refugee calling behavior as compared to Turkish calling behavior for the 24-hour interval of an average weekday and weekend. We also examine Fridays separately, as for Syrians, Friday is the main day off, unlike for Turks whose weekend is Saturday and Sunday.

[7] The number of refugees increased from 2,880,325 to 3,049,879 between January and June 2017. It further increased to 3,381,005 by December 2017 as per statistics provided by the Turkish government.

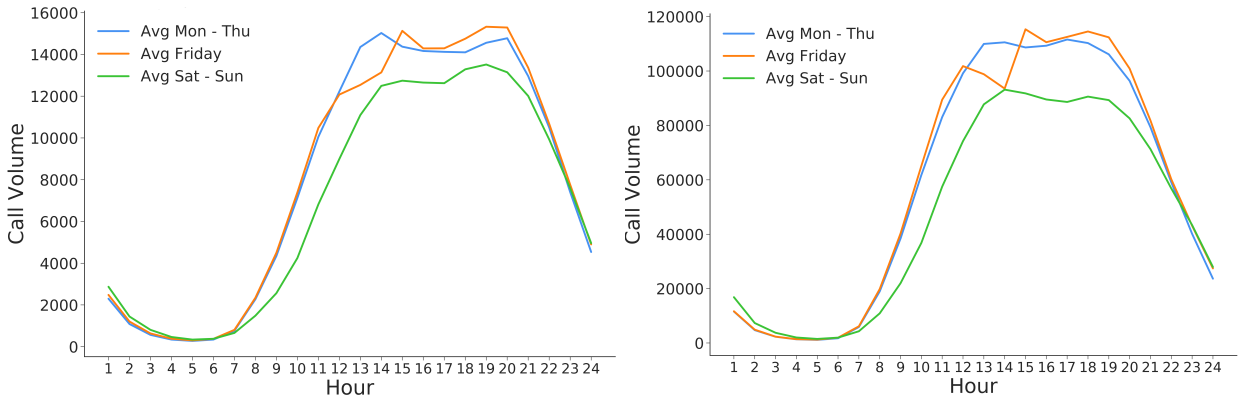


Figure 3: Call Volume for Refugees (left) and Turks (right) for average Weekday versus Weekend

Table 2: Summary Table: Refugee and Turk T-test Results

Value	Comparison	Effect	Significance Level
Turk Call Volume	Weekday vs Friday	More calls on Friday	1%
Turk and Refugee Call Volume	Weekend vs Weekday, Friday	More calls on Weekday, Friday	1%
Refugee Call Volume	Ramadan vs Non-Ramadan	More calls during Non-Ramadan	1%
Refugee Call Volume	Camp Refugees vs Non-Camp Refugees	Higher for Camp Refugees	1%
Refugee Call Volume	Ashura, Muharram, and Mawlid vs Average	Higher than average	1%
Camp Refugee Call Volume	Ashura, Muharram, Mawlid, and Eid al Adha vs Average	Higher than average	1%
Turk Call Volume	Eid al Fitr vs Average	Lower than average	5%

As per the figures above, both Syrians and Turks appear to follow similar calling patterns except Friday. We test to see if any observed differences on the graphs are also statistically significant using a paired t-test (see Appendix Table 16). The observations used for the t-tests in this part are the average number of calls/SMS per hour, averaged over all. This way, both groups start the next the same volume, 100. Syrian refugees appear to have adjusted to calling patterns in Turkey as both Syrians and Turks tend to make more calls on weekdays (1061 and 10136) and Fridays (1138 and 11056) than on weekends (same for SMS; 98 and 1686 more messages on weekdays, 356 and 12833 more messages on Fridays). Refugees, however, are less likely to make phone calls (-153 for weekdays, -179 for Fridays, and -42 for weekends) and more likely to use SMS (2.92 for weekdays, and 3.45 for weekends) than Turks on all days, indicating their lower spending capacity as compared to the host population and their potential use of alternate communication such as in-person interactions or web related communication.

To compare refugee and Turkish calling behaviors more directly, we conduct a paired t-test for the two ethnic groups in Appendix Table 17. As expected, due to population size, the total number of calls is much higher for Turks throughout the day. For this reason, we adjust for the total number of calls in the two groups by establishing the number of calls in the first hour, 1-2 am, as the baseline level (setting it to 100), and rescaling the calls in the remaining 23 hours as a percentage of the baseline. Even after setting the baseline level to 100 for both refugees and Turks, we still find that refugees are in general less likely to make phone calls than Turks on all days, indicating the lower

spending capacity of refugees as compared to Turks. We do not observe this negative gap in SMS volume, however. On the contrary, based on a baseline level of 100, the adjusted SMS volume is higher among Syrians than among Turks, with a 7% gap on weekdays and 15% gap on weekends (see Appendix Table 23). These findings suggest that SMS might serve as a substitute for calls for some refugees and that they may turn to alternate means of communication.

2.5 Calling Behavior in Individual-Level CDRs

Next, we focus on Dataset 2 to get at the trends of communication between refugees and the host community, by looking at the volume of calls between refugees and Turks during 2017. [8]First, we examine the aggregate level of Turkish and refugee calls broken down by received or initiated by refugees. Then we see how many calls were among refugees, how many between Turks and refugees, and how many among Turks, along with the information on who initiated them. Unsurprisingly, the majority of calls is in-group calls among Turks, and only 0.05% of calls from Turks are directed toward refugees. On the other hand, most of the calls made by refugees are to Turks (89.78%), whereas only 10.22% of refugee calls are directed towards other refugees. Of the 2,710,716 unique Turkish callers in Dataset 2, 17,330 make calls to refugees, while 2,693,386 do not. On the other hand, 565,437 refugees call Turks and only 22,921 do not. This suggests that refugees may use alternate communication methods with other refugees, such as the internet or in-person communication. Using the daily number of calls per individuals, we examine how these calling patterns vary by hour of the day, by day of the week and by month, both for incoming and outgoing calls.

We show calling behavior in two ways. First, we break down the calls by call type, categorizing them into four groups: Refugee-to-Turk, Refugee-to-Refugee, Turk-to-Refugee, and Turk-to-Turk. The number of calls is lowest in the Turk-to-Refugee and Refugee-to-Refugee call categories. The average number of calls made to refugees is lower than those to Turks, a trend that may result from their lower proportion in the population or lack of integration. Even when we look at who refugees call, the number of calls to refugees remains limited, and it is hard to draw any implications without controlling for demographic variables.[9] Yet, a small part of calls made to Turks appear to be work-related since they happen on weekdays with a dramatic drop over the weekend. Appendix Figures 20 and 21 suggest that the hourly, weekly, and monthly trends overlap.

Second, we break down the calls by caller type, categorizing them again in four groups: Refugees who contact Turks, refugees who do not contact Turks, Turks who contact refugees, Turks who do not contact refugees. A comparison of these four groups reveals interesting trends. Turks contacting refugees make more calls than Turks not contacting refugees. At the busiest time of day, the former group makes around 7 calls, as compared to the latter's 3.5 (right graph in Appendix Figure 20). This suggests that Turks contacting refugees are generally more networked people. If this gap between the two types of Turks is due to work-related reasons, it should disappear at weekends, but it persists (right graph in Figure 21). This further suggests that Turks contacting refugees are generally people who make more calls for social, rather than work-related reasons. Similarly, refugees not contacting Turks are on average individuals making fewer calls, suggesting that those who engage in inter-group communication are more networked and of higher spending power.

In this descriptive analysis, and the regression in the coming section, we treat an overtime increase in the percentage of refugee-to-Turk calls and Turk-to-refugee calls as indicative of increased

[8]This dataset contains 42,053 antennas spread across Turkey's regions, with over 14 thousand antennas in the Marmara region, roughly half that in the Central region, and over 6 thousand in the Aegean region. The rest are distributed among the Mediterranean, Black Sea, Southeast, and East regions. The histogram in the Appendix shows the distribution of antennas across Turkey's 81 provinces.

[9]We trust that this difference in relative call volume is not due to the way the data was sampled in Dataset 2, but rather is a reflection of calling patterns in the populations considered.

integration, and treat rates of incoming and outgoing calls of refugees to other refugees as in-group socialization. To enable an annual comparison between in-group and across-group calls, we set the initial value for all groups to the same level. We thus adjust for the average daily number of calls among the four call types, by establishing the number of calls in the first biweekly period, the start of the year, as the baseline level (setting it to 100). Rescaling the calls in the remaining 24 periods, the values in the rest of the periods are shown as a percentage of the baseline. The adjusted values are in Appendix Figure 23. Even after setting the baseline level to 100 for different types of calls or callees, we do not find a significant annual increasing trend in the refugee-to-Turk or Turk-to-refugee calls, as compared to refugee-to-refugee and Turk-to-Turk calls.

3 Individual-Level Regression Analysis

The UN Refugee Agency estimates that it takes displaced persons an average of fifteen years to determine whether to return back home, stay in their host country, or migrate elsewhere (Cupolo 2017). During this long displacement cycle, asylum seekers face a slew of issues on the path to integration—from accessing basic rights and services, to finding jobs, to adjusting to new lifestyles and cultures. Displaced for nearly seven years, the over three and a half million Syrian refugees currently in Turkey have had their needs shift from emergency response to programs focused on their integration (Broomfield 2016). This main part of our report focuses on questions derived from the academic literature on social integration and from qualitative and survey accounts on Syrian refugees in Turkey. In line with the potential factors of integration discussed in the literature, we first look at the importance of socio-economic factors. Next, we turn to the effect of service and welfare-related factors, including refugee-targeted programs, and close with a look at spatial factors affecting integration. We estimate the effect of all these different factors in a single regression analysis.

To merge the individual-level data with the antenna- and district-level measures that we calculated and the district-level administrative data obtained from official sources, we identify a home location, or "home antenna" for each caller id, as determined by the antenna the caller uses most frequently. We define the district in which the antenna is located as the "home district". Since Dataset 2 allows for fine-grained mobility patterns by antenna, we can identify a plausible home location as well as patterns of movement around that home-location during the day and during different days of the week, albeit only for two weeks for each randomly selected user. Seeing in which district or antenna of Turkey the caller resides allows us to see the effect of various characteristics of the district on social integration.

We use Dataset 2 to calculate regression estimates. Since our data has a pooled cross-sectional nature, in the design we use a multilinear regression with province and period dummies (unless stated otherwise), where each two-week period of Dataset 2 corresponds to a time unit and each observation corresponds to one individual. We also control for the unobserved time effect using linear or unit-specific time trends. The standard errors (SEs) are clustered both by province and by period. To measure our independent variables, we use individual, as well as antenna or district-level measures. To account for confounding factors, we add a number of control variables. The specification of the OLS regression model we present in the graphs is in the Appendix.

A refugee's level of social integration is proxied by the ratio of calls made by refugees to Turks over the total number of calls made (in %). To create the measure, we first subset Dataset 2 to those observations where the caller is a refugee and then calculate the percentage of inter-group calls over all the calls made by the user. We also run the same analysis for the subsample where the caller is a Turkish citizen, by operationalizing the dependent variable (DV) as the percentage of inter-group calls made by Turks. Although both measures rest on a similar idea, whether the inter-group call is

made by a refugee or by a Turk gives us different pieces of integration. Although both types of calls involve inter-group contact, the proportion of inter-group calls made by a refugee gives us a measure of the refugee's ability to interact with Turks, while the proportion of inter-group calls made by a Turkish citizen can be a good proxy for how open the Turkish citizen is to interacting with refugees. In addition to this conceptual difference, it allows us to investigate the impact of a Turkish citizen's home location on calling patterns, in addition to analyzing the impact of a refugee's home location. The fact that a refugee calls a Turk does not mean that the call will be reciprocated by Turks in the neighborhood. Whether integration, as measured by two different types of calling direction, gives conflicting findings is important for policy. All data sources and variables we employ are listed in the Appendix.

D4R data should be approached with caution in terms of the representativeness of the data as the market share of Turk Telekom shows fluctuation across provinces. In addition, since we combine individual-level call data with district-level administrative data, we also need to omit individuals whose home antenna lacks geospatial location information. For these reasons, the probability of sampling does not perfectly overlap with the underlying demographic distribution by province. To address that concern, we calculate post-stratification weights by province [10] and use these weights in our individual-level regression.

3.1 Socio-Economic Factors

Difficulty securing employment and ensuring self-sufficiency is considered a hurdle to integration (Burchett and Matheson 2010; Bansak et al. 2016), and economic insecurity has been shown to drive prejudice against immigrants and refugees (Scheve and Slaughter 2001). In line with this literature, there have been some indications of frustration among local residents in Turkey, who have blamed refugees for an array of economic challenges associated with hosting them including dwindling employment opportunities, rising housing prices, and business competition (Erdogan 2015; Getmansky et al. 2018). Lazarev and Sharma (2017) find that, though references to religion can positively predispose Turkish respondents to the plight of Syrian refugees as it concerns Muslim refugees in a Muslim host nation, this effect disappears as soon as there is a reference to the costs of refugee presence in the country. According to a comprehensive and representative survey conducted with refugee populations in Turkey, some portion of refugees agree with the view that the adverse effects from Syrian arrivals include impacts on housing prices and rents (41%) and job opportunities and wages (21%) (AFAD 2017:99). This line of research also contends that low levels of education are a powerful predictor of anti-immigration sentiments, as low education makes citizens more vulnerable to risks such as unemployment and low wage (Kitschelt 1997; Cavaille and Marshall 2018). Survey answers from Turkish citizens, however, point to an opposite effect, showing that less educated Turks may be more tolerant of refugees (KONDA 2016).

The sources of anti-immigrant attitudes are not limited to economic and material reasons, nor do they go away when economic conditions improve. Ethnic and religious differences, coupled with negative ethnic stereotypes, play a role in anti-immigrant and anti-refugee sentiment (Burns and Gimpel 2000; Hainmueller and Hopkins 2013). In fact, refugee country of origin proves more important for naturalization success than other applicant traits including language skills, levels of integration and economic status (Hainmueller and Hangartner 2013). Gadarian and Albertson (2014) identify that increased anxiety and stress towards refugees creates a self-fulfilling bias towards seeing refugees as a threat. Only according to 18% of refugees, though, religious differences in religious life seem to be an obstacle for refugee integration, while cultural differences (44%), differences in social life (40%),

[10] Let p_A be the probability of being from province A in the population, and s_A the probability of being from province A in the sample, the weight for province A is calculated as p_A/s_A.

ethical differences (29%) play a more important role (AFAD 2017:95). Nevertheless, identity-related differences aren't the only factors that affect levels of social integration. Ample evidence in the literature suggests that citizens' immediate exposure to heightened refugee flows leads to difficulties with social integration (Hangartner et al. 2017). Consistent with this literature, Getmansky et al. (2018) find that Turkish host citizens who have higher levels of interactions with refugees on a daily basis are more likely to perceive them as a threat and express negative views about them.

Support for parties with anti-immigrant messages also creates problems for integration. Recent research has looked at the effect of refugee flows on vote shares for extreme right parties (Steinmayr 2016; Dustmann et al. 2016; Dinas et al. 2017) and suggests that negative attitudes toward immigrants are associated with host citizens on the political right (Karreth et al. 2015). However, the attitude of parties against refugees doesn't necessarily align with this categorization on the right-left dimension. While the leading conservative party is more tolerant and open, the leftist-secular and the extreme-rightist voters prefer a more restricted policy. Getmansky et al. (2018) find that partisan identification matters, with supporters of the governing party being less likely to perceive refugees as a threat. Relevant social factors include the religious background of refugees, local voting patterns, and the degree of refugee flows to the area (as proxied by the proportion of refugee population over the total population). Given that we do not have district-level data on unemployment and housing prices, relevant economic factors included in the analysis are the district's economic activity level (as proxied by the number of ATMs), individual-level spending power (as proxied by the total number of outgoing calls), and illiteracy rates. We add controls for district population and refugee population because a larger population size implies a higher probability of a given individual to meet someone from the other ethnic group. We also control whether the district is an urban center or not.

In the absence of data on background characteristics of refugees, we rely on the district-level data and CDRs to determine to what extent local socio-economic factors might track levels of social integration. To identify antennas with Shia, Sunni, or Friday prayer activity, we use CDRs from religious holidays (See the Appendix for more detail.). Data on vote shares come from official election statistics, data on district-level refugee population come from the Ministry of Interior, and population data come from the Turkish Census. Data on ATMs come from official government web pages, the total number of outgoing calls from CDRs, and illiteracy rates from official education statistics.

While all specifications, including those with individual fixed effects, linear time trends, and unit specific time trends, give similar estimates, we interpret the relatively more conservative two-way fixed effects unless otherwise stated (Table 32, Column 4).

First, we look at the subgroup where the caller is a refugee, and the integration measure is the percentage of calls made to Turks. In line with the literature above, we expect that identity differences, the potential 'burden' of the refugee, and the size of the economic pie shared by local population to have a negative effect on social integration. Accordingly, Sunni activity, the incumbent party's vote share, and individual spending power of the refugee should have positive coefficients, while the estimates for Shia activity, local economic activity, and proportion of refugee population should have negative signs. In our results, we find no effect for Sunni or Shia activity, incumbent party vote share, individual spending power, or proportion of refugee population. From the economic factors, only the level of economic activity in the district has a statistically significant effect. In line with our expectations, the direction of the effect is negative. Accordingly, the percentage of inter-group calls shows a 13 percentage point (p.p.) decrease for a 10 percent increase in the number of ATMs, suggesting that integration is lower in areas with higher economic activity. The population control is positive, as expected. This result may suggest that refugees' integration is harder in places where the economic pie shared by local population is large. Refugees in those places

might be unemployed in high numbers and thus excluded from the formal economic landscape, and consequently, they are not reaping the benefits of residing in a location with higher economic activity. Looking at the control variables, the percentage of inter-group calls shows a 2.3 p.p. increase for a one percent increase in district population. This probably results from the fact that refugees living in districts with high a population of Turkish citizens are more likely to know and interact with any Turkish citizen.

Next, we turn to the subgroup where the caller is a Turkish citizen and the integration measure is the percent of calls made to refugees. When we measure integration by the percent of inter-group calls made by Turks, we find that Turks with higher spending power are less likely to call refugees. In addition, Turks living in places with lower education levels are relatively more likely to call refugees. In both estimates, the effect size is minuscule. For an average Turk, who makes a total of 50 calls in two weeks, this means an only 0.01 p.p. decrease in the proportion of calls to refugees. With other predictor variables held constant, a one percent increase in the illiteracy rate is associated with a 0.02 p.p. increase in the proportion of calls to refugees. This shows that, in contrast with the literature but in line with the survey answers from Turkish citizens, more educated Turkish neighborhoods appear to be less open to refugees.

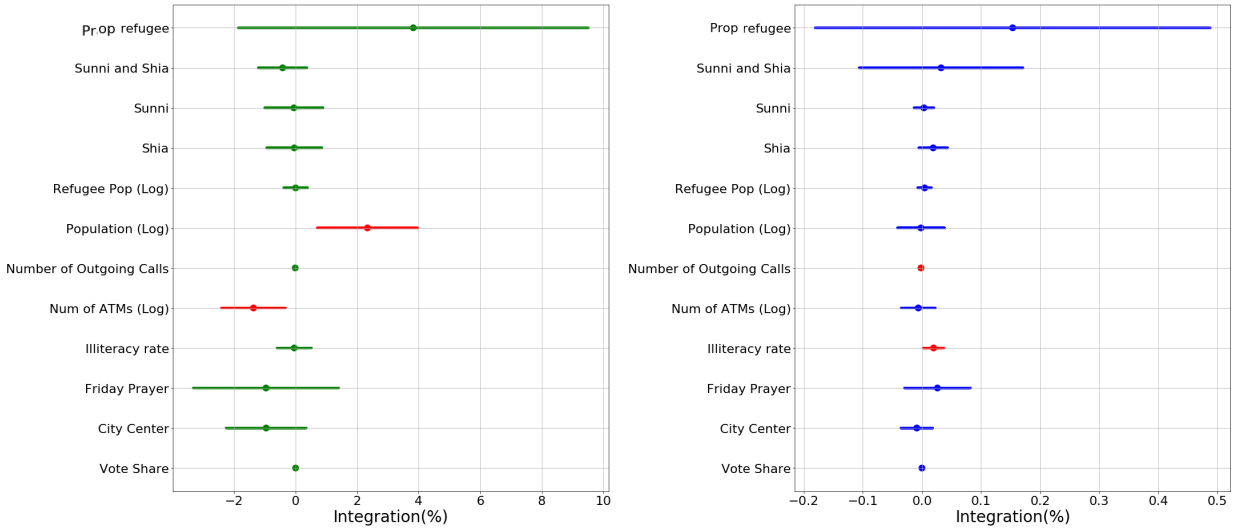


Figure 4: Confidence Intervals for Potential Socio-Cultural Factors - Integration measured by refugees' calls to Turks (left) and by Turks' calls to refugees (right), with significant results in red

3.2 Welfare-Related Factors

Next, we examine the welfare-related factors associated with refugee integration. Given that financial struggles (72%) are listed as one of the most salient obstacles to integration by refugees in Turkey, we expect to find a positive association between services and welfare resources available to refugees and integration. In the literature, the proponents of the contact theory contend that contact among beneficiaries of facilities such as schools and health clinics can produce positive effects on integration (Allport, 1954; Paluck and Green, 2009; Pettigrew and Tropp, 2006). On the other hand, competition over local resources such as public services may limit access or cause frustration among local residents, negatively impacting the degree of refugee integration (Esses et al. 2001). Our data on the number of education and health institutions allows us to see whether the presence of public services positively correlates with refugee integration. To take into account the amount of extra demand over public services, we also look at the effect of how the impact of public services

interacts with the number of refugees. According to survey evidence, a substantial proportion of refugee population (37%) also received aid from non-governmental organizations (AFAD 2017:67), which suggests that another welfare resource for refugees and immigrants is charity foundations.

We use CDRs along with information on geolocated educational and health facilities, as well as religious or non-religious *waqfs* (charity foundations that provide social assistance to those in need) across Turkey to examine whether the availability of such facilities and resources is associated with higher levels of integration. We also analyze whether the availability of such facilities and institutions is positively correlated with the degree of refugee *access* to services. Our district-level data on health facilities and public schools come from official government websites. Data on religious or non-religious waqfs is coded based on information from the Directorate of Foundations.

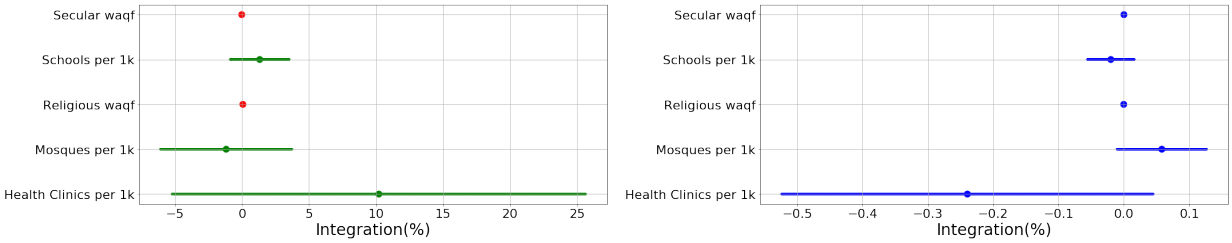


Figure 5: Confidence Intervals for Potential Welfare-Related Factors - Integration measured by refugees' calls to Turks (left) and by Turks' calls to refugees (right), with significant results in red

In line with the literature above, we expect the amount of welfare-related services and resources to be positively correlated with social integration. Accordingly, the number of education and health facilities, mosques, and waqfs should have positive coefficients. We first look at the subgroup where the caller is a refugee, and the integration measure is the percent of calls made to Turks. The results suggest that the number of health, school, or religious facilities, on average, does not have a significant association with integration (Appendix Table 32). The number of health facilities is positively correlated with integration only in districts with low refugee population. In a district with 100 refugees, for example, each additional health facility (per 1k residents) is associated with a 19 point increase in the percentage of calls made to Turks. In districts with larger refugee populations, this positive effect disappears (Figure 5). The effects of schools and mosques are insignificant across all levels of refugee population.

These results lend support to the view that excessive competition over local resources can have a negative impact on integration. Consistent with qualitative reports, the high number of Syrian refugees in some areas might have overwhelmed the capacity of local health institutions that lack requisite personnel and equipment, casting refugees in a negative light (Ergin 2016). On the other hand, the reason for the lack of an effect from school facility access 32) might be that refugee children oftentimes attend specific sessions in schools. Another potential reason is that many Syrian refugee children get employed in the informal sector instead of continuing their education due to financial obstacles. Similarly, the reason why we do not find a significant association between the availability of mosques and integration 32) may be that refugees and Turkish citizens might be visiting specific mosques.

Whether the availability of waqfs has a positive effect on integration or not changes based on the affiliation of waqfs, or more specifically, on whether they have a religious character or not. Specifically, each additional religious waqf in the district is associated with an increase of 0.065 p.p. in the proportion of calls to Turks, while the number of secular waqfs is associated with a 0.017 p.p. decrease. If we had not controlled for vote share or number of mosques in the district, this might have simply reflected the more secular nature of the district. It is known that parties

supported by secular Turks, such as CHP and IYI, oppose refugee presence in Turkey along with the overall Syria politics of the government (Sputniknews Turkish 2018). But since we also control for the governing party's vote share and number of mosques, which both capture the conservativeness effect, it could be argued that religious waqfs are more successful in having a tangible effect on refugee lives due to the stronger grassroots nature of such organizations, (Kuran 2001, Buğra and Keyder 2006). This finding is consistent with an array of qualitative sources suggesting a notable amount of monetary and charity assistance towards Syrian refugees originating from waqfs (IHH 2018; Turkish Diyanet Vakfı 2016). Secular waqfs, on the other hand, seem to have a minuscule, yet negative effect. Further research is needed to find the mechanism underlying this negative effect.

When measuring integration by the percent of inter-group calls made by Turks, we do not find any statistically significant association with the district level availability of health and education services. Given the null findings with the alternative DV, the availability of welfare-related services appears to be more associated with whether refugees will be more likely to integrate and connect with Turks, rather than to whether Turks will be more likely to connect with refugees or their view of refugees. The mechanism through which public services affect social integration is an important factor to keep in mind when pushing for specific policies on refugee integration.

We also look at the effect of two social assistance programs targeting refugees in a separate panel data analysis. As discussed in detail in Appendix B.7, we exploit regional variation in the timing of the launch and in the proportion of beneficiaries to see whether access to such services and social inclusion facilitates integration. We find no statistically significant effect for such programs. Part of the lack of significance of this type of assistance could be that they only target the weakest and economically most marginal subset of the refugee population (i.e., there is an underlying selection effect biasing the results against integration) and/or that the amounts of aid given are actually quite small if one takes into account overall levels of need.

3.3 Spatial Factors

In the final part of our analysis of sources of refugee integration, we look at the relationship between spatial factors and integration. Given that the literature on spatial factors linked to social integration is limited, we use Ridge and Lasso models to determine the spatial factors of importance. Ridge and Lasso models use an objective function that minimizes the MSE with a penalty term, which leads many covariates to be assigned a coefficient of zero (or shrink toward zero, in Ridge). The penalty parameter is selected using a 10-fold cross-validation. According to the Ridge and Lasso estimates, the centrality of the home antennas among Turks or refugees; whether the home district is at the Syrian or Greek maritime border or at districts with refugee camps; and whether or not the refugee visits border or coastal areas are the most important spatial factors associated with integration. In our main linear regression analysis, we only used those factors selected by the Lasso model.

The full list of spatial factors considered in the Lasso and Ridge models are a) the proximity of a refugee's residence to borders (to Syria and Greece) or to refugee camps, b) patterns of travel to borders (to Syria and Greece) or camps, c) across-district, within-district, and overall mobility of refugees, and d) centrality of the districts among Turks and refugees. Using Dataset 2, we calculate a number of metrics related to spatial mobility, including measures of how many different districts and provinces the caller visits; the proportion of calls made within the home district and home province, used as a proxy for within-district and within-province mobility; whether or not the caller visits a district with refugee camps, coastal departure points to Greece, or border crossing points to Syria; whether or not the user lives in a district with refugee camps, coastal departure points to Greece, or border crossing points to Syria; and the total spatial area the caller covered over the time period, as measured by the trapezoid area that encompasses all points the refugee visited.

To create the centrality measure, we take advantage of the network structure of Dataset 1 and calculate the degree centrality of antennas, or the number of other antennas that each antenna is connected to with outgoing refugee calls, grouped by cell tower location.[11] The highest degree of a cell phone tower node in the refugee call network is 2,547, shown in appendix Figure 25, while the mean is 151. There are 1,179 tower locations with degree one. To have a better understanding of how the districts with high centrality look, we consider the relative degree centrality of antennas in Istanbul, shown in Appendix Figure 25, as well as across the country, shown in Appendix Figure 24. Appendix Figure 25 shows the antenna with the highest degree, as well as the location of all of its adjacent antennas. We use degree centrality to see how the level of connectedness of an antenna as a node in the network might differ for Syrians or Turks and correlate with levels of social integration. Following the approach in previous parts of the regression analysis, we match this centrality information with individual-level CDR data based on the caller's "home antenna", i.e., where the caller makes the highest number of calls.

The Ridge and Lasso estimates are very similar to one another, suggesting that the centrality of the antenna (where the caller lives) among Turks or refugees is the most important factor associated with refugee integration (see Appendix). Other estimates that do not shrink toward zero are whether or not the caller resides in/visits a district with refugee camps, coastal departure points to Greece, or border crossing points to Syria, as well as overall mobility, as measured by the log of the trapezoid area visited.

We use the Ridge and Lasso regression models not only to estimate the importance of specific spatial factors for the outcome, but also to select the variables to be included in the regression analysis. Our regression analysis first looks at the subgroup where the caller is a refugee, and the integration measure is the percentage of calls made to Turks. Our results rely on the same regression that we use in the previous parts. A one percentage point (also the mean value for the centrality index) increase in the centrality among refugees (where the highest centrality degree is 11.2%) is associated with a 1.28 point decrease in the percentage of calls made to Turks. Of course, the direction of causality is not certain here, because those less integrated might also be people who prefer to live in places more popular for refugees, rather than Turks, in the first place. We also find that refugees living in the camp or coastal districts are less likely to call Turks, and those visiting these districts are more likely to do so. Living in a camp district is associated with a decrease of around 1.6 p.p. in the proportion of inter-group calls from refugees, while living in a coastal district is associated with a decrease of 3.2 p.p. Visiting a camp or coastal district, on the other hand, is associated with an increase of 0.9 and 1.8 p.p. in the proportion of inter-group calls from refugees. This implies that refugees visiting camps and coasts might actually be better integrated while living at the Turkish-Greek border or in camps lowers degrees of integration. Multiple factors can be at the root of this correlation. Syrians that visit camp or coastal districts - probably to visit their friends or relatives or business partners - might be those refugees with larger social networks, and refugees with larger social networks are also expected to be better integrated. On the other hand, the negative correlation between residing in a coastal district and integration is also reasonable, considering that refugees living in coastal areas might be there with the intent to depart Turkey and cross to Greece and might have fewer incentives to integrate. Several existing studies suggest that coastal districts serve as launching points for migration to Europe, with anecdotal evidence on refugees reaching coastal districts with the intent to cross into Greece and what that ordeal entails (UNHCR 2015). In a similar fashion, people living in camps might be those who stayed there due to their hesitation or inability to live in non-camp settings.

[11] We note that only a subset of the total call volume of Dataset 1 is used for this network analysis because 90.8% of the call volume is listed as having an unknown base station for either incoming or outgoing location.

Next, we look at the subsample where the caller is a Turkish citizen and the integration measure is the percent of calls made to refugees. The geographical and spatial mobility trends of Turks give us additional hints regarding factors associated with integration. First, mirroring the findings where we measure the integration by calls made by refugees, an increase in the centrality of an antenna among Turks is associated with a decrease in the proportion of calls made to refugees. A one percentage point increase in the centrality index is associated with a 0.02 p.p. decrease in the percentage of calls made to refugees. On the other hand, a one percentage point increase in the centrality index for refugees is associated with a 0.04 p.p. increase in the percentage of calls made to Turks. In short, being positioned near an antenna with high centrality among the other group leads to better integration in terms of the number of inter-group calls made by an individual, both for Turks and refugees. In addition, Turks that visit camp and coastal districts are more likely to call refugees. While it is not surprising that Turkish citizens visiting camps interact with refugees at a higher proportion, Turkish citizens that visit coasts might be those with larger and more diverse social networks. Finally, Turkish citizens living in border, camp, and coastal districts are significantly less likely to call refugees despite the high refugee population in these areas. While there is no sufficient data to reveal the mechanism underlying the negative impact of coastal districts, the negative impact of border and camp districts is consistent with the literature that shows dense exposure to refugee influx leads to difficulties with social integration.

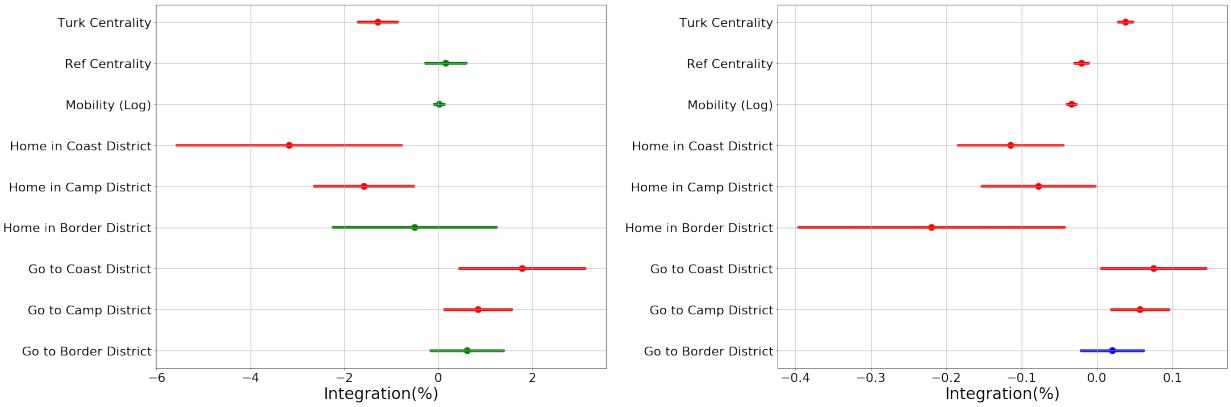


Figure 6: Confidence Intervals for Mobility-Related Factors - integration measured by refugees' calls to Turks (left) and Confidence Intervals for Mobility-Related Factors - integration measured by Turks' calls to refugees (right), with significant results in red

4 Over-time Movement in a Year

Prior scholarship suggests that family and home-country networks, as well as linguistic and cultural ties, predict where refugees choose to settle. However, advances in communication technologies, including low-cost cell connections, social media, and mapping apps, have transformed how refugees engage with information and make these decisions. Refugees increasingly rely on these tools to choose migration destinations, maintain close ties with co-ethnics, access public services in host countries, and interact with formal authorities or with informal brokers (Mandic, 2017). With the new methods of communication and information technologies, refugees might learn where to best utilize public services, educational opportunities, and labor market openings that host societies offer. We thus revisit core debates in the migration policy literature to better understand what informs contemporary refugee movements.

We use Dataset 3 to track each user throughout the year, but location information is available

only on the district level, as per the structure of the data. This does not allow us to identify a specific home-location for the user or see how a user moves within a district during his daily activities. Nevertheless, it allows us to see whether a user stays mostly within a specific district boundary, which we define as the "home district", and whether the home district changes across time. In this part, we use a panel design with individual-level fixed effects, where each observation corresponds to one refugee. Since the main independent variables are the characteristics of the district the refugee decides to live in, due to the use of fixed effects, the individuals that stay in the same district throughout 2017 drop from the model. The dependent variable is a dummy, where it takes 1 if the refugee moves to a new district, and 0 otherwise. Exploiting the longitudinal structure of the data, we determine the home district on a monthly basis for all individuals. If there is a home district that is different from the previous month's home district, we code the current home district as a receiving district and the previous home district as a sending district. The goal of the design is to see what kind of district characteristics increase the likelihood of being a receiving district.

Using our rich district-level data, we can see whether being a receiving district is associated with any of the following characteristics: high refugee population, which would indicate a desire to be close to other Syrian refugees; being in an urban center, which may indicate a desire to integrate in places with higher job opportunities; having higher levels of economic activity, as proxied by the number of ATMs in the district. We can also see whether there is more movement toward districts on the Aegean coast, which could suggest a desire to migrate to Europe; districts in close distance to the camps, which in turn might suggest a willingness to spend time with in-camp refugees; and being close to a border crossing point with Syria, which could suggest movement in and out of Syria

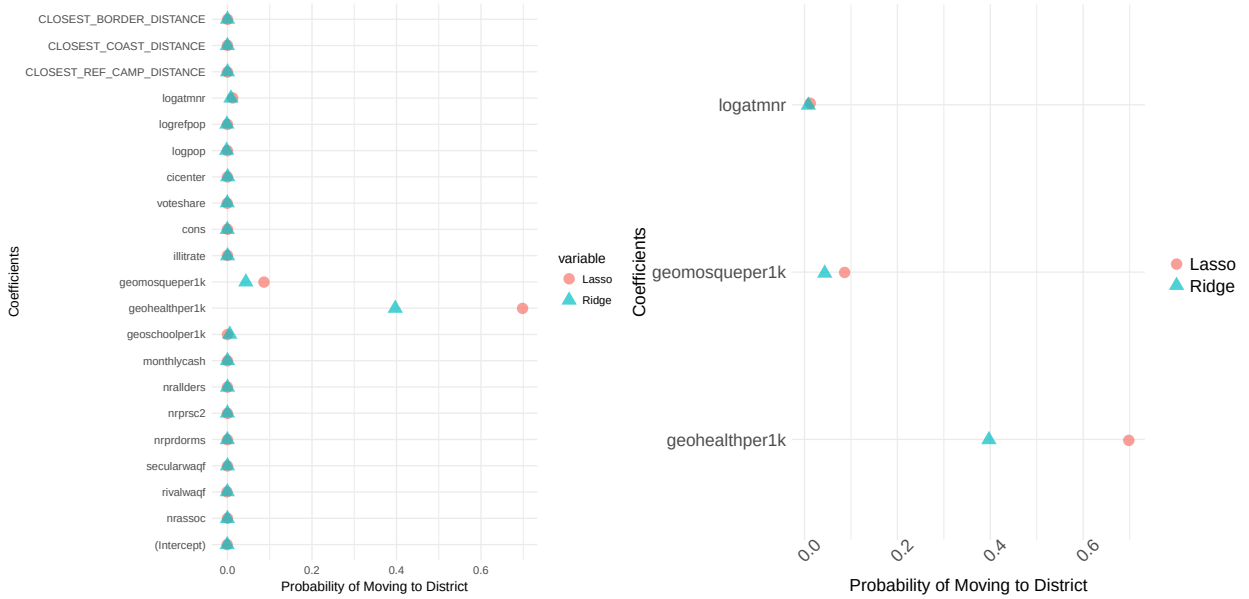


Figure 7: Lasso and Ridge Estimates for Spatial Factors Associated with Across- District Movements, all covariates included (left) and near-zero estimates excluded (right)

To see which variables are most important in determining whether a district is a receiving district or not, we first use a Ridge and a Lasso regression. We use demeaned data to isolate individual-specific unobserved characteristics. Since it is well known that refugees follow seasonal trends (such as harvests) for job opportunities, we first compared whether movements in harvest periods and non-harvest periods are statistically different from one another. This confirmation of a hypothesis

also serves as a sanity check for the mobility analysis. Given that movements in the harvest period were significantly more likely ($p < 0.1$) to be in provinces with the most seasonal workers (CSGB, 2018), we exclude the movements in the harvest period (May- September) from our analysis. The two different regularization techniques give very similar estimates to one another. Accordingly, the per capita number of health clinics in the district appears to be the most important factor for a receiving district. The other variables with non-zero estimates are the number of mosques and economic activity level of the district, as proxied by the number of ATMs.

To validate our findings, we use an OLS regression model with individual-level fixed effects, thereby keeping time-invariant individual-level unobserved factors constant. The OLS results show that each additional health clinic (per 1k residents) is associated with a 60 p.p. increase in the likelihood of whether a refugee will move to the district or not. The number of mosques, on the other hand, points to a 10 p.p. increase in the probability of moving to a district for each additional mosque (per 1k residents) (See Table 38 in the Appendix). Finally, a 10 percent increase in the number of ATMs in the district increases the likelihood to move to that district by 40 p.p. There might be several reasons as to why schools, unlike health clinics, do not appear to be a determinant of whether the refugee will move to a district or not. First, since refugee children only attend a subset of schools, the overall number of schools is not a good indicator for whether a district is appealing in terms of education services. Second, school quantity might not inform refugees' preferences, since time invested in education, unlike health, has returns only in the long run. Finally, the supply of schools might be seen as exceeding the sufficient threshold in all districts in Turkey, while the health service distribution changes greatly across districts; there are some regional hubs that offer much better health service in terms of quality and capacity, rendering health investments a reason to move to a new district.

When interpreting these results, one should note that due to the use of fixed effects in the model, our sample excludes refugees that did not change their home. In this line, the findings above suggest that a potential explanation for the long-term over-time movement of refugees, if any, is to a great extent motivated by the availability of welfare resources or employment opportunities. Given the significant effect of the number of mosques, social and cultural considerations appear to be a consideration as well. Distance to the Syrian border, refugee camps, or distance from the Greek coast do not appear to be associated with refugee movement. This additional finding reinforces the idea that when refugees move, they are motivated by the availability of resources as they intend to settle and integrate in Turkey rather than leave the country.

5 Summary and Policy Implications

Among the socio-economic factors associated with social integration, only economic activity level, as proxied by the number of ATMs in the district, seems to have a statistically and substantively significant effect on integration. Areas with higher economic activity suggest significantly lower levels of integration. This finding highlights the fact that refugees cannot partake in the formal economic landscape, either because they are unemployed in high numbers or because they are part of the shadow economy. It could also suggest that environments that have more resources to share are not particularly welcoming to refugees. Initiating employment and social inclusion projects in economically active regions, with the goal to create more inclusive economic environments for refugees, might be an investment that could lead to better integration in the long run.

Among the welfare-related factors, only health clinics and Islamic waqfs seem to be positively correlated with refugee integration. In the effect of health clinics, the size of the refugee population in the district seem to affect how conducive it is to integration. The number of health facilities is positively correlated with integration only in districts with very low refugee population, suggesting

that it may be easier for refugees in such districts to get the requisite information and healthcare services without burdening the local system. A policy implication of these results is that the state capacity for healthcare services needs to be strengthened in places with large refugee populations. The positive correlation between Islamic waqfs and social integration seems to be due to the material aspect of charity they offer, as our model controls for district-level social characteristics such as vote share and the number of mosques. This is not a surprising finding since there is substantial evidence showing the role of Islamic waqfs in service provision. The state may watch and adopt the good practices of such waqfs in service provision.

Multiple spatial factors appear to be associated with social integration. If a refugee lives in a place with high centrality among refugees, or if a Turkish citizen in a place with high centrality among Turks, her integration with the other group is lower. Consistent with this finding, if a Turkish citizen lives in a place with high centrality among refugees, her acceptance of refugees is higher. This finding might be a point of consideration in planning refugee housing and settlement projects. Another set of findings with critical policy implications is that where a refugee or Turkish citizen resides influences her social integration dramatically. For example, refugees and Turkish citizens residing in camp districts are less likely to call people from the other group, as measured by the proportion of intergroup calls over the total number of calls made. Similarly, refugees and Turkish citizens residing in districts neighboring Greek islands are less likely to call people from the other group. While selection bias might play a role for refugees, as refugees living in these areas might be less willing or able to integrate, less openness among Turks in camp or coastal districts may point to the negative effect of problems resulting from the volume of refugee inflows. Potential projects on the social inclusion of refugees may prioritize these regions.

Lastly, when the long-term over-time movement of refugees is analyzed, the availability of rather scarce welfare resources such as health clinics, rather than schools, arises as the most important factor. The economic activity level of the district, as well as social factors such as the abundance of religious facilities, also appears to determine refugees' decisions to move. This finding, consistent with the previous findings, points to the vitality of health clinics and economic environment to the integration and welfare of refugees.

6 Reference List

AFAD: Field Survey on Demographic View, Living Conditions, and Future Expectations of Syrians in Turkey (2017).

Allport, G. W.: The nature of prejudice. Basic books (1979).

Baldo, N. and P. Closas.: Disease outbreak detection by mobile network monitoring: A case study with the d4d datasets. NetMob D4D Challenge. pp. 1 (2013).

Bansak, K., J. Hainmueller, and D. Hangartner. How economic, humanitarian, and religious concerns shape European attitudes toward asylum seekers. Science. (2016).

Betts, Alexander.: Institutional proliferation and the global refugee regime. Perspectives on Politics. 7, 1, 53-58 (2009).

Blondel V. D., A. Decuyper, and G. Krings.: A survey of results on mobile phone datasets analysis. EPJ Data Science. 4, 1, 10 (2015).

Bohmer, Carol, and Amy Shuman: Rejecting refugees: Political asylum in the 21st century. Routledge (2007).

BTK Bilgi Teknolojileri ve İletişim Kurulu: Türkiye Elektronik Haberleşme Sektörü Pazar Verileri Raporu (2017).

Bugnion, Pascal: gmaps. GitHub repository. URL: https://github.com/pbugnion/gmaps. (accessed 6.10.2018) (2014).

Buğra, A., and Çağlar Keyder. The Turkish welfare regime in transformation. Journal of European social policy, 16(3), 211-228.(2006).

Burchett, Nicole, and Ruth Matheson: The need for belonging: The impact of restrictions on working on the well-being of an asylum seeker. Journal of Occupational Science. 17, 2, 85-91 (2010).

Burns, Peter and James G Gimpel: Economic insecurity, prejudicial stereotypes, and public opinion on immigration policy. Political Science Quarterly. 115, 2, 201–225 (2000).

Carlson, Melissa, Laura Jakli, and Katerina Linos: Rumors and Refugees: How Government-Created Information Vacuums Undermine Effective Crisis Management. International Studies Quarterly. Forthcoming (2017).

Cavaille, Charlotte and John Marshall: Education and Anti-immigration Attitudes: Evidence from Compulsory Schooling Reforms Across Western Europe. American Political Science Review, forthcoming.

Collyer, Michael: When do social networks fail to explain migration? Accounting for the movement of Algerian asylum-seekers to the UK. Journal of Ethnic and Migration Studies. 31, 4, 699-718 (2005).

Dancygier, R. M., & Laitin, D. D.: Immigration into Europe: Economic discrimination, violence, and public policy. Annual Review of Political Science., 17, 43-64 (2014).

Dinas, Elias, Konstantinos Matakos, Dimitrios Xefteris and Dominik Hangartner: Waking up the Golden Dawn: Does Exposure to the Refugee Crisis Increase Support for Extreme-Right Parties? Working Paper (2017).

Dustmann, Christian, Kristine Vasiljeva and Anna Piil Damm: Refugee Migration and Electoral Outcomes. The Rockwool Foundation Research Unit. Study Paper 111 (2016).

EDAM Center for Economics & Foreign Policy Studies: Reaction Mounting Against Syrian Refugees in Turkey. Public Opinion Surveys of Turkish Foreign Policy. 2014, 1 (http://edam.org.tr/en/reaction-mounting-againstsyrian-refugees-in-turkey/) (2014).

Erdoğan, M. M. Türkiye'deki Suriyeliler: Toplumsal kabul ve uyum. İstanbul Bilgi Üniversitesi Yayınları. (2015).

Erkan, E. Suriyeli Göçmenler ve Dini Hayat: Uyum, Karşılaşma, Benzeşme Gaziantep Örneği. İlahiyat Akademi Dergisi (Gaziantep Üniversitesi İlahiyat Fakültesi), 3(4), 1-35.

Esses, V. M., Dovidio, J. F., Jackson, L. M., & Armstrong, T. L.: The immigration dilemma: The role of perceived group competition, ethnic prejudice, and national identity. Journal of Social issues., 57(3), 389-412 (2001).

Fisk, Kerstin: Refugee geography and the diffusion of armed conflict in Africa. Civil Wars. 16, 3, 255–275 (2014).

Gadarian, Shana Kushner and Bethany Albertson: Anxiety, immigration, and the search for information. Political Psychology. 35, 2, 133–164 (2014).

Getmansky, Anna, Tolga Sinmazdemir and Thomas Zeitzoff: Refugee Influxes, Xenophobia, and Domestic Conflict: Evidence from a Survey Experiment in Turkey. forthcoming Journal of Peace Research.

Gundogdu D., O. D. Incel, A. A. Salah, and B. Lepri: Countrywide arrhythmia: emergency event detection using mobile phone data. EPJ Data Science. 5, 1, 25 (2016).

Hagberg, Aric A., Schult, Daniel A. and Pieter J. Swart (2008) Exploring network structure, dynamics, and function using NetworkX. Proceedings of the 7th Python in Science Conference (SciPy2008). Gael Varoquaux, Travis Vaught, and Jarrod Millman (Eds) 11–15.

Hainmueller, Jens, and Dominik Hangartner: Who Gets a Swiss Passport? A Natural Experiment in Immigrant Discrimination. American Political Science Review. 107, 1, 159-187 (2013).

Hainmueller, Jens and Daniel J Hopkins: Public attitudes toward immigration. Annual Review of Political Science. 17, 1, 225–249 (2014).

Hainmueller, Jens, Dominik Hangartner, and Giuseppe Pietrantuono: Naturalization fosters the long-term political integration of immigrants. PNAS. 112, 41, 12651-12656 (2015).

Hainmueller, Jens, Dominik Hangartner and Giuseppe Pietrantuono: Catalyst or Crown: Does Naturalization Promote the Long-Term Social Integration of Immigrants? American Political Science Review. 111, 2, 256-276 (2017).

Hagen-Zanker, Jessica, Martina Ulrichs, Rebecca Holmes, and Zina Nimeh: Cash Transfers for Refugees. Overseas Development Institute (2017).

Hunter, J. D.: Matplotlib: A 2D graphics environment. Computing In Science & Engineering. 9, 3, 90-95 (2007).

IHH İnsani Yardım Vakfı: Suriye Faaliyet Raporu (2012-2018). (2018).

Karreth, J., S. P. Singh and S. M Stojek. Explaining attitudes toward immigration: the role of regional context and individual predispositions. West European Politics, 38(6), 1174-1202. (2015).

Kitschelt, Herbert.: 1997. The Radical Right in Western Europe: A Comparative Analysis. Ann Arbor: University of Michigan Press, Ann Arbor (1997).

Konda Barometer: "Perceptions toward Syrian Refugees". Available at: http://konda.com. (2016).

Kuran, T. The provision of public goods under Islamic law: Origins, impact, and limitations of the waqf system. Law and Society Review, 841-898. (2001).

Lazarev, Egor and Kunaal Sharma:Brother or burden: An experiment on reducing prejudice toward Syrian Getmansky et al. 15 refugees in Turkey. Political Science Research and Methods. 5, 2, 201–219 (2017).

Lima A., M. De Domenico, V. Pejovic, and M. Musolesi: Disease containment strategies based on mobility and information dissemination. Scientific reports. 5, p. 10650 (2015.).

Mandic, Danilo: Trafficking and Syrian refugee smuggling: evidence from the balkan route. Social Inclusion. 5, 2, 28-38. doi:10.17645/si.v5i2.917 (2017).

Mari L., M. Gatto, M. Ciddio, E. D. Dia, S. H. Sokolow, G. A. De Leo, and R. Casagrandi: Big-data-driven modeling unveils country-wide drivers of endemic schistosomiasis. Scientific Reports. 7, 1, 489 (2017).

Martinez-Cesena E. A., P. Mancarella, M. Ndiaye, and M. Schlapfer: Using mobile phone data for electricity infrastructure planning. arXiv preprint arXiv:1504.03899 (2015).

McKinney, Wes: Data Structures for Statistical Computing in Python. Proceedings of the 9th Python in Science Conference. 51-56 (2010).

Oliphant, Travis E.: A guide to NumPy. Trelgol Publishing (2006).

Orhan, Oytun and Sabiha Senyucel Gundogar: Effects of the Syrian Refugees on Turkey. Center for Middle Eastern Strategic Studies (ORSAM), Ankara. (http://www.orsam.org.tr/files/ Raporlar/rapor195/195eng.pdf) (2015).

Paluck, E. L., & Green, D. P.: Prejudice reduction: What works? A review and assessment of research and practice. Annual review of psychology., 60, 339-367 (2009).

Pérez, Fernando and Brian E. Granger: IPython: A System for Interactive Scientific Computing. Computing in Science & Engineering. 9, 21-29 (2007).

Pettigrew, T. F., & Tropp, L. R.: A meta-analytic test of intergroup contact theory. Journal of personality and social psychology., 90(5), 751 (2006).

Pokhriyal N. and D. C. Jacques: Combining disparate data sources for improved poverty prediction and mapping. Proceedings of the National Academy of Sciences. 114, 46, E9783 (2017).

Refugee Connectivity: A Survey of Mobile Phones, Mental Health, and Privacy at a Syrian Refugee Camp in Greece. Harvard Humanitarian Initiative [WWW Document], n.d. URL /publications/refugee-connectivity-survey-mobile-phones-mental-health-and-privacy-syrian-refugee-camp (accessed 9.4.18).

Salah, A.A., Pentland, A., Lepri, B., Letouzé, E., Vinck, P., de Montjoye, Y.A., Dong, X. and Dagdelen, Ö.: Data for Refugees: The D4R Challenge on Mobility of Syrian Refugees in Turkey. arXiv preprint arXiv:1807.00523 (2018).

Scheve, Kenneth F and Matthew J Slaughter: Labor market competition and individual preferences over immigration policy. Review of Economics and Statistics. 83, 1, 133–145 (2001).

Schweitzer, Robert; Shelley Perkoulidis, Sandra Krome, Christopher Ludlow and Melanie Ryan: Attitudes towards refugees: The dark side of prejudice in Australia. Australian Journal of Psychology. 57, 3, 170–179 (2005).

Shaver, Andrew and Yang-Yang Zhou: Questioning refugee camps as sources of conflict. Working paper (https://scholar-dev.princeton.edu/sites/default/files/ISADraft_0. pdf) (2015).

Sniderman, Paul M., Louk Hagendoorn and Markus Prior: Predisposing Factors and Situational Triggers: Exclusionary Reactions to Immigrant Minorities. American Political Science Review. 98, 1, 35-49 (2004).

Steinmayr, Andreas: Exposure to Refugees and Voting for the Far-Right: (Unexpected) Results from Austria. Working Paper 9790 IZA (2016).

Trestian R., P. Shah, H. Nguyen, Q.-T. Vien, O. Gemikonakli, and B. Barn: Towards connecting people, locations and real-world events in a cellular network. Telematics and Informatics. 34, 1, 244 (2017).

Tomaszewski, B: Geographic information systems (GIS) for disaster management. CRC Press (2014).

Tompkins A. M. and N. McCreesh: Migration statistics relevant for malaria transmission in Senegal derived from mobile phone data and used in an agent-based migration model. Geospatial health. 11, 1s (2016).

Türkiye Diyanet Vakfı: Suriye Raporu. (2014).

UNHCR: Syrian Refugee Arrivals in Greece. (2015)

van Rossum, G.: Python tutorial. Centrum voor Wiskunde en Informatica (CWI). Technical Report CS-R9526 (1995).

Veness, C.: Calculate distance and bearing between two latitude/longitude points using haversine formula in javascript. URL: http://www.movable-type.co.uk/scripts/latlong.html (accessed 6.29.2018) (2010).

Waskom, Michael; Olga Botvinnik, Paul Hobson, John B. Cole, Yaroslav Halchenko, Stephan Hoyer, Dan Allan, et al.: seaborn: v0.5.0 (Version v0.5.0). Zenodo. (http://doi.org/10.5281/zenodo.12710) (2014).

APPENDIX

A Antenna-Level CDRs

A.1 Refugee Concentration

Table 3: Provinces with High Syrian Refugee Concentration

	Province	Syrian Refugees outside Camps	Syrian Refugees in Camps	Total
1	İstanbul	483490	0	483490
2	Şanlıurfa	319522	104809	424331
3	Hatay	369898	18374	388272
4	Gaziantep	293531	37880	331411
5	Adana	159214	555	159769
6	Mersin	149563	0	149563
7	Kilis	92017	33651	125668
8	Bursa	110889	0	110889
9	İzmir	110656	0	110656
10	Mardin	91909	2919	94828
11	Kahramanmaraş	73819	18359	92178
12	Ankara	76130	0	76130
13	Konya	75185	0	75185
14	Kayseri	60342	0	60342
15	Osmaniye	34625	10480	45105
16	Kocaeli	33375	0	33375
17	Diyarbakır	30195	0	30195
18	Adıyaman	16974	9532	26506
19	Malatya	12195	10077	22272
20	Batman	20010	0	20010

Source: Republic of Turkey Ministry of Interior, 2017

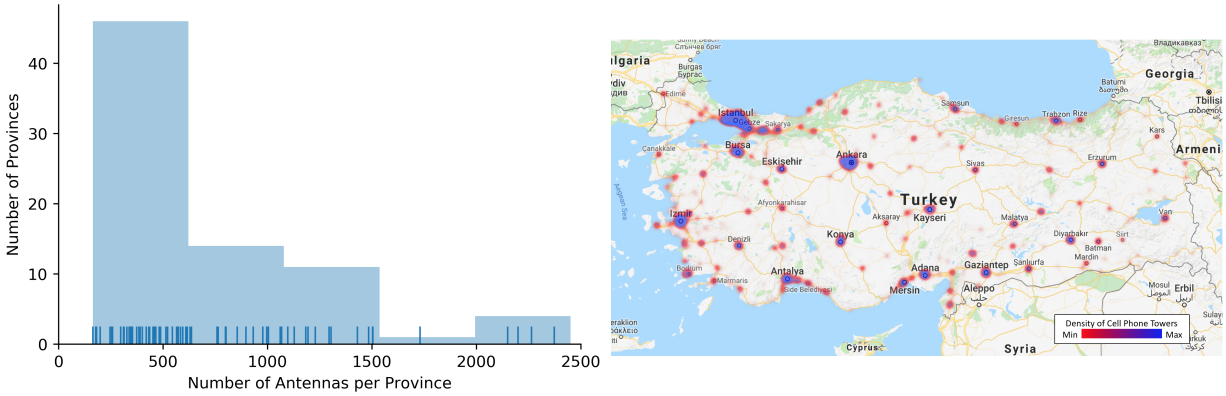


Figure 8: Geographic Spread of Antennas

A.2 Camp, Coast, and Border Points

Table 4: Camp Locations and Assigned Camp Antennas

	Province	District	Camp	Latitude Longitude	Area (sq. m)	Population (07/17)	Nr of Assigned Antennas
1	Hatay	Altınözü	Altınözü	36.184620, 36.364610	115515	8085	3
2	Hatay	Yayladağı	Yayladağı	35.906908, 36.055004	35513	3785	2
3	Hatay	Antakya	Apaydın	36.235153, 36.351624	112436	5054	2
4	Hatay	Yayladağı	Güveççi	35.892969, 36.171710	70733	1498	1
5	Gaziantep	İslahiye	İslahiye	36.983674, 36.620043	134890	17368	6
6	Gaziantep	Karkamış	Karkamış	36.875717, 38.028828	161018	6270	1
7	Gaziantep	Nizip	Nizip 1	37.047227, 37.895094	435805	9750	5
8	Gaziantep	Nizip	Nizip 2	37.043118, 37.904091	435805	4322	5
9	Şanlıurfa	Ceylanpınar	Ceylanpınar	36.814825, 39.923316	601476	21455	8
10	Şanlıurfa	Akçakale	Akçakale	36.753323, 38.953342	653064	29370	10
11	Şanlıurfa	Harran	Harran	36.873547, 38.931628	291803	13536	1
12	Şanlıurfa	Suruç	Suruç	37.042836, 38.483849	1000000	25310	9
13	Kilis	Merkez	Öncüpınar	36.645826, 37.083076	412400	13818	5
14	Kilis	Elbeyli	Elbeyli Beşiriye	36.662121, 37.361855	394500	19402	7
15	Mardin	Midyat	Midyat	37.411587, 41.391273	210000	2857	1
16	Kahramanmaraş	Dulkadiroglu	Merkez	37.442186, 37.016455	476055	18352	7
17	Osmaniye	Merkez	Cevdetiye	37.135980, 36.206350	200270	12706	5
18	Adıyaman	Merkez	Merkez	37.546573, 38.229345	508615	9464	4
19	Adana	Sarıçam	Sarıçam	37.042786, 35.490087	757011	555	1
20	Malatya	Battalgazi	Beydağı	38.341454, 38.154092	250000	10035	4

Sources: AFAD (2017); Republic of Turkey Ministry of Interior (2017)

Figure 9: Refugee Camp Images from Google Maps. Left to right, from top: Adana, Saricam; Adiyaman, Merkez; Gaziantep, Islahiye; Gaziantep, Karakamis; Gaziantep, Nizip 1-2; Hatay, Altinozu; Hatay, Apaydin; Hatay, Guvecci; Hayat, Yayladagi

Figure 10: Refugee Camp Images from Google Maps. Left to right, from top: Kahramanmaras, Merkez; Kilis, Elbeyli Besiriye; Kilis, Oncupinar; Malatya, Beydagi; Mardin, Midyat; Osmaniye, Cevdetiye; Sanliurfa, Akcakale; Sanliurfa, Ceylanpinar; Sanliurfa, Suruc; Sanliurfa, Harran

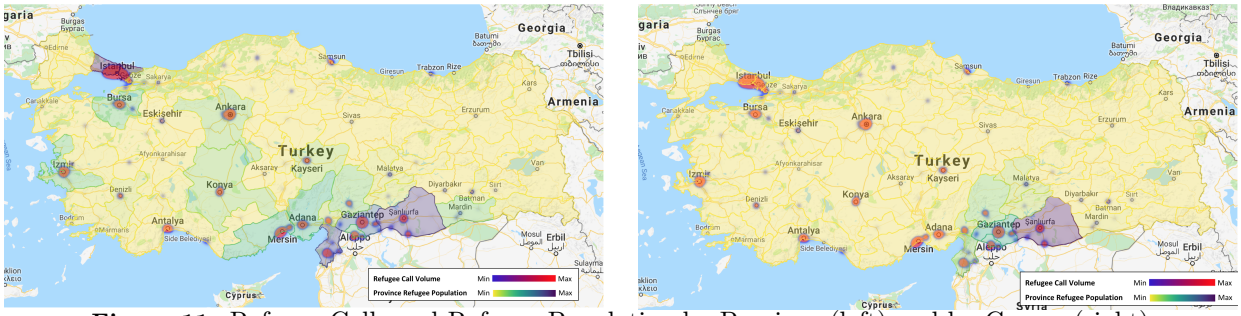


Figure 11: Refugee Calls and Refugee Population by Province (left) and by Camps (right)

Table 5: Departure Points to Greek Islands and Assigned Antennas

	Village	District	Latitude Longitude	Nr of Assigned Antennas
1	Hisar	Didim	37.353640, 27.194439	2
2	Güzelçamlı	Kuşadası	37.689677, 27.030615	1
3	Yavansu	Kuşadası	37.826442, 27.242757	1
4	Küçükköy	Ayvalık	39.273557, 26.609972	1
5	Koyun Evi	Ayvacık	39.459095, 26.178191	1
6	Şehit Mehmet	Çeşme	38.290036, 26.232871	1
7	Salman	Karaburun	38.573686, 26.360541	2
8	Bademli	Dikili	39.023625, 26.795487	3
9	Atatürk	Seferihisar	38.029160, 26.866164	3
10	Akyarlar	Bodrum	36.957397, 27.286430	1
11	Cumali	Datça	36.748807, 27.437015	1
12	Gümüşlü	Bodrum	37.063727, 27.231726	3

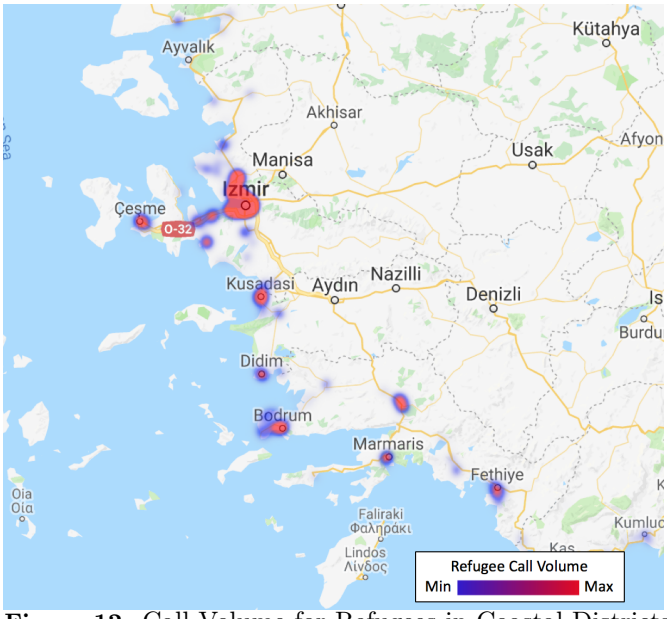


Figure 12: Call Volume for Refugees in Coastal Districts

Table 6: Border Crossing Points to Syria and Assigned Antennas

	Crossing Point	Nearest District or Town	Latitude Longitude	Nr Antennas Assigned	Status
1	Aşşağıpulluyazı-Ein al-Bayda	Aşşağıpulluyazı	35.861159, 36.17472	2	Restricted
2	Güveççi-Kherbet Eljoz Bridge	Güveççi	35.888083, 36.171253	2	Restricted
3	Cilvegözü-Bab al-Hawa	Reyhanli	36.233275, 36.680922	9	Open
4	Bükülmez-Atmeh	Bükülmez	36.309626, 36.657794	3	Restricted
5	Öncüpınar-Bab al-Salam	Kilis	36.634401, 37.085608	6	Open
6	Karkamış-Jarabulus	Karkamış	36.831056, 37.996371	9	Open

Table 7: Antennas Assigned to Both Border Crossing Points and Refugee Camps

	BTS ID	Latitude, Longitude	Province	District	Camp Name	Crossing Point
1	5228841	36.645826, 37.083076	Kilis	Merkez	Öncüpınar Konteynerkenti	Öncüpınar-Bab al-Salam
2	5228848	36.645826, 37.083076	Kilis	Merkez	Öncüpınar Konteynerkenti	Öncüpınar-Bab al-Salam

A.3 Sectarian and Friday Prayer Activity

To predict Sunni refugee concentration, we do the same analysis for Mawlid-al-Nabawi. Among the approximately 30,000 antennas for which we have complete data, we find that in 2,306 antennas there is more refugee call traffic during Ashura, and in 1727 antennas during Mawlid-al-Nabawi (p-value < 0.1). Figure 14 shows antenna locations by estimated sect for Istanbul, and Figure 15 shows the same for Şanlıurfa district. The total call volume on these holidays as compared to the corresponding regular day (a regular Friday, in this case) is presented in Figure 13. The percentage of antennas within the district with a considerably higher amount of Ashura (for Shi'a) and Mawlid (for Sunni) call volume is calculated for all districts and the top 10 districts of each category are presented in Table 9. We also use the data on SMS volumes to derive the same measures for Friday prayer activity areas and check the correlation of the obtained dummy indicators with the original measures. We find a correlation of 0.11.

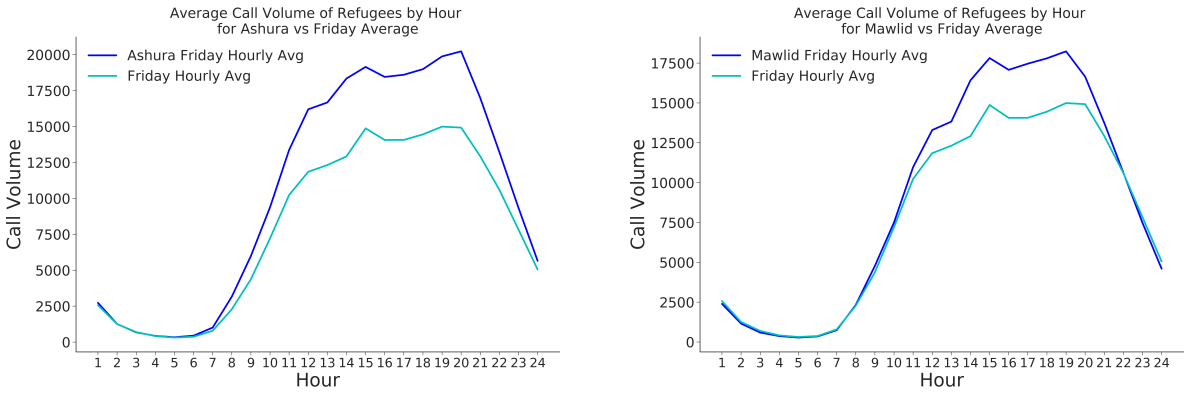


Figure 13: Volume of Refugee Calls on Syrian Religious Holidays

Table 8: Summary Table: Holiday T-test Results

Value	Comparison	Effect	Nr of Antennas	Significance Level
Mawlid Call Volume	Refugee vs Turk	Higher for Refugees	1,727	10%
Ashura Call Volume	Refugee vs Turk	Higher for Refugees	2,306	10%
Friday Prayer Activity, 11 am - 12 pm	Refugee vs Turk	Higher for Refugees	500	10%

We also look at the concentration of Friday prayer activity.In 500 of around 27,000 antennas for which we have complete data, refugee calls are significantly higher between 11am-12pm, at the 10% significance level. The correlation coefficient for the number of Friday prayer antennas and number of refugees per district across all districts is 0.34 based on June 2017 refugee population statistics. Table 28 shows the number of Friday prayer antennas and approximate refugee population by district for the top 20 districts by number of Friday antennas.

Table 9: Percentage of Antennas with a Higher Call Volume during Ashura, Mawlid, or Friday Prayer Time

Sunni Concentration	Pct.	Shia Concentration	Pct.	Friday Prayer Concentration	Pct.
Afyonkarahisar Başmakçı	100.00	Çorum Uğurludağ	100.00	Bursa Harmancık	100.00
Balıkesir Havran	100.00	Konya Çeltik	100.00	Kilis Polateli	50.00
Uşak Karahallı	100.00	Şanlıurfa Harran	66.67	Burdur Yeşilova	33.33
Burdur Ağlasun	50.00	Denizli Beyağaç	50.00	Şanlıurfa Harran	33.33
Muğla Kavaklıdere	50.00	Denizli Kale	50.00	Trabzon Düzköy	33.33
Sakarya Taraklı	50.00	Giresun Piraziz	50.00	Kilis Musabeyli	25.00
Gaziantep Karkamış	44.44	Kastamonu Devrekani	50.00	Şanlıurfa Akçakale	21.43
Afyonkarahisar Çay	33.33	Hatay Antakya	46.21	Kırklareli Vize	20.00
Bolu Mengen	33.33	Gaziantep Karkamış	44.44	Konya Altınekin	20.00
Çanakkale Bozcaada	33.33	Hatay Altınözü	43.75	Mersin Mut	20.00

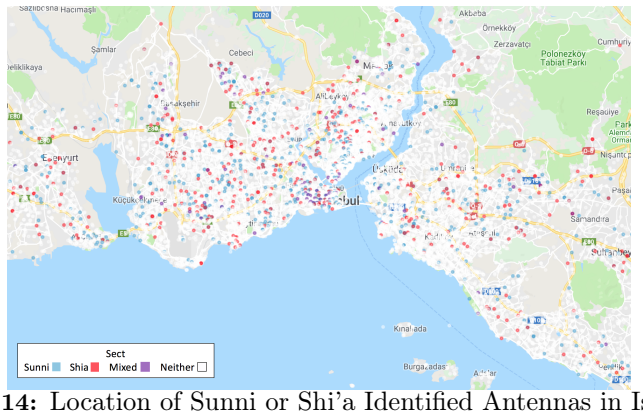


Figure 14: Location of Sunni or Shi'a Identified Antennas in Istanbul

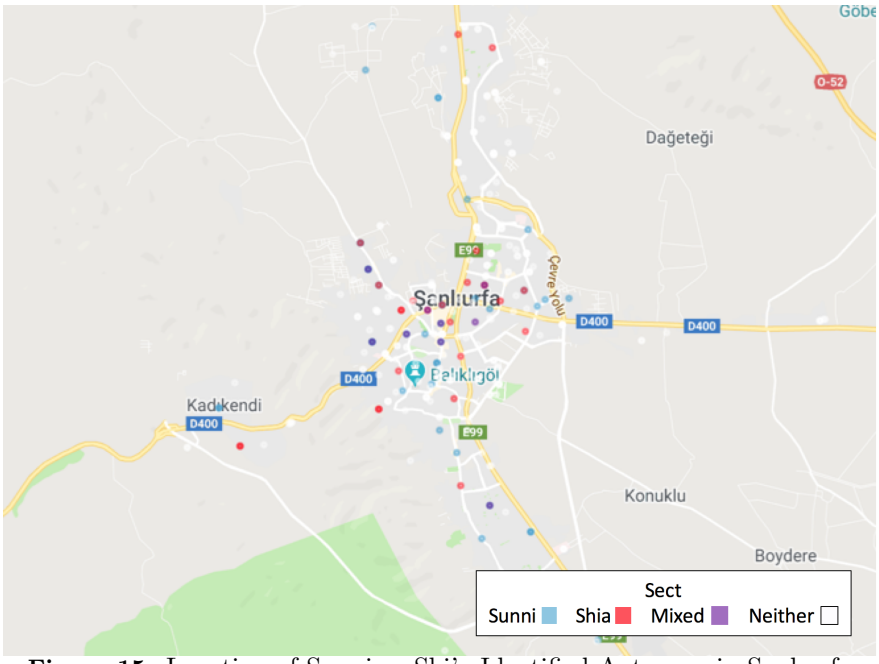


Figure 15: Location of Sunni or Shi'a Identified Antennas in Şanlıurfa

A.4 Refugee Call and SMS Volumes by Population

Table 10: Daily Refugee Call Volumes and Refugee Population by Province

	Dependent variable: Avg Daily Nr of Calls	
	All Refugee Population	Camp Population
Refugee Population	0.063***	0.037***
	(0.007)	(0.008)
Constant	−81.667	−77.537
	(623.921)	(121.372)
Observations	81	19
R^2	0.541	0.536
Adjusted R^2	0.535	0.508
Residual Std. Error	5,212.486 (df = 79)	287.8 (df = 17)
F Statistic	92.985*** (df = 1; 79)	19.6 *** (df = 1; 17)

Source: Republic of Turkey Ministry of Interior (July 2017)

*p<0.1; **p<0.05; ***p<0.01

Table 11: Daily Refugee SMS Volumes and Refugee Population by Province

	Dependent variable: Avg Daily Nr of SMS	
	All Refugee Population	Camp Population
Refugee Population	0.015***	0.016***
	(0.001)	(0.004)
Constant	116.361	−34.688
	(132.220)	(65.229)
Observations	81	19
R^2	0.615	0.440
Adjusted R^2	0.610	0.407
Residual Std. Error	1,102.931 (df = 79)	154.652 (df = 17)
F Statistic	125.974*** (df = 1; 79)	13.354*** (df = 1; 17)

Source: Republic of Turkey Ministry of Interior (July 2017)

*p<0.1; **p<0.05; ***p<0.01

Table 12: Daily Call Volume (Avg per Antenna) and Refugee Flows at Focal Coastal Departure Points

	Dependent variable:
Arrivals	0.002***
	(0.000)
Constant	0.932***
	(0.033)
Island Fixed Effect	✓

Note:Least trimmed squares regression with an α of 0.9

*p<0.1; **p<0.05; ***p<0.01

Table 13: Daily SMS Volume (per Antenna) and Refugee Flows at Focal Coastal Departure Points

	Dependent variable:
SMS	−0.054
	(0.361)
Constant	16.778***
	(3.700)
Island Fixed Effect	✓
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 14: Summary Statistics by Province

	Daily Nr of Calls		Refugee Population	
	General	In Camps	General	In Camps
Min.	19	3	32	555
1st Qu.	150	118	572	5662
Median	296	192	2145	12710
Mean	2156	382	35560	12260
3rd Qu.	869	466	14340	17860
Max.	64990	1376	458600	29370

Source: Republic of Turkey Ministry of Interior (July 2017)

Table 15: SMS Summary Statistics by Province

	Daily Nr of SMS		Refugee Population	
	General	In Camps	General	In Camps
Min.	5	2	39	555
1st Qu.	50	17	732	5662
Median	113	52	2624	12706
Mean	674	165	37653	12263
3rd Qu.	412	279	14959	17860
Max.	12814	547	485227	29370

Source: Republic of Turkey Ministry of Interior (July 2017)

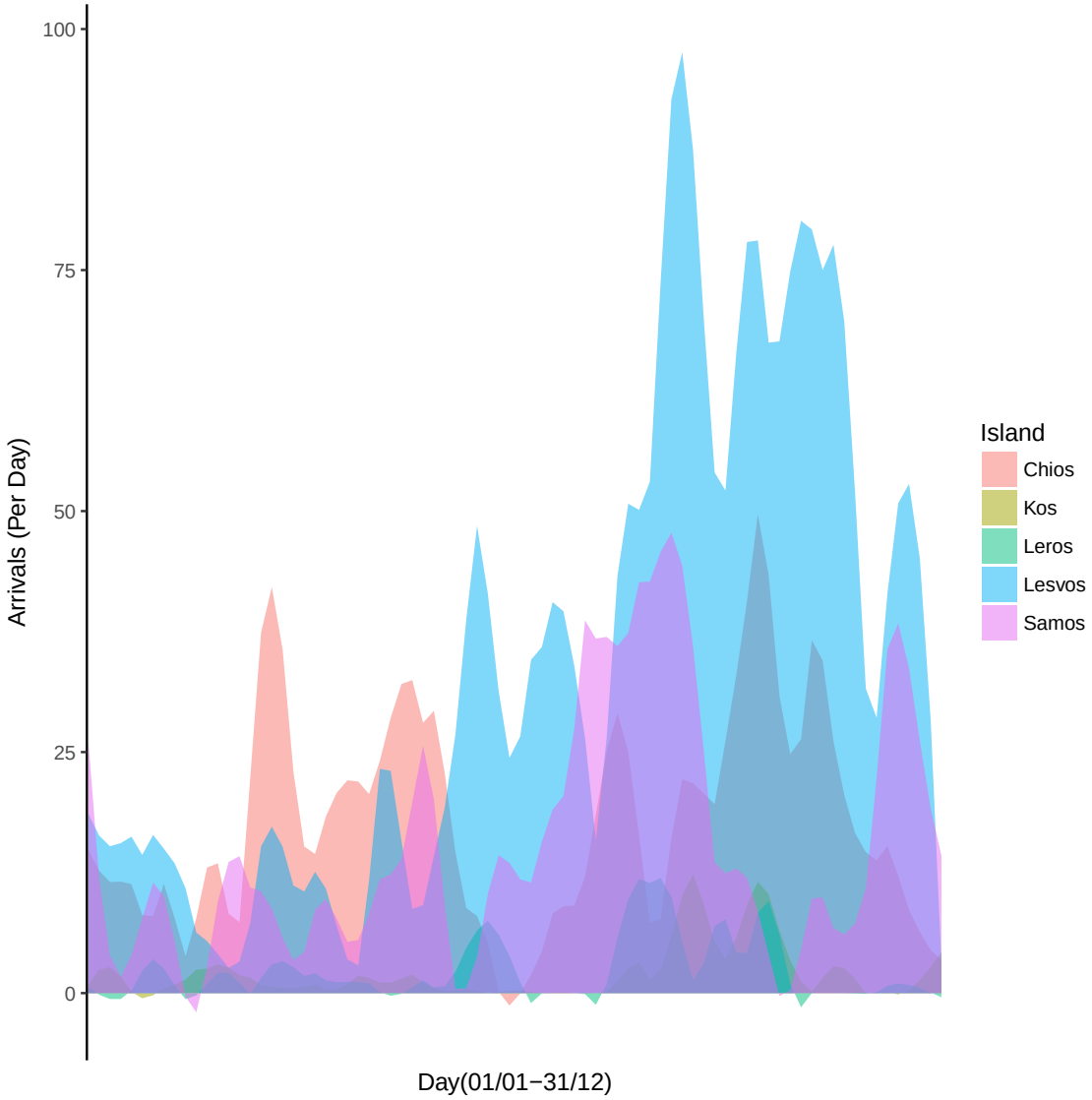


Figure 16: Daily Refugee Flows from Focal Coastal Departure Points.

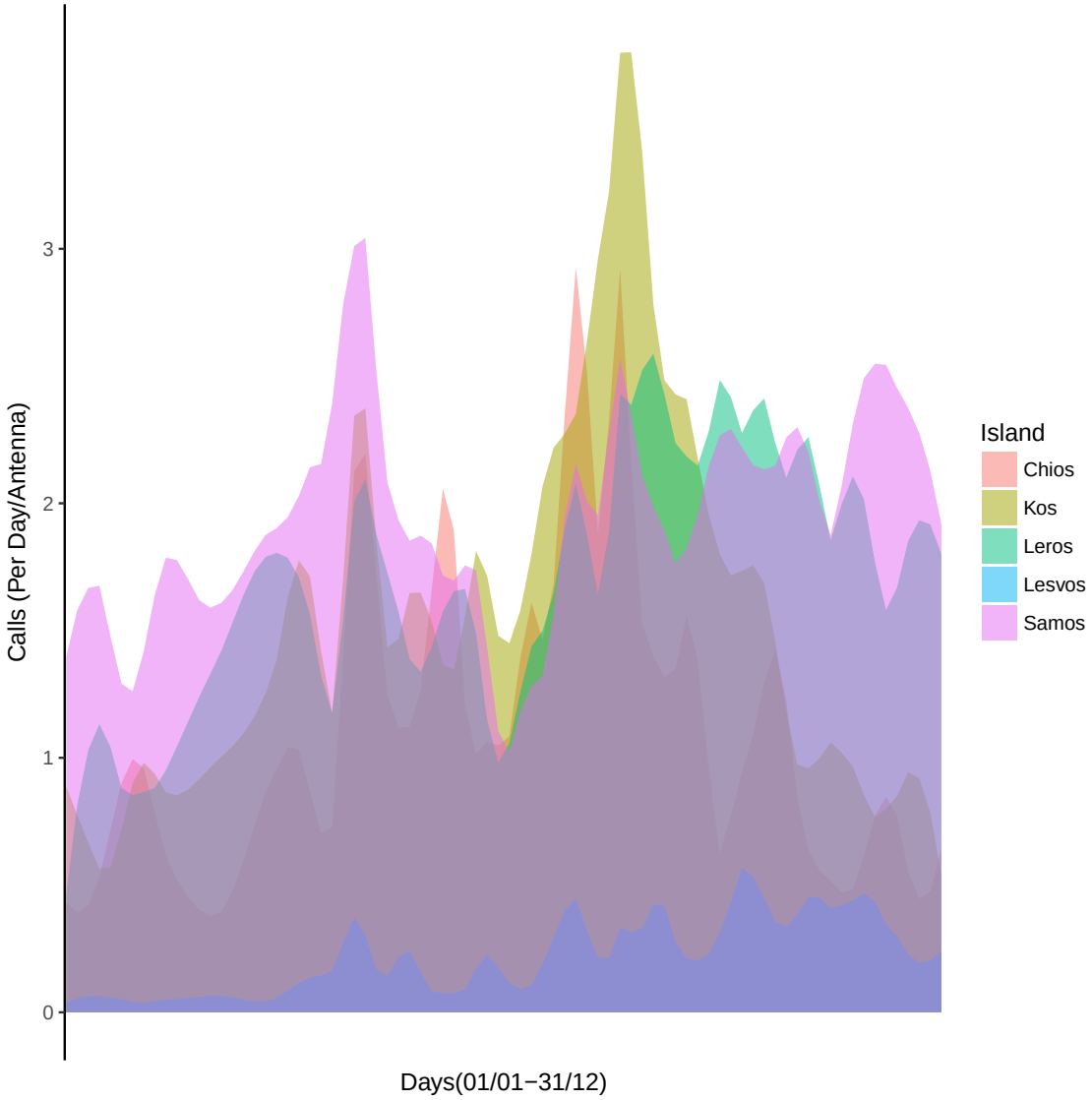


Figure 17: Daily Call Volume (per Antenna) at Focal Coastal Departure Points.

A.5 Calling Behavior on Religious Holidays

We also look at the average daily calling patterns of Syrians and Turks during religious holidays such as the holy month of Ramadan (26 May -24 June 2017) which are strikingly similar for Turks and Syrians. The average Ramadan weekday does not seem to deviate much from a regular weekday with the main difference being a steep drop in the breaking of the fast which happens at sunset. This sharp decline in calls, which pick up again between 8 pm - 9 pm before they drop again, is also present during Ramadan Fridays and weekends. Ramadan Fridays show a somewhat higher call volume among Ramadan days for both refugees and Turks (unlike a regular Friday that is higher than a regular weekend but about the same as a regular weekday), with a distinct decrease in slope in calls during Friday prayer and a dip during the breaking of the fast. Comparing the two groups, we see decreased calling on Ramadan weekends for Turks, as compared to regular weekends, presumably as people opt to sleep longer during fasting hours (see Figure 18).

We test to see if any of the visual differences are also statistically significant using a paired t-test (Appendix Table 18). Similarly, the observations used in the t-tests in this part are the average

number of calls/SMS per hour, averaged over all antennas and over the whole year. Both groups tend to make fewer calls on weekends than on weekdays and Fridays. Turkish calling behavior differs from that of Syrians during Ramadan, as weekdays and Fridays become statistically distinguishable from each other for Turks, with more calls made on Friday than a regular weekday, while they show no statistically different patterns for refugees. This finding might indicate a potential Friday socialization pattern for Turks during Ramadan that is not there for refugees. The patterns are similar in the paired t-tests that employ SMS volume, although, among refugees, the significance of across-day differences disappears (Appendix Table 24). This might be due to the already limited use of SMS.

There is also a statistically significant difference in calling behavior among refugees in camps versus those outside camps. While there is a general decrease in total refugee call volume during Ramadan (Appendix Table 20), this reduction appears to be stronger for the in-camp population, with lower call volume inside-camps than among refugees outside camps (Appendix Table 19). This general decrease in call volume during Ramadan also seems to have closed the gap between refugee calls at the Greek maritime coast and the rest of the country, possibly because refugee flows toward Europe appear to not deviate much during Ramadan. A final interesting finding that shows significant deviation from non-Ramadan months is that during Ramadan, the calls at border crossing points tend to be lower than the calls in non-border areas on weekends, pointing to a decrease in border crossing traffic during the holy month. Ramadan calling patterns generally overlap with SMS volume patterns, with the exception of border crossings, where the call volume gets lower than the non-border call volume, while the border SMS volume tends to get higher (Appendix Table 25).

We also look at the distribution of calls on other important Muslim religious holidays such as Eid al Fitr, the occasion marking the end of Ramadan and Eid al Adha, the occasion marking the culmination of the annual pilgrimage to Mecca, celebrated by all Muslims irrespective of sect. There are, however, certain days that are primarily important for Sunnis or Shiites, such as Mawlid-al-Nabawi, the birthday of the Prophet Mohammed, for Sunnis, and Ashura, the commemoration of the death of Imam Hussein for Shiites.

We check to see if calling behavior is significantly different during these holidays by running paired t-tests at the antenna level, comparing the call volume of refugees against that of Turks. As reflected in Appendix Table 21, t-test results suggest that on all religious holidays but Eid al Fitr, refugees tend to call other people at higher rates compared to regular weekdays or weekends. The finding is similar for the refugee population in camps. On the other hand, during the two main religious holidays, Eid al Fitr and Eid al Adha, Turks make significantly fewer calls than normal, suggesting that cultural norms of communication on religious holidays are different for refugees than for Turks as reflected in their calling behavior. This could be because of differences in levels of religiosity between refugees and Turks; and/ or differences in the way religiosity and religious holidays are celebrated in the respective cultures. The analysis of SMS volume is less informative since SMS volume for Turks between religious days and regular days becomes statistically indistinguishable, while for Syrians we find mostly statistically significant mean differences between regular days and religious days, with the direction of the gap consistent with that of their call volumes (Appendix Table 27).

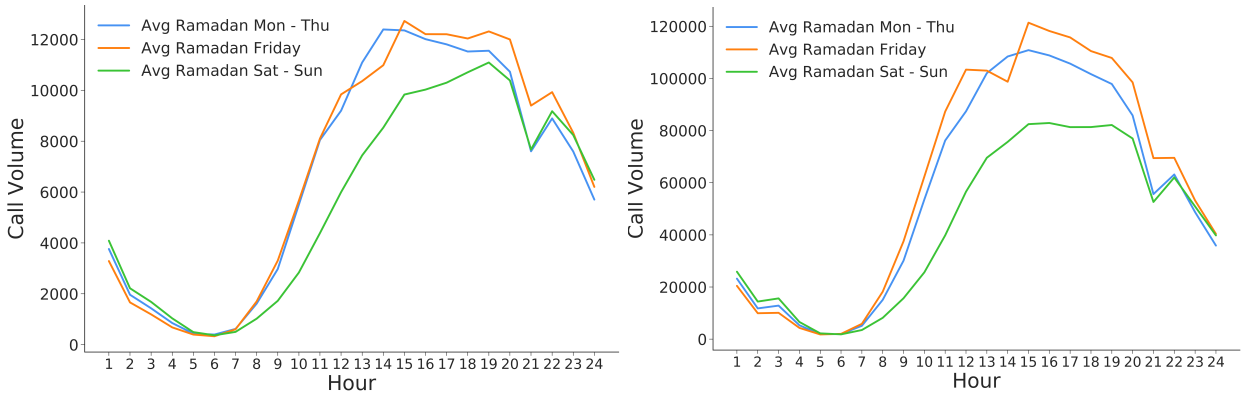


Figure 18: Call Volume for Refugees (left) and Turks (right) for Ramadan Weekday vs Weekend

A.6 Calling Behavior T-Tests

Table 16: Refugee and Turkish Call Patterns Across-Day Comparisons

		Refugees		Turks	
	day	pvalue	Tstat	pvalue	Tstat
1	Weekday-Friday	5.60e-01	-77.28	3.91e-01	-920.27
2	Weekday-Weekend	4.19e-04	1061.02	1.51e-04	10136.37
3	Friday-Weekend	1.02e-04	1138.30	1.68e-04	11056.64

Table 17

Refugee vs. Turkish

	day	pvalue	Tstat
1	Weekday	2.55e-05	-153.18
2	Friday	1.33e-05	-179.37
3	Weekend	4.44e-04	-42.36

Camp vs. Non-Camp

	day	pvalue	Tstat
1	Weekday	1.41e-05	115.75
2	Friday	1.12e-04	83.77
3	Weekend	1.58e-06	126.47

Inclusive Camp vs. Non-Camp

	day	pvalue	Tstat
1	Weekday	1.48e-06	125.62
2	Friday	3.32e-06	89.31
3	Weekend	1.16e-06	105.19

At Border vs. Not At Border

	day	pvalue	Tstat
1	Weekday	1.62e-05	137.43
2	Friday	8.00e-01	-3.70
3	Weekend	2.24e-06	337.56

Inclusive At Border vs. Not At Border

	day	pvalue	Tstat
1	Weekday	2.01e-06	104.36
2	Friday	2.16e-05	72.96
3	Weekend	2.80e-06	122.70

On Coast vs Not On Coast

	day	pvalue	Tstat
1	Weekday	9.36e-04	-196.23
2	Friday	1.81e-05	-248.19
3	Weekend	5.40e-04	-152.16

Inclusive On Coast vs Not On Coast

	day	pvalue	Tstat
1	Weekday	1.69e-05	-107.94
2	Friday	4.47e-05	-80.89
3	Weekend	1.87e-04	-54.94

Table 18: Refugee and Turkish Call Patterns Across-Day Comparisons during Ramadan

		Refugees		Turks	
	day	pvalue	Tstat	pvalue	Tstat
1	Weekday-Friday	1.11e-01	-224.86	8.05e-04	-5029.42
2	Weekday-Weekend	3.75e-03	992.24	3.14e-04	12311.73
3	Friday-Weekend	3.32e-04	1217.10	7.43e-05	17341.15

Table 19

Refugee vs. Turkish

	day	pvalue	Tstat
1	Weekday	2.62e-05	-65.06
2	Friday	2.44e-05	-89.99
3	Weekend	9.88e-05	-30.85

Camp vs. Non-Camp

	day	pvalue	Tstat
1	Weekday	4.26e-01	-10.82
2	Friday	1.59e-02	-40.26
3	Weekend	1.56e-02	-25.41

Inclusive Camp vs. Non-Camp

	day	pvalue	Tstat
1	Weekday	4.65e-02	-17.39
2	Friday	2.08e-02	-22.10
3	Weekend	9.52e-03	-16.23

At Border vs. Not At Border

	day	pvalue	Tstat
1	Weekday	5.21e-03	47.24
2	Friday	1.60e-05	501.31
3	Weekend	6.73e-03	-38.63

Inclusive At Border vs. Not At Border

	day	pvalue	Tstat
1	Weekday	7.62e-05	-52.71
2	Friday	1.93e-05	-68.14
3	Weekend	1.10e-03	-25.10

On Coast vs Not On Coast

	day	pvalue	Tstat
1	Weekday	8.76e-03	-23.40
2	Friday	4.24e-01	-11.10
3	Weekend	7.10e-01	9.79

Inclusive On Coast vs Not On Coast

	day	pvalue	Tstat
1	Weekday	4.66e-06	68.02
2	Friday	1.00e-04	56.39
3	Weekend	2.17e-05	64.70

Table 20: Ramadan versus Non-Ramadan Call Patterns of Refugees

	day	pvalue	Tstat
1	Weekday	1.20e-03	-1355.87
2	Friday	6.11e-04	-1208.29
3	Weekend	1.79e-03	-1287.10

Table 21: Refugee and Turkish Call Patterns on Religious Days Compared to Regular Days

		All Refugees		In Camp Ref.		Turks	
	day	pvalue	Tstat	pvalue	Tstat	pvalue	Tstat
1	Ashura	5.18e-06	2463.74	2.95e-05	97.29	8.05e-02	1719.96
2	Muharram	3.01e-06	2592.28	5.42e-06	141.37	3.34e-03	2900.96
3	Mawlid	1.44e-03	1047.11	2.62e-03	65.24	2.12e-02	-2267.75
4	EidAlFitr1	4.73e-02	-1844.66	3.31e-01	-53.74	1.53e-02	-17035.26
5	EidAlFitr2	2.06e-03	-4579.04	9.05e-03	-173.69	6.81e-03	-27730.30
6	EidAlFitr3	2.45e-04	-3955.40	1.80e-03	-136.33	3.96e-04	-30474.53
7	EidAlFitr4	9.40e-03	-2685.79	2.93e-03	-142.48	6.02e-03	-21400.81
8	EidAlAdha1	8.46e-06	2511.32	3.47e-06	176.85	3.13e-02	-5439.88
9	EidAlAdha2	1.10e-07	1741.99	2.00e-06	133.12	3.87e-01	-2635.92
10	EidAlAdha3	8.54e-06	1629.34	1.68e-05	130.60	1.36e-01	-1707.34
11	EidAlAdha4	1.51e-05	1423.92	2.01e-05	115.72	5.02e-03	-4030.22
12	EidAlAdha5	3.03e-01	334.07	5.82e-03	57.60	1.51e-04	-13190.84

A.7 SMS Behavior T-Tests

Table 22: Refugee and Turkish SMS Patterns Across-Day Comparisons

		Refugees		Turks	
	day	pvalue	Tstat	pvalue	Tstat
1	Weekday-Friday	1.19e-06	-258.18	6.98e-03	-11146.92
2	Weekday-Weekend	1.39e-01	98.28	4.08e-02	1686.53
3	Friday-Weekend	7.11e-04	356.46	5.65e-03	12833.45

Table 23

Refugee vs. Turkish

	day	pvalue	Tstat
1	Weekday	4.68e-03	7.68
2	Friday	1.47e-01	-21.63
3	Weekend	9.08e-05	15.29

Camp vs. Non-Camp

	day	pvalue	Tstat
1	Weekday	3.30e-02	-8.60
2	Friday	2.41e-07	-31.23
3	Weekend	1.83e-05	-16.65

Inclusive Camp vs. Non-Camp

	day	pvalue	Tstat
1	Weekday	6.52e-01	0.77
2	Friday	3.86e-05	-9.39
3	Weekend	2.54e-03	-6.32

At Border vs. Not At Border

	day	pvalue	Tstat
1	Weekday	9.64e-03	-20.36
2	Friday	1.24e-02	34.70
3	Weekend	5.09e-05	74.77

Inclusive At Border vs. Not At Border

	day	pvalue	Tstat
1	Weekday	4.21e-02	-6.20
2	Friday	8.70e-06	-22.06
3	Weekend	1.62e-03	-8.21

On Coast vs Not On Coast

	day	pvalue	Tstat
1	Weekday	6.48e-01	-11.69
2	Friday	1.24e-02	34.70
3	Weekend	1.17e-01	49.30

Inclusive On Coast vs Not On Coast

	day	pvalue	Tstat
1	Weekday	6.38e-01	1.85
2	Friday	7.21e-01	7.18
3	Weekend	3.16e-01	-9.20

Table 24: Refugee and Turkish SMS Patterns Across-Day Comparisons during Ramadan

		Refugees		Turks	
	day	pvalue	Tstat	pvalue	Tstat
1	Weekday-Friday	5.81e-01	54.51	3.01e-02	-6354.60
2	Weekday-Weekend	9.69e-01	4.01	6.08e-02	3063.45
3	Friday-Weekend	6.17e-01	-50.50	1.72e-02	9418.05

Table 25

Refugee vs. Turkish

	day	pvalue	Tstat
1	Weekday	1.20e-02	2.92
2	Friday	2.58e-02	-17.51
3	Weekend	3.45e-09	7.22

Camp vs. Non-Camp

	day	pvalue	Tstat
1	Weekday	1.66e-04	-14.99
2	Friday	3.92e-05	-13.21
3	Weekend	3.80e-04	-13.21

Inclusive Camp vs. Non-Camp

	day	pvalue	Tstat
1	Weekday	1.39e-04	-8.88
2	Friday	6.88e-01	0.66
3	Weekend	8.59e-03	-6.83

At Border vs. Not At Border

	day	pvalue	Tstat
1	Weekday	8.60e-01	2.75
2	Friday	8.34e-06	-49.09
3	Weekend	3.60e-04	281.36

Inclusive At Border vs. Not At Border

	day	pvalue	Tstat
1	Weekday	1.37e-01	-5.30
2	Friday	5.25e-02	8.50
3	Weekend	3.18e-02	-8.91

On Coast vs Not On Coast

	day	pvalue	Tstat
1	Weekday	8.76e-03	-23.40
2	Friday	4.24e-01	-11.10
3	Weekend	7.10e-01	9.79

Inclusive On Coast vs Not On Coast

	day	pvalue	Tstat
1	Weekday	4.66e-06	68.02
2	Friday	1.00e-04	56.39
3	Weekend	2.17e-05	64.70

Table 26: Ramadan versus Non-Ramadan SMS Patterns of Refugees

	day	pvalue	Tstat
1	Weekday	4.21e-04	-1078.64
2	Friday	1.11e-05	-1391.33
3	Weekend	1.15e-04	-984.37

Table 27: Refugee and Turkish Call Patterns on Religious Days Compared to Regular Days

		All Refugees		Turks	
	day	pvalue	Tstat	pvalue	Tstat
1	Ashura	2.79e-05	1362.27	1.66e-01	1279.53
2	Muharram	6.34e-05	5290.05	3.33e-01	938.95
3	Mawlud	9.92e-03	313.42	1.70e-04	-6073.97
4	EidAlFitr1	1.97e-04	-1950.72	5.01e-02	-8433.47
5	EidAlFitr2	5.51e-01	-427.23	3.15e-01	15479.34
6	EidAlFitr3	5.01e-08	-2008.09	3.11e-05	-12938.53
7	EidAlFitr4	2.60e-02	-563.01	3.25e-01	-2992.05
8	EidAlAdha1	4.85e-01	183.80	9.18e-01	252.64
9	EidAlAdha2	1.43e-03	1009.44	4.11e-02	17154.93
10	EidAlAdha3	5.62e-01	86.41	7.80e-01	-260.77
11	EidAlAdha4	3.49e-02	648.19	5.26e-01	980.11
12	EidAlAdha5	8.67e-02	934.58	7.97e-01	769.41

Table 28: Friday Prayer Antennas and Refugee Population by District for Top 20 Districts

District	Province	Nr Friday Prayer Antennas	District Population (approx)
Cankaya	Ankara	13.0	1342.0
Uskudar	Istanbul	12.0	2270.3
Yenimahalle	Ankara	12.0	5186.9
Sisli	Istanbul	10.0	9720.8
Kepez	Antalya	10.0	205.4
Osmangazi	Bursa	10.0	34370.3
Selcuklu	Konya	9.0	22071.7
Umraniye	Istanbul	9.0	14437.0
Fatih	Istanbul	8.0	29214.8
Buca	Izmir	8.0	11136.7
Esenyurt	Istanbul	8.0	40207.5
Inegol	Bursa	8.0	11025.4
Bakirkoy	Istanbul	8.0	2689.9
Bornova	Izmir	7.0	21649.7
Basaksehir	Istanbul	7.0	25009.0
Haliliye	Sanliurfa	6.0	101608.1
Seyhan	Adana	6.0	77294.1
Nilufer	Bursa	6.0	3270.3
Yildirim	Bursa	6.0	48645.3
Besiktas	Istanbul	5.0	106.9

B Individual-Level CDRs

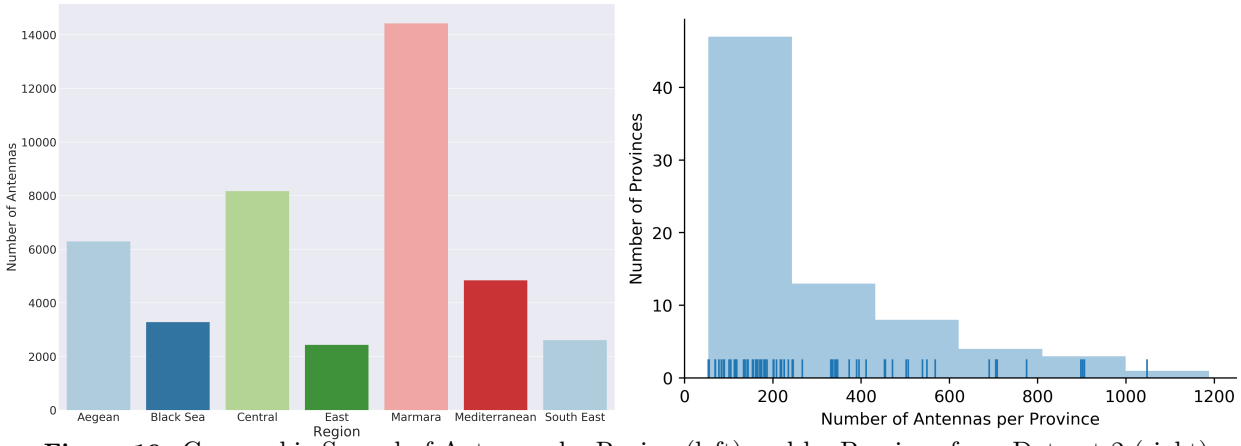


Figure 19: Geographic Spread of Antennas by Region (left) and by Province from Dataset 2 (right)

B.1 Individual-Level Tests - Trends Over Time

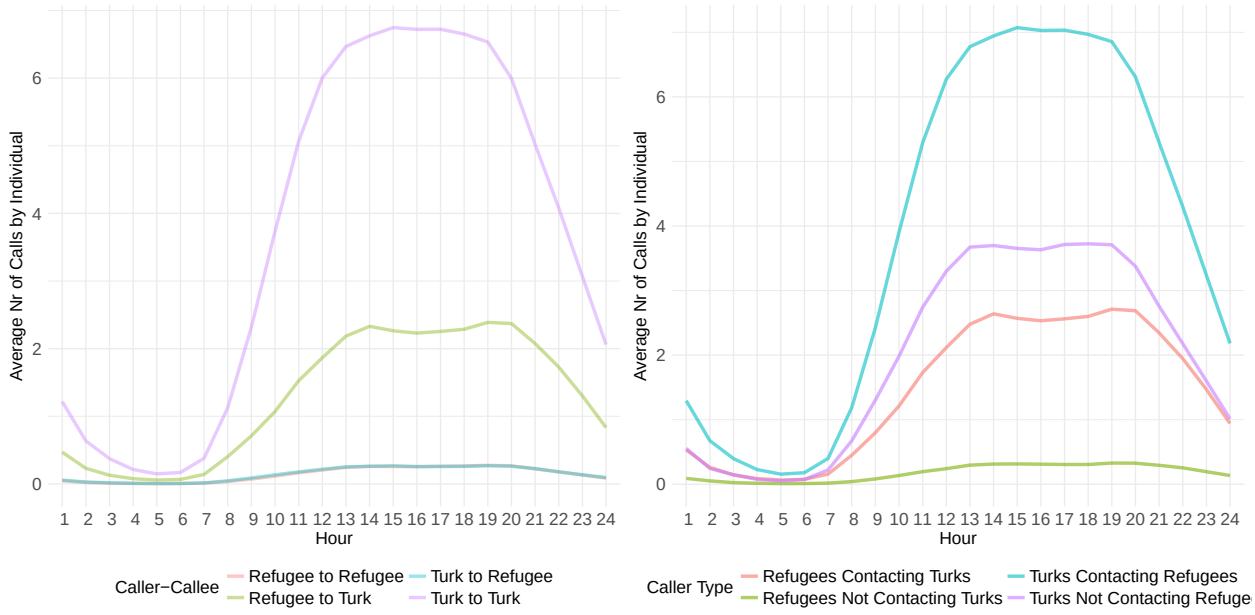


Figure 20: Hourly Call Patterns by Call Type (left) and by Caller Type (right)

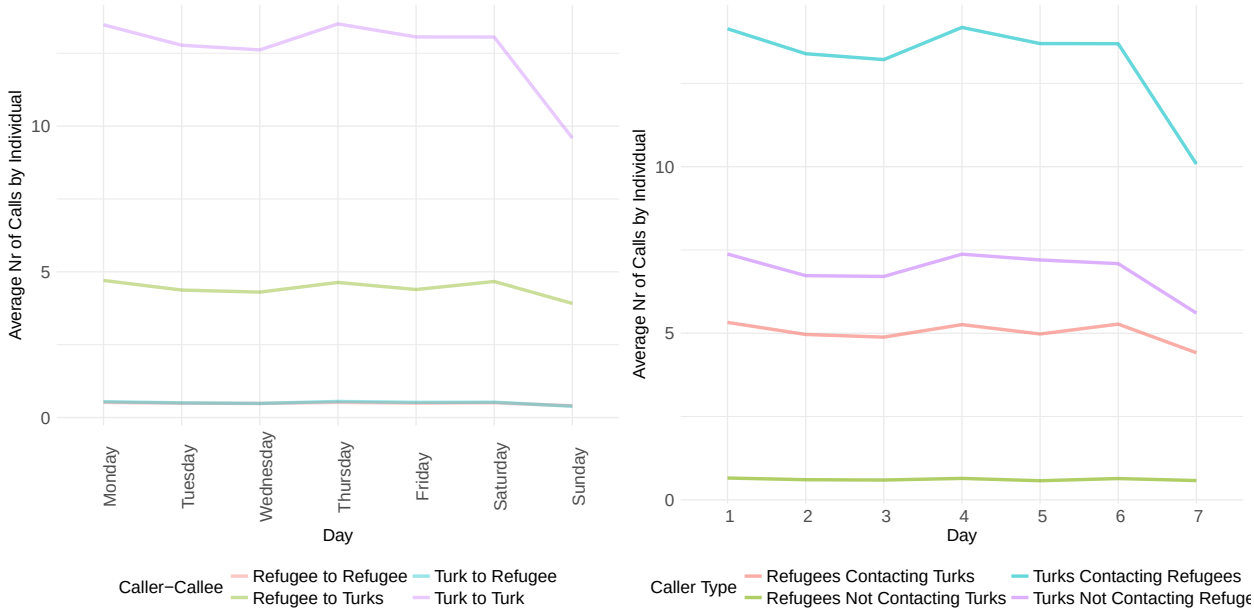


Figure 21: Weekly Call Patterns by Call Type (left) and by Caller Type (right)

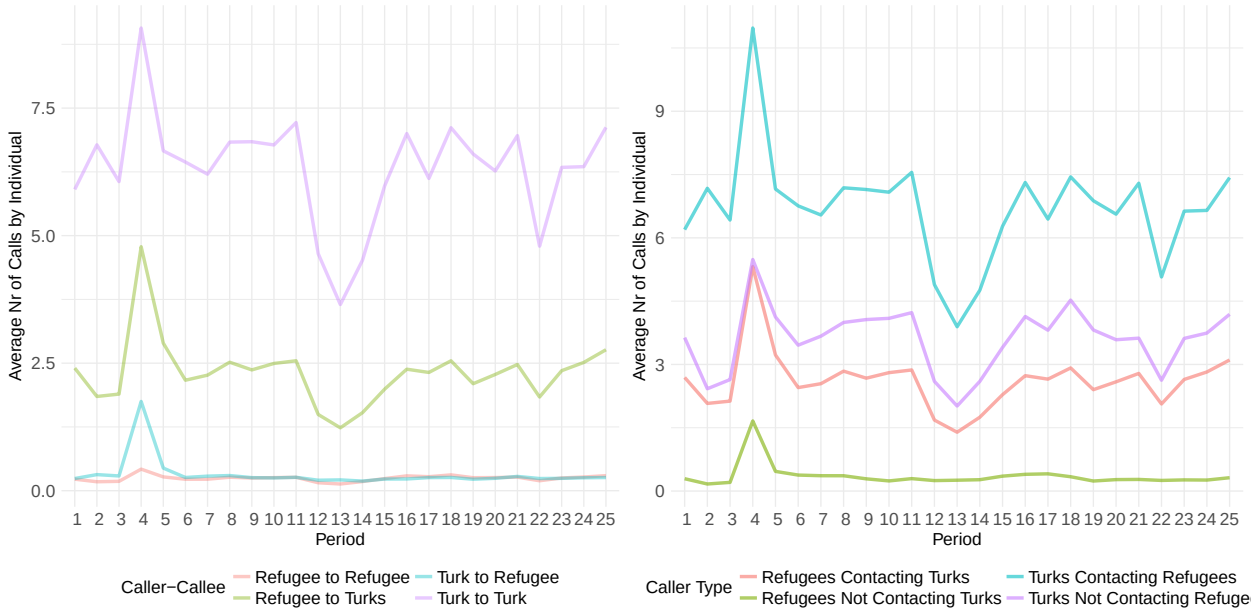


Figure 22: Biweekly Call Patterns by Call Type (left) and by Caller Type (right)

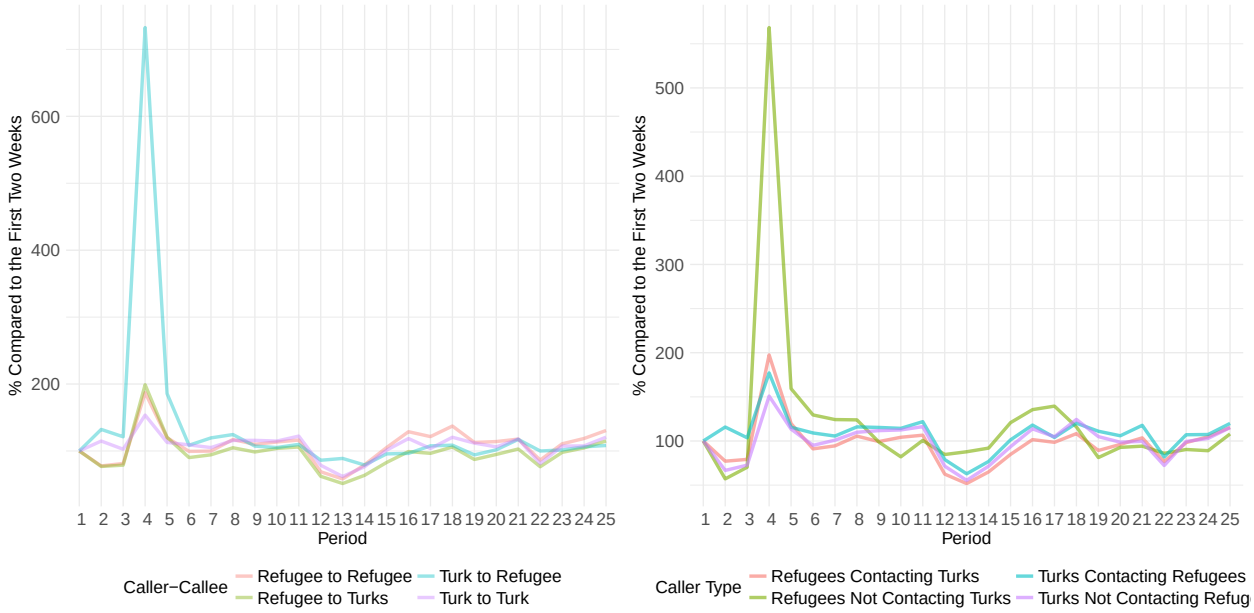


Figure 23: Biweekly Call Patterns by Call Type (left) and by Caller Type (right)

	Refugee To Turk vs Ref.	Turk To Turk vs Ref.
p-value	0.12	0.31
t-Statistic	-1.60	-1.03

Table 29: Aggregate Call Statistics - Dataset 2

Caller:	Turkish		Refugee	
Callee:	Turkish	Refugee	Refugee	Turkish
Nr of Calls	129764748	60981	1994463	17524998
Pct Within Caller Group	99.95	0.05	10.22	89.78
Pct Within Total	86.89	0.04	1.34	11.73

B.2 Specification

$$y_i = \alpha_0 + \beta x_i + \gamma w_a + \delta z_d + \theta_p + \eta_t + \epsilon_i$$

where α_0 is the intercept for the base category, θ_p is the intercept for province-level dummies, and η_t is the intercept for the time-level dummies (one for each of the 25 periods in Dataset 2). y_i denotes the integration measure. x_i is a vector of individual-level variables, w_a is a vector of antenna-level variables, and z_d is a vector of district-level variables. Finally, ϵ_i captures the individual-specific residual. In addition, all the standard errors are clustered at the province- and time-level.

B.3 Variable Names

Individual-level Covariates:

- `integrationRefCaller` : Percentage of calls made to Turks over the total number of calls made by refugees

- integrationTurkCaller : Percentage of calls made to refugees over the total number of calls made by Turks

- N_Outgoing_Calls : The volume of outgoing calls, as a proxy of spending power (i.e. we presume that individuals that call a lot have higher spending power)

- HOME_IN_CAMP_DIST : A dummy indicating whether the home district is a district with a refugee camp

- HOME_IN_COAST_DIST : A dummy indicating whether the home district is at the Greek-Turkish maritime border

- HOME_IN_BORDER_DIST : A dummy indicating the home district involves a Syrian border crossing point

- GO_TO_CAMP_DIST : A dummy that shows whether the person visited a refugee camp district

- GO_TO_COAST_DIST : A dummy that shows whether the person visited a district at the Greek-Turkish maritime border

- GO_TO_BORDER_DIST : A dummy that shows whether the person went to a district with a Syrian border crossing point

- districtperc : Percentage of calls made via an antenna in the home district (versus other districts)

- provinceperc : Percentage of calls made via an antenna in the home province (versus other provinces)

- N_District : The number of districts visited

- N_Provinces : The number of provinces visited

- logmobility : The (log of) trapezoid area in meters that takes into account the farthest points the refugee visited

- NUMBER_OF_SCHOOL_VISITS : The number of days on which a call was made during the school pick up or drop off hours (weekdays 6-9 am, 2-6pm)

- NUMBER_OF_MOSQUE_VISITS : The number of days on which a call was made before, during, and after the Friday prayer (11am-2 pm)

- CLOSE_SCHOOL_Dist : Distance between the home antenna and the closest school (Source: Ministry of Education)

- CLOSE_MOSQUE_Dist : Distance between the home antenna and the closest mosque (Source: Directorate of Religious Affairs)

- CLOSE_SCHOOL_Dist : Distance between the home antenna and the closest school visited (Source: Ministry of Education)

- CLOSE_MOSQUE_Dist : Distance between the home antenna and the closest mosque visited (Source: Directorate of Religious Affairs)

District-level Covariates:

- `geoschoolper1k` : Number of schools in the home district per 1k residents (Source: Ministry of Education)
- `geohealthper1k` : Number of health clinics or hospitals in the home district per 1k residents (Source: Ministry of Health)
- `geomosqueper1k` : Number of mosques in the home district per 1k residents (Source: Directorate of Religious Affairs)
- `Islamicwaqf` : Number of religious-oriented waqfs in the home district per 1k residents (Source: Directorate of Foundations)
- `secularwaqf` : Number of non religious-oriented waqfs in the home district per 1k residents (Source: Directorate of Foundations)
- `kizilayperperson` : Number of Kizilay cards (cash transfers/grocery credits) per refugee in the home district (Source: the Red Crescent - Kızılay)
- `afadperperson` : Number of AFAD cards (grocery credits) per refugee in the home district (Source: the Disaster and Emergency Management Authority - AFAD)
- `nrprsch` : Number of private schools in the home district (Source: Ministry of Education)
- `nrprdorms` : Number of private dorms in the home district (Source: Ministry of Education)
- `nrtutoring` : Number of private tutoring centers in the home district (Source: Ministry of Education)
- `logrefpop` : Log of refugee population in the district (Source: the Ministry of Interior)
- `logpop` : Log of Turkish population in the district (Source: Turkish Statistical Institute)
- `prop_refpop` : Proportion of refugee population to total population (in percent)
- `logatmnr` : Log of number of ATMs in the district as a proxy for district wealth (Source: Banking Regulation and Supervision Agency)
- `illitrate` : Illiteracy rate in the district (Source: TURKSTAT)
- `voteshare` : Percentage of the incumbent party votes in the home district (Source: Turkish Statistical Institute)
- `cicenter` : Whether the home district is urban center or rural center. (Coded from relevant legislations)

Antenna-level Covariates:

- `Centrality_Turk` : Degree centrality of antenna for network of calls made by Turks (Dataset 1)

- `Centrality_Ref` : Degree centrality of antenna for network of calls made by refugees (Dataset 1)

- `shia` : Shia concentration (Dataset 1)

- `sunni` : Sunni concentration (Dataset 1)

- `fri` : Friday prayer activity concentration (Dataset 1)

- `mixed` : Mixed sectarian behavior(Dataset 1)

B.4 Summary Statistics

Table 30: Summary Statistics for Individual-Level Regression Analysis (Refugee Caller Subset)

Statistic	N	Mean	St. Dev.	Min	Max
integrationRefCaller	277,721	89.991	20.143	0.000	100.000
N_Outgoing_Calls	277,721	36.848	41.465	1	771
refpop	277,721	35,755.440	45,218.610	0.000	187,465.600
pop	277,639	399,936.600	245,522.100	2,318	913,715
monthlykizilay	277,721	9,569.527	15,925.190	0	85,345
monthlyafa	277,721	34.808	466.371	0.000	23,131.000
monthlycash	277,721	9,604.335	15,923.950	0.000	85,345.000
shia	271,290	0.248	0.432	0	1
sunni	269,272	0.168	0.374	0	1
mixed	273,233	0.077	0.267	0	1
fri	267,218	0.026	0.159	0	1
HOME_IN_CAMP_DIST	277,721	0.108	0.310	0	1
HOME_IN_COAST_DIST	277,721	0.007	0.084	0	1
HOME_IN_BORDER_DIST	277,721	0.029	0.168	0	1
GO_TO_CAMP_DIST	277,721	0.133	0.339	0	1
GO_TO_COAST_DIST	277,721	0.012	0.111	0	1
GO_TO_BORDER_DIST	277,721	0.044	0.204	0	1
provinceperc	277,721	95.980	11.620	0.000	100.000
districtperc	277,721	82.343	21.023	0.000	100.000
Islamicwaqf	277,671	12.515	23.460	0	96
secularwaqf	277,671	32.133	71.951	0	542
nrprdorms	277,671	15.801	18.849	0	120
nrprsch	277,671	46.642	49.174	0	305
nrtutoring	277,671	37.845	50.538	0	370
geoschoolper1k	277,639	0.491	0.554	0.122	4.844
geohealthper1k	277,639	0.090	0.040	0.017	0.863
geomosqueper1k	277,639	0.304	0.250	0.012	2.604
illitrate	277,639	1.266	0.904	0.221	8.689
cons	277,639	84.055	3.158	66.667	93.671
cicenter	277,639	0.835	0.371	0	1
totalatmnr	277,671	283.331	230.618	0	1,412
voteshare	277,639	50.948	15.428	0.108	92.419
N_Visits_CLOSEST_FRI_MOSQUE	277,721	0.711	0.966	0	13
CLOSE_SCHOOL_Dist	277,719	Inf.000		0.005	Inf.000
CLOSE_MOSQUE_Dist	277,719	1.176	7.466	0.0003	459.315
N_Provinces	277,721	1.244	0.848	1	39
N_Districts	277,721	2.799	2.806	1	78
N_Visits_CLOSEST_WKDY_SCHOOL	277,721	1.859	2.704	0	52
N_Visits_CLOSEST_FRI_MOSQUE.1	277,721	0.711	0.966	0	13
CENTRALITY_T	277,391	0.012	0.010	0.00005	0.079
CENTRALITY_R	277,190	0.014	0.015	0.0001	0.112

Table 31: Summary Statistics for Individual-Level Regression Analysis (Turkish Caller Subset)

Statistic	N	Mean	St. Dev.	Min	Max
integrationTurkCaller	1,351,536	0.184	1.507	0.000	100.000
N_Outgoing_Calls	1,351,536	50.823	43.561	1	827
refpop	1,351,536	28,987.830	42,819.650	0.000	187,465.600
pop	1,350,554	422,138.400	250,377.400	1,666	913,715
monthlykizilay	1,351,536	7,360.167	13,961.540	0	85,345
monthlyafa	1,351,536	11.850	297.554	0.000	23,131.000
monthlycash	1,351,536	7,372.017	13,961.180	0.000	85,345.000
shia	1,324,870	0.151	0.358	0	1
sunni	1,313,507	0.115	0.319	0	1
mixed	1,332,418	0.035	0.184	0	1
fri	1,314,319	0.023	0.150	0	1
HOME_IN_CAMP_DIST	1,351,536	0.036	0.186	0	1
HOME_IN_COAST_DIST	1,351,536	0.006	0.076	0	1
HOME_IN_BORDER_DIST	1,351,536	0.004	0.060	0	1
GO_TO_CAMP_DIST	1,351,536	0.074	0.262	0	1
GO_TO_COAST_DIST	1,351,536	0.016	0.124	0	1
GO_TO_BORDER_DIST	1,351,536	0.011	0.106	0	1
provinceperc	1,351,536	94.801	12.494	0.000	100.000
districtperc	1,351,536	79.006	20.862	0.000	100.000
Islamicwaqf	1,351,035	8.434	14.410	0	96
secularwaqf	1,351,035	22.433	58.097	0	542
nrprdorms	1,351,035	13.358	15.785	0	120
nrprsch	1,351,035	53.421	51.008	0	305
nrtutoring	1,351,035	37.932	46.074	0	370
geoschoolper1k	1,350,554	0.452	0.483	0.122	4.844
geohealthper1k	1,350,554	0.089	0.042	0.017	1.200
geomosqueper1k	1,350,554	0.256	0.205	0.012	2.749
illitrate	1,350,554	1.160	0.905	0.221	14.042
cons	1,350,554	84.432	2.998	52.381	93.769
cicenter	1,350,554	0.793	0.405	0	1
totalatmnr	1,351,014	272.894	211.688	0	1,412
voteshare	1,350,554	51.342	14.848	0.108	95.721
N_Visits_CLOSEST_FRI_MOSQUE	1,351,536	1.043	1.071	0	11
CLOSE_SCHOOL_Dist	1,351,516	Inf.000		0.004	Inf.000
CLOSE_MOSQUE_Dist	1,351,516	1.240	11.873	0.0001	459.315
N_Provinces	1,351,536	1.340	0.924	1	31
N_Districts	1,351,536	3.422	3.177	1	82
N_Visits_CLOSEST_WKDY_SCHOOL	1,351,536	2.419	3.237	0	60
N_Visits_CLOSEST_FRI_MOSQUE.1	1,351,536	1.043	1.071	0	11
CENTRALITY_T	1,349,523	0.013	0.010	0.00005	0.079
CENTRALITY_R	1,331,543	0.007	0.009	0.0001	0.112

B.5 Degree Centrality

Figure 24: Map of Degree Centrality of Antennas in Refugee Call Network in Turkey

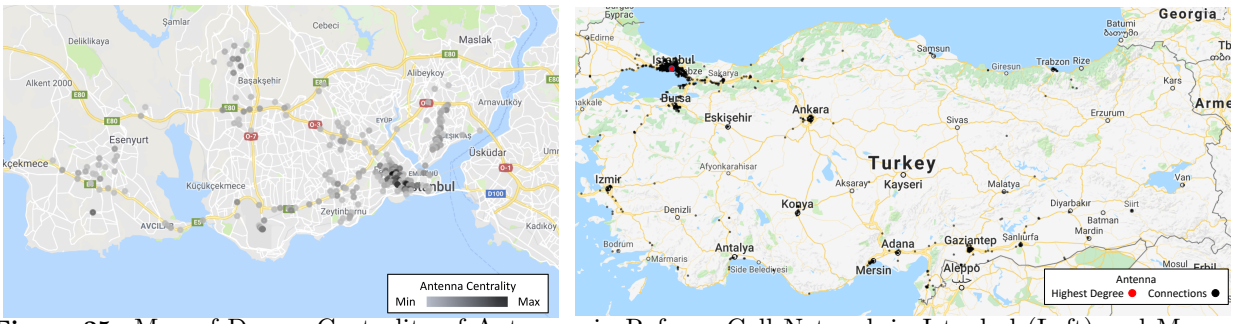


Figure 25: Map of Degree Centrality of Antennas in Refugee Call Network in Istanbul (Left) and Map of Antenna with Highest Degree and Adjacent Antenna Locations (right)

B.6 Results

Table 32: Effect of Socioeconomic, Welfare-related and Spatial Factors on Refugee Integration (Calls Made by Refugees)

	Dependent variable:						
	% of inter-group calls						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
N_Outgoing_Calls	−0.011	−0.010	−0.010	−0.009	−0.009	−0.009	−0.009
	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)
CENTRALITY_T	0.163	0.156	0.147	0.167	0.158	0.155	0.166
	(0.224)	(0.223)	(0.224)	(0.223)	(0.225)	(0.226)	(0.223)
CENTRALITY_R	−1.283***	−1.289***	−1.286***	−1.281***	−1.280***	−1.276***	−1.280***
	(0.203)	(0.202)	(0.198)	(0.216)	(0.216)	(0.218)	(0.218)
GO_TO_COAST_DIST	1.660***	1.603***	1.587**	1.796***	1.798***	1.775***	1.794***
	(0.605)	(0.607)	(0.626)	(0.686)	(0.684)	(0.678)	(0.684)
GO_TO_BORDER_DIST	0.570*	0.577*	0.609**	0.622	0.626	0.632	0.622
	(0.304)	(0.303)	(0.310)	(0.398)	(0.400)	(0.403)	(0.399)
GO_TO_CAMP_DIST	0.861**	0.867**	0.864**	0.856**	0.873**	0.871**	0.857**
	(0.341)	(0.337)	(0.346)	(0.367)	(0.365)	(0.373)	(0.369)
HOME_IN_BORDER_DIST	−0.617	−0.527	−0.519	−0.496	−0.628	−0.454	−0.479
	(0.911)	(0.907)	(0.896)	(0.891)	(0.880)	(0.871)	(0.884)
HOME_IN_CAMP_DIST	−1.606***	−1.605***	−1.594***	−1.577***	−1.507***	−1.384**	−1.583***
	(0.506)	(0.506)	(0.500)	(0.543)	(0.568)	(0.546)	(0.552)
HOME_IN_COAST_DIST	−3.086**	−3.038**	−2.916**	−3.172***	−3.208**	−3.243**	−3.156**
	(1.227)	(1.227)	(1.227)	(1.224)	(1.257)	(1.327)	(1.238)
geoschoolper1k	1.505	1.466	1.431	1.325	2.467	1.465	1.285
	(1.032)	(1.030)	(1.021)	(1.123)	(1.584)	(1.183)	(1.160)
geohealthper1k	8.978	9.458	9.395	10.212	10.049	32.407***	10.170
	(8.466)	(8.430)	(8.457)	(7.877)	(8.121)	(10.859)	(7.939)
geomosqueper1k	−0.974	−1.084	−1.155	−1.177	−0.923	−0.667	0.887
	(2.356)	(2.322)	(2.327)	(2.505)	(2.581)	(2.515)	(7.241)
Islamicwaqf	0.064***	0.065***	0.066***	0.065***	0.063***	0.061***	0.068***
	(0.017)	(0.017)	(0.017)	(0.019)	(0.020)	(0.020)	(0.022)
secularwaqf	−0.017***	−0.017***	−0.017***	−0.017**	−0.017**	−0.017**	−0.017**
	(0.006)	(0.006)	(0.006)	(0.007)	(0.007)	(0.007)	(0.007)
shia	−0.157	−0.110	−0.109	−0.040	−0.044	−0.046	−0.041
	(0.312)	(0.316)	(0.325)	(0.458)	(0.459)	(0.458)	(0.457)
sunni	−0.062	−0.007	−0.020	−0.053	−0.051	−0.043	−0.052
	(0.514)	(0.510)	(0.502)	(0.486)	(0.484)	(0.488)	(0.488)
mixed	−0.397	−0.363	−0.283	−0.420	−0.425	−0.435	−0.421
	(0.353)	(0.352)	(0.354)	(0.404)	(0.402)	(0.406)	(0.405)
fri	−1.066	−1.041	−1.025	−0.959	−0.957	−0.950	−0.959
	(1.133)	(1.159)	(1.170)	(1.207)	(1.206)	(1.208)	(1.208)
voteshare	0.002	0.002	0.001	0.002	0.002	0.002	0.002
	(0.014)	(0.014)	(0.014)	(0.012)	(0.012)	(0.011)	(0.012)
logrefpop	−0.006	0.006	−0.002	0.003	0.052	0.253	0.047
	(0.222)	(0.224)	(0.223)	(0.198)	(0.234)	(0.193)	(0.285)
logpop	2.346***	2.358***	2.346***	2.337***	2.343***	2.337***	2.324***
	(0.785)	(0.784)	(0.782)	(0.834)	(0.842)	(0.871)	(0.844)
prop_refpop	4.045	3.863	3.887	3.827	4.572*	3.903	3.833
	(2.855)	(2.857)	(2.851)	(2.905)	(2.575)	(2.968)	(2.888)
logmobility	0.028	0.029	0.028	0.027	0.027	0.027	0.027
	(0.054)	(0.055)	(0.055)	(0.055)	(0.055)	(0.055)	(0.055)
logatmnr	−1.329***	−1.346***	−1.334***	−1.367**	−1.361***	−1.339**	−1.370**
	(0.416)	(0.416)	(0.413)	(0.532)	(0.528)	(0.540)	(0.537)
illitrate	−0.081	−0.059	−0.026	−0.041	−0.038	−0.054	−0.032
	(0.373)	(0.380)	(0.383)	(0.285)	(0.304)	(0.308)	(0.311)
cicenter	−1.098*	−1.011	−0.920	−0.955	−1.020	−1.008	−0.942
	(0.641)	(0.637)	(0.631)	(0.665)	(0.683)	(0.648)	(0.685)
period_trend		−0.077***	0.010***				
		(0.018)	(0.003)				
geoschoolper1k:logrefpop					−0.142		
					(0.167)		
geohealthper1k:logrefpop						−2.866**	
						(1.286)	
geomosqueper1k:logrefpop							−0.222
							(0.784)
Constant	68.188***	69.018***	67.785***	69.577***	69.057***	67.325***	69.352***
	(9.175)	(9.157)	(9.237)	(9.255)	(9.222)	(9.977)	(9.242)
Observations	261,198	261,198	261,198	261,198	261,198	261,198	261,198
R^2	0.024	0.025	0.026	0.027	0.027	0.027	0.027
Adjusted R^2	0.024	0.025	0.026	0.027	0.027	0.027	0.027
Residual Std. Error	18.262 (df = 261091)	18.256 (df = 261090)	18.248 (df = 261010)	18.237 (df = 261067)	18.237 (df = 261066)	18.236 (df = 261066)	18.237 (df = 261066)
F Statistic	61.766*** (df = 106; 261091)	63.060*** (df = 107; 261090)	37.735*** (df = 187; 261010)	56.298*** (df = 130; 261067)	55.905*** (df = 131; 261066)	56.073*** (df = 131; 261066)	55.877*** (df = 131; 261066)

Note: *p<0.1; **p<0.05; ***p<0.01

Table 33: Effect of Socioeconomic, Welfare-related and Spatial Factors on Refugee Integration (Calls Made by Turks)

	Dependent variable:						
	% of inter-group calls						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
N_Outgoing_Calls	−0.001***	−0.001***	−0.001***	−0.002***	−0.002***	−0.002***	−0.002***
	(0.0002)	(0.0002)	(0.0002)	(0.0003)	(0.0003)	(0.0003)	(0.0003)
CENTRALITY_T	−0.021***	−0.021***	−0.021***	−0.020***	−0.020***	−0.020***	−0.020***
	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
CENTRALITY_R	0.037***	0.037***	0.038***	0.038***	0.040***	0.040***	0.038***
	(0.006)	(0.006)	(0.006)	(0.005)	(0.005)	(0.005)	(0.005)
GO_TO_COAST_DIST	0.081**	0.081**	0.081**	0.076**	0.075**	0.076**	0.075**
	(0.033)	(0.033)	(0.033)	(0.036)	(0.036)	(0.036)	(0.036)
GO_TO_BORDER_DIST	0.018	0.018	0.016	0.021	0.019	0.018	0.021
	(0.018)	(0.018)	(0.017)	(0.021)	(0.021)	(0.021)	(0.021)
GO_TO_CAMP_DIST	0.054***	0.054***	0.054***	0.057***	0.056***	0.057***	0.057***
	(0.018)	(0.018)	(0.017)	(0.020)	(0.020)	(0.020)	(0.020)
HOME_IN_BORDER_DIST	−0.212**	−0.212**	−0.209**	−0.219**	−0.223***	−0.222***	−0.219**
	(0.089)	(0.089)	(0.088)	(0.090)	(0.084)	(0.073)	(0.091)
HOME_IN_CAMP_DIST	−0.074*	−0.075*	−0.075*	−0.077**	−0.081**	−0.081**	−0.077**
	(0.039)	(0.039)	(0.039)	(0.038)	(0.038)	(0.035)	(0.038)
HOME_IN_COAST_DIST	−0.110***	−0.110***	−0.109***	−0.114***	−0.110***	−0.110***	−0.114***
	(0.031)	(0.031)	(0.030)	(0.036)	(0.035)	(0.034)	(0.036)
geoschoolper1k	−0.019	−0.019	−0.017	−0.020	−0.065*	−0.028	−0.019
	(0.019)	(0.019)	(0.020)	(0.018)	(0.035)	(0.018)	(0.019)
geohealthper1k	−0.221	−0.221	−0.224	−0.239*	−0.189	−0.556***	−0.234
	(0.146)	(0.146)	(0.146)	(0.145)	(0.159)	(0.153)	(0.148)
geomosqueper1k	0.057*	0.057*	0.059*	0.058*	0.053	0.047	0.044
	(0.033)	(0.033)	(0.033)	(0.035)	(0.034)	(0.035)	(0.074)
rivalwaqf	−0.0001	−0.0001	−0.0001	−0.0001	−0.00001	−0.0001	−0.0002
	(0.0003)	(0.0003)	(0.0003)	(0.0003)	(0.0003)	(0.0003)	(0.0003)
secularwaqf	0.0002	0.0002	0.0001	0.0002	0.0001	0.0002*	0.0002
	(0.0002)	(0.0002)	(0.0002)	(0.0001)	(0.0002)	(0.0001)	(0.0001)
shia	0.019	0.019	0.018	0.019	0.019	0.019	0.019
	(0.013)	(0.013)	(0.013)	(0.012)	(0.013)	(0.012)	(0.012)
sunni	0.002	0.002	0.002	0.003	0.004	0.004	0.003
	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)
mixed	0.033	0.033	0.034	0.032	0.033	0.032	0.032
	(0.071)	(0.071)	(0.071)	(0.071)	(0.071)	(0.071)	(0.071)
fri	0.023	0.023	0.024	0.026	0.026	0.026	0.026
	(0.027)	(0.027)	(0.028)	(0.029)	(0.028)	(0.029)	(0.029)
voteshare	0.0001	0.0001	0.0001	0.0001	0.0002	0.0003	0.0001
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
logrefpop	0.005	0.005	0.004	0.005	−0.001	−0.006	0.004
	(0.006)	(0.006)	(0.006)	(0.006)	(0.007)	(0.008)	(0.006)
logpop	0.001	0.001	0.002	−0.001	0.002	0.008	−0.001
	(0.020)	(0.020)	(0.020)	(0.020)	(0.020)	(0.020)	(0.020)
prop_refpop	0.148	0.148	0.149	0.154	0.148	0.177	0.151
	(0.163)	(0.163)	(0.163)	(0.171)	(0.172)	(0.191)	(0.172)
logmobility	−0.032***	−0.032***	−0.032***	−0.033***	−0.033***	−0.033***	−0.033***
	(0.002)	(0.002)	(0.002)	(0.003)	(0.003)	(0.003)	(0.003)
logatmnr	−0.009	−0.009	−0.010	−0.006	−0.008	−0.011	−0.006
	(0.014)	(0.014)	(0.014)	(0.015)	(0.014)	(0.014)	(0.014)
illitrate	0.021**	0.021**	0.021**	0.020**	0.021**	0.021**	0.020**
	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)
cicenter	−0.010	−0.010	−0.010	−0.009	−0.006	−0.005	−0.009
	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)	(0.013)	(0.014)
period_trend		−0.00002	−0.001***				
		(0.0005)	(0.0002)				
geoschoolper1k:logrefpop					0.007		
					(0.005)		
geohealthper1k:logrefpop						0.087**	
						(0.042)	
geomosqueper1k:logrefpop							0.002
							(0.008)
Constant	0.243	0.243	0.244	0.303	0.287	0.249	0.301
	(0.198)	(0.201)	(0.197)	(0.205)	(0.202)	(0.207)	(0.205)
Observations	1,276,195	1,276,195	1,276,195	1,276,195	1,276,195	1,276,195	1,276,195
R^2	0.028	0.028	0.028	0.030	0.030	0.030	0.030
Adjusted R^2	0.028	0.028	0.028	0.030	0.030	0.030	0.030
Residual Std. Error	1.308 (df = 1276088)	1.308 (df = 1276087)	1.308 (df = 1276007)	1.306 (df = 1276064)	1.306 (df = 1276063)	1.306 (df = 1276063)	1.306 (df = 1276063)
F Statistic	345.377*** (df = 106; 1276088)	342.149*** (df = 107; 1276087)	199.449*** (df = 187; 1276007)	306.863*** (df = 130; 1276064)	304.729*** (df = 131; 1276063)	305.054*** (df = 131; 1276063)	304.523*** (df = 131; 1276063)
Note:							*p<0.1; **p<0.05; ***p<0.01

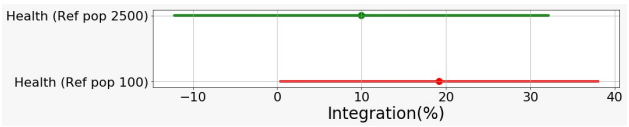


Figure 26: Effect of Health Center Availability Across Different Refugee Population Sizes, with significant factors in red

B.7 Social Assistance Policies

Recent evidence on Syrian refugees suggests that cash transfers to refugees reduce barriers to accessing basic services and employment, in addition to having long-term implications for economic and social outcomes (Hagen-Zanker et al. 2017). To see whether this has a positive effect on social integration, we look at the effect of two policies: The Emergency Social Safety Net (ESSN), implemented by the Red Crescent (Kizilay), which offers a monthly cash allowance in the form of debit cards, known as Kizilay cards, to about 1.1 million refugees within or outside-camps that are most in need. The second project is the AFAD card project implemented by the Disaster and Emergency Management Authority (AFAD), which provides refugees living in camps a card with money they can use to buy food in designated stores.

We exploit regional variation in the timing of project launch and the proportion of beneficiaries to see whether access to such services and social inclusion facilitates integration. We examine how the aggregate number of program beneficiaries (per capita) in a given district affects integration in that district. The main independent variable, number of aid card users (per refugee) is a time-varying indicator that varies at the district level, is available by month. Thus this multi-period design includes 12 periods (month). In the design, we use district and month dummies to control for time-invariant district-level characteristics and time shocks, and each observation corresponds to one individual. For the dependent variable, integration, we look both at the percentage of calls made to Turks by refugees and at the percentage of calls made to refugees by Turks.

For districts with no refugee camps, we only look at the first program, the effect of Kizilay cards, since AFAD cards are only distributed in camp locations. In camp districts, while the direction of the effect of social assistance appears negative, statistical significance is not robust across different specifications. In non-camp districts, in addition to the null finding across different specifications, the direction of the effect is not consistent either. The findings don't differ when we measure integration by the number of inter-group calls made by Turks, instead of those made by refugees. Although the effect of Kizilay cards appears to be positive this time, it's not robust across different specifications (see Appendix Table 36). We see no substantial change in the results with sampling weights.

Part of the lack of significance for this type of assistance could be due to the fact that they only target the weakest and economically most marginal subset of the refugee population (i.e., there is an underlying selection effect biasing the results against integration) and that the amounts given are actually quite small if one takes into account overall levels of need. Specifically, the Kizilay card provides a 120 TL cash per family member, which amounts to only 1/3rd of the minimum wage for a four-member family, while the AFAD card provides, even less in the form of cash deposits worth 100 TL per family. The mean number of Kizilay cards provided in a district is 4,060, as compared to 687 for AFAD cards.

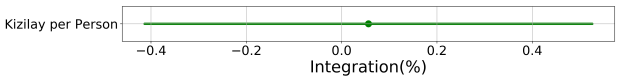


Figure 27: Confidence Intervals for Factors Associated with Social Assist. Policies (integration measured by refugees' calls to Turks)

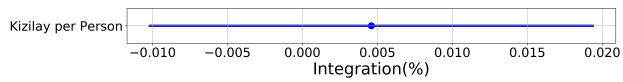


Figure 28: Confidence Intervals for Factors Associated with Social Assist. Policies (integration measured by Turks' calls to refugees)

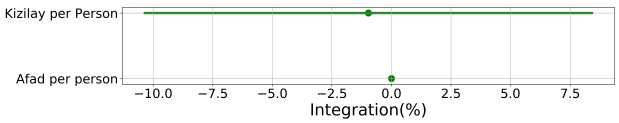


Figure 29: Camp Districts: Confidence Intervals for Factors Associated with Social Assist. Policies (integration measured by refugees' calls to Turks)

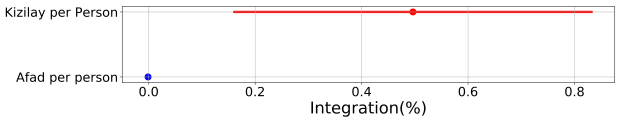


Figure 30: Camp Districts: Confidence Intervals for Factors Associated with Social Assist. Policies (integration measured by Turks' calls to refugees)

Table 34: Within-Camp Effect of Social Assistance Policies on Refugee Integration- (Calls Made By Refugees)

	Dependent variable:			
	% of inter-group calls			
	(1)	(2)	(3)	(4)
kizilayperperson	-10.326^{***}	0.233	-0.050	-0.545
	(3.467)	(5.422)	(3.182)	(6.545)
afadperperperson	0.010	0.008^{*}	-0.213^{**}	-0.008
	(0.008)	(0.005)	(0.101)	(0.033)
period_trend		-0.142^{**}	-0.211	
		(0.068)	(0.174)	
Constant	88.017^{***}	89.670^{***}	94.661^{***}	88.604^{***}
	(0.287)	(0.945)	(2.503)	(0.910)
Observations	29,996	29,996	29,996	29,996
R^2	0.015	0.015	0.003	0.022
Adjusted R^2	0.012	0.013	0.003	0.019
Residual Std. Error	25.077 (df = 29924)	25.069 (df = 29923)	25.193 (df = 29990)	24.996 (df = 29900)
F Statistic	6.314^{***} (df = 71; 29924)	6.533^{***} (df = 72; 29923)	20.755^{***} (df = 5; 29990)	7.065^{***} (df = 95; 29900)
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 35: Outside-Camp Effect of Kizilay Kards on Refugee Integration (Calls Made By Refugees)

	Dependent variable:			
	% of inter-group calls			
	(1)	(2)	(3)	(4)
kizilayperperson	−0.204	0.411**	1.100***	0.570***
	(0.382)	(0.175)	(0.087)	(0.162)
HOME_IN_BORDER_DIST	−5.497***	−5.917***	3.937***	−5.126
	(0.067)	(0.109)	(0.335)	
period_trend		−0.075***	−0.056	
		(0.021)	(0.050)	
Constant	100.000***	101.492***	90.373***	101.344***
	(0.000)	(0.417)	(0.847)	(0.049)
Observations	247,675	247,675	247,675	247,675
R^2	0.041	0.042	0.006	0.044
Adjusted R^2	0.038	0.039	0.006	0.041
Residual Std. Error	17.093 (df = 246939)	17.088 (df = 246938)	17.377 (df = 247669)	17.064 (df = 246915)
F Statistic	14.415*** (df = 735; 246939)	14.596*** (df = 736; 246938)	301.563*** (df = 5; 247669)	15.116*** (df = 759; 246915)
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 36: Within-Camp Effect of Kizilay Kards on Refugee Integration - (Calls Made By Turks)

	Dependent variable:			
	% of inter-group calls			
	(1)	(2)	(3)	(4)
kizilayperperson	−0.091	0.302**	−0.018	0.245
	(0.085)	(0.126)	(0.133)	(0.177)
afadperperperson	−0.002	−0.001	−0.004	−0.002
	(0.002)	(0.002)	(0.003)	(0.002)
period_trend		−0.006**	−0.0004	
		(0.002)	(0.003)	
Constant	0.010	0.008	0.251***	0.004
	(0.010)	(0.007)	(0.060)	(0.033)
Observations	48,500	48,500	48,500	48,500
R^2	0.007	0.007	0.001	0.008
Adjusted R^2	0.004	0.004	0.001	0.005
Residual Std. Error	1.321 (df = 48339)	1.321 (df = 48338)	1.323 (df = 48494)	1.321 (df = 48315)
F Statistic	2.086*** (df = 160; 48339)	2.171*** (df = 161; 48338)	8.858*** (df = 5; 48494)	2.198*** (df = 184; 48315)
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 37: Outside-Camp Effect of Kizilay Kards on Refugee Integration (Calls Made By Turks)

	Dependent variable:			
	% of inter-group calls			
	(1)	(2)	(3)	(4)
kizilayperperson	−0.014	0.019	−0.025**	0.018
	(0.019)	(0.021)	(0.012)	(0.025)
period_trend		−0.002**	−0.003**	
		(0.001)	(0.001)	
Constant	0.112***	0.138***	0.217***	0.144***
	(0.001)	(0.011)	(0.028)	(0.014)
Observations	276,077	276,077	276,077	276,077
R^2	0.011	0.011	0.0002	0.011
Adjusted R^2	0.010	0.010	0.0002	0.010
Residual Std. Error	1.435 (df = 275888)	1.435 (df = 275887)	1.442 (df = 276072)	1.435 (df = 275864)
F Statistic	15.963*** (df = 188; 275888)	15.987*** (df = 189; 275887)	16.047*** (df = 4; 276072)	14.557*** (df = 212; 275864)
Note:				*p<0.1; **p<0.05; ***p<0.01

C Long-Term Movement (Dataset 3)

Table 38: District-Level Factors and Receiving Refugees

	Dependent variable:
	receiving_or_sending
geoschoolper1k	−0.021***
	(0.005)
geohealthper1k	0.294***
	(0.106)
logrefpop	−0.002
	(0.002)
logpop	−0.020***
	(0.005)
Constant	0.000
	(0.001)
Observations	83,658
R^2	0.001
Adjusted R^2	0.001
Residual Std. Error	0.276 (df = 83653)
F Statistic	26.804*** (df = 4; 83653)
Note:	*p<0.1; **p<0.05; ***p<0.01

C.1 Lasso and Ridge Models on Spatial Factors Associated with Integration and Over-Time Movement

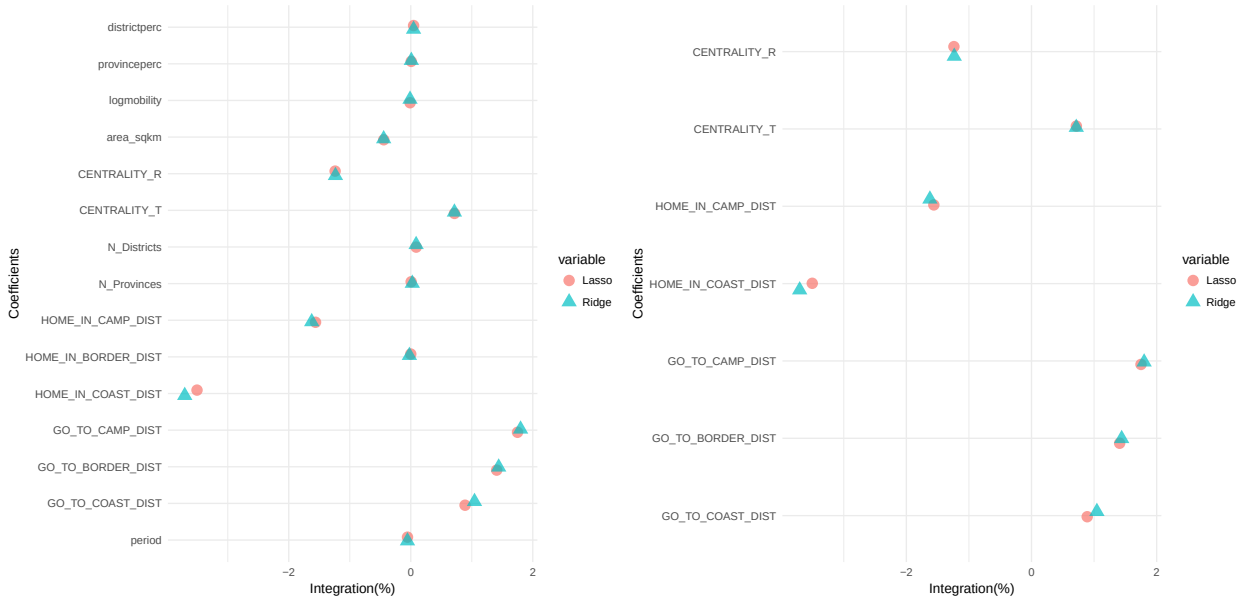


Figure 31: Lasso and Ridge Estimates for Spatial Factors Associated with Integration (as measured by refugee calls to Turks), all covariates included (left) and near-zero estimates excluded (right)

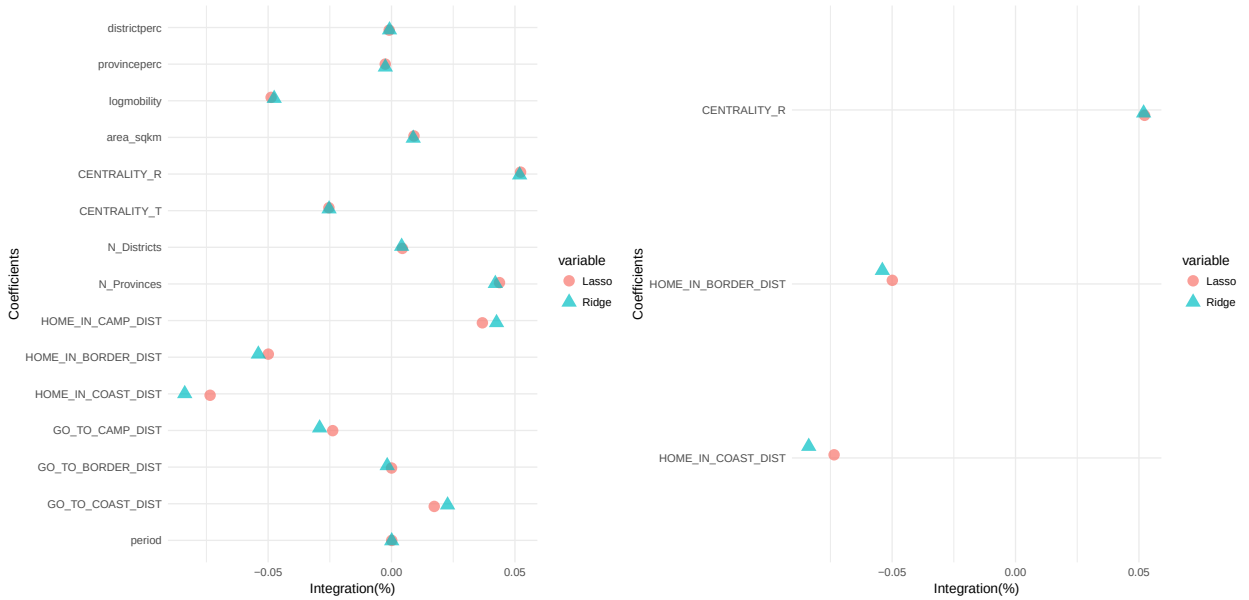


Figure 32: Lasso and Ridge Estimates for Spatial Factors Associated with Integration (as measured by refugee calls to Turks), all covariates included (left) and near-zero estimates excluded (right)

Optimizing the Access to Healthcare Services in Dense Refugee Hosting Urban Areas: A Case for Istanbul

M. Tarik Altuncu[1,2] m.altuncu@imperial.ac.uk, Ayse Seyyide Kaptaner[3] ayseyyide@gmail.com, and Nur Sevencan[1] nur.sevencan@trtworld.com

[1] TRT World, Ahmet Adnan Saygun Cad. No:83, Istanbul, Turkey
[2] Imperial College London, SW7 2AZ, London, United Kingdom
[3] Birkbeck University of London, WC1E 7HX, London, United Kingdom

Abstract. With over 3.5 million refugees, Turkey continues to host the worlds largest refugee population. This introduced several challenges in many areas including access to healthcare system. Refugees have legal rights to free healthcare services in Turkeys public hospitals. With the aim of increasing healthcare access for refugees, we looked at where the lack of infrastructure is felt the most. Our study attempts to address these problems by assessing whether Migrant Health Centers locations are optimal. The aim of this study is to improve refugees' access to healthcare services in Istanbul by improving the locations of health facilities available to them. We used call data records provided by Turk Telekom.

Keywords: refugees· public health · access · migration · healthcare · D4R · CDR

1 Introduction

With over 3.5 million refugees, Turkey continues to host the worlds largest refugee population 21. This introduced several challenges in many areas including health sector. Refugees have legal rights to free health service in Turkeys public hospitals 10. In addition to free health services at public hospitals, 178 Migrant Health centers (MHCs) have been opened in Turkey to fulfill the healthcare needs of refugees and solve overcrowding at public hospitals 1. With the aim of increasing healthcare access of refugees, we looked at where the lack of infrastructure is felt the most. Therefore, we focus on Istanbul, Turkey's biggest city in terms of population and urbanization, which also happens to be the city that hosts the largest refugee population in Turkey. 20 MHCs have been opened in Istanbul.

Despite the immense efforts for the integration of refugees into Turkey's healthcare system, language barriers 7, 8, registration problems 8, 6, navigation of the system 8, overcrowding 16, 19, 12, refugees lack of knowledge about the services available to them 10, and lack of translators 10 remain as the main

challenges. The studies cited above point to challenges refugees face in Turkey in general. There are not extensive studies focusing on refugees' health access in Istanbul.

In order to check whether Turkey-wide problems facing refugees also apply to refugees in Istanbul, we obtained data from Sultanbeyli Municipality's Refugees Association through Syrian Coordination Center Software (SUKOM). According to the data SUKOM provided, only 670 (50%) of the refugees who reached out to the Association made appointments for the local public hospital in Sultanbeyli. For instance, out of the 1332 refugees who contacted the association to request language assistance for making a hospital appointment, 360 refugees requested an appointment at a public hospital in Pendik, a further away district in addition to other appointment requests made for hospitals in other districts such as Uskudar, Umraniye, Atasehir, Levent, etc. Making appointments in districts other than Sultanbeyli can be explained by these two scenarios which point to either to overcrowding or lack of language support:

1. Not all the refugees who reach out to SUKOM for language assistance reside in Sultanbeyli, thus what we describe as further away hospitals could be their local hospitals or hospitals closer to them
2. Overcrowding at the Sultanbeyli public hospital can be expected since it is the district with one of the highest refugee population in Istanbul.

501 out of 639 refugees who made their appointment between February 2018 and September 2018 through the association were also accompanied by a translator from the association during their hospital visit. This illustrates a lack of language support at hospitals.

Our study attempts to address the problems of lack of language support and overcrowding by assessing whether the Migrant Health Centers locations are optimal. The aim of this study is to increase the access of refugees to healthcare services in Istanbul by improving the locations of health facilities available to them by using the cellular network usage data provided by Turk Telekom.

We tested if Migrant Health Center (MHC) locations are optimal. We suggested optimal locations for Migrant Health Centers.

We considered two different methods of optimization for MHC locations : distance minimization and duration minimization. According Our results, the locations we suggest are more cost effective than their current locations.

2 Methodology

Call Detail Records(CDR) are being used by researchers for different purposes such as urban planning, anomaly detection, and understanding peoples behaviour under certain circumstances. A small portion of the literature also consists of network science approach trying to model complex social systems in 22, and understanding peoples movement patterns for transportation planning 9. Here we use a similar data source to optimise access to healthcare services of refugees living in Istanbul, Turkey.

2.1 Datasets

We use the mobile telecommunication datasets provided by Turk Telekoms Data for Refugees competition 18. Whole range of datasets consist geographical lookup tables for each cell tower (BTS) and other geographical identifiers, call and text communications among each BTSs, call and text activity details of a sampled set of subscribers with their connected cell tower, refugee statuses of callee and callers, and another dataset of a larger and more comprehensive sample of users but with reduced resolution of details. In this study, we only use call activity details of refugees (we will refer this data as *CDR* or *CDR data*) together with their corresponding geographical locations of cell towers (we will refer this data as *BTS records*).

Although a more comprehensive information has been published by the organizers of the competition 18, we provide a short description of the datasets below. The BTS records contains each Turk Telekom cell towers' geographical coordinates, city and borough details along with unique IDs which we can use to join with the CDR data in order to detect geographical coordinates of calling activities.

The CDR data also contains the following information;

- caller ID with two categories (refugee or non-refugee),
- whether the receiving-party of the call is a refugee or non-refugee,
- whether the call is an inbound or an outbound call with reference to the caller ID,
- the timestamp for the call made,
- the ID of base stations to which the caller is connected at the beginning of call.

2.2 Computational Approach and Data Pipeline

Our methods create a data pipeline starting with the detection of the residential location for each refugee subscriber provided in the CDR. We derive the residential location for a subscriber based on their night time calling activities on the CDR. Aggregating the residential locations of these refugees in the form of total number of refugee residents per cell tower enables us to determine refugee densities at each site, and to illustrate this on a Voronoi region[4] map 5 of Istanbul. As there exists an excessive number of cell towers in the city, we cluster them on the basis of their proximity to each other and the number of refugees residents in their respective Voronoi regions. The centers of these clusters constitute the centers of residential regions for refugees. We obtained the distance and the duration of using public transit among residential region centers from a commercial API. Then, we computed the optimum locations for the refugee health centers. The rest of this section provides step by step detailed information about the pipeline.

[4] Voronoi regions are polygonal regions constructed by unit areas that have the same base station as the nearest one 22.

Data Adaptation We made some changes in the datasets before starting computations. Along with some warnings mentioned in 18, we experienced some other problems and inconsistencies within the datasets. For the purpose of reproducibility of the results, we describe the chages we made on the data.

BTS Records: At first we dropped the BTS records with no geographical coordinates provided due to our need for higher geographical precision than city level resolution. Then we converted degree, minutes, seconds (DMS) syntax to latitute and logitude based coordinates. Further, we merged some BTS records because they were either too close to each other to distinguish in terms of the region they cover or on exactly the same coordinates. We used DBSCAN 11 algorithm with epsilon value of 0.0005, and using the euclidean distance metric to measure the distance between geographical coordinates. While joining the BTS with CDR, we used inner join method in order to avoid records with lack of either call or geographical details. Finally, we discovered some BTSs had wrong city records. For instance, some BTSs that were supposed to be in the city of Bolu had geographical coordinates which were actually in Istanbul. Hence, we had to create polygons stating boundaries of Istanbul to filter down the BTS records to our region of interest instead of using the provided city data in BTS records. We also use the same method to determine whether the cell tower is located in European or Asian side of Istanbul. We will use this feature in Observation Weighted Clustering of Cell Towers section. We mirrored all changes and filters in BTS records to the CDR data using the *'SITE_ID'* field.

CDR: We filtered CDR data to cover only the voice calls made by the refugees and made in Istanbul using the merged BTS coordinates. To note, we did not rely on the *'CALL_TYPE'* flag on any part of our analyses because we saw that many users are making either only inbound or only outbound calls according to this feature. We believe that this phenomena is probably a result of sampling made by the data provider, and does not reflect the subscribers' behaviour.

Computing the Number of Refugee Residents Detection of a user's residential region via cellular network usage is an ongoing attempt as various form of approaches can be found in the literature. In 20, authors attempt to detect home location based on activities made nearest to sleeping period which is approximated using the inactive time periods per user, whereas in 13 a small subset of labelled data was used to train a logistic regression model to detect important sites based on the activities made during *'home hours'*, which they define as being between 7 PM and 7 AM. Although the limitations of such methods apply to many cell-phone users having unusual call and text patterns or low usage at individual level, it still provides a quantitatively reliable measure at aggregate level when there is enough data.

Similar to 13, we make an assumption that people are active at their homes during the nighttime, but we selected a narrower time interval which is between 11 PM and 8 AM. Among all users who have ever registered in Istanbul, 62% of

all subscribers and 54% of refugees have at least one call records in this time interval. Using the number of calls which the refugees make during the nighttime, we extract the frequencies of cell towers being used per refugee subscriber. These frequencies can be regarded as the probabilities of corresponding user's residential address is in the area covered by the cell towers. To detect the boundaries of regions served by cell towers, we calculate Voronoi regions for cell towers and refer to these regions as *cell tower areas*.

Since we are interested in finding the refugee distribution in Istanbul, for each cell tower area we sum up the probabilities of refugees living there. This new way of representing the same computation gives us the advantage of mapping refugee density using choropleth maps. However, this map creates a very fine resolution of Istanbul since we have more than 4 thousand cell tower locations based on our BTS dataset. Because the scope of our objective is finding the optimal placement for MHCs, we will not use this resolution in our analyses.

Observation Weighted Clustering of Cell Towers To obtain coarser grain representation of the Voronoi cell regions, we apply k-means clustering 4 to each cell tower. Although this algorithm is directly applicable to spatial positions in two dimensions, our cell towers are not identical in terms of their inertia because they have varying number of refugees. In line with our objective of providing higher resolution for refugee dense regions but coarser resolution for the rest, we weighed each cell tower's location with its corresponding number of refugee residents. To compute this, we modified the input to k-means algorithm in a similar fashion described in the referenced blog post 3.

For k-means clustering algorithm one has to provide the number of clusters as *a priori*. For our case, we set total number of clusters to 200. This is an intuitively selected quantity which is intended to be small enough to be comparable to the number of MHCs while being large enough to maintain enough resolution for refuge-dense regions

We have an additional concern about the geographical boundary of the Bosporus which divides Istanbul into two continents. This natural boundary creates bottle-necks in the transportation networks which local people mostly avoid although there are multiple transportation options that could potentially ease the commute only from specific locations. Therefore, we know that although the two coasts of the Bosporus are very close, we have to avoid grouping regions from different continents together as the transportation between the two is generally more time-consuming. Therefore, we divide the cell towers into two sets for Europe and Asia. Then we separate the parameter k into two parts based on the proportions of refugee call activities from each side of the Bosporus using the CDR data. Since the Asian side consists of only the 35% of all calling activities, we assign 70 clusters to Asian side. We assign the rest 130 clusters to European side.

After running two k-means clustering algorithm for two sides, we assign our residential region centers to the central points of k-means clusters. We create another set of Voronoi cell regions using residential region centers and refer

them as *residential regions*. Lastly we add up the refugee residents in all the cell towers in each cluster and plot them in Figure 1 as choropleth map.

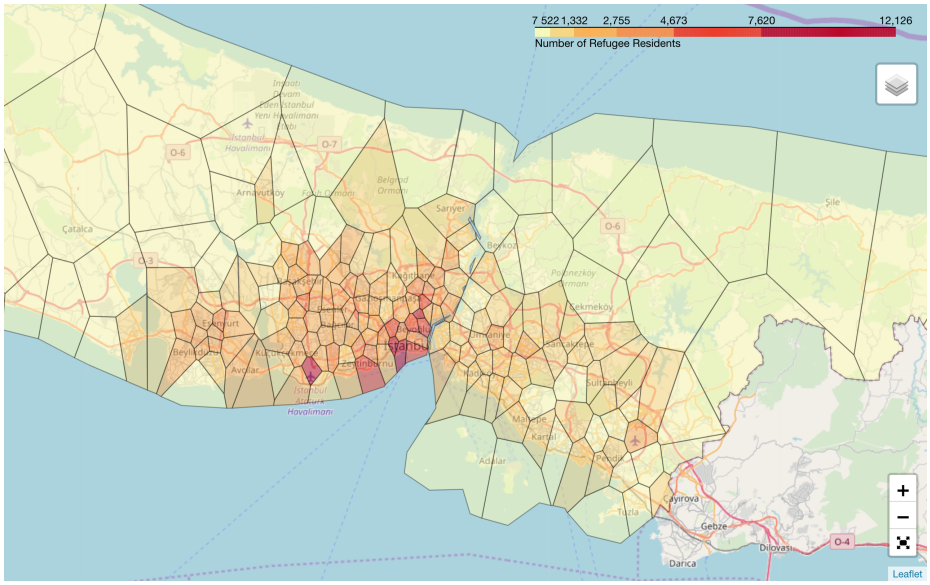


Fig. 1. Map shows the number of refugee residents per residential region in Istanbul based on the number of refugee residents metric which we computed using the night time calling activity of refugees in CDR data.

Transit Costs Between Clusters Having 200 clusters as residential regions and their centers, we must obtain the average costs of public transit among these locations before running any optimization. For that, we used Google Cloud Platform's Distance Matrix API[5]. We requested its estimated time and distance costs between pairs of all residential region centers at 10 AM with local time on December 10^th, 2018 via public transit. We converted the results into two matrices where D contains distances in meters and T the transit duration in seconds. For some cells that Google failed to provide, we used the mean value of the rest of the matrix.

Optimizing the Access to MHCs We applied a multi-facility location optimization linear programming model based on the problem described and solved in 17. The problem considers selecting m locations from n candidates, considering

[5] Details about the Distance Matrix API could be obtained from https://developers.google.com/maps/documentation/distance-matrix/start

either the distance or the duration matrix between all pairs of n candidate locations with the objective function of minimizing the total travel costs of various number of people initially located in n locations.

The equation for this problem is given below;

$$\begin{aligned} &\underset{x}{\text{minimize}} \sum_{i=1}^{n}\sum_{j=1}^{n} a_i \cdot d_{ij} \cdot x_{ij} \\ &\text{subject to} \sum_{j=1}^{n} x_{ij} = 1,\ i = 1, 2, \ldots, n, \\ &\qquad x_{jj} > x_{ij},\ i = 1, 2, \ldots, n, j = 1, 2, \ldots, n, i \neq j, \\ &\qquad \sum_{j=1}^{n} x_{ii} = m \\ &\qquad x_{ij} > 0,\ i = 1, 2, \ldots, n, j = 1, 2, \ldots, n. \end{aligned} \tag{1}$$

Where a_i is the number of people initially located at point i, d_{ij} represents the distance or transit duration matrices for travelling from point i to point j, and x_{ij} is the decision variable which we optimize using PuLP optimization package 15 in Python. The solution matrix X assigns $n \times n$, and assignment is indicated by the entry 1 for each row on its assigned column and 0 for the rest. Array input a consists of the number of refugee residents for all residential regions. We set n to 200 as we have 200 candidate locations, and m to 20 because Istanbul currently hosts 20 MHCs at the time of our analysis. The optimization has been run twice because we found optimum MHC locations with both using the distance D and transit duration T matrices. Figure 3 and Figure 4 shows the produced results respectively.

3 Results

Our results are also plotted on an interactive map[6]. On the map, the dropdown menu displays the following layers: Cell Tower, Cell Tower Areas, Residential Regions, Residential Region Centers, Current MHC Locations and MHC Locations Optimized by Distance/Duration. *Cell Towers* are where calls and texts are first registered. *Cell Tower Areas* shows the regions the cell towers cover. As explained in Computational Approach and Data Pipeline section in detail, we grouped similar cell towers based on their proximity to each other and where refugees make phone calls during the night time. The centers of these cell tower regions are referred to as *Residential Region Centers* which later constituted the Voronoi cells which are referred to as *Residential Regions* in our study.

[6] The interactive map is published on `http://bit.ly/refugee_map`

4 Discussion

The availability of migrant health centers improves refugees access to healthcare. We focused on location optimization because healthcare literature finds negative correlation between the distance travelled to hospitals and health outcomes 14.

4.1 Proposing Optimal Locations for Migrant Health Centers

CDR data provided by Turk Telekom enabled us to create an alternative residential map of refugees. Refugee demographics are not registered in a census data. Instead, each municipality keeps refugees' information in their database. Their address changes might not be recorded especially given that refugees move often as they settle down in the host cities. By taking this into account, we initially took the locations where refugees spend the most time (their mode location) as their departure points for hospitals. However, we saw that it does not really match with the refugee distribution in Istanbul as seen in the study by 2

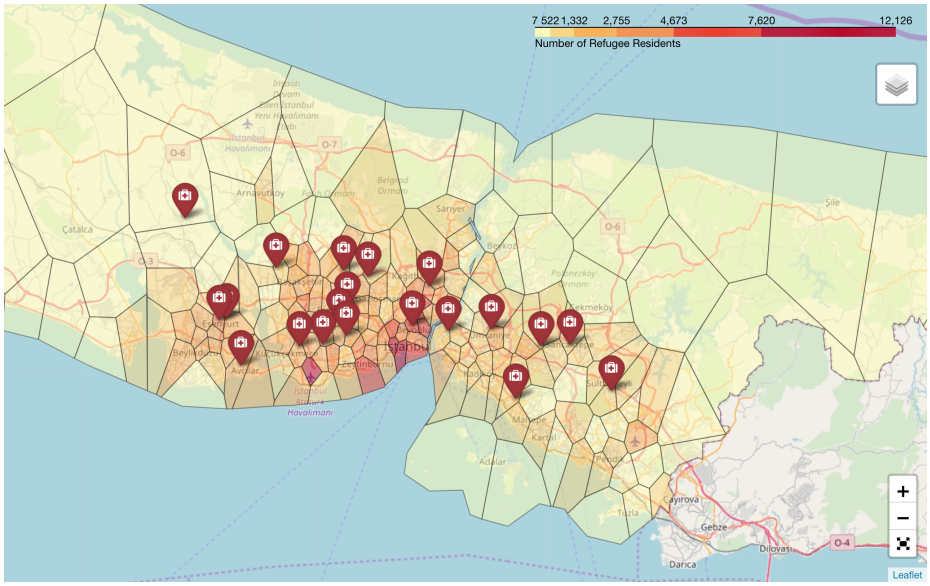


Fig. 2. Map with red pins showing the locations of current MHCs over the choropleth which illustrates the density of refugee residents at different districts in Istanbul.

Our results suggest that the mode locations are more likely to be places where refugees socialize and spend their daytime regularly, such as the Aksaray district. Thus, using their home address is a better proxy for the access available in each municipal district. However, modes we identified through our study can be used to find optimum locations for other public services, such as social integration

centers as well as employment and education facilities. Our method of estimating refugees home addresses using CDR data is explained in detail in our methods section. We calculated optimal locations for the Migrant Health Centers (shown with blue pins on Figure 3) and compared them with the existing MHC locations (indicated by red pins on Figure 2). The MHC locations our optimization suggest are not in exact coordinates, they can be positioned at any central place around the pinned region.

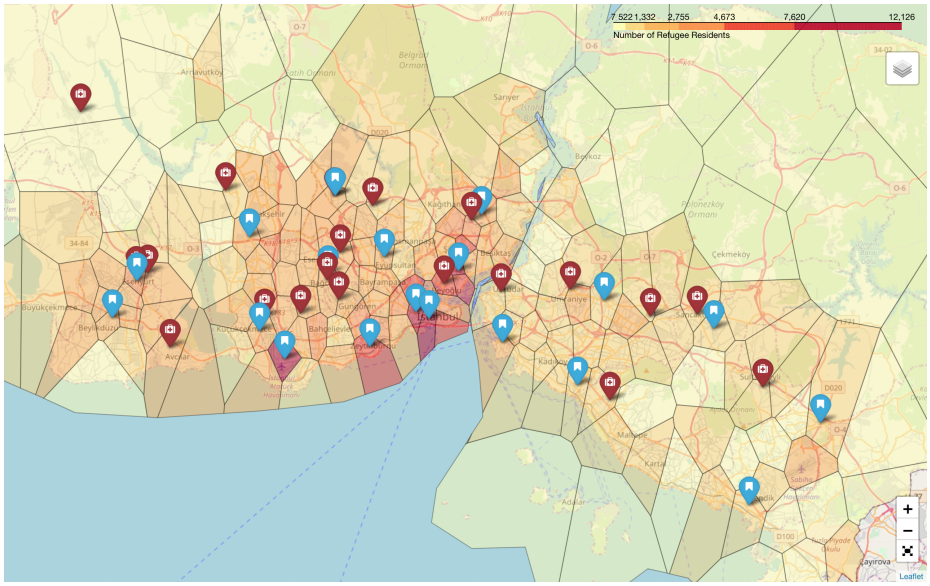


Fig. 3. Map with all current MHCs (red pins) and optimized MHC locations based on travel distances (blue pins) over choropleth map with the refugee residential densities.

As can be seen on Figure 3, the locations we suggested are sometimes very close to the already existing MHCs. Notable differences can be better observed on Figure 5). For example, our optimization suggests MHCs in Yeilky near Ataturk Airport (lower left) and near the Sabiha Gokcen Airport(middle right). We believe this is caused by the continuous call traffic in the airports that are mistakenly identified as refugee residences by our resident detection method. Another noticeable difference between optimal and current locations can be observed around historical peninsula(Fatih district) on Figure 5 . Currently, there is no MHC in the area, whereas our optimization suggests locating three MHCs there. We believe this to be a crucial insight since MHCs can provide refugees with primary and secondary healthcare services without facing language barriers and overcrowding at public hospitals.

Travelling costs in Table 1 are the sum of travelling distances and durations for all refugees in the case that each refugee visits their nearest MHC averaged by

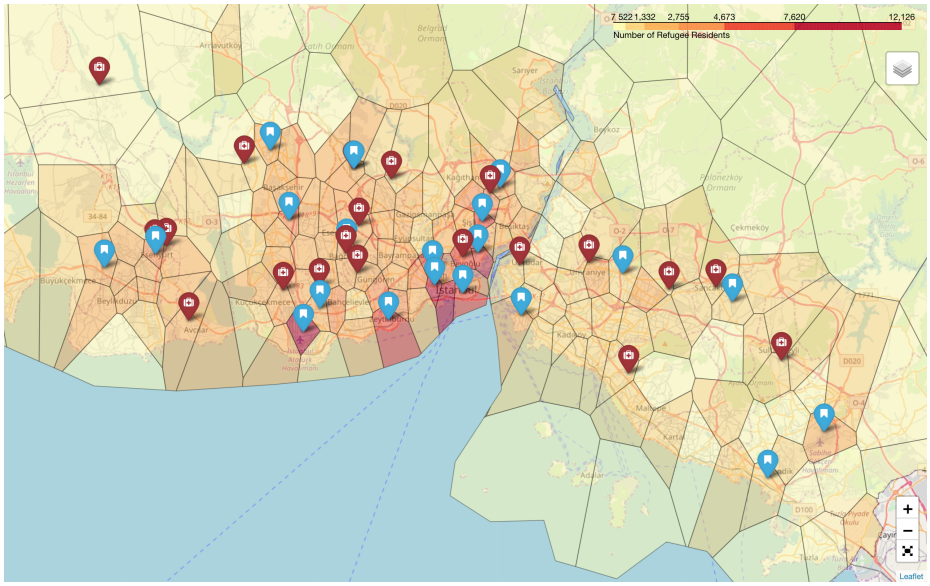


Fig. 4. Map with all current MHCs (red pins) and optimized MHC locations based on travel durations (blue pins) over choropleth map with the refugee residential densities.

the total number of refugees who we can detect their residencies. Cost of travel distances and durations are shown to represent an average cost for one refugee to access their nearest MHC for three cases: Current MHC locations, distance based optimized MHC locations and duration based optimized MHC locations. Besides the current MHC locations, we present the overall performance of our suggested MHC locations which have been optimized to minimize the travel cost on the basis of distance and duration.

Overall, we found that the average travelling cost to MHCs is 5.9 km for one person and it takes approximately 26 minutes in the present case. The locations we proposed based on refugees activities from CDR data cuts the travel distance down to as little as 3.6 km in distance based optimization, and travelling time to as little as 18 minutes for duration based optimization.

Table 1. Cost of travel distances and durations for current MHC locations and optimized MHC locations (on the basis of distance and duration). All distances are provided in kilometers and durations in minutes.

MHC Locations	Travel Distance	Travel Duration
Current	5.9	26
Optimized (Distance)	3.6	20
Optimized (Duration)	4.4	18

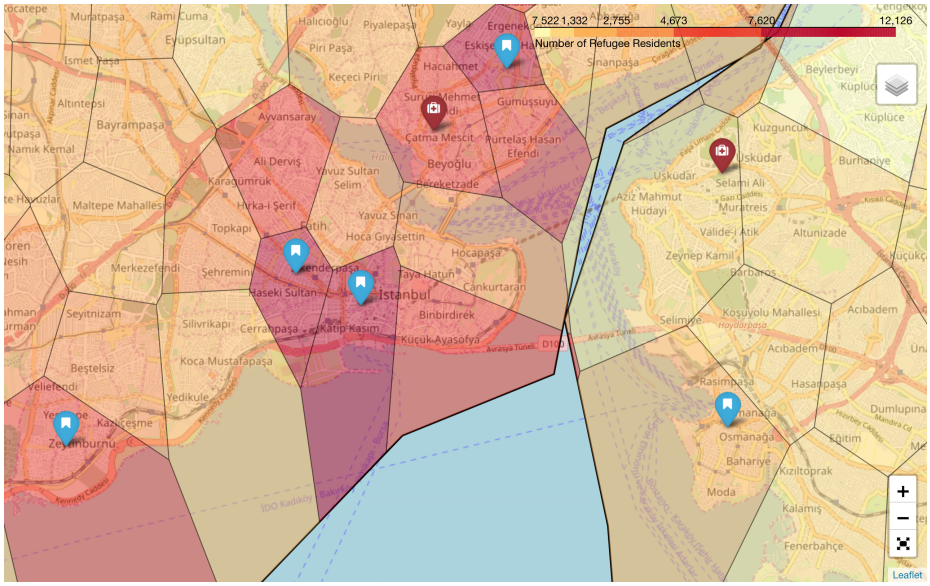


Fig. 5. A close-up view of central Istanbul where optimized locations based on travel distances (blue pins) show particular districts where new MHCs can be opened

5 Conclusion

One of our biggest challenges during this study was the lack of health care service data available for public use. Another major limitation we can define for this study is the representativeness of the sample. The CDR data shows that 85% of the total activities of refugees in Istanbul are in Europe, while 15% is in Asia. We suppose that this discrepancy could be because of the data's inherent bias which is that the refugee population which uses Turk Telekom does not represent the whole refugee population.

Our paper proposes the following:

- Opening Migrant Health Centers by optimizing travel time, in Istanbuls case in districts such as Zeytinburnu, Fatih, Kadikoy, and as shown in Figure 5
- Our study focuses on improving access to healthcare services for refugees in Istanbul. However, our approach is applicable and scalable to other cities.
- When it is not possible to open a new MHC or move existing MHCs to more optimal locations, translation services that will help refugees with making appointments and communicating in hospitals can be offered near the optimal locations we proposed.
- We strongly encourage that refugees are directed to public hospitals instead of research or university hospitals that are already overcrowded.

References

Halk Sağlğ Genel Müdürlüğü, https://hsgm.saglik.gov.tr/tr/

Akgün, G., Atlar, A.B., Balyemez, S., Çalşkan, .O., Çekiç, T.., Çlgn, K., Çolak, N., Doğançayr, C.M., Koca, A., Kurtarr, E., Kutluca, A.K., Pak, E.., Ölmez, M., Sungur, C., Ykc, A., Ylmaz, E.: Timeline of Syrian Refugees [In Turkish]. In: Çekiç, T.., Pak, E.., Ylmaz, E. (eds.) Istanbul Kent Almanag 2015, pp. 37–47. TMMOB Şehir Planclar Odas stanbul Şubesi, stanbul (2016)

Anderson, C.: Clustering the US population: observation-weighted k-means (2016), https://towardsdatascience.com/clustering-the-us-population-observation-weighted-k-means-f4d58b370002

Arthur, D., Vassilvitskii, S.: K-means++: The Advantages of Careful Seeding. In: Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms. pp. 1027–1035. SODA '07, Society for Industrial and Applied Mathematics, Philadelphia, PA, USA (2007), http://dl.acm.org/citation.cfm?id=1283383.1283494

Aurenhammer, F.: Voronoi diagrams—a survey of a fundamental geometric data structure. ACM Computing Surveys **23**(3), 345–405 (9 1991). https://doi.org/10.1145/116873.116880, http://portal.acm.org/citation.cfm?doid=116873.116880

Bahadir, H., Ucku, R.: A descriptive field research: Syrian Refugee Women's Reproductive Health, Izmir. [In Turkish]. In: ŞAŞMAZ, C.T. (ed.) Book of Proceedings of 18th National Public Health Congress. pp. 1008–1009. HASUDER Yaynlar, Konya (2015), http://halksagligiokulu.org/anasayfa/components/com_booklibrary/ebooks/18_UHSK_KONGRE_KITABI.pdf

Baş, D., Arkant, C., Muqat, A., Arafa, M., Sipahi, T., Eskiocak, M.: The Circumstances of Syrian Refugees in Edirne. [In Turkish]. In: ŞAŞMAZ, C.T. (ed.) Book of Proceedings of 18th National Public Health Congress. pp. 214–215. HASUDER Yaynlar, Konya (2015), http://halksagligiokulu.org/anasayfa/components/com_booklibrary/ebooks/18_UHSK_KONGRE_KITABI.pdf

Bilecen, B., Yurtseven, D.: Temporarily protected Syrians access to the healthcare system in Turkey: Changing policies and remaining challenges. Migration Letters **15**(1), 113–124 (1 2018), https://ideas.repec.org/a/mig/journl/v15y2018i1p113-124.html

Demissie, M.G., Antunes, F., Bento, C., Phithakkitnukoon, S., Sukhvibul, T.: Inferring origin-destination flows using mobile phone data: A case study of Senegal. In: 2016 13th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON). pp. 1–6 (6 2016). https://doi.org/10.1109/ECTICon.2016.7561328

Diker, E.: Bibliographies on Syrian Refugees in Turkey: Health (2018), https://mirekoc.ku.edu.tr/wp-content/uploads/2016/05/Mirekoc_Bibliographies_on_Syrian_Refugees_in_Turkey_30.07.2018-WEB.pdf

Ester, M., Kriegel, H.P., Sander, J., Xu, X.: A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise (1996), https://dl.acm.org/citation.cfm?id=3001507

Gulacti, U., Lok, U., Polat, H.: Emergency department visits of Syrian refugees and the cost of their healthcare. Pathogens and Global Health **111**(5), 219–224 (2017). https://doi.org/10.1080/20477724.2017.1349061, https://doi.org/10.1080/20477724.2017.1349061

Isaacman, S., Becker, R., Cáceres, R., Kobourov, S., Martonosi, M., Rowland, J., Varshavsky, A.: Identifying Important Places in People's Lives from Cellular Network Data. In: Lyons, K., Hightower, J., Huang, E.M. (eds.) Pervasive Computing. pp. 133–151. Springer Berlin Heidelberg, Berlin, Heidelberg (2011)

Kelly, C., Hulme, C., Farragher, T., Clarke, G.: Are differences in travel time or distance to healthcare for adults in global north countries associated with an impact on health outcomes? A systematic review (11 2016). https://doi.org/10.1136/bmjopen-2016-013059, https://bmjopen.bmj.com/content/6/11/e013059.citation-tools

Mitchell, S., Consulting, S.M., Dunning, I.: PuLP: A Linear Programming Toolkit for Python (2011)

Ozdogan, H.K., Karateke, F., Ozdogan, M., Satar, S.: Syrian refugees in Turkey: effects on intensive care. Lancet (London, England) **384**(9952), 1427–8 (10 2014). https://doi.org/10.1016/S0140-6736(14)61862-6, http://www.ncbi.nlm.nih.gov/pubmed/25390324

ReVelle, C.S., Swain, R.W.: Central Facilities Location. Geographical Analysis. **2**(1), 30–42 (1970)

Salah, A., Pentland, A., Lepri, B., Letouzé, E., Vinck, P., de Montjoye, Y., Dong, X., Dağdelen, .: Data for Refugees: The D4R Challenge on Mobility of Syrian Refugees in Turkey (7 2018), https://arxiv.org/abs/1807.00523http://d4r.turktelekom.com.tr/presentation/project-evaluation-committee

Savas, N., Arslan, E., nand, T., Yeniçeri, A., Erdem, M., Kabacaoğlu, M., Peker, E., Alşkn, .: Syrian refugees in Hatay/Turkey and their influence on health care at the university hospital. Int J Clin Exp Med **9**(9), 18281–18290 (2016), http://www.ijcem.com/files/ijcem0027188.pdf

Tongsinoot, L., Muangsin, V.: Exploring Home and Work Locations in a City from Mobile Phone Data. In: 2017 IEEE 19th International Conference on High Performance Computing and Communications; IEEE 15th International Conference on Smart City; IEEE 3rd International Conference on Data Science and Systems (HPCC/SmartCity/DSS). pp. 123–129 (12 2017). https://doi.org/10.1109/HPCC-SmartCity-DSS.2017.16

UNHCR: UNHCR Turkey: Key Facts and Figures. Tech. Rep. November, UNHCR (2016), https://data2.unhcr.org/en/documents/details/52526

Zhi-Dan, Z., Hu, X., Shang, M.S.: Uncovering individual and collective human dynamics from mobile phone records. J. Phys. A: Math. Theor **41**, 224015 (2008). https://doi.org/10.1088/1751-8113/41/22/224015, http://iopscience.iop.org.iclibezp1.cc.ic.ac.uk/article/10.1088/1751-8113/41/22/224015/pdf

Abbreviations MHC: Migrant Health Center BTS: Base Transceiver Station CDR: Call Detail Record SUKOM: Syrian Coordination Center Software

Authorś Contributions: Muhammed Tark Altuncu conducted the computational research. Nur Sevencan and Aye Seyyide Kaptaner were responsible for the social science research and obtaining additional data. All authors analysed the data and wrote the manuscript.

Acknowledgements: We thank the rest of the members of TRT World's team for the Data for Refugees challenge: Basri Ciftci, Berk Baytar, Soud Hyder, and Yasin Sancaktutan. We also thank Mehmet Efe Akengin for providing intellectual support, and Hamza Osmanogullari for logistic support.

Social Integration of Syrian Refugees: Some Insights from Call Detail Record Datasets

Nuran Bayram-Arli[1[0000-0001-5492-184X]], Fatih Cavdur[2[0000-0001-8054-5606]], Mine Aydemir[1[0000-0003-3276-8148]], Fadime Aksoy[3[0000-0003-0211-8304]], Asli Sebatli[2[0000-0002-9445-6740]]

[1] Bursa Uludag University, Department of Econometrics, Bursa, Turkey
[2] Bursa Uludag University, Department of Industrial Engineering, Bursa, Turkey
[3] Bandirma Onyedi Eylul University, Department of Econometrics, Balikesir, Turkey
fatihcavdur@uludag.edu.tr

Abstract. Data for Refugees Turkey is a big data challenge aiming at providing anonymized mobile phone usage data to researchers for the purposes of providing better living conditions for the Syrian refugees in Turkey. The challenge invites researchers to create solutions on one of the five prioritized subject areas. Our research team focuses on the social integration of Syrian Refugees as one of these areas and presents some insights from Call Detail Record datasets provided for the challenge. This study presents a summary of our work. As detailed in the study, we first import the datasets to a sophisticated database server for being able to perform data operations efficiently. Our resulted database includes the tables of the three main datasets (i.e., the datasets about antenna traffic, fine-grained mobility and coarse-grained mobility) as well as the ones about base station locations, district locations, city mapping and district mapping. We then design and implement several queries for gathering social integration information of the refugees. The results of the queries are analyzed to gain some insights.

Keywords: Social Integration, Migration, Refugees, Mobility, Call Detail Record, Data Processing, Big Data.

1 Introduction

As a result of the ongoing humanitarian crisis in Syria which was sparked in 2011, the number of Syrian refugees hosted in Turkey has reached up to 3.1-3.2 million. 4% of Turkey's population is now made up of Syrian refugees. This massive migration wave has been one that is hard to contain. Although 26 camps with a total of 270,000 beds were built on Turkey's border with Syria, the refugees started to move into the country from border areas. Today, only 8% of Syrian refugees live in refugee camps in Turkey while the others started a new life in several cities (UNHCR, 2018).

Turkey currently is the country which hosts the highest number of refugees. Syrian refugees in Turkey have become an issue which needs attention with social, political, economic dimensions, not to mention their integration and security. Being able to

ensure successful integration of the refugees will contribute to a rich and multicultural society in the medium and long-term. The improved social relationship in the society will lead to political and economic cooperation between the refugees and the local residents. Therefore, the case of Syrian refugees can be explored as a case of social integration (ORSAM, 2015).

Data for Refugees (D4R) Turkey is a big data challenge aiming at providing anonymized mobile phone usage data to researchers for the purposes of providing better living conditions for the Syrian refugees in Turkey. The challenge invites researchers to create solutions on one of the five prioritized subject areas. Our research team focuses on the social integration of Syrian Refugees as one of these areas and presents some insights from the Call Detail Record (CDR) datasets provided for the challenge. This study presents a summary of our work.

As detailed in the following sections of the study, we first import the CDR datasets to a sophisticated database server for being able to perform data operations efficiently. Our resulted database includes the tables of the three main datasets (i.e., the datasets about antenna traffic, fine-grained mobility and coarse-grained mobility) as well as the ones about base station locations, district locations, city mapping and district mapping. We then design and implement several queries for gathering social integration information of the refugees. The results of the queries are analyzed to gain some insights.

The organization of the study is as follows. In the next section, we summarize some basic concepts about social integration. In Section 3 a brief description of the D4R datasets is presented. Section 4 describes the methodology and Section 5 presents our findings. Final remarks are presented in the last section of the study.

2 Social Integration

Human beings have long been migrating from one location to another. Migration, in the modern sense, has started in the 19th century which in turn made it necessary for legislation relevant regulations governing the immigrants turning into a political concept (Torpey, 2000). One of the common measures taken by nations in the face of "intense migration" is the social, political, economic and cultural integration of the immigrants into the host society. The purpose of integration is to ensure that the immigrants become a part of the host society without discrimination and enjoy equal rights as well as to allow them to create a shared-culture combining their own culture and the culture of the host society. Integration can be defined as the process of becoming an accepted member of the society.

Multiculturalism has found its reflections in the integration of immigrants. Multiculturalism helps a society become heterogeneous given that it was homogeneous once, and the nations which take the most immigrants respect immigrants' needs and allow them to honor their culture as they consider multiculturalism as a value added to their own culture (Entzinger, 2000). According to Lockwood (1964) there are two types of integration: "social integration" and "system integration." System integration refers to a collaborative approach between institutions, mechanisms and corporations

available in a society, while social integration refers to the inclusion of individuals into the system, interpersonal relationships and the individuals' attitude towards the society.

Esser (1999) identifies four dimensions of social integration: acculturation, placement, interaction, and identification. Acculturation is the process by which an individual acquires the knowledge, cultural standards and skills needed to interact and communicate successfully in a society. Placement refers to an individual gaining a position in society. Interaction refers to the formation of relationships while identification refers to an individual's identification with a social system, both intellectually and emotionally.

According to Heckmann and Schnapper (2003), immigrant integration is a specialized form of social integration and it must be explored in terms of structural, cultural, interactive and identification integration. Structural integration is the acquisition of rights and the access to position and status in the core organs of the host society. Among these core organs are education, health, social security, labor market, economy and politics. Cultural integration is the ability of the immigrants to claim rights and assume positions in their new society if they acquire the core skills of that culture and society. Cultural integration does not necessarily mean that immigrants will have to give up the culture of their home country. Multiculturalism approach can be an asset both for the immigrant and for the host society. Interactive integration means the acceptance and inclusion of immigrants in the primary relationships and social networks of the host society and among the indicators of interactive integration are social networks, friendships, partnerships, marriages and membership in voluntary organizations. Identity integration is the feeling of belonging a person may develop later in the integration process as a result of participation and acceptance.

Social integration of refugees into to the host community is not an easy process which takes time and requires a detailed analysis (Strang and Ager, 2010; Spencer, 2011, Yildiz and Uzgoren, 2016). It is shown in some studies that the integration process might be quite efficient for the refugees and the host community in areas where the population densities are relatively low and the same linguistic origin is shared by the refugees and the host community (Leach, 1992; Bakewell, 2000, 2002). Some studies show that the refugees and the members of the host community have different social relationships among themselves in different types of social networks which might make contributions to the process of social integration (Campbell and Lee, 1992; Atfield et al., 2007; Porter et al., 2008; Robinson, 2010; Cheung and Philimore, 2013). On the other hand, providing special areas for the refugees might prevent them from integrating into the host community (Bosswick and Heckmann, 2006; Bayram et al., 2009; Platts-Fowler and Robinson, 2015; Sonmez, 2016).

3 Datasets

This section presents a brief description of the D4R data. For more details, the interested reader can refer to the study of Salah et al. (2018). The D4R data are collected from 992,457 customers of Turk Telekom, consisting of 184,949 customers tagged as

refugees and 807,508 Turkish citizens (Salah et al., 2018). In terms of the gender distribution of the customers, it is noted that 75% of the refugee-tagged customers are recorded as male, and 25% as female. The same gender distribution is preserved for the customers of Turkish citizens in the D4R data (Salah et al., 2018).

It is noted that, compared to the others, significantly more refugees live in some cities in Turkey. In this study, we consider nine of such cities (from now on, referred as major cities) with respect to its refugee population presented in the study of Salah et al. (2018) for the sake of consistency as they are based on the D4R data. In addition, the cities in the country have a number of differences in terms of some metrics such as social structure, ethnic background, economic status etc. Accordingly, the reflection of the level of social integration of the refugees might vary depending on the city of interest. Therefore, it is necessary to explore each city individually instead of generalizing the situation to a nationwide level.

A list of these cities with the numbers and percentages of customers in them is given in Table 1. More detailed statistics are presented by the Turkish Statistical Institute recently (TUIK, 2018). Similarly, the customers of Turkish citizens are also sampled mainly from the same cities. Each city is also represented with its city code for convenience as we use city codes rather than their names in the figures presented in the study for readability. A map of the distribution of the refugees in most of these cities (including Adana and Kilis, excluding Ankara) based on the data of the Ministry of Interior - Directorate General of Migration Management is shown in Figure 1 (Salah et al., 2018).

Table 1. The distribution of customers tagged as refugees and their registered locations. Numbers rounded to the third significant digit (Salah et al., 2018).

City	Number of Customers	Percentage of Customers
34-Istanbul	84,173	45.511
27-Gaziantep	14,898	8.055
35-Izmir	10,425	5.637
63-Sanliurfa	9,701	5.245
33-Mersin	9,660	5.223
31-Hatay	7,024	3.798
06-Ankara	5,580	3.017
42-Konya	4,718	2.551
16-Bursa	3,479	1.881
Outside Turkey	2,902	1.569
Other	32,440	17.540

In addition to the other datasets of Base Station Locations, District Locations, City Mapping and District Mapping, there are three main datasets provided by Turk Telekom for the challenge about Antenna Traffic (Dataset-1), Fine-Grained Mobility (Dataset-2) and Coarse-Grained Mobility (Dataset-3) for the year of 2017 as detailed in the following paragraphs.

3.1 Dataset-1: Antenna Traffic

The first dataset provided by Turk Telekom includes one-year site-to-site traffic on an hourly basis. This dataset contains the traffic between each site for a year. Calls between Turk Telekom customers and other service providers only have information about the Turk Telekom side. For each record, total number and duration of calls are recorded in an aggregated fashion. The dataset is split into voice and SMS partitions where each file contains the voice or SMS data of a month of the year of 2017 (Salah et al., 2018).

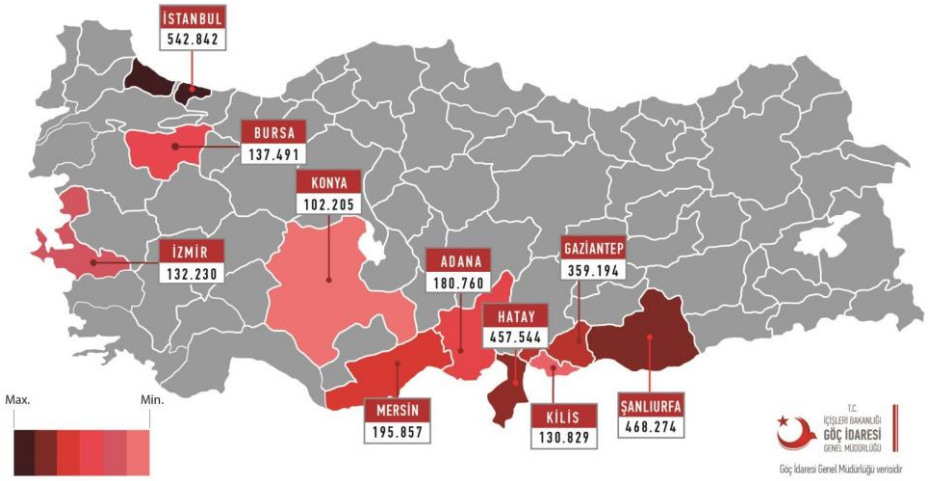


Fig. 1. Distribution of refugees in the country based on the data of the Ministry of Interior, Directorate General of Migration Management (Salah et al., 2018).

3.2 Dataset-2: Fine-Grained Mobility

The dataset contains cell tower identifiers used by a group of randomly chosen active users to make phone calls and send texts. The data are timestamped and a particular group of users is observed for a period of two weeks. At the end of the two-week period, a fresh sample of active users is drawn at random. The users are represented by random digits in the dataset without any personal information. New random identifiers are chosen for every two-week period to protect privacy which, on the other hand, makes it impossible to analyze the mobility of a particular user. Missing antenna locations are represented with the codes of -99 or 9999. This dataset is also separated into voice and SMS partitions. It is further divided into files containing incoming (in) and outgoing (out) calls to deal with large files, resulting in four files per 15-day period (Salah et al., 2018).

3.3 Dataset-3: Coarse-Grained Mobility

In this dataset, the trajectories of a randomly selected subset of users are provided for the entire observation period with reduced spatial resolution. The country is divided into the electoral prefectures, and for each call record, only the prefecture information is provided. The IDs are randomly assigned, and it is noted that two different users may have the same ID in Dataset-2 and Dataset-3. The dataset is also split into incoming (in) and outgoing (out) calls to deal with large files (Salah et al., 2018).

3.4 Other Datasets: Base Station & District Locations and District & City Mappings

There are a few more datasets provides such as Base Station Locations, District Locations, District Mapping and City Mapping. The cell tower (i.e., base station) locations are provided in a text file. In some rare cases, the precise location information of the base station is missing, only the city is indicated. The interpretation of the latitude and longitude follows degree, minutes, seconds (DMS) syntax (Salah et al., 2018). Another file that contains district coordinates with the fields of city, district, population, 2D and DMS coordinates is also provided to disambiguate the base stations (Salah et al., 2018). Finally, each of the remaining two files contains information about district mapping and city mapping, respectively.

4 Methodology

This section details the methodology used in the study. We use a sophisticated database server (Microsoft SQL Server) to perform all data operations considering some issues of efficiency, performance and robustness. The first step thus is importing all plain text data files to the server. As summarized in the previous section and detailed in the study of Salah et al. (2018), in addition to two other datasets of Base Transceiver Station and District Locations, there are three main datasets provided by Turk Telekom for the challenge about Antenna Traffic (Dataset-1), Fine-Grained Mobility (Dataset-2) and Coarse-Grained Mobility (Dataset-3) for the year of 2017. It is also noted that the first datasets are further partitioned into files of SMS or voice data. Our database, as a result, contains the corresponding tables representing these five datasets as follows:

1. TableDS1SMS for Antenna Traffic (Dataset-1) SMS data.
2. TableDS1Voice for Antenna Traffic (Dataset-1) Voice data.
3. TableDS2SMS for Fine-Grained Mobility (Dataset-2) SMS data.
4. TableDS2Voice for Fine-Grained Mobility (Dataset-2) SMS data.
5. TableDS3 for Coarse-Grained Mobility (Dataset-3) data.
6. TableBSL for Base Station Locations data.
7. TableDL for District Locations data.
8. TableDM for District Mapping data.
9. TableCD for City Mapping data.

Although each of the Base Station Locations, District Locations, District Mapping and City Mapping datasets is provided in a single plain text file, it is noted that the three main datasets are originally split into smaller files representing a partition of the corresponding dataset ("Dataset 1 SMS 2017XX.txt", for instance, contains the SMS data for month XX). The total number of such files for the first three datasets is 148 (24 for Antenna Traffic (Dataset-1), 100 for Fine-Grained Mobility (Dataset-2) and 24 Coarse-Grained (Dataset-3)) to deal large files. As a result, for Antenna Traffic Dataset (Dataset-1), for instance, instead of a single table (Table_DS1_SMS), there are 12 tables for Antenna Traffic (Dataset-1) SMS data. We use some SQL procedures to combine such files into a single larger file. Note that while combining files, we also store the name of the split file (i.e., the SourceTableName field of each of the combined tables) so that we do not lose any information of the datasets.

After the multiple tables of the datasets of Antenna Traffic (Dataset-1), Fine-Grained Mobility (Dataset-2) and Coarse-Grained Mobility (Dataset-3) are combined, we then design and run several queries to retrieve the desired records in the database. The queries run on the three datasets are expressed in their respective subsections. Several generic queries designed in the study are detailed in the corresponding subsections for the datasets of Antenna Traffic (Dataset-1), Fine-Grained Mobility (Dataset-2) and Coarse-Grained Mobility (Dataset-3), respectively. Note that the queries are presented in their generic forms and each of them is run as many times as the total number of parameter settings it has. In general, we apply the following general rules in query designs:

1. All unknown locations are ignored in all queries (i.e., sites with IDs of -99 and 9999)
2. All unknown callers are ignored in all queries (i.e., callers with prefixes of 3).

The details of the queries are presented as much as possible considering the page limitation of the report in the following subsections for the datasets of Antenna Traffic (Dataset-1), Fine-Grained Mobility (Dataset-2) and Coarse-Grained Mobility (Dataset-3), respectively, along with their relationships with the other datasets (Base Station & District Locations and District & City Mappings) where necessary. Although all queries are implemented using the Structured Query Language (SQL), we use a schematic representation to explain the queries rather than the SQL codes as detailed in the following subsections.

4.1 Dataset-1: Antenna Traffic

There are two different types of queries for the dataset of Antenna Traffic (Dataset-1) in terms of their hierarchical levels. The queries of the first type retrieve the records in all cities in the country for a particular type of activity (i.e., SMS, voice, in, out) whereas the queries of the second type produce detailed records in a particular city.

In this subsection, we present the details of SMS-In and Voice-In queries for the dataset of Antenna Traffic (Dataset-1). The interested reader can derive the corresponding queries of SMS-Out and Voice-Out in a similar manner.

Figure 2 shows the schematic representation of the query that lists the total number of messages and total number of refugee messages all cities receive. The results are grouped by cities in descending order.

We can explain the structure of a schematic query using the query shown in Figure 2 since we have a similar structure for all of the schematic query representations in the study. It is noted from the figure that the main table (i.e., TableDS1SMS) is located on the left-hand-side whereas we see the auxiliary table(s) (i.e., TableBSL) on the right-hand-side of the figure. In the middle section of the figure, we see the query name (above) and its brief description (below). We further note that the red fields in a particular table are used for establishing the relationship between the tables (i.e., join operation). Additionally, the black and grey ones represent the fields that are returned and not returned, respectively. We note that some criteria are applied on the blue-colored fields such as retrieval of the records satisfying certain conditions. It is also noted that query parameter(s) are presented in brackets in the query description.

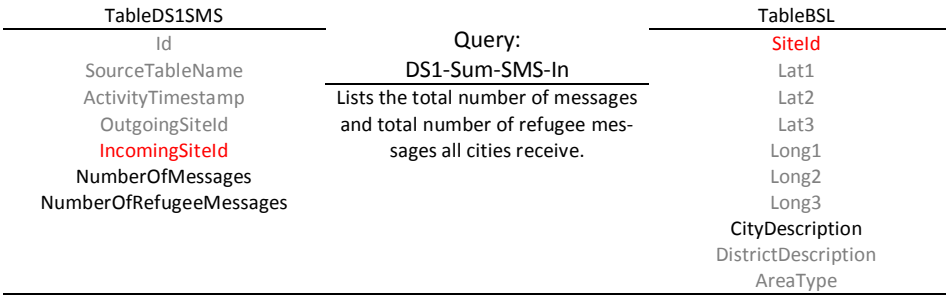


Fig. 2. Schematic description of query DS1-Sum-SMS-In.

Similar interpretations are also valid for all schematic query representations in this study. Figure 3 shows the second query for the dataset of Antenna Traffic (Dataset-1) that lists all messages *a city* receives. This query results a detailed list of messages for a city.

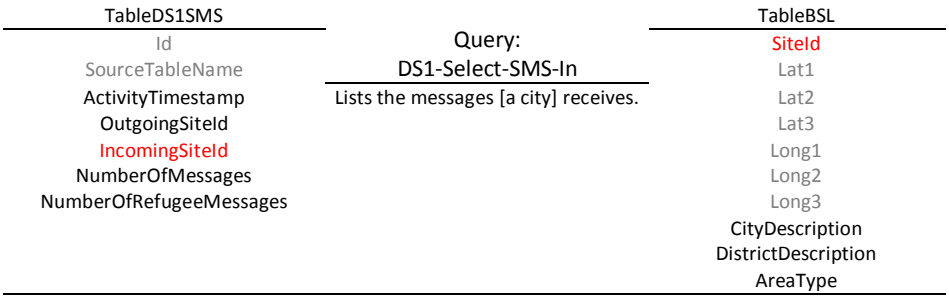


Fig. 3. Schematic description of query DS1-Select-SMS-In.

Note that the previous queries are related to SMS messages. The remaining queries presented in this subsection are the voice counterparts of the first two queries. Figure

4 represents the query that lists the total number of calls and call durations *a city* receives. The final query presented in this subsection lists all calls all cities receive as shown in Figure 5.

We note that two other queries designed and run in the study which belongs to this subsection are not presented in the report to keep the current length of the study. On the other hand, these queries can be easily derived from the ones presented in this subsection. In other words, the queries presented in this subsection all retrieve incoming records (i.e., the queries DS1-Sum-SMS-In, DS1-Select-SMS-In, DS1-Sum-Voice-In and DS1-Select-Voice-In). Similarly, the interested reader can easily derive the remaining queries we implement as the queries DS1-Sum-SMS-Out, DS1-Select-SMS-Out, DS1-Sum-Voice-Out and DS1-Select-Voice-Out, respectively.

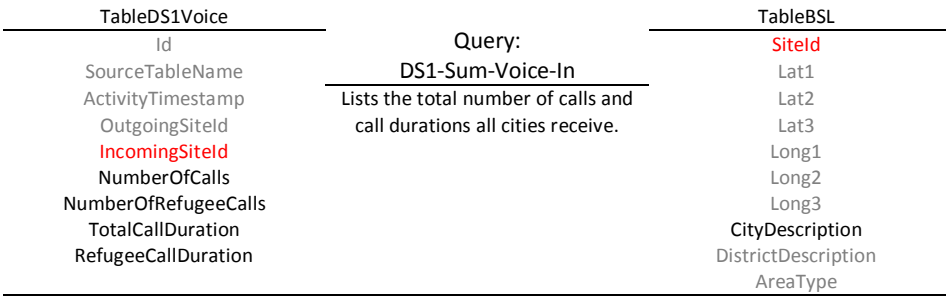


Fig. 4. Schematic description of query DS1-Sum-Voice-In.

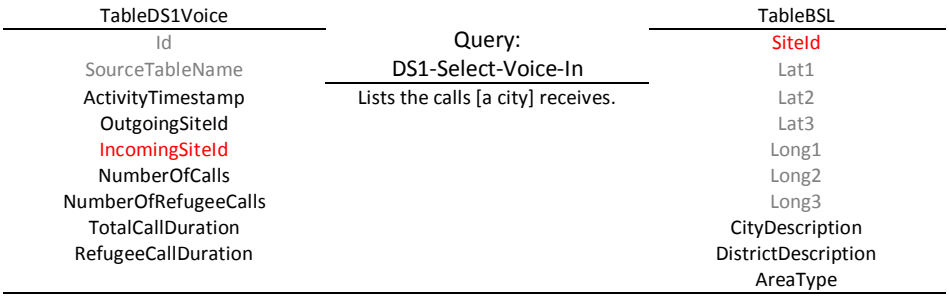


Fig. 5. Schematic description of query DS1-Select-Voice-In.

4.2 Dataset-2: Fine-Grained Mobility

The queries of this subsection are implemented to retrieve the records of refugee customers. Two of the queries implemented for the dataset of Fine-Grained Mobility (Dataset-2) are presented in this subsection. The first one in Figure 6 represents the query that retrieves the total number of messages all refugee customers in *a city* receive. The query can be used to analyze the mobility of any refugee customer in the dataset by retrieving the detailed records of the corresponding person.

The second query of this subsection lists the messages of a customer in terms of the number of SMS messages the customer receives. The query can be used to ana-

lyze the mobility of the most active refugee customers in terms of the number of SMS messages they receive as shown in Figure 7. As expressed in the previous subsection, some other queries implemented in the study which belong to this subsection are not presented here to keep the current length of the study. We again note that these queries which are not presented here can easily be derived by the interested reader.

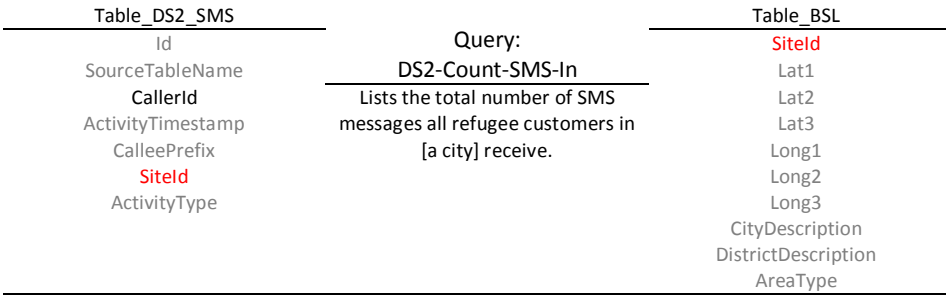


Fig. 6. Schematic description of query DS2-Count-SMS-In.

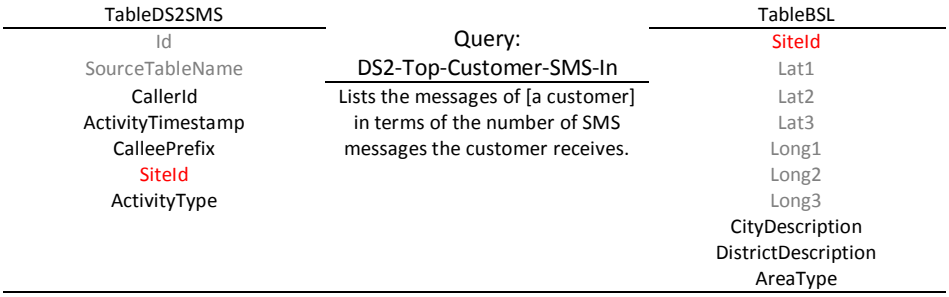


Fig. 7. Schematic description of query DS2-Top-Customer-SMS-In.

In other words, the queries presented in this subsection are all related to incoming SMS messages (i.e., DS2-Count-SMS-In and DS2-Top-Customer-SMS-In) which can be easily modified to obtain the voice counterparts of the queries as DS2-Count-Voice-In and DS2-Top-Customer-Voice-In since the table designs both for SMS and voice are the same. Similarly, the outgoing versions of the queries for messages and calls can also be derived as DS2-Count-SMS-Out and DS2-Top-Customer-SMS-Out for the SMS and DS2-Count-Voice-Out and DS2-Top-Customer-Voice-Out for the voice version, respectively.

4.3 Dataset-3: Coarse-Grained Mobility

In this subsection, we also present two of the queries designed for the dataset of Coarse-Grained Mobility (Dataset-3) and left the derivation of the remaining queries to the reader. Figure 8 represents the query that lists the total number of mobility activities of refugees all cities receive. This query can be used to analyze the general mobility activities of the refugee customers in a city.

The second query shown in Figure 9 lists the total number of mobility activities of refugees all districts of *a city* receives. Similarly, this query can be used to analyze the general mobility activities of the refugee customers in the districts of a city.

We finally note that the queries presented in this subsection are only related to incoming mobility activities (i.e., DS3-Count-In and DS3-Detailed-Count-In). Although some other queries implemented in the study which belong to this subsection are not presented here to keep the current length of the study, the corresponding queries can easily be derived by the interested reader for listing the outgoing mobility activities as DS3-Count-Out and DS3-Detailed-Count-Out, respectively.

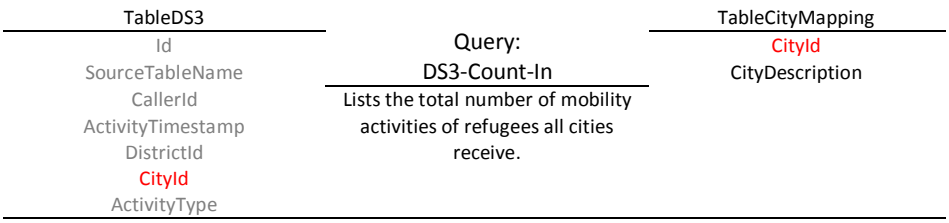


Fig. 8. Schematic description of query DS3-Count-In.

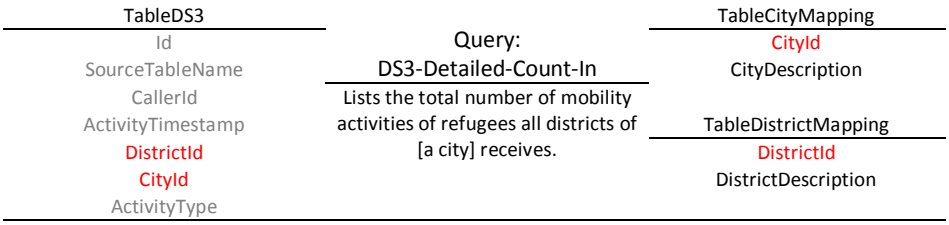


Fig. 9. Schematic description of query DS3-Detailed-Count-In.

5 Findings

This section summarizes our findings by implementing the queries of the previous section. We present our finding in the following three subsections summarizing the results obtained from the datasets of Antenna Traffic (Dataset-1), Fine-Grained Mobility (Dataset-2) and Coarse-Grained Mobility (Dataset-3), respectively. We aim at presenting our findings by summarizing the results using appropriate visualizations rather than presenting all details to keep the current length of the study.

5.1 Dataset-1: Antenna Traffic

This subsection presents our findings about the dataset of Antenna Traffic (Dataset-1) by implementing the queries in the corresponding subsection (Section 4.1) of the previous section. As explained in Section 4.1, these queries are designed and implemented to analyze the mobility activities of the refugees both in each of the major cities and between them. Using the corresponding queries, we retrieve the mobility activi-

ties of the refugees in each of the major cities and between them. The distributions of different types of mobility activities in the major cities along with their refugee percentages are shown in Figure 10.

Similarly, we present the distributions of different types of mobility activities between 34-Istanbul (as an example major city with the largest number of refugees) and the other major cities in Figure 11. Additionally, we also construct the corresponding distributions for the eight other major cities; however, they are not presented in the study to keep its current length.

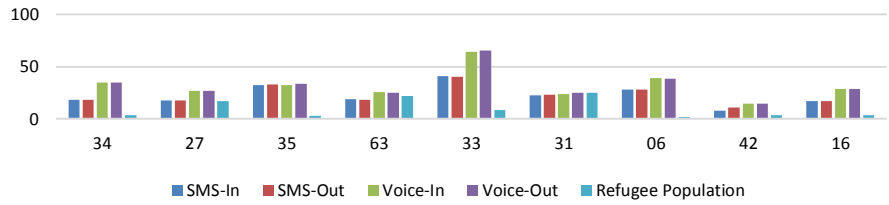


Fig. 10. Distributions of different types of mobility activities in major cities with refugee populations (percentage)

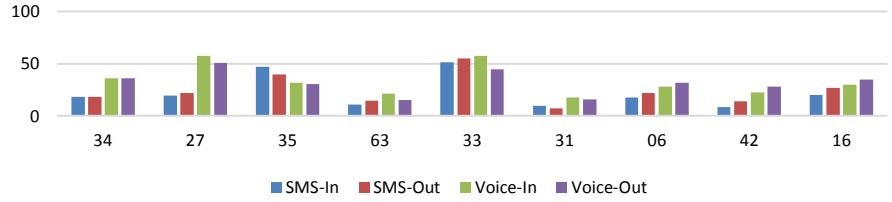


Fig. 11. Distributions of different mobility activities between Istanbul and other major cities (percentage)

In Figure 10, we note that although the percentages of refugee populations in 27-Gaziantep, 63-Sanliurfa and 31-Hatay are the highest, the mobility activities in these cities are lowest compared to the activity levels in other major cities with respect to their refugee population percentages. We can consider 31-Hatay, as an extreme case for instance, with its 24.69% refugee population. It is noted that the percentage of both incoming and outgoing calls as well as incoming calls are below its refugee population percentage (i.e., 22.60%, 23.14% and 23.77%, respectively) whereas the percentage of outgoing calls is slightly higher than the refugee percentage of the city (i.e., 24.91%). On the other hand, significantly higher levels of mobility activities of the refugees in other cities are observed (i.e., in 34-Istanbul, 35-Izmir, 33-Mersin, 06-Ankara, 42-Konya and 16-Bursa). Among these cities, although the highest levels are observed in 33-Mersin, a relative comparison with respect to the refugee populations in the cities shows that the level of the mobility activities of the refugees in 06-Ankara

is the highest (i.e., its refugee population of 1.37% whereas it is 8.28% in 33-Mersin). It might be due to the fact that 06-Ankara is the capital where bureaucratic procedures could be completed faster and some organizations (such as the agencies of the United Nations (UN), the International Labour Organization (ILO) and the International Monetary Fund (IMF)) are located in this city. Nevertheless the overall high levels of mobility activities in Mersin might be related to the commercial activities the refugees are involved in the city (ORSAM, 2015).

As an example major city with the largest number of refugees, the distributions of different types of mobility activities between 34-Istanbul and the other major cities are shown in Figure 11. Similar graphs for other major cities are also available on request; however, they are not presented in the study to keep its current length. An interesting result observed in Figure 11 that 34-Istanbul has the highest levels of interaction in terms of percentages of both incoming and outgoing messages with 33-Mersin whereas the percentages of both incoming and outgoing calls between 34-Istanbul and 27-Gaziantep are the higher than the other interactions 34-Istanbul.

We also illustrate the mobility activities of the refugees using network models. In Figure 12, a network view of refugee mobility between the nine major cities considered in this study is shown. The network is positioned on the map based on the data of the Ministry of Interior - Directorate General of Migration Management presented earlier in Figure 1. In the network, vertices represent the cities. An arc from city *i* to city *j* shows the total number of messages (call durations) from city *i* to city *j*. Similarly, loops show the total number of messages (call durations) within the corresponding cities. Vertices are approximately positioned on the regions of the cities they represent using the latitude / longitude information. It is noted that the network represent both mobility activities of messages and calls since both activities are observed in any combination of city pairs resulting a complete graph as seen in Figure 12.

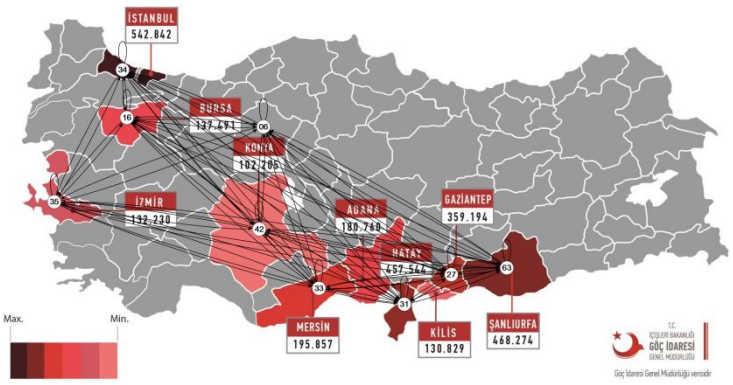


Fig. 12. A network view of refugee mobility activities between the major cities

The connectivity of the network show the existing mobility (both message and call) interactions between all city pairs considered in the study. The degree distribution of the network (i.e., the total number / duration of incoming-outgoing-all messages /

calls) is presented in Table 2. It is noted from the table that, 34-Istanbul, probably due to its highest refugee population compared to other major cities, mostly has the highest traffic except in terms of the duration of the incoming calls where 27-Gaziantep leads the list. On the other hand, 42-Konya has the least total number / duration of messages / calls for both incoming and outgoing mobility activities. Another interesting observation is as follows. Some cities receive more than they send (i.e., 31-Hatay and 33-Mersin and 42-Konya) while some others send more than they receive (i.e., 34-Istanbul and 35-Izmir) for both type of mobile activities which are not valid for the remaining major cities (i.e., more incoming messages-less outgoing calls in 06-Ankara and less incoming messages-more outgoing calls in 16-Bursa, 27-Gaziantep and 63-Sanliurfa).

Table 2. Distributions of incoming and outgoing number of messages and duration of calls for major cities

City	SMS-In	SMS-Out	SMS-All	Voice-In	Voice-Out	Voice-All
34-Istanbul	6,359	6,588	12,947	7,040,267	9,436,907	16,477,174
27-Gaziantep	4,670	5,207	9,877	7,950,883	6,524,413	14,475,296
35-Izmir	2,307	2,361	4,668	2,041,374	2,543,774	4,585,148
63-Sanliurfa	1,780	1,807	3,587	1,954,102	1,442,382	3,396,484
33-Mersin	2,746	2,355	5,101	3,473,612	3,194,241	6,667,853
31-Hatay	1,809	1,706	3,515	1,617,006	1,526,674	3,143,680
06-Ankara	1,785	1,593	3,378	1,889,349	1,899,711	3,789,060
42-Konya	848	650	1,498	1,397,362	1,106,435	2,503,797
16-Bursa	1,808	1,845	3,653	2,362,662	2,052,080	4,414,742

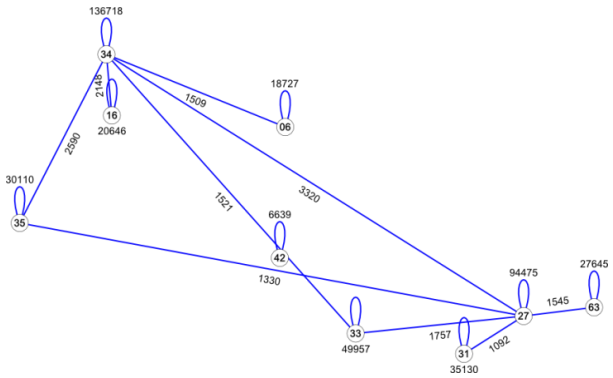


Fig. 13. An undirected graph representing the refugee mobility (SMS) between the major cities.

As noted from the network representation, although all major cities are connected to each other in both message (SMS) and call (voice) networks, their strengths or degrees of connectivity are different. An undirected SMS network is shown in Figure 13 where the background map is removed to increase its readability. In this network, the edges are only placed between two cities if the total number of messages between them in both directions is greater than 1,000 messages. Similarly, Figure 14 represents

an undirected network of call activities. In this network, the edges are only placed between two cities if the total duration of calls between them in both directions is greater than 1,000,000 time units. The central positions of 34-Istanbul and 27-Gaziantep are noted in both networks with significantly more connections compared to the other cities in the networks. As stated in many studies, this situation (having more connections in social networks) has a positive effect on social integration (Campbell and Lee, 1992; Atfield et al., 2007; Cheung and Phillimore, 2013). It is also noted that the mobility activities of 42-Konya are below the thresholds as seen in both networks.

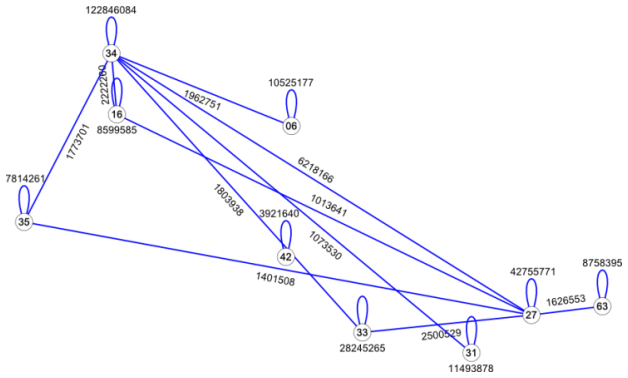


Fig. 14. An undirected graph representing the refugee mobility (voice) between the major cities

5.2 Dataset-2: Fine-Grained Mobility

This subsection presents our findings about the dataset of Fine-Grained Mobility (Dataset-2) by implementing the queries in the corresponding subsection (Section 4.2) of the previous section. As explained in Section 4.2, these queries are designed and implemented to analyze the mobility activities of the refugees in the major cities in more detail. Using the corresponding queries, we retrieve the records of the refugees with high mobility activities (i.e., top 10 refugees with the highest level of activities in each major city). The results are summarized in Figure 15 and Figure 16 for incoming and outgoing messages and in Figure 17 and Figure 18 for incoming and outgoing calls, respectively. We note two classifications in the figures in terms of time of week (i.e., as Weekday and Weekend) and customer type (i.e., as Turkish Citizen and Refugee). We also analyze the distributions of messages and calls with respect to the geographical regions (i.e., cities) and note that most of the messages (98% of all messages on average) and calls (99% of all calls on average) are within the same city.

Figure 15 (Figure 16) shows the distribution of the incoming (outgoing) number of messages of the refugees with high mobility activities in percentages, respectively. First of all, the distributions of both incoming and outgoing number of messages follow a similar trend in the major cities.

In terms of the customer type classification, we note that the refugees with the highest level of activities in each major city mostly send messages and make calls to Turkish Citizens. They mainly receive messages and calls from Turkish Citizens also. These results might represent the high interactions between the refugees with high mobility activities and Turkish Citizens.

On the other hand, we note that the highest degree of interactions between the refugees themselves are observed in 33-Mersin for incoming messages and in 31-Hatay for outgoing messages as 19% and 23%, respectively.

We finally note that although the refuges with high mobility activities mostly interact with Turkish Citizens, in general, the durations of the calls they make to other refugees is higher than the durations of the calls they receive from other refugees as noted from the distributions of incoming and outgoing calls where the highest percentage of outgoing calls is observed in 16-Bursa as 28% as noted in Figure 17 and Figure 18.

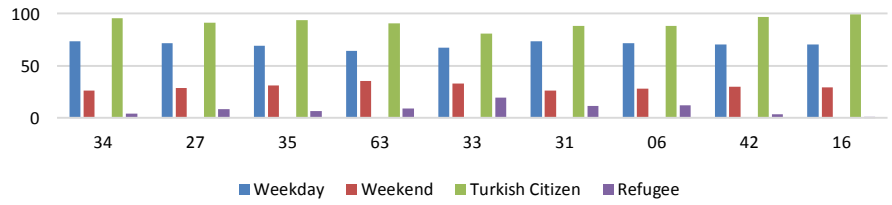


Fig. 15. Distribution of the incoming number of messages of the refugees with high mobility activities (percentage)

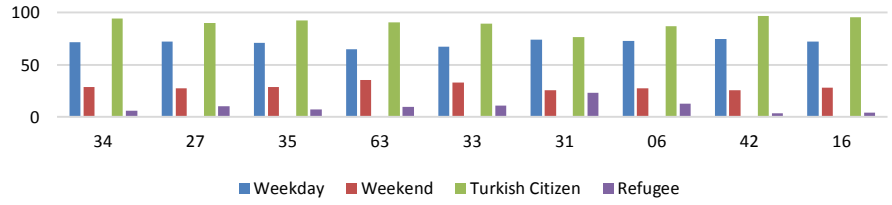


Fig. 16. Distribution of the outgoing number of messages of the refugees with high mobility activities (percentage)

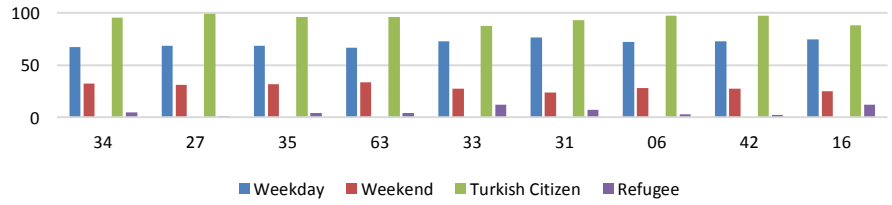


Fig. 17. Distribution of the incoming call durations of the refugees with high mobility activities (percentage)

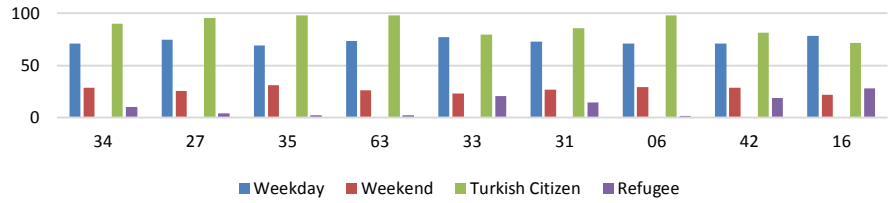


Fig. 18. Distribution of the outgoing call durations of the refugees with high mobility activities (percentage)

5.3 Dataset-3: Coarse-Grained Mobility

This section presents our findings about the dataset of Coarse-Grained Mobility (Dataset-3) by implementing the queries in the corresponding subsection (Section 4.3) of the previous section. As explained in Section 4.3, these queries are designed and implemented to analyze the mobility activities in the districts of the major cities. We aim at summarizing the results using a compact representation as in Figure 19.

Figure 19 shows the distribution of the mobility activities in the districts of the nine major cities. We present the incoming and outgoing mobility activity percentages on the left-hand-side and right-hand-side of the figure, respectively. At the top of the graph, we note the city codes in the same order presented throughout the study. The percentages in each city are presented in descending order as also defined using a color-coded representation where the districts colored as darker-red (darker-blue) have higher (lower) levels of mobility activities. We do not define the districts explicitly defined; however, aim at presenting the distribution of the mobility activities over the districts of each city.

It is noted that the distributions of the mobility activities in the nine major cities are different although we do not observe a significant difference in terms of the incoming and outgoing activities. 34-Istanbul represents a different case compared to the situations in other major cities as it seems that the refugees are more homogenously-distributed among its districts. In other words, the highest level of activities are observed in its districts is at most 20% (20% incoming and 19% outgoing, both in the Fatih district) whereas the remaining activities are distributed among its other districts with a maximum of 7% incoming and 6% of outgoing levels of activities. The homogeneity observed in 34-Istanbul might be a result of the certain characteristics of the city (Kaya and Kirac, 2016; Kaya, 2017). As stated in some studies, this homogeneity together with the dynamic demographics, multicultural environment and socioeconomic opportunities of the region might indicate a high level of social integration (Bosswick and Heckmann, 2006; Strang and Ager, 2010).

The same level of incoming mobile activities (20%) is observed in the Bornova district of 35-Izmir. On the other hand, including 35-Izmir, in each of the other cities

(the cities other than Istanbul), a few districts include an important level of mobility activities as noted with the red-colored districts in Figure 19.

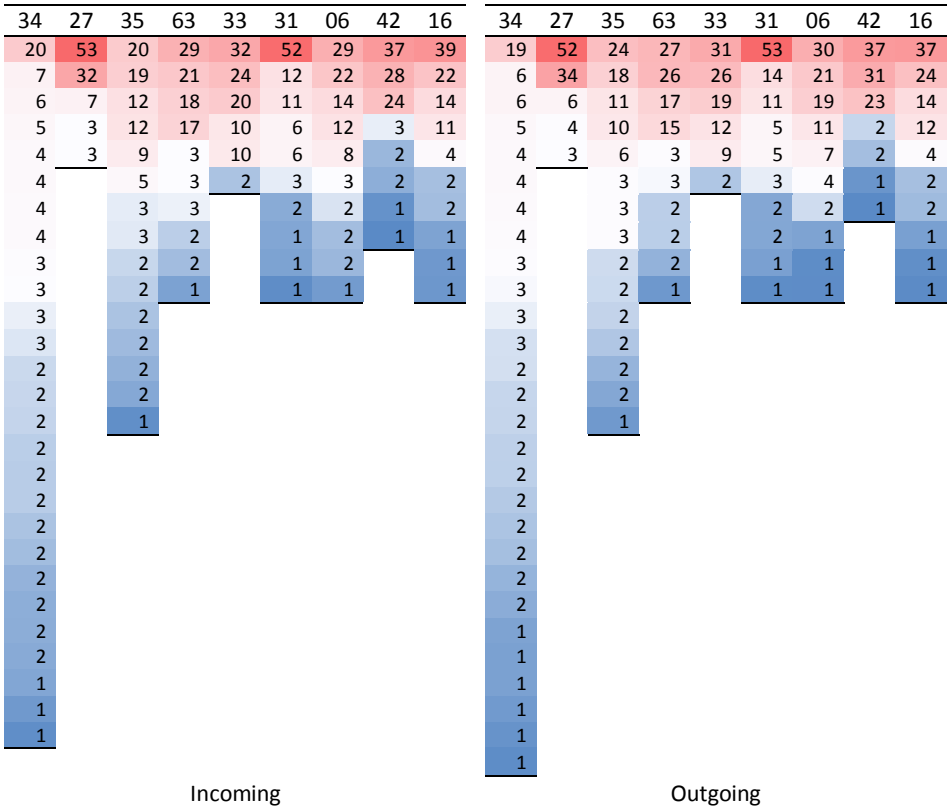


Fig. 19. Distribution of the refugee mobility activities on the districts of the major cities represented in percentages. The quantities less than 1% are not shown.

It is noted that more than half of the all mobile activities are performed in the top-three districts of each city whereas the total activity level of the top-three district of Istanbul is only 33% (20% + 7% + 6%). 27-Gaziantep and 31-Hatay represent the other extreme cases since more than half of the activities (53% in the Sahinbey district of 27-Gaziantep and 52% in the Antakya district of 31-Hatay, respectively) are observed in a single district. The potential for spatial isolation might create negative effects on the population and social integration process (Bosswick and Heckmann, 2006; Sonmez, 2016).

6 Conclusions

Data for Refugees (D4R) Turkey is a big data challenge aiming at providing anonymized mobile phone usage data to researchers for the purposes of providing better

living conditions for the Syrian refugees in Turkey. Our research team focuses on the social integration of Syrian Refugees as one of the prioritized subject areas and presents some insights from Call Detail Record (CDR) datasets provided.

There are three main datasets provided for the challenge (i.e., the datasets about Antenna Traffic, Fine-Grained Mobility and Coarse-Grained Mobility) in addition to the ones about base station locations, district locations, city mapping and district mapping. We first import the datasets to a sophisticated database server for being able to perform data operations efficiently and then design and implement several queries for gathering social integration information of the refugees. The results of the queries are analyzed to gain some insights.

We present the queries and their results in the corresponding subsections of Section 4 (Methodology) and Section 5 (Findings) for the main datasets, respectively. The queries designed for the dataset about Antenna Traffic (Dataset-1) are basically used to obtain the mobility activities in the major cities considered in the study. Similarly, we analyze the individual mobility activities using the queries presented for the dataset of Fine-Grained Mobility (Dataset-2). Finally, the queries for Coarse-Grained Mobility (Dataset-3) are designed and implemented to analyze the mobility activities in the districts of the major cities. The results are summarized and presented in the study which could be used by decision makers to understand if the Syrian refugees in Turkey could integrate to their host communities and discuss the potential strategies to help them in the integration process.

It is important for refugees to have access to public sphere where they can establish relationships with the members of the host community to contribute to the social integration process. Planning some activities, such as festivals or celebrations, to ensure that the refugees leave the areas where they live together with the other refugees might help them to integrate to their host community. Different types of sociocultural events to bring refugees and the members of the host community together could be organized so that they could spend time together. Matching refugees with the members of the host community in such events might be also be considered. Educational activities, especially for the younger refugees, might be very useful for the overall social integration process.

References

1. Atfield, G., Brahmbhatt, K., O'Toole, T.: Refugee Council and University of Birmingham Refugees' Experiences of Integration (2007).
2. Bakewell, O.: Refugees and Local Hosts: a Livelihood Approach to Local Integration and Repatriation. Insights – Development Research (2002).
3. Bakewell, O.: Uncovering Local Perspectives on Humanitarian Assistance and its Outcomes. Disasters 24(2): 103–116 (2000).
4. Bayram, N., Nyquist, H., Thorburn, D., Bilgel, N.: Turkish immigrants in Sweden: Are they integrated?. International Migration Review 43(1), 90–111 (2009).
5. Bosswick, W., Heckmann, F.: Integration of migrants: Contribution of local and regional authorities. European Foundation for the Improvement of Living and Working Conditions 11, (2006).

6. Campbell, K. E., Lee, B. A.: Sources of personal neighbor networks: social integration, need, or time?. Social forces 70(4), 1077–1100 (1992).
7. Cheung, S. Y., Phillimore, J.: Social networks, social capital and refugee integration. Report for Nuffield Foundation: London (2013).
8. Entzinger, H.: The Dynamics of Integration Policies: A Multidimensional Model. In: Koopmans, R., Statham, P. (eds.) Challenging Immigration and Ethnic Relations Politics: Comparative European Perspectives, pp. 97–118. Oxford University Press (2000).
9. Esser, H.: Die Situationslogik ethnischer Konflikte. Zeitschrift fur Soziologie 28(4), 245–262 (1999).
10. Heckmann, F., Schnapper, D. (eds.) : The integration of immigrants in European societies: National differences and trends of convergence. Vol. 7. Lucius & Lucius, Stuttgart (2003).
11. Kaya, A., Kirac, A.: Vulnerability Assessment of Syrian Refugees in Istanbul (2016).
12. Kaya, A.: Istanbul as a Space of Cultural Affinity for Syrian Refugees. Southeastern Europe 41(3), 333–358 (2017).
13. Leach, M.: Dealing with Displacement: Refugee–Host Relations, Food and Forest Resources in Sierra Leone Mende Communities during the Liberian Influx, 1990–1991. IDS Research Report no. 22 Brighton: IDS, University of Sussex, Institute of Development Studies, UK (1992).
14. Lockwood, D.: Social integration and system integration. Explorations in social change, 244–257 (1964).
15. ORSAM-Center for Middle Eastern Studies: The Economic Effects of Syrian Refugees on Turkey: A Synthetic Modelling, (2015).
16. Platts-Fowler, D., Robinson, D.: A place for integration: refugee experiences in two English cities. Population, Space and Place 21(5), 476–491 (2015).
17. Porter, G., Hampshire, K., Kyei, P., Adjaloo, M., Rapoo, G., Kilpatrick, K.: Linkages between livelihood opportunities and refugee–host relations: learning from the experiences of Liberian camp-based refugees in Ghana. Journal of refugee studies 21(2), 230–252 (2008).
18. Robinson, D.: The neighbourhood effects of new immigration. Environment and Planning A 42(10), 2451–2466 (2010).
19. Salah, A.A., Pentland, A., Lepri, B., Letouze, E., Vinck, P., de Montjoye, Y.A., Dong, X. Dagdelen, O.: Data for Refugees: The D4R Challenge on Mobility of Syrian Refugees in Turkey. arXiv preprint arXiv:1807.00523, 1–13 (2018).
20. Sonmez, M., E.: Spatial distribution and futurity of Syrian refugees in the city of Gaziantep. In: International Geography Symposium, Ankara (2016).
21. Spencer S.: Integration: Policy Primer. Oxford: COMPAS (2011).
22. Strang, A., Ager, A.: Refugee integration: Emerging trends and remaining agendas. Journal of Refugee Studies 23(4), 589–607 (2010).
23. Torpey, J. C.: The invention of the passport: surveillance, citizenship and the state. 1st edn. Cambridge University Press, UK (2000).
24. TUIK-Turkish Statistical Institute: TurkStat, International Migration Statistics, 2017, (2018).
25. UNHCR-The United Nations High Commissioner for Refugees Homepage, http://www.unhcr.org/, last accessed 2018/08/19.
26. Yildiz, A., Uzgoren, E.: Limits to temporary protection: non-camp Syrian refugees in Izmir, Turkey. Southeast European and Black Sea Studies 16(2), 195–211 (2016).

Refugee Integration in Turkey: A Study of Mobile Phone Data for D4R Challenge

Ismail Uluturk[1†], Ismail Uysal[1], and Onur Varol[2†]

[1] RFID Lab for Applied Research, University of South Florida, Tampa, FL
{uluturki,iuysal}@usf.edu
[2] Center for Complex Network Research, Northeastern University, Boston, MA
ovarol@northeastern.edu

Abstract. One of the defining crises of our age is the unprecedented number of refugees, caused by social upheaval and political conflicts all around the world. In this work, individual and aggregated mobility patterns are studied to provide insights on social integration and unemployment issues faced by refugees in Turkey, utilizing mobile phone datasets from a national mobile carrier, provided by the Data for Refugees(D4R) organization. Aggregated base station traffic data and coarse-grained user mobility data is used to analyze and compare mobile phone usage and movement patterns of refugee and non-refugee groups respectively. Results show that mobile phone traffic of refugee users is not as spread out geographically, and there exists a significant number base stations where at least 90% of the traffic involves refugees. Analysis of mobility data also shows that movement of refugee users is confined into smaller areas and is not as spread out while they also move more frequently with shorter distance steps compared to non-refugee users. We comment that this distinction in mobile phone usage and movement patterns, as well as their geographical segregation may indicate a lack of suitable employment opportunities outside certain central locations which also drives a lack of integration to local social life for refugees.

Keywords: social integration · unemployment · human mobility · spatial networks · time series

1 Introduction

Unprecedented numbers of displaced refugees, caused by social upheavals and political conflicts in recent history have overwhelmed the support systems in place for refugees and is considered to be one of the most important humanitarian crises of our times. The situation is especially critical for minors, as their experiences can have a direct influence over their mental health and educational development [9, 21]. While significant work is being done by governments and non-governmental organizations to alleviate some of these problems [17], the current state of technology allows for additional unique opportunities for determining the shortcomings and improving conditions not possible in the past.

Extensive mobile and smart phone usage by refugees for coordination and communication among themselves as well as people they have left behind has been consistently reported by officials and volunteers in the field. In fact, it has even been reported that most refugees ask for charging and data services for their mobile phones before food or water, showing that they consider their mobile phones a vital tool for survival [22]. This phenomena allows for a wealth of data, uniquely presented by the current state of communication technologies, that can be leveraged.

In this work, we focus on two main issues: *social integration* and *unemployment*. We have identified these issues as our main focus, given the nature and limitations of the provided dataset. We have utilized the base station traffic data (`Dataset1`) for studying communication patterns, and coarse-grained mobility data (`Dataset3`) for studying individual and aggregate movement patterns, with the goal of identifying markers of *social integration* and *unemployment*.

2 Related Work

Human migration, both globally and locally, has been a common topic of study for scientists in various fields [8, 4]. However, recent improvements in communication technologies and availability of mobile phone data have enabled studying movement of individuals in urban areas. Recent studies on individual mobility data show that most individuals have a predictable mobility pattern, where they travel between a relatively small number of locations regularly, such as their workplaces and accommodations [16, 20, 6].

Previous efforts on analyzing mobility patterns for past Data for Development(D4D) challenges point out the importance of mobile phone data [5, 14]. Researchers have also investigated the relationship between mobile phone usage and regional economic development [13, 19]. Information on mobility data has also been used for predicting human behavior [3] and daily pulse of cities [1, 2], studying transformation in metropolitan cities [12], and development of vaccination strategies for disease prevention [11, 10, 23].

3 Dataset

Dataset is provided by Data for Refugees: The D4R Challenge on Mobility of Syrian Refugees in Turkey. It includes anonymized mobile phone data of both refugee and non-refugee user samples. Data is provided in 3 distinct datasets and further details on the dataset can be found in the paper published by the organizers [18].

- `Dataset1`: Antenna traffic captures one year site-to-site traffic information on an hourly intervals.
- `Dataset2`: Fine grained mobility dataset contains usage information for randomly sampled accounts. Accounts selected for analysis resampled in every two weeks to prevent security concerns.

– `Dataset3`: Coarse grained mobility information is one of the most valuable resources to be able to study mobility of individuals. Trajectories of randomly selected refugees and non-refugees contain locations of serving base station and time of record for 50,000 individuals.

3.1 Limitations

Dataset provided includes data collected from only one major cellular carrier. Furthermore, three different identifiers are used to flag a user as a refugee when they are registering; having an ID number given to refugees and foreigners in Turkey, registering with a Syrian passport, or using special tariffs reserved for refugees. All three of these methods are noisy, and there is no guarantee that a user flagged as a refugee is actually one. This implies that statistics of this dataset may not necessarily generalize to national statistics and care should be taken when inferring results from them.

There are significant gaps in the temporal data as well. Dates with missing data for `Dataset3` can be seen in Fig. 1. As a result, first 6 months for `Dataset3` is discarded for reliable analysis of temporal features. This brings the total number of entries from 66,000,731 to 49,830,623, which is still sufficiently large for large scale data analysis.

`Dataset2` provides incoming and outgoing traffic data for individual users. However, traffic data is captured from different subsets of users for incoming and outgoing traffic. This means that it is not possible to study both incoming and outgoing traffic patterns for individual users in this dataset.

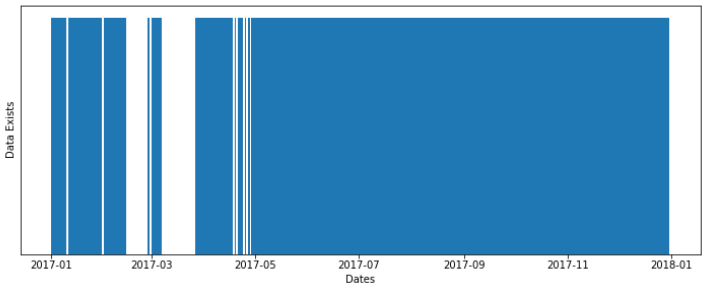


Fig. 1. Visual representation of `Dataset3`, showing dates where data exists or is missing.

4 Results

4.1 Antenna Traffic Analysis

Monthly base station traffic information is valuable for studying overall usage patterns of groups. For all the analyses found in this section, total activity between source and target stations over the entire range of observation period is aggregated together in pairs. Pairs of stations that has less than 10 activity records between them are filtered out since the effects of these stations could be considered negligible.

Locations with most and least refugee activity is identified by the fraction of refugee traffic over all traffic between every pair of stations. A histogram of the refugee traffic fractions is presented in Fig. 2. It is observed that there are base station pairs where at least 90% of the total traffic involves refugees. Bimodal shape of the distribution indicates existence of majority refugee and non-refugee station pairs, which may be a marker of geographical segregation. Official refugee accommodation centers can be one explanation for the existence of these refugee dominated base station pairs. The locations of Disaster and Emergency Management Presidency (AFAD) Temporary Protection Centers[3] (TPC) are show in Fig. 4, together with the base station pairs that fall within top and bottom %10 percent on Fig. 2. It is observed that the TPCs are not sufficient to explain all of these refugee dominated pairs, especially in central and western parts of the country.

Distances between base stations in these refugee and non-refugee dominated station pairs are another topic of interest. Fig. 3 shows the distribution of inter-pair distances of base station pairs that are dominated by refugee and non-refugee activity respectively. It can be observed from Fig. 3 that the base station pairs dominated by refugee activity tend to be closer to each other compared to their non-refugee dominated counterparts.

Geographical spread of majority refugee and non-refugee traffic can also be observed by visualizing the traffic between base stations over a map. Fig. 4 shows the traffic between the base stations that fall within the top and bottom 10% of Fig. 2, respectively. Only the voice call counts data is displayed on the maps for the sake of brevity, however both voice call duration and SMS data also show very similar results. It can be observed from Fig. 4 that, base station pairs with refugee dominated traffic are less spread out geographically, and are mainly focused between certain centers of dense activity. On the other hand, non-refugee dominated cellular traffic appears to be spread out much more evenly.

Since Istanbul is one of the largest metropolitan areas in the world, it merits a closer look. Fig. 5 displays the same data as Fig. 4, only within the city limits of Istanbul. It can be observed that, refugee dominated traffic is similarly less spread around. This divide can most easily be observed on the Asian side of the city, where refugee dominated traffic is lesser than its non-refugee dominated counterpart.

[3] Compiled on December 6th, 2018 from reports found in: https://www.afad.gov.tr/tr/2374/Barinma-Merkezlerinde-Son-Durum

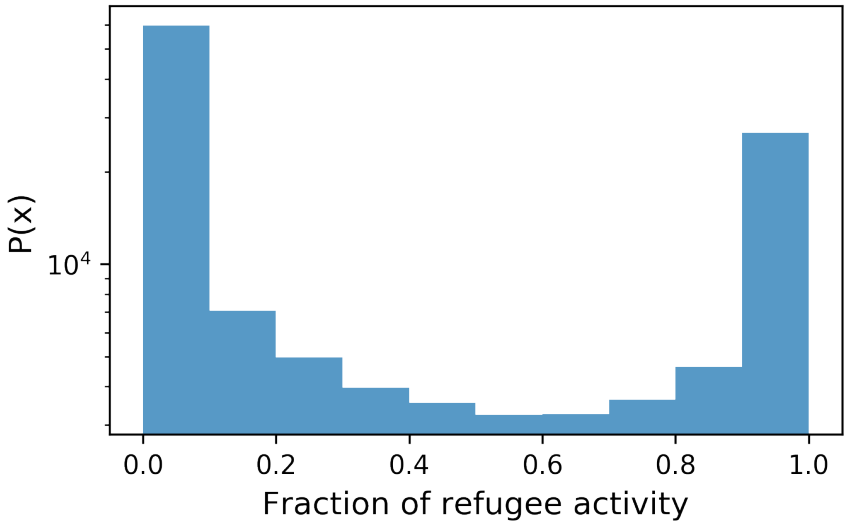


Fig. 2. Histogram of refugee activity density for pairs of stations.

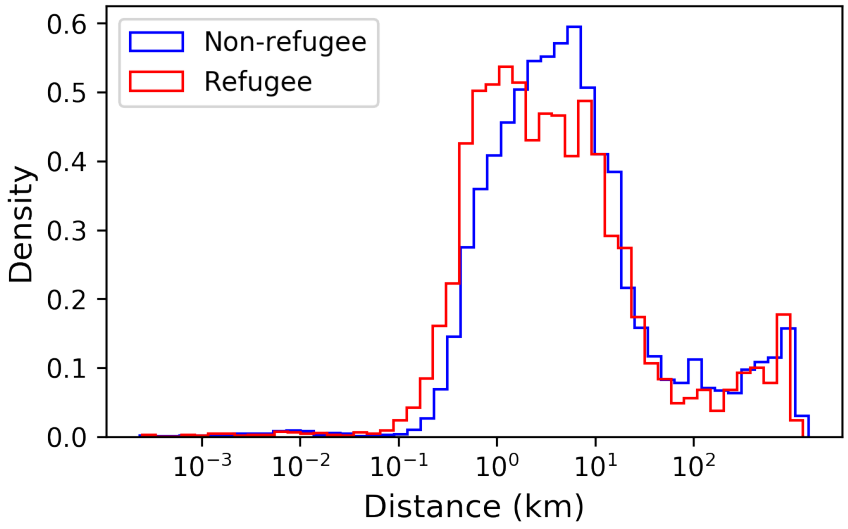


Fig. 3. Distance distribution of refugee and non-refugee dominated base station pairs.

This difference between mobile phone usage behavior and geographical spread of mobile phone traffic suggests a lack of social integration for refugee users. Fig. 4 shows that refugees are not in contact with a significant portion of the nation.

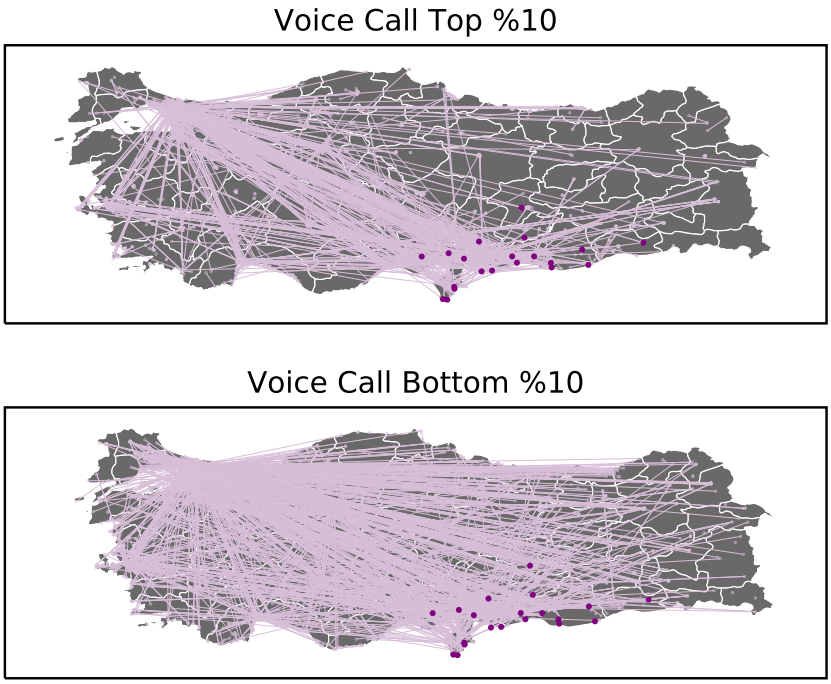


Fig. 4. Voice call traffic between base station pairs that fall within the top and bottom 10% of Fig. 3. Locations of AFAD Temporary Accommodation Centers are marked with purple hexagons.

This, together with the bimodal nature of Fig. 2 also suggests existence of spatial segregation, which will be further investigated in the following sections.

4.2 Movement Analysis from Coarse-grained Mobility

Coarse-grained mobility data from `Dataset3` provides high temporal resolution and continuity for each user. Therefore, it is especially suitable for the analysis of individual movement patterns over time.

High-order Mobility Network Analysis A user mobility network for both refugee and non-refugee users are constructed. Traditionally, mobility networks are constructed with a first-order assumption. This means that, probability of a user moving to a certain node is assumed to be only dependent on the current node they are in. It has been shown that this single order assumption is lacking in representing complex mobility data [25]. Therefore, a variable order high-order network is generated for analysis in this section, using `BuildHON+`[4] algorithm [24].

[4] Software implementation: https://github.com/xyjprc/hon

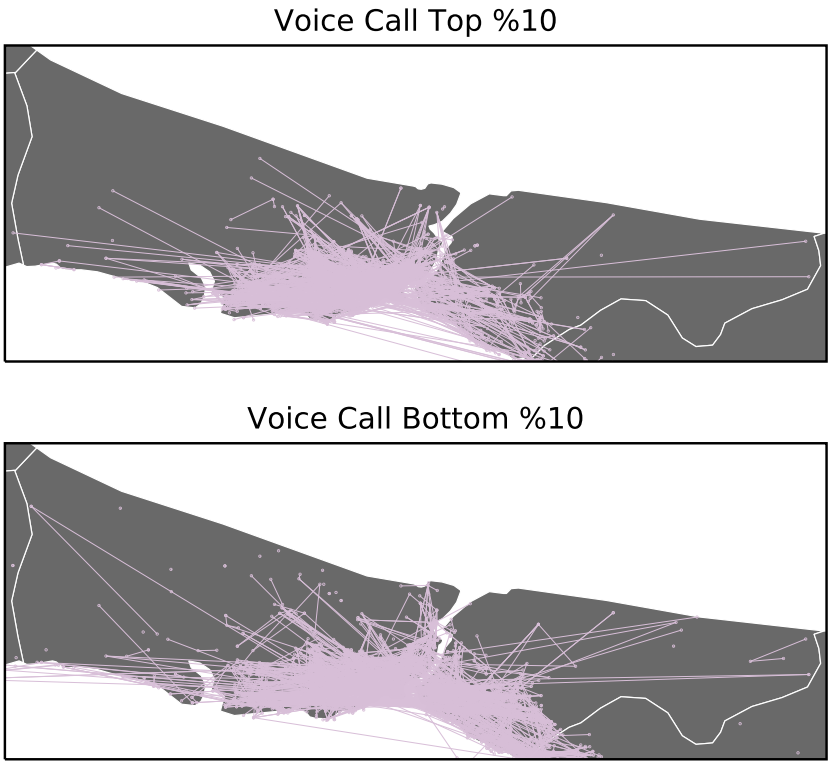


Fig. 5. Voice call traffic between base station pairs that fall within the top and bottom 10% of Fig. 2, confined within the city limits of Istanbul.

Pagerank [15] is a widely used algorithm that is used to rank the importance of a node in a network, calculated based on incoming and outgoing links to the node. Intuitively, pagerank of a node in a mobility network will show how many paths include that node, and how often users take these paths. A node with a high pagerank in a mobility network can be viewed as a hub that is visited by users from all over. Some good examples of high pagerank nodes will be dense residential areas and public transportation transfer stations.

A visualization of district rankings over the generated variable high-order user mobility networks, based on pagerank values, can be seen in Fig. 6. There are some interesting observations that can be made based on this figure. First of all, mobility network of refugee users is more sparse, meaning that refugee movement is confined to a more limited area, compared to non-refugee users mobility network. There is a large amount of area that is not covered by the refugee users mobility. Furthermore, regions of importance appear to be more spread out in the non-refugee users mobility network, while important regions are more focused in smaller regions for refugee users. Finally, while non-refugee

users mobility network has several distinct centers of high importance, network for refugee users is dominated by İstanbul. Last two points merit a closer look into the districts within İstanbul, which is one of the worlds largest metropolitan areas.

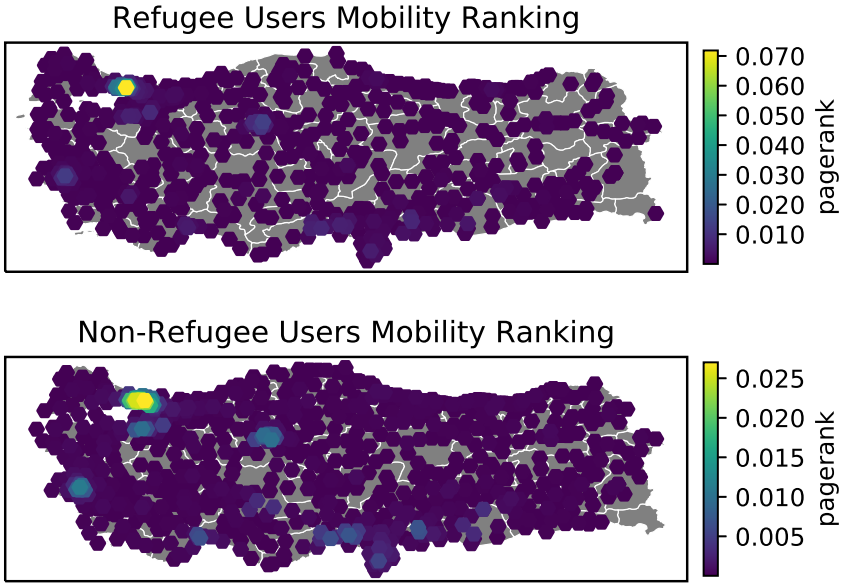


Fig. 6. Visualization of district rankings based on pagerank values over high-order user mobility networks.

It can be seen from Fig. 7 that, even within the city limits of İstanbul, mobility of refugee users is still confined to focused and denser regions, as opposed to the almost evenly spread out non-refugee users. For example, there is very little movement by refugee users on the Anatolian side of the İstanbul compared to non-refugee users. It can be inferred that, while refugees and non-refugees share the same urban public spaces, they are segregated in their use of these spaces. This suggests that a significant portion of the refugee users are not integrated in the local social life. It can also be an indicator of differing employment statistics and employment opportunities available to refugee users.

Inter-event Analysis Data from incoming and outgoing traffic is combined for movement analysis, and every time a location change for a user occurs it is marked as a transition event. A normalized histogram for the transition event count of each user can be seen in Fig. 8. This figure shows that a refugee user is more likely to have a smaller number of transition events compared to a non-refugee user.

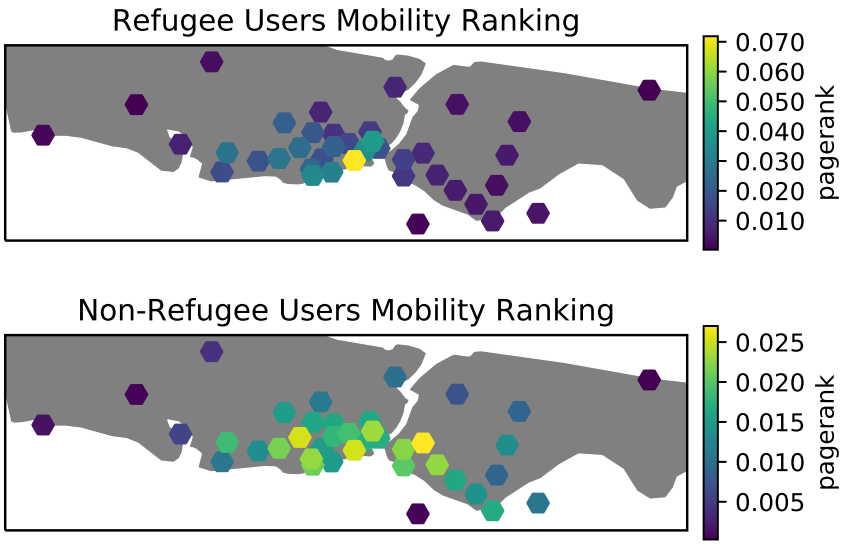


Fig. 7. A closer look into the districts within the city limits of Istanbul from Fig. 6.

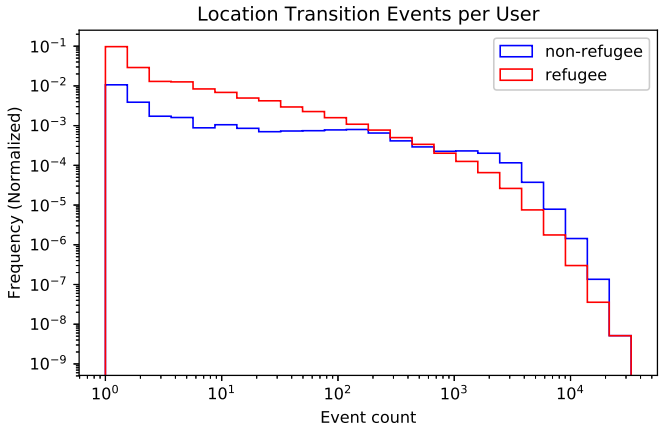


Fig. 8. Histogram of transition events on log-log scale, marking every time a user changes locations.

The average time each user waits before a new movement event occurs can also be enlightening. A normalized histogram for inter-event time averaged for each user is provided in Fig. 9. This figure shows that a refugee user is more likely spend a shorter amount of time at each location compared to a non-refugee user, moving more frequently. Fig. 9 also shows a significant number of refugee

users, with a mean inter-event time under 10^3 seconds, or 17 minutes. This can be explained by spending time in moving vehicles or between regional borders, where transition-events can occur more frequently or with smaller movements.

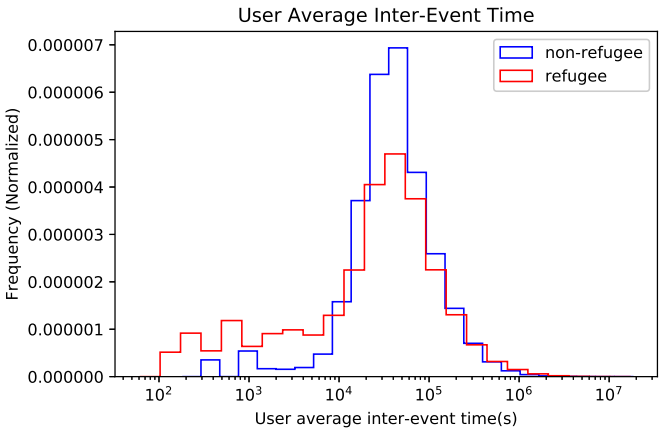


Fig. 9. Histogram of the average time elapsed between subsequent transition events for each user.

Distance traveled between each transition event will help paint a more complete picture of the user movement, when combined with the previous results. A histogram of average distance traveled by each user is given in Fig. 10. This figure shows that a refugee user is more likely to travel shorter distances between transition events compared to a non-refugee user.

Finally, we will take a look at the size of the overall movement regions of each user. Movement bounding-boxes are computed by calculating the smallest rectangle that encompasses the center coordinates of every region each user has visited over the observation period. The geographical distance between diagonal corners of the bounding box is used as an indicator of the size of the overall movement regions for each user. A histogram of this bounding-box diagonal distances can be seen in Fig. 11. This figure shows that a refugee user is more likely to travel within a smaller region compared to a non-refugee user, confirming the previous observations.

Looking at the results in this subsection, it will be possible to infer that refugee users move more frequently in a smaller regions via short distances. Individual movement is expected to be significantly regular, where a person frequently returns to a small number of anchor locations, such as work and home [7]. These observations can suggest that refugee users are more likely to be unemployed, or hold irregular jobs that require them to frequent larger than usual number of locations.

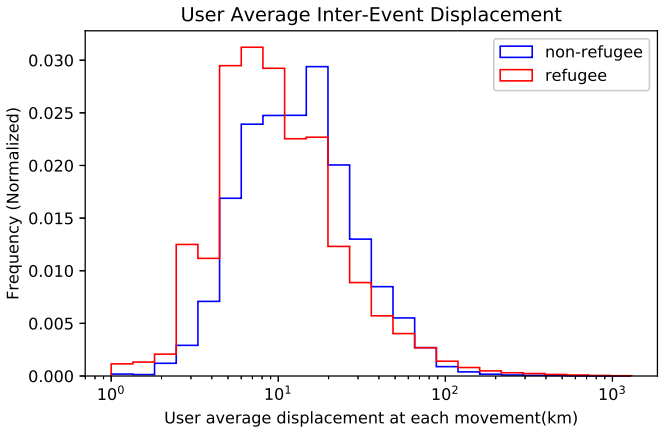


Fig. 10. Histogram of mean distance traveled between subsequent transition events for each user.

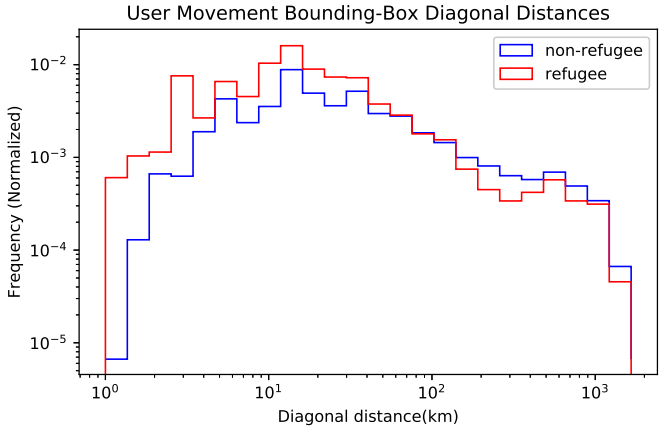


Fig. 11. Histogram of mean distance traveled between subsequent transition events for each user.

5 Discussion and Suggestions

In this work, mobile phone data belonging to refugee and non-refugee users have been analyzed with a focus on discovering individual and aggregate mobility and connectivity patterns. Since connectivity is linked with employment and social life, and human mobility is very regular and dependent on a few frequented locations, such as places of work and residence, any irregularities on these patterns can be interpreted as indicators of employment and integration issues.

Connectivity analysis showed a divide between base station pairs that are servicing majority refugee and non-refugee traffic. It is also found that the distance of refugee connections tends to be shorter. Visualization over a map also showed a much smaller geographical spread of refugee traffic compared to its non-refugee counterpart. These observations suggest that refugee and non-refugee users are segregated geographically and refugee users do not interact with a significant portion of the nation.

User mobility network analysis yielded the results that refugee users move around in smaller regions, and movement of refugee users is less spread out compared to non-refugee users. Focusing on a metropolitan area, Fig. 7 showed that refugee movement is confined in a smaller and denser area. This supports that, while the urban areas and public spaces are available to both groups, use of these spaces is segregated. Furthermore, results from inter-event analysis also support these findings, as they suggest that refugee users move more frequently in smaller regions and with shorter distance steps.

Combining results from these analyses, it is possible to infer that refugee users are largely not integrated into the local social life of the regions they reside in. It is likely that, this integration issue is driven by the socioeconomic conditions and especially availability of employment options, which focuses refugee users into areas where there is work available for them. Providing suitable employment opportunities outside the current centers of focused refugee activity can encourage and speed up the distribution of refugees more evenly within the nation and facilitate their integration to the local societies.

6 Future Work

Our current study captures information obtained from D4R dataset alone. Due to privacy concerns and sampling limitations, information available for individual mobility patterns is limited and it is not possible to construct communication networks of users from the call records since the the resolution of the records is on base station level. In the future, we plan to do a deeper study of temporal data while improving and supporting our findings with surveys, refugee statistics, integrated GIS information, and social data.

We are interested in discovering if the groups are also segregated temporally, underlying social structures, and any prevalent migration paths and patterns that might exist. We are working on quantifying the access to quality public transportation by each group, as well as their access to points of service through public transportation. We are also investigating activity sets and spaces for each group to quantify and compare exploration-exploitation behaviors, which may be an indicator towards social behaviour of users. We'll be more than happy to share our future findings with the organizers and the scientific community at large with an enhanced understanding of observed patterns of mobility and communication.

References

1. Andris, C., Bettencourt, L.M.: Development, information and social connectivity in côte divoire. Infrastructure Complexity **1**(1), 1 (2014)
2. Barlacchi, G., De Nadai, M., Larcher, R., Casella, A., Chitic, C., Torrisi, G., Antonelli, F., Vespignani, A., Pentland, A., Lepri, B.: A multi-source dataset of urban life in the city of milan and the province of trentino. Scientific data **2**, 150055 (2015)
3. Bianchi, F.M., Rizzi, A., Sadeghian, A., Moiso, C.: Identifying user habits through data mining on call data records. Engineering Applications of Artificial Intelligence **54**, 49–61 (2016)
4. Black, R., Adger, W.N., Arnell, N.W., Dercon, S., Geddes, A., Thomas, D.: The effect of environmental change on human migration. Global environmental change **21**, S3–S11 (2011)
5. Blondel, V.D., Esch, M., Chan, C., Clérot, F., Deville, P., Huens, E., Morlot, F., Smoreda, Z., Ziemlicki, C.: Data for development: the d4d challenge on mobile phone data. arXiv preprint arXiv:1210.0137 (2012)
6. Candia, J., González, M.C., Wang, P., Schoenharl, T., Madey, G., Barabási, A.L.: Uncovering individual and collective human dynamics from mobile phone records. Journal of physics A: mathematical and theoretical **41**(22), 224015 (2008)
7. Gonzalez, M.C., Hidalgo, C.A., Barabasi, A.L.: Understanding individual human mobility patterns. nature **453**(7196), 779 (2008)
8. Greenwood, M.J.: Human migration: Theory, models, and empirical studies. Journal of regional Science **25**(4), 521–544 (1985)
9. Keles, S., Friborg, O., Idsøe, T., Sirin, S., Oppedal, B.: Depression among unaccompanied minor refugees: The relative contribution of general and acculturation-specific daily hassles. Ethnicity & health **21**(3), 300–317 (2016)
10. Lima, A., De Domenico, M., Pejovic, V., Musolesi, M.: Disease containment strategies based on mobility and information dissemination. Scientific reports **5**, 10650 (2015)
11. Lima, A., De Domenico, M., Pejovic, V., Musolesi, M.: Exploiting cellular data for disease containment and information campaigns strategies in country-wide epidemics. arXiv preprint arXiv:1306.4534 (2013)
12. Liu, F., Janssens, D., Cui, J., Wang, Y., Wets, G., Cools, M.: Building a validation measure for activity-based transportation models based on mobile phone data. Expert Systems with Applications **41**(14), 6174–6189 (2014)
13. Mao, H., Shuai, X., Ahn, Y.Y., Bollen, J.: Mobile communications reveal the regional economy in côte divoire. Proc. of NetMob (2013)
14. de Montjoye, Y.A., Smoreda, Z., Trinquart, R., Ziemlicki, C., Blondel, V.D.: D4d-senegal: the second mobile phone data for development challenge. arXiv preprint arXiv:1407.4885 (2014)
15. Page, L., Brin, S., Motwani, R., Winograd, T., et al.: The pagerank citation ranking: Bringing order to the web (1998)
16. Pappalardo, L., Simini, F., Rinzivillo, S., Pedreschi, D., Giannotti, F., Barabási, A.L.: Returners and explorers dichotomy in human mobility. Nature communications **6**, 8166 (2015)
17. Regional Refugee & Resilience Plan 2018-2019 in Response to the Syria Crisis: 3RP Regional Strategic Overview 2018-19. Tech. rep., The 3RP (2017)
18. Salah, A.A., Pentland, A., Lepri, B., Letouzé, E., Vinck, P., de Montjoye, Y.A., Dong, X., Dağdelen, Ö.: Data for refugees: The d4r challenge on mobility of syrian refugees in turkey. arXiv preprint arXiv:1807.00523 (2018)

19. Šćepanović, S., Mishkovski, I., Hui, P., Nurminen, J.K., Ylä-Jääski, A.: Mobile phone call data as a regional socio-economic proxy indicator. PloS one **10**(4), e0124160 (2015)
20. Simini, F., González, M.C., Maritan, A., Barabási, A.L.: A universal model for mobility and migration patterns. Nature **484**(7392), 96 (2012)
21. Sirin, S.R., Rogers-Sirin, L.: The educational and mental health needs of Syrian refugee children. Migration Policy Institute Washington, DC (2015)
22. The GSM Association: The Importance of Mobile for Refugees: A Landscape of New Services and Approaches. Tech. rep., The GSM Association (2017), https://www.gsma.com/mobilefordevelopment/wp-content/uploads/2017/02/The-Importance-of-mobile-for-refugees_a-landscape-of-new-services-and-approaches.pdf
23. Tompkins, A.M., McCreesh, N.: Migration statistics relevant for malaria transmission in senegal derived from mobile phone data and used in an agent-based migration model. Geospatial health **11**(1 Suppl), 408 (2016)
24. Xu, J., Saebi, M., Ribeiro, B., Kaplan, L.M., Chawla, N.V.: Detecting anomalies in sequential data with higher-order networks. arXiv preprint arXiv:1712.09658 (2017)
25. Xu, J., Wickramarathne, T.L., Chawla, N.V.: Representing higher-order dependencies in networks. Science advances **2**(5), e1600028 (2016)

Measuring Segregation of Syrian Refugees via Mobile Call Detail Records

Fatih ULUDAĞ[1*], H. Eray ÇELİK[1], Serbest ZİYANAK[2], Murat CANAYAZ[3], and Fikriye ATAMAN[4]

[1] Department of Econometrics, Van Yüzüncü Yıl University, [2] Van Vocational High School Computer Programming, [3] Department of Computer Engineering, [4] Department of Informatics
{fatihuludag, ecelik, sziyanak, mcanayaz, fataman}@yyu.edu.tr

Abstract. Segregation is one of the biggest obstacles to political and socio-economic development. After the war that started in 2011, many Syrian people had to leave their countries. Turkey hosts largest refugee population in world. Syrian refugees in Turkey have now become a part of social life. Measuring the segregation of Syrian refugees will help policies for legislators and decision-makers. Within the context of the "Data for Refugees" organized by Turk Telekom, the segregation was calculated with the mobile call detail records provided to the researchers. Dissimilarity Index (D) and Modified Isolation of Index (MII) were used to measure the segregation. The resulting values are shown on the map of Turkey. According to the Kruskal-Wallis test statistically significant results were found between geographical regions ($\chi^2 = 14.98, p = 0.02, df = 6$).

Keywords: Segregation, Call Detail Records, Big Data, Syrian Refugees

1 Introduction

Spatial segregation is described as measurement of living in different areas of a city or how separate more groups live from each other [1]. Segregation causes inequality of social groups in accessing to services and job opportunities and also, they are exposed to marginalization, becoming impoverished, violence, and being alienated from society [2]. It is proved by the studies that spatial segregation has negative

effects on socio-economic and political developments of minority groups and the group is exposed to inequality on supply of public goods and sharing of social capital [3, 4, 5, 6].

The Arab Spring Process showing up in the later 2010 caused serious political crisis in Arab countries. Syria is one of the leading countries which was influenced from Arab Spring. A as a result of Arab Spring a big civil war started in Syria and this war has still continued. Because of this continuing civil war, approximately 12.5 million Syria citizens has become refugees and this number is increasing day by day. Syria has the longest border on Turkey with its 911 km length. Turkey is the most favorite country in the world for Syria refugees with the reasons like geographical closeness, cultural reasons, affinity relations coming from history, open door policy of political authority etc. According to the report of United Nations Refuge Organization, at the end of 2016, Turkey is the country which hosted the most refugees in the world over three years.

The essential data for the studies of segregation is generally obtained from population censuses and also, negotiations, questionnaires and trip information can be used for specifying the segregation. To be able to perform these measurements, mobile call detail records has a great potential. Recording the details by the third party automatically, taking them from quite big sample [7], being efficient in terms of time and cost gain the upper hand.

The aim of this study is to identify the segregation of Syrian refugees via mobile call detail records which are provided by Turk Telekom. Inadequacy originating from traditional data sources causes limited study at this field.

2 Theoretical Background

Being able to perform a quantitative measurement of segregation composes the main theme of the segregation literature. From 1940s, in order to be able to measure the segregation, various segregation indexes have been suggested and the features of these are discussed [8, 9, 10, 11, 12, 13]. Segregation has different dimensions and suggested indexes are used with the aim of evaluating these dimensions. Massey and Denton (1988) studying on various segregation measurement have identifed that segregation has five dimensions. Evenness which shows the degree of dispersion of a group to the quarters homogeneously, exposure which indicates the degree of potential interaction with the other groups, concentration demonstrating the physical area taken by a group, centralization indicating the proximity of a group to the centre of the city and clustering which demonstrates the tendency of the people in the same group to live at the close neighborhood. They put forward that evenness is the most important dimension among the others and Dissimilarity Index developed by Duncan and Duncan (1955) is the most successful one at measuring evenness dimension [14].

D Dissimilarity index is the most popular index proposed by Duncan and Duncan in 1955 and it is frequently used by many researchers as it can be easily calculated.

Dissimilarity index indicates the percentage of group that needs to be replaced in order for the urban population to have a uniform distribution. Index has values between 0 and 1. While 0 value indicates completely integration, 1 value does completely segregation.

Dissimilarity index is used for calculating the segregation between two groups. Beyond the dissimilarity index, other indexes such as P exposure index [15], Gini Index [15] , Theory Index [15], and Atkinson Index [16] are used.

These indexes are used with the aim of measuring the segregation between two groups. To measure the segregation for more than two groups, segregation indexes are suggested [17, 18, 19, 20].

3. Syrian Refugees in Turkey

The war breaking out in 2011 at Syria and still continuing today has made Syria citizen refugee position. In April of the same year, tent cities are established for refugees who started to come to Turkey in districts of Hatay Yayladağı, Altınözü and Reyhanlı. According to the report of Turkish Republic Ministry of Interior Immigration Authority Directorate- General, the number of registered Syrian refugees residing in Turkey is 3.552.303. In the early days of the war, 3.355.575 refugees who are placed in the temporary refugee centers live in the center of population out of the camp. Syrian refugees were seen as guests by Turks at the beginning, but then because of the extended war they have been started to be seen permanent component of the society and also it is estimated from the studies that even if the wars ends in Syria, a great percentage of them will not turn back to their country [21, 22, 23].

The refugees spreading to many cities of Turkey bring problems with themselves such as social, economic, health, education etc. The distribution of Syrian refugees to the cities is shown in the Figure-1.

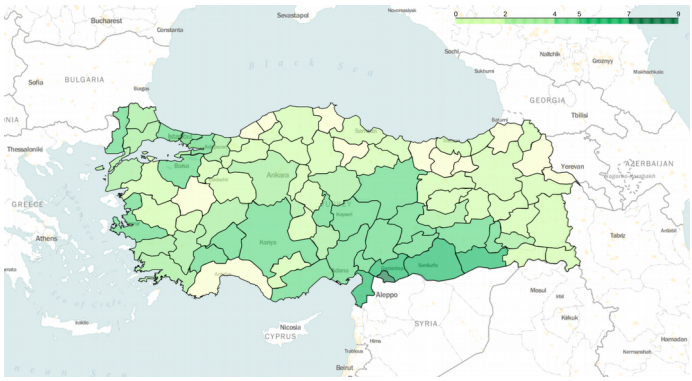


Fig-1 Distribution of Syrian refugees by city

As seen on the heat map, there is a dense refugee rate in the cities of Syria border. The population of refugee spreading from Syria border to central Anatolia live densely in developed industrial zones like Kocaeli, Bursa, Konya and metropolises like İstanbul, Ankara and İzmir. The population of refugees in some cities is more than a lot of city population in Turkey.

Recent developments in Syria show that the war will continue quite a while. As long as the war continues, the number of refugees who had to migrate to Turkey are increases.

Turkey admitted Syrian refugees with open door policy of political authority without an extensive integration plan. Even though the level of acceptance of Turkish society is high, hate crimes and racist discourse against to Syrian refugees are increasing gradually. The burdens of refugees to the economy, growing unemployment increase the risk of conflict between Turkish society and refugee society.

4. Materials and Methods

4.1 Data

In this study mobile call detail records data was used by Turk Telekom for the contest which is named data for refuges [24]. The data which was provided by Turk Telekom- comprises of mobile call records that are belonged Turkish citizens and refugees gathered in between December 2017 and January 2017. The data set that was prepared for D4R contest was obtained from 992.457 Turk Telekom clients in total and 184.949 of these were labeled as refuges [24]. The data consists of three parts. Regional distribution aerial locations in Turkey is given in Figure-2.

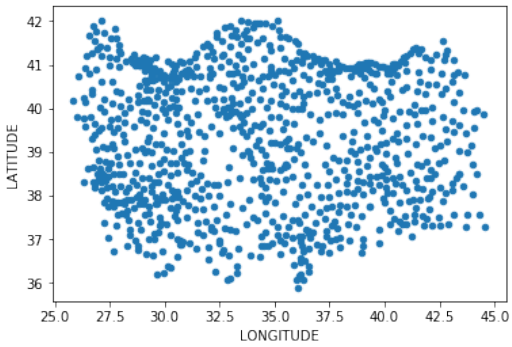


Fig-2 Geographical distribution of antenna positions

4.1.1 Dataset - 1 Antenna Traffic

In this data set, from area to area traffic is given. Each row corresponds to one record in data set. In each row, time stamp, the identity of area taking and incoming calls, total numbers of all the calls between two calls, the number of calls originating from the refugees, the information belongs to total time of all the calls between two areas are located [24].

4.1.2 Dataset – 2 Fine Grained Mobility

This data set contains call tower identifiers used by randomly selected active users to make phone calls and send text messages. Data is time stamped and randomly selected users have been examined for two weeks. At the end of the biweekly period, a new sample has been chosen from among the active users. In order to protect personal information, random identifiers have been assigned to users. In each row; caller id, time stamp, prefix of the called, the id of the area recording the call and information of the caller type are available in this data set [24].

4.1.3 Dataset – 3 Coarse Grained Mobility

In this data set, monitoring is given which consists of randomly selected 50.000 non-refugee users and randomly selected 50.000 refugees users via reduced spatial resolution for the whole observation process. Spatial resolution is given at district-based instead of aerial identifiers at this data set. There are caller identification, time stamp and the information of district id in this data set which has been taken from 481 districts [24].

4.2. Method

The most used index developed by Duncan and Duncan to measure the segregation of evenness dimension is Dissimilarity index (D). Dissimilarity Index identifies the minority population needs to be displaced to get homogeneous distribution of population in a city. Suggested in 1955, the index

$$D=\frac{1}{2}\sum_{i=1}^{n}\left|\frac{b_i}{B}-\frac{w_i}{W}\right|$$

was given with this formula. Here; w_i i. means the number whites in the area , W means total number of white in the area, b_i i. means the number of blacks living in the area and B means the number of whites living in that city [10].

Another index used in the study is the modified isolation index (MII), which gives a measure of the exposure dimension of the group that is discussed in a given area and shows the likelihood of interaction with other group members. Modified index of isolation

$$MII = \left(\sum_{i=1}^{n} \left[\left(\frac{w_i}{W} \right) * \left(\frac{w_i}{b_i} \right) \right] - \left(\frac{W}{B} \right) \right) / \left(1 - \frac{W}{B} \right)$$

is given with this formula [25, 26]. The purpose of using this index is that the index uses relative magnitudes instead of absolute magnitudes of groups.

In this study, w_i means the number of calls made by the refugees in i. area and b_i means the number of calls made by the non-refugees.

The use of mobile phones varies considerably in different groups, for example in different age categories, in different socio-economic groups [27]. One of the weaknesses of CDR data is its dependence on mobile cell phone use [7]. It is possible to overcome these weaknesses with long-term big data.

5. Results

In this paper, Dataset - 2 Fine Grained Mobility dataset was used provided by Turk Telekom to obtain the results. In this dataset a unique id was given to every user instead of the phone number. The 10-digit ID numbers start with 1 for refugees, 2 for non-refugees and 3 for unknowns. A sample data row looks like this

```
CALLER_ID,TIMESTAMP,CALLEE_PREFIX,SITE_ID,CALL_TYPE
1100140407,02-01-2017 16,2,5059421,2
1100140407,02-01-2017 16,2,5191433,2
```

In order to calculate the segregation indexes given in the method section, incoming and outgoing voice data are used. With an aim to determine the number of calls taken place among the refugees, call records beginning with the CALLER_ID column 1 with a CALLEE_PREFIX value of 1 were counted. Similarly , to determine the number of calls taken place among the non-refugee users, number of lines beginning with the CALLER_ID column 2 with a CALLEE_PREFIX value of 2 were found.

Data about the antenna positions of the calls is in the SITE_ID column. SITE_ID is combined with the data in Base_Station_Location.txt file to determine the locations of mobile call records. Dissimilarity Index is calculated by dividing the data into administrative or population areas. In this paper, the antenna positions are used as the unit area. The dissimilarity index values are shown below in the heat map. Dark green colors correspond to high dissimilarity index values.

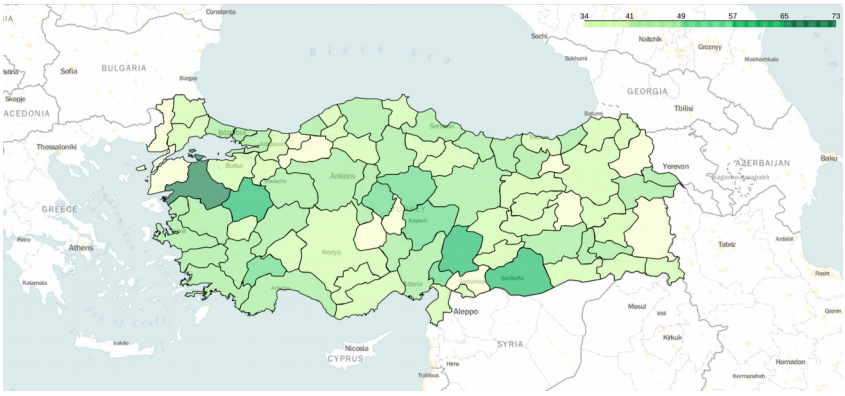


Fig-3 Dissimilarity index results by cities

The highest value with 0.74 was found in Balıkesir province. In addition, D value in Kütahya (0.71), Şanlıurfa (0.68), Kahramanmaraş (0.67), Osmaniye (0,65) is high. The lowest value with 0.34 was found in Bayburt province. Aksaray (0.34), Gaziantep (0,35) are the cities that D values is low. In Istanbul, Ankara and İzmir, the D value is around 50 percent. The D values are 0.51, 0.55 and 0.55, respectively. The obtained values are given in Appendix-1.

For the normality test of D index The Shapiro-Wilk test was used. D index does not have normal distribution according to the results of this test ($W = .95$, $p < .05$). The Kruskal-Wallis test was used to determine whether there was a statistical difference between the geographical regions. According to the test results, statistically significant results were found between geographical regions ($\chi^2 = 14.98, p = 0.02, df = 6$).

The modified index of isolation values are shown below in the heat map.

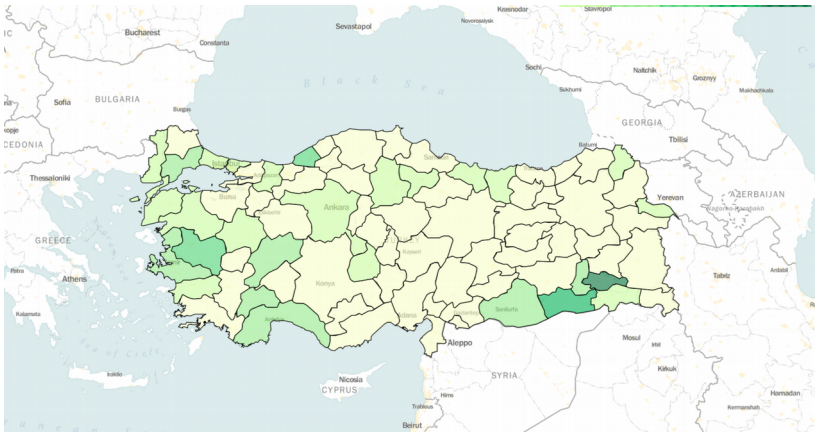


Fig-4 Modified index of isolation results by cities

The highest value with 0,53 was found in Osmaniye province. Also, MII value in Denizli (0.44), Kilis (0.38), Rize (0.38) is high. The lowest value with 0.002 was found in Bayburt province. The obtained values are given in Appendix-2 According to Shapiro-Wilk test result MII does not have normal distribution (W=.176, p< 0.05), therefore classic statistical methods such as one-way ANOVA can not be performed. In this study we use the Kruskal-Wallis test alternative the one-way ANOVA test that is preffered to use non-normality exist in the data. After performing the Kruskal-Wallis test, results show that there is no evidence to reject the null hypotesis stated as the median value of the seven geographical areas are not differ each other, i.e.

$H_0 : M_1 = M_2 = ... = M_7$ where M_i denotes the median value of the $i-th$ geographical area. In other words there is no statistically significant difference among the median value of the geographical regions ($\chi^2 = 10.966, p = 0.089, df = 6$); see the Table-2

6. Discussion

In this study, segregation of Syrian refugees in Turkey were examined. For this purpose, the dissimilarity index which is the most used index in the literature and the modified index of isolation are used. Results are shown on the heat map. Survey data, population data are generally used for such studies. We used mobile call detail records in this study to calculate segregation with known indexes in the literature. Similar calculations can be made in the future with other indexes used in the literature. Even segregation can be calculated on a personal basis.

The high level of segregation also increases the risk of intergroup conflict. For this reason, public administrations should increase their security measures especially in cities with high segregation value. As a short-term solution according to a similar study conducted at the neighborhood level, refugee resettlement can be considered by public administrations.

For many years, it is foreseen that Syrian refugees will continue to reside in Turkey. For this reason, in the long term the programs that will contribute to the integration of Syrian refugees should be included in the education system.

References

1. Massey, D. S., & Denton, N. A. (1988). The dimensions of residential segregation. Social forces, 67(2), 281-315.

2. Caldeira, T. P. (2000). City of walls: crime, segregation, and citizenship in São Paulo. Univ of California Press.

3. Collins, W. J., & Margo, R. A. (2000). Residential segregation and socioeconomic outcomes: When did ghettos go bad?. Economics Letters, 69(2), 239-243.

4. Cutler, D. M., & Glaeser, E. L. (1997). Are ghettos good or bad?. The Quarterly Journal of Economics, 112(3), 827-872.

5. Trounstine, J. (2016). Segregation and inequality in public goods. American Journal of Political Science, 60(3), 709-725.

6. Uslaner, E. M. (2012). Segregation and mistrust: Diversity, isolation, and social cohesion. Cambridge University Press.

7. Silm, S., & Ahas, R. (2014). The temporal variation of ethnic segregation in a city: Evidence from a mobile phone use dataset. Social Science Research, 47, 30-43.

8. Bell, W. (1954). A probability model for the measurement of ecological segregation. *Social Forces*, *32*(4), 357-364.

9. Cowgill, D. O., & Cowgill, M. S. (1951). An index of segregation based on block statistics. *American Sociological Review*, *16*(6), 825-831.

10. Duncan, O. D., & Duncan, B. (1955). A methodological analysis of segregation indexes. *American sociological review*, *20*(2), 210-217.

11. Jahn, J. A. (1950). The measurement of ecological segregation: Derivation of an index based on the criterion of reproductibility. *American Sociological Review*, *15*(1), 100-104.

12. Jahn, J., Schmid, C. F., & Schrag, C. (1947). The measurement of ecological segregation. *American Sociological Review*, *12*(3), 293-303.

13. Cortese, C. F., Falk, R. F., & Cohen, J. K. (1976). Further considerations on the methodological analysis of segregation indices. *American sociological review*, 630-637.

14. Wong, D. W. (2005). Formulating a general spatial segregation measure. *The Professional Geographer*, *57*(2), 285-294.

15. Bell, W. (1954). A probability model for the measurement of ecological segregation. *Social Forces*, *32*(4), 357-364.

16. Atkinson, A. B. (1970). On the measurement of inequality. *Journal of economic theory*, *2*(3), 244-263.

17. Reardon, S. F., & Firebaugh, G. (2002). Measures of multigroup segregation. *Sociological methodology*, *32*(1), 33-67.

18. Morgan, B. S. (1975). The segregation of socio-economic groups in urban areas: a comparative analysis. *Urban Studies*, *12*(1), 47-60.

19. Sakoda, J. M. (1981). A generalized index of dissimilarity . *Demography*, *18*(2), 245-250.

20. Jargowsky, P. A. (1996). Take the money and run: Economic segregation in US metropolitan areas. *American sociological review*, 984-998.

21. Sirkeci, I. (2017). Turkey's refugees, Syrians and refugees from Turkey: a country of insecurity. *Migration Letters*, *14*(1), 127-144.

22. Orhan, Oytun, Şenyücel, Gündoğar Sabiha (2015). Suriyeli Sığınmacıların Türkiye'ye Etkileri, ORSAM, Ankara

23. Güçtürk, Y. (2014, February). İnsanlığın kaybı: Suriye'deki iç savaşın insan hakları boyutu. SETA.

24. Salah, A.A., Pentland, A., Lepri, B., Letouzé, E., Vinck, P., de Montjoye, Y.A., Dong, X. and Dağdelen, Ö., 2018. Data for Refugees: The D4R Challenge on Mobility of Syrian Refugees in Turkey. arXiv preprint arXiv:1807.00523.

25. Johnston, R., Poulsen, M., & Forrest, J. (2011). Evaluating changing residential segregation in Auckland, New Zealand, using spatial statistics. Tijdschrift voor economische en sociale geografie, 102(1), 1-23.

26. Marcińczak, S., Musterd, S., & Stępniak, M. (2012). Where the grass is greener: social segregation in three major Polish cities at the beginning of the 21st century. European Urban and Regional Studies, 19(4), 383-403.

27. Wei, R., & Lo, V. H. (2006). Staying connected while on the move: Cell phone use and social connectedness. New Media & Society, 8(1), 53-72.

Appendix-1 Dissimilarity Index Values

The Marmara	D	The Aegean	D	The Black Sea	D	The Eastern Anatolia	D
Edirne	0,37	İzmir	0,55	Rize	0,55	Ağrı	0,55
Kırklareli	0,51	Manisa	0,50	Trabzon	0,46	Ardahan	0,43
Tekirdağ	0,45	Aydın	0,50	Artvin	0,46	Bingöl	0,39
İstanbul	0,52	Denizli	0,54	Sinop	0,46	Bitlis	0,42
Kocaeli	0,49	Kütahya	0,71	Tokat	0,46	Elazığ	0,47
Yalova	0,55	Afyonkarahisar	0,56	Çorum	0,50	Erzincan	0,45
Sakarya	0,48	Uşak	0,44	Amasya	0,44	Erzurum	0,44
Bilecik	0,43	Muğla	0,55	Samsun	0,53	Hakkari	0,48
Bursa	0,47			Zonguldak	0,55	Iğdır	0,52
Balıkesir	0,74			Bolu	0,36	Kars	0,38
Çanakkale	0,38			Düzce	0,36	Malatya	0,43
The Central Anatolia	**D**	**The Mediterranean**	**D**	Karabük	0,43	Muş	0,46
Aksaray	0,34	Adana	0,52	Bartın	0,45	Tunceli	0,49
Ankara	0,55	Osmaniye	0,65	Kastamonu	0,53	Van	0,39
Çankırı	0,43	Antalya	0,53	Bayburt	0,34	Şırnak	0,49
Eskişehir	0,51	Burdur	0,57	Giresun	0,49		
Karaman	0,47	Hatay	0,48	Gümüşhane	0,53	**The Southeastern Anatolia**	**D**
Kırıkkale	0,47	Isparta	0,48	Ordu	0,52	Adıyaman	0,48
Kırşehir	0,47	İçel	0,48			Batman	0,51
Konya	0,45	Kahramanmaraş	0,67			Diyarbakır	0,54
Nevşehir	0,40					Gaziantep	0,36
Niğde	0,54					Kilis	0,58
Sivas	0,52					Mardin	0,45
Yozgat	0,57					Siirt	0,54
Kayseri	0,57					Şanlıurfa	0,68

Appendix – 2 Modified Index of Isolation Values

The Marmara	MII	The Aegean	MII	The Black Sea	MII	The Eastern Anatolia	MII
Edirne	0,02	İzmir	0,12	Rize	0,38	Ağrı	0,04
Kırklareli	0,22	Manisa	0,10	Trabzon	0,27	Ardahan	0,09
Tekirdağ	0,03	Aydın	0,08	Artvin	0,02	Bingöl	0,00
İstanbul	0,06	Denizli	0,44	Sinop	0,09	Bitlis	0,01
Kocaeli	0,06	Kütahya	0,11	Tokat	0,04	Elazığ	0,07
Yalova	0,22	Afyonkarahisar	0,09	Çorum	0,23	Erzincan	0,05
Sakarya	0,08	Uşak	0,05	Amasya	0,16	Erzurum	0,06
Bilecik	0,03	Muğla	0,28	Samsun	0,21	Hakkari	0,02
Bursa	0,04			Zonguldak	0,12	Iğdır	0,13
Balıkesir	0,01			Bolu	0,03	Kars	0,10
Çanakkale	0,02			Düzce	0,01	Malatya	0,11
The Central Anatolia	**MII**	**The Mediterranean**	**MII**	Karabük	0,09	Muş	0,00
Aksaray	0,03	Adana	0,07	Bartın	0,06	Tunceli	0,07
Ankara	0,12	Osmaniye	0,53	Kastamonu	0,14	Van	0,03
Çankırı	0,08	Antalya	0,07	Bayburt	0,00	Şırnak	0,06
Eskişehir	0,09	Burdur	0,14	Giresun	0,09		
Karaman	0,12	Hatay	0,04	Gümüşhane	0,03	**The Southeastern Anatolia**	**MII**
Kırıkkale	0,05	Isparta	0,07	Ordu	0,10	Adıyaman	0,13
Kırşehir	0,07	İçel	0,09			Batman	0,14
Konya	0,02	Kahramanmaraş	0,19			Diyarbakır	0,07
Nevşehir	0,04					Gaziantep	0,03
Niğde	0,07					Kilis	0,39
Sivas	0,22					Mardin	0,10
Yozgat	0,17					Siirt	0,24
Kayseri	0,20					Şanlıurfa	0,22

Reaching all children: A data-driven allocation strategy of educational resources for Syrian refugees.

Suad AlDarra[1], Laura Alessandretti[2], Lorenzo Lucchini[3,4], and Elisa Omodei[1]

[1] UNICEF Office of Innovation, 3 United Nations Plaza, New York, NY 10017, USA
[2] Technical University of Denmark, DK-2800 Kgs., Lyngby, Denmark
[3] University of Trento, Sommarive 9, I-38123, Trento, Italy
[4] FBK, Sommarive 18, I-38123, Trento, Italy

Abstract. The number of Syrian refugees in Turkey has reached 3.4 millions at the end of 2017, of which almost 50% are children under 18 years old. The Turkish government, Unicef and the EU are allocating resources towards their education. Despite this effort, Syrians schooling rates remains under 60%. According to field studies, one of the obstructions to Syrian childrens education is the lack of statistics about school-age children in Turkey. Difficulties in data collection include the fact that families relocate frequently within Turkey, and that the total Syrian population keeps growing. In this report, we use mobile phone data to estimate the spatial distribution of the Syrian population and its changes across 2017. We show that, during 2017, cities close to the Syrian border and large cities have experienced the largest population increase. At the same time, less popular cities have experienced the largest percentage increase. We found that cities near the Syrian border have experienced a migration wave during the summer of 2017, followed by a constant population growth. On the other hand, during the same period, large cities underwent a drop in population. The distribution and changes of the Syrian refugee population can proxy well those of school-age children. Hence, the results of our study can potentially inform targeted interventions in the different Turkish provinces to address the needs of Syrian children refugees.

Keywords: Education · Refugees · CDR

1 Introduction

Over the past years, Turkey has been a haven for over 3 million Syrian refugees, one of the largest refugees population in the world. With half of the migrants being children and numbers resuming to increase, Unicef, the government of Turkey and other partners have launched a campaign titled 'No Lost Generation' [1]. The campaign aims to prevent the loss of an entire generation of Syrian children and guarantee them the right to education. Despite the efforts of the Turkish Ministry of National Education (MoNE) to facilitate Syrian children

enrollment in public schools and to create temporary education centers staffed with Syrian teachers, refugee children needs are still hard to meet. MoNE estimates that around 400,000 Syrian children are out of schools (see fig. 1) due to barriers including forced children labour, lack of Identification Documents (IDs), language challenges, and access to transportation [9, 12].

According to [12], reporting a field study on the education of Syrian children in Turkey, another issue is the lack of data that could be used for monitoring Syrian children access to school. The Provincial/District Directorates of National Education (MEM) officials interviewed in the study could not provide precise statistics about Syrian children and their schooling rates. The study suggests that more effort should be devoted to identify unschooled children.

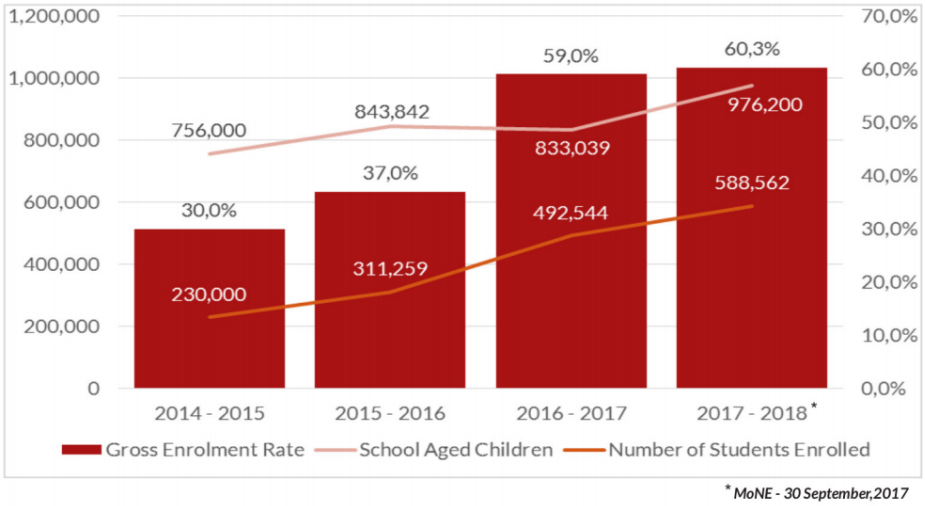


Fig. 1. Gap between the number of school aged Syrian children and number of children enrolled in school. The red bars show the schooling rate of Syrian children in Turkey, the pink line shows the number of school-age children and the orange line shows the total number of Syrian children enrolled in schools.

Data around refugees is hard to collect in real time due to reasons including high cost and limited resources. Data collection is also hindered by the size of the Syrian refugee population, its diversity, and the movements of refugees across the country [6]. At the same time, recent research has shown that good estimations of population densities can be produced at national scales, from alternative sources including mobile phone call data [7].

In this report, we analyze anonymized and aggregated Call Phone Records (CDR) of refugees shared by Turk Telekom, to estimate the changes in the spatial distribution of refugees in Turkey over 2017. Our results can be beneficial for short and long term allocation of educational resources.

2 Results

According to the UNHCR data portal on the Syrian refugee situation in Turkey ([13], see also appendix A.1), the number of Syrian refugees in Turkey has been growing over 2017 (see fig. 2), from ~ 2.8 millions in January up to ~ 3.4 millions in December 2017. Likewise, the CDR volume from Syrian refugees customers registered by Turk Telekom has been increasing. The data, collected by the Government of Turkey, show that 45% of the refugees are under 18 years, 31 % are under 12 and 15% are under 5 years old [13]. In the following, we refer to school aged children as to children between 5 and 18.

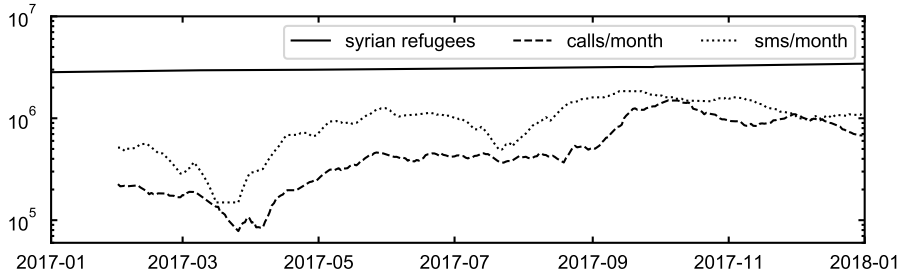


Fig. 2. Growing Syrian refugees population in Turkey. The number of Syrian refugees in Turkey [13] (plain line), the number of outgoing calls (dashed line) and sms (dotted line) from Syrian refugees over a rolling window of 30 days.

In March 2017, refugees were located predominantly in Istanbul ($\sim 100/Km^2$ individuals) and the provinces close to the Syrian border (Hatay, Kilis, Gaziantep, Sanliurfa), where the density of refugees is higher than $20/Km^2$ individuals (see fig. 3-A). The density of Turkish population and distance from Syria alone explain most of the variation in the data (see fig. 3-B). A significant number of Syrian refugees is located in provinces with high density of Turkish population, indicating important centers and cities in Turkey. Provinces close to the Syrian border are also more densely populated.

Data collected by MoNe in March 2017 [12] reveals that, not surprisingly, there is strong correlation between the number of school age refugees and total refugees per province (see fig. 4A). However, the proportion between the two fluctuates between 25% and 38% (see fig. 4C). Discarding these differences, in the following, we consider the distribution of all Syrian refugees as a proxy for the distribution of school-age children across the country.

2.1 From CDR to population density estimates

In this section, we present the method followed to estimate refugee population density over time from CDR data.

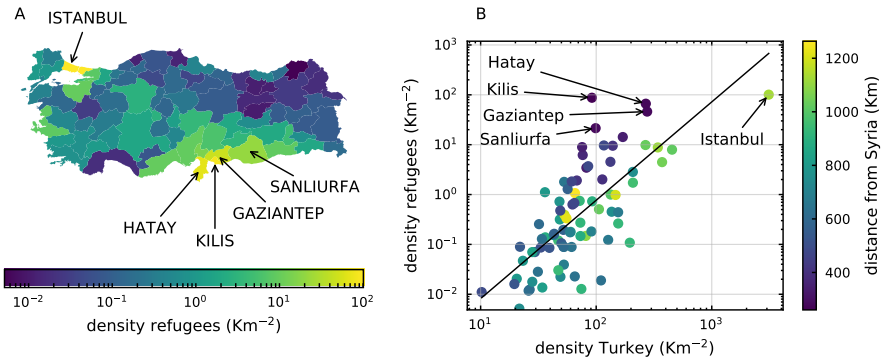


Fig. 3. Factors explaining the distribution of Syrian refugees in Turkey. A) The number of Syrian refugees per Km^2 in the various Turkish provinces (data from March 2017. Provinces with more than 20 refugees per Km^2 are indicated [10]. B) The density of Syrian refugees vs the density of Turkish in all provinces. Colors indicate the distance from Syria in Km. Data is fitted with a model of type $\rho_S = a \cdot \rho_T^b$, where ρ_S is the density of Syrian refugees, ρ_T is the density of Turkish population, A and B are reported in the legend. We report also the error $RMSE$ and the R^2 coefficient of the linear model (note that these are the errors on the linear fit of log-transformed variables). Data from [3] and [10].

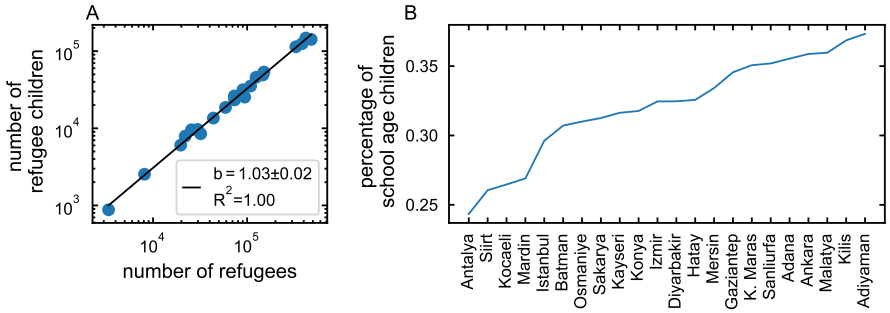


Fig. 4. Number of Syrian school age refugees. A) The number of school age Syrian children refugees per province vs the total number of Syrian refugees (blue dots, data from [12]). The black line indicates a power law fit with exponent $b = 1.03 \pm 0.02$ and $R^2 = 1$. B) The proportion of school age Syrian refugees per province. Only the 22 most populated provinces are shown.

Our analysis is based on dataset D1, providing Turk Telekom antenna traffic for refugees and non-refugees (see appendix A.1) over a year. We find the area covered by each antenna using Voronoi tessellation (see appendix A.2 and fig. 10). This allows to find the overlap between antennas and the 81 administrative Turkish provinces (see appendix A.1)

We follow the method developed in [7], that has proven to successfully estimate population densities from CDR data. In [7], the authors show that the relation between the population density ρ_c within a given region c and the night-time call density σ_c within the same region can be modelled as a power-law: $\rho_c = \alpha \cdot \sigma_c^{\beta}$.

First, we fit the model. Since the official population estimates, provided for the *Data4Refugee* challenge [10] are collected in March 2017, we estimate the parameters α and β using CDR data from the same month. In our framework, $\rho_c(t) = N(t)/A(c)$, where $N(t)$ is the number of night-time CDR (8 pm to 7 am) in the month preceding time t, and $A(c)$ is the area of c in Km^2.

We fit four different models, considering incoming and outgoing calls or sms. Using a Markov chain Monte Carlo (MCMC) approach (see appendix A.3), we find the values of α and β (see fig. 5) together with the estimate of the error of the model. The Root Mean Squared Error (RMSE) of the model ranges between 9.88 and 12.67, while R^2 ranges between 0.26 and 0.75. For comparison, we run the same method on data for the Turkish population (see appendix A.4 and fig. 11). In this case, better performance (in terms of normalized RMSE) is achieved.

In the following, we consider the model based on outgoing call density (which achieves the best performance). We use the model to estimate the density $\rho_c(t)$ of individuals in space over 2017. Estimations are computed every day over a rolling window of 30 days. We account for the fact that the number of Turk Telekom users fluctuates in time (see fig. 2), by normalizing every month for the total number of refugees reported in [13]. Note that the fluctuations observed for refugees and Turkish correlate, suggesting they are driven by seasonality in calling behaviour (e.g. less calls during the summer).

2.2 Evolution of the Syrian distribution in time.

The analysis reveals changes in the distribution of users between the beginning and the end of 2017 (see fig. 6A). Regions most populated by refugees in the beginning of 2017 have attracted most of the incoming flux during the year (see fig. 6A). The largest percentage change in 2017 (among the top 20 provinces) is observed in the region of Osmaniye (see fig. 6B).

We further investigate the relation between the median density of Syrian refugees before March 2017 and after December 2017 (fig. 7A). We find that the final density ρ_f of refugees grows sublinearly with the initial density ρ_i as $\sim \rho_i^{0.94}$, and, as a consequence, the percentage change in density decreases as a function of the initial density (fig. 7B). This implies that areas with lower density in March 2017 are those that, in relative terms, attract more refugees, whereas areas that already have high refugee density have a lower relative change in refugee population. As a robustness test, we have verified that these results hold

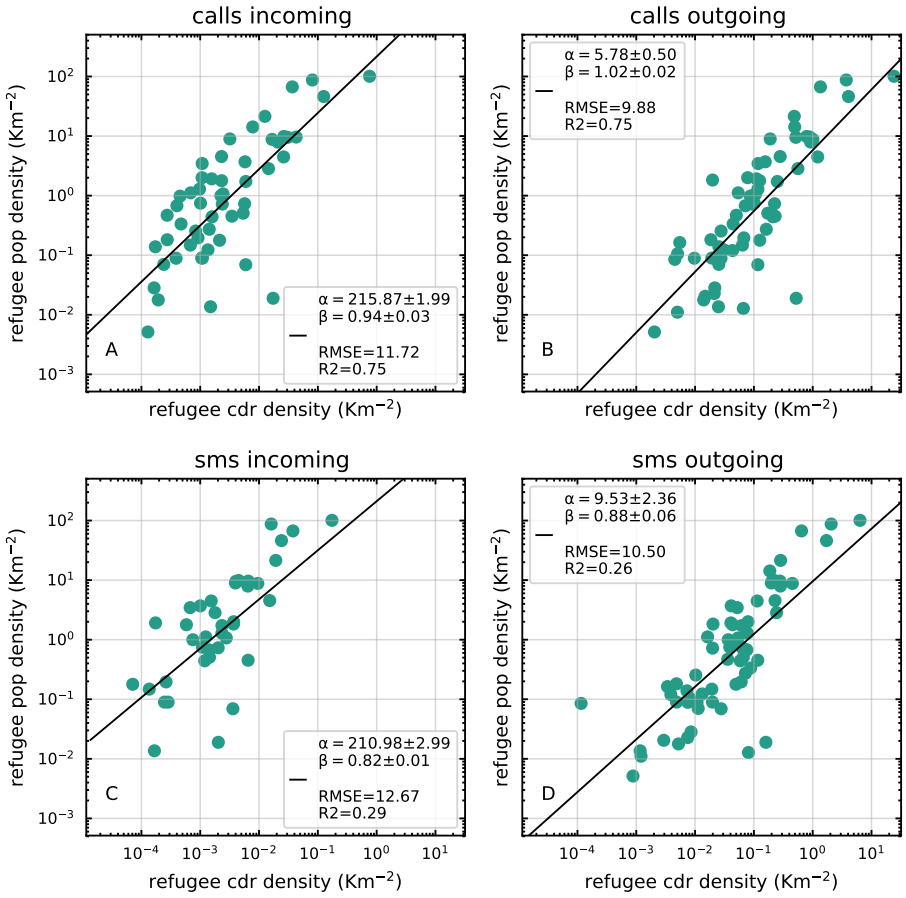


Fig. 5. Inferring population densities from CDR data. The density of Syrian refugees per province as a function of the night-time call density computed over the month of March 2017 (circles). The inferred relation $\rho_c = \alpha \cdot \sigma_c^{\beta}$ between night-time call density σ_c and population density ρ_c(line). The parameter α and β, together with the coefficient of determination R^2 and the RMSE are indicated in the legend. Subplots display data for incoming calls (A), outgoing calls (B), incoming sms (C), outgoing sms (D).

when we consider CDR instead of population densities (fig. 12) and provinces instead of cells (fig. 13). Note also that this effect is not observed for the Turkish population.

Finally, we consider the evolution of the 20 top provinces over time (see fig. 8). For regions relatively close to the Syrian border (see fig. 8A), we observe a similar pattern: a wave between May and August 2017 is followed by a constant increase that continues until the end of the year. For larger cities including Istanbul, Bursa

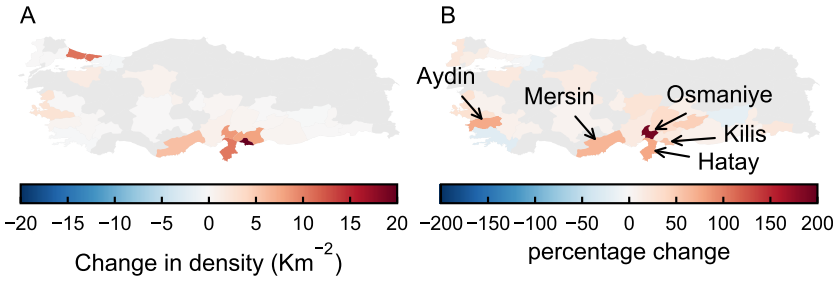


Fig. 6. Change in population distribution. A) The estimated absolute difference in the density of Syrian Refugees between March and December 2017. B) The estimated change in density of Syrian Refugees between March and December 2017. Here, we show only the 20 top provinces in terms of total number of refugees.

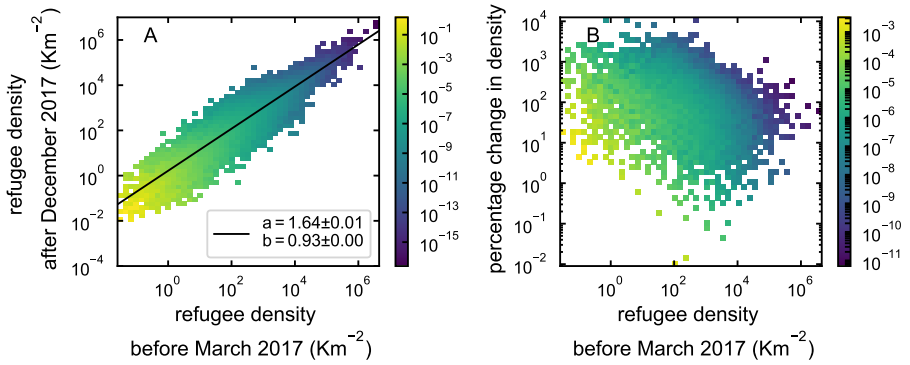


Fig. 7. Relation between initial and final density. (A) Density of refugees before December 2017 vs the density after March 2017 for all antennas. The full line shows a model $\rho_f = a \cdot \rho_i^b$ where ρ_f and ρ_i are the final and initial densities, respectively. The linear model of the log-transformed variables has root mean squared error $RMSE = 0.47$ and coefficient of determination $R^2 = 0.92$. (B) Percentage change in density between March and December 2017 vs the density in March 2017. The two quantities are negatively correlated with Spearman coefficient $S = -0.34$, $p < 0.01$.

and Ankara (see fig. 8B), the refugee population goes instead through a decrease between May and August 2017, and stabilizes after August. Finally, a set of 4 cities displays different evolution patterns. Among them, the city of Osmaniye experiences the largest relative increase in population.

2.3 Distribution of schools in Turkey.

Based on a report by UNICEF Turkey and UNHCR, by December 2017, the enrollment of Syrian children in schools has increased to 61.8 % [2]. Turkey gov-

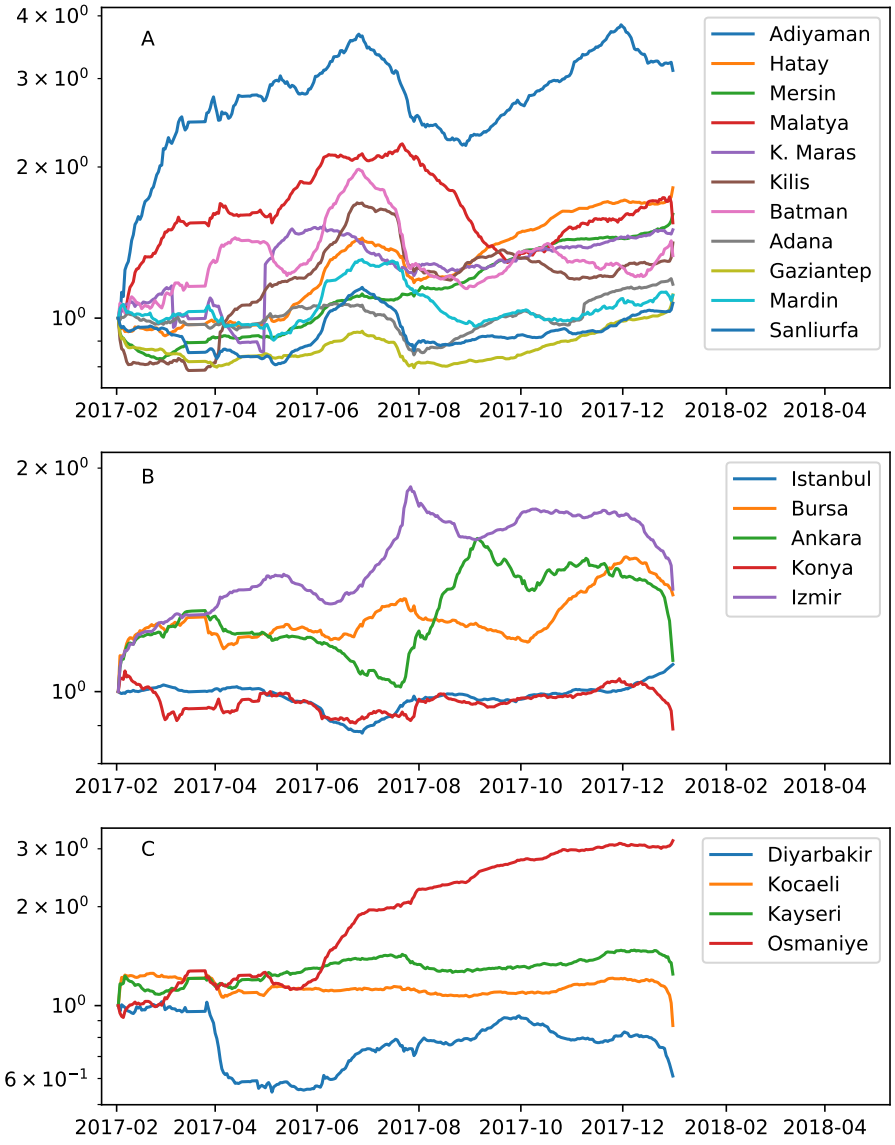


Fig. 8. Evolution of the Syrian population in the top 20 provinces. Normalized number of refugees over time for (A) Provinces where a peak in the number of refugees is observed between May and August 2017. (B) Provinces such that a decrease in number of refugees is observed between May and August 2017. (C) Other provinces. The number is normalized by the value obtained for February 2017.

ernment along with partners have made efforts to cover the educational gap such

as waving fees for Syrian students, offering cash assistance to stop child labour, providing scholarships and increasing the number of education 'interventions' as shown in fig. 9. The provinces with the highest interventions are Istanbul and the Syrian borders regions which matches the same regions highlighted in fig. 6, with the most increased change in density.

However, it is important to notice that several provinces with high number of Syrian refugee received fewer interventions than others with lower number levels. An important example is given by the province of Ankara, even if the number of refugees was less than 100.000, the number of interventions (32) exceeded the number of interventions of other more refugee-populated provinces, e.g. Bursa.

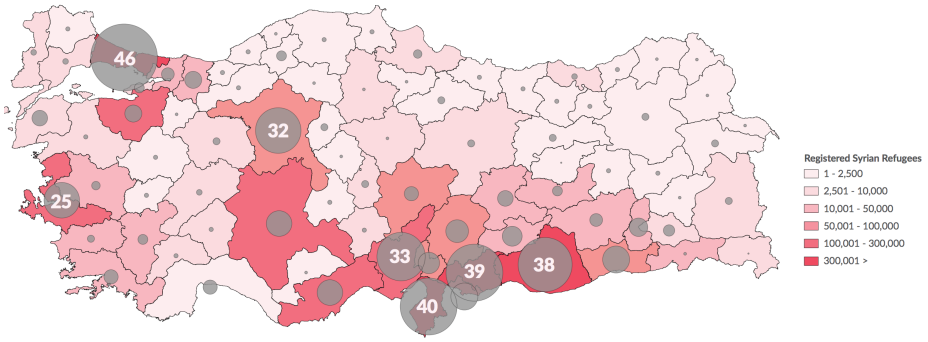


Fig. 9. Number of registered Syrian refugees and number of education interventions by province - UNHCR

3 Conclusions

We have extracted information on the evolution of the Syrian refugee population in Turkey, using data on the outgoing calls effectuated by refugees in Turkish territory. Given the observed correlation between the number of refugees and the number of refugee children in a given province, our results can help estimating the number of school-age children and its evolution across time. We have shown that refugees are established in areas close to the Syrian borders (e.g.: Gaziantep) and in high density areas (e.g.: Istanbul), matching the figures provided by UNHCR [13] (see fig. 9). We have found that areas with low refugee density display the highest percentage change over 2017 (see fig. 6). Possibly, this is the result of relocation policies proposed by the Government of Turkey. In certain areas, such as Antalya, we have observed that the number of refugees has overall decreased during 2017. Antalya is known as a touristic area where refugees were not allowed to work or reside, but, on the other hand, it is an active smuggling point to Europe due to its location. The decrease in density could be explained by the

fact that less smuggling take place towards the end of the year, when it is colder and more stormy, or due to increased security measures. We have shown that, in regions close to the Syrian border the refugee population has experienced a peak during the summer 2017, followed by a constant increase until the end of the year. Instead, in large cities such as Istanbul, Bursa and Ankara, the number of refugees estimated by CDR has dropped during the summer and stabilized between August and September. Optimized targeted intervention to help refugees in the integration and education process could be designed based on the result of our study. In particular, the provinces that can be identified as those addressing lower density chances (as depicted in fig. 6) and with a higher number of refugees (as shown in fig. 9) might be places in which more stable resolving and longer term policies can be proposed. In contrast, those with sharper changes in refugee density are suggested to be those with higher needs of fast and specific (place-dependent) interventions.

A Appendix

A.1 Data description

CDR data (Turk Telecom dataset D1) Data provides the number of outgoing and incoming calls/sms per antenna, for refugee and non-refugee individuals. It is aggregated in temporal bins of 1 hour and it includes data from 53842 different antennas. No data is stored if no call were performed in a specific temporal bin from (outgoing call) or to (incoming call) a set of antennas for that set of antennas in that temporal bin. The dataset is divided in 12 different files roughly aggregated by month. From each file we discarded the information about the outgoing/incoming calls whose temporal bin did not belong to the same month as labeled in the file name. The 12 files together span a time period of one year, from 1 January 2017 to 31 December 2017. We report that for several days CDR data is missing for almost the entire set of antennas. In our analysis we considered only days for which at least one call was recorded for at least one antenna.

Administrative borders data Administrative regions border in Turkey are obtained from [4]. Here, we consider the first administrative level (provinces) as the highest level for which educational policies in Turkey can be proposed with different intensities, depending on the needs of the single province.

Groundtruth data In order to estimate the refugee population in the different provinces during the whole period for which CDR data were available during the challenge, we used the "*Refugees per city(March 2017*" dataset provided by the organizers [10]. The dataset contains the number of Syrian refugees during March 2017 at the province level together with the Turkish population is provided [10]. These data was used as groundtruth to estimate the parameters of the model.

For a temporally longer comparison we used a more coarse grained dataset: the number of refugees in Turkey with a weekly update for all the 2017. This data was collected by UNHCR and the Government of Turkey and it is available at [13]. With this information we were able to compare the fluctuations of the refugee population over time, the growth in number, and the internal migrations.

A.2 Voronoi tessellation

We find the areas covered by each antenna using Voronoi tessellation [8] (using the quickhull algorithm [5]). The typical area covered by an antenna is $0.36Km^2$ (median). 50% of the antennas cover an area between $0.07Km^2$ and $11Km^2$.

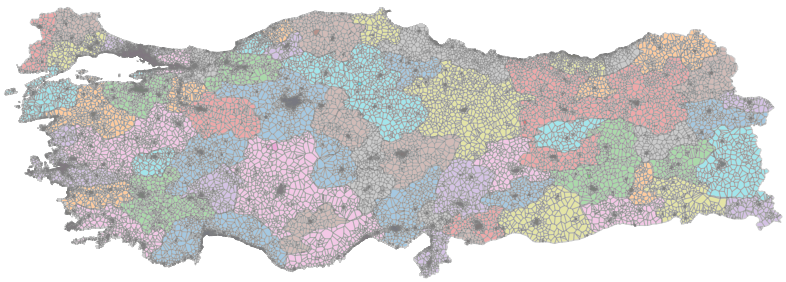


Fig. 10. Voronoi Tessellation. The area covered by each antenna in Turkey based on Voronoi tessellation. Colors correspond to different provinces.

A.3 Parameter estimation

As discussed above, we adopt the model from [7] to infer the Syrian refugee population density both at the province level and at the antenna level. However, due to the limited availability of Syrian refugee demographic data in Turkey, the estimation of the parameters of the model is performed at the province level using only CDR data during March 2017. The model proposed by Deville is summarized by the formula

$$\rho_c = \alpha \cdot \sigma_c^{\beta}, \tag{1}$$

where where ρ_c is the population density of Syrian refugees, σ_c is the density of call performed by the refugees in a certain time window. In our work we evaluated the density by computing the number of calls per squared kilometer at the level of the Voronoi cells. We then averaged over the province weighting the contribution of each Voronoi cell with the overlapping area between the cell and the province. We also want to stress the importance of the size of the

time window to study the temporal evolution of the population. In this work we trained the model averaging the number of call performed during March 2017 on a daily basis. By doing so we are able to estimate the population (and its time evolution) for all the days for which we have CRD data. The call density we use for the evaluation, σ_c, is the average number of call per day and square kilometer for each different province of Turkey.

To estimate the parameter we adopt a Markov chain Monte Carlo approach (MCMC). This technique is based on a Bayesian approach to the parameter estimation problem. MCMC requires some prior hypotheses on the distribution of the variable of the systems. In particular, we assumed a normal prior distribution for the two free parameters of the model, namely α and β. We assumed the posterior distribution of the dependent variable, ρ_c, to be normal whose width is modeled by a third parameter, σ, that encodes the information about the goodness of the fit and the precision of the estimate given by the model. σ is assumed to be an half-normal with standard deviation of 0.25. The estimate is done using PyMC3, a library for Bayesian inference written in Python [11]. To achieve a higher precision in the estimate, and to speed up the process we tackle the problem by estimating the parameters of the linear model obtained by computing the natural logarithm of eq. (1). This step is also suggested as a better performing solution in [7]. The results obtained for the parameter evaluation, using the procedure shown here, are presented and discussed in section 2.1.

A.4 Results for Turkish population

To compare the result obtained for the Syrian refugees, we also run the model using as training set the Turkish population data. The approach used is the same as the on discussed in the previous appendix. However, the parameter estimation for the Turkish population shows a better performance of the model. The higher number of call available for the Turkish population translate in a higher stability of the density of calls in the different provinces and over the time. This is mainly caused by the difference in the user number between Turks and Syrian refugees but possibly also more complicated social and anthropological motivation can be involved. in particular the different values of the exponent, β, shows that a simple difference in the amplitude does not explain the behavior of the two populations. In fig. 11 we show the results of the evaluation process for the Turk population. In the box are reported the estimates for the parameters of the model and the root mean square error and the coefficient of determination.

A.5 Robustness tests

The results presented hold also when considering the number of calls instead of density (see fig. 12) and the cells rather than provinces (see fig. 13). However, the RMSE increase while the coefficient of determination decrease. This is motivated, as for the Syrian refugees model, by the reduced records of incoming calls ans SMS with respect to the outgoing calls. In a similar way, we noticed that the

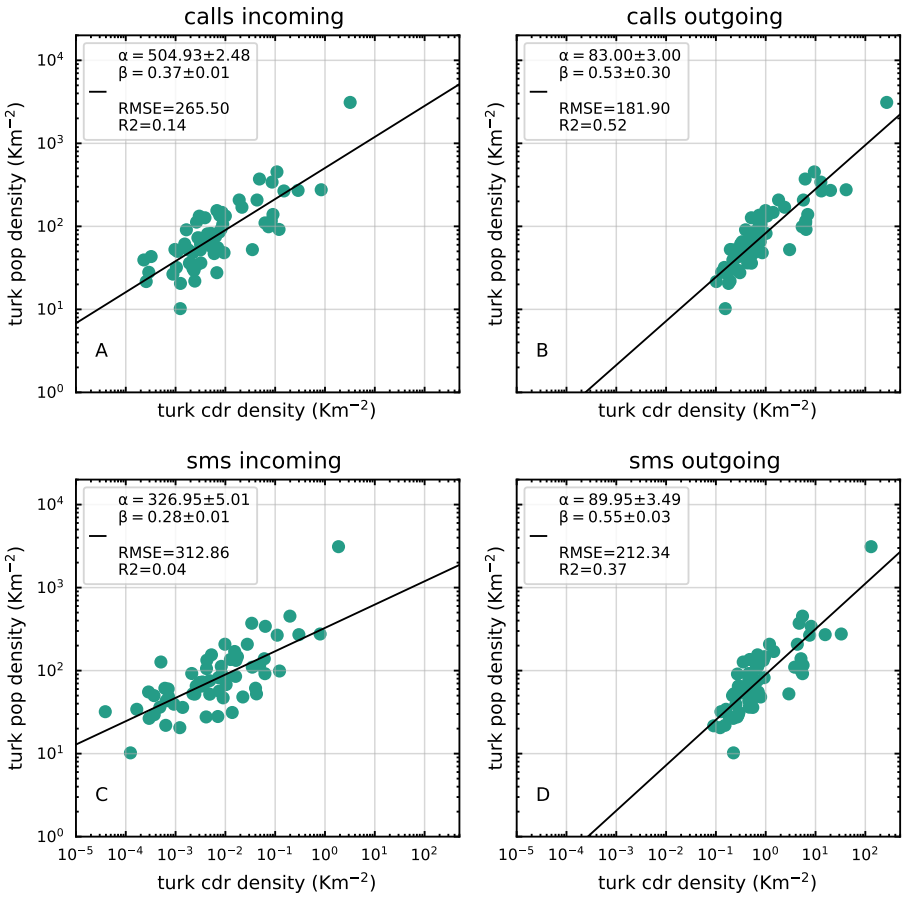


Fig. 11. Inferring population densities from CDR data (turkish population). The density of turkish per province versus the night-time call density computed over the month of March 2017 (circles). The inferred relation$\rho_c = \alpha \cdot \sigma_c^{\beta}$ between night-time call density σ_c and population density ρ_c(line). The parameter α and β, together with the coefficient of determination R^2 and the RMSE are indicated in the legend. Subplots display data for incoming calls (A), outgoing calls (B), incoming sms (C), outgoing sms (D).

number of outgoing SMS is smaller than the number of outgoing calls, motivating the more stable result for the model using the outgoing call dataset.

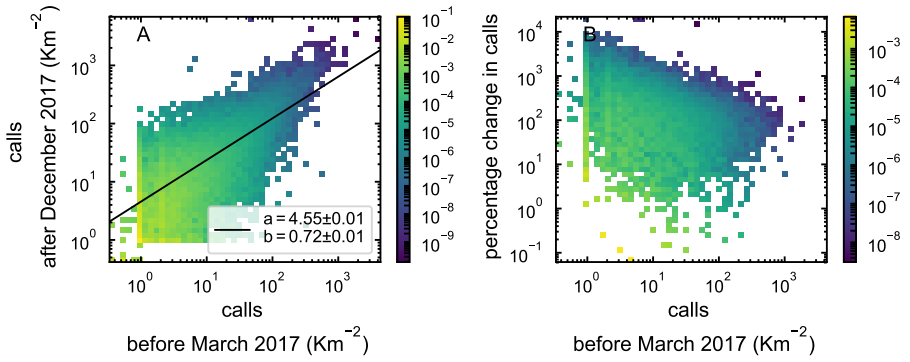


Fig. 12. Relation between initial and final number of calls. (A) Number of calls issued by refugees before December 2017 vs after March 2017 for all Turkish provinces. The full line shows a model $\rho_f = a \cdot \rho_i^b$ where ρ_f and ρ_i are the final and initial densities, respectively. The linear model of the log-transformed variables has root mean squared error $RMSE = 0.44$ and coefficient of determination $R^2 = 0.66$. (B) Percentage change in number of calls between March and December 2017 vs the number of outgoing calls in March 2017. The two quantities are negatively correlated with Spearman coefficient $S = -0.42$, $p < 0.01$.

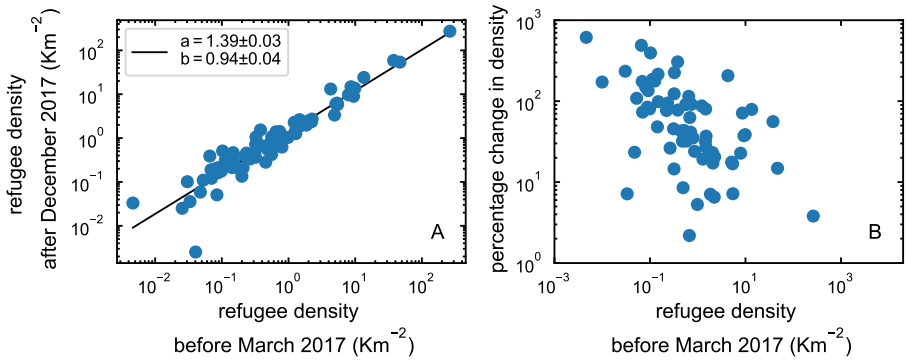


Fig. 13. Relation between initial and final number of calls. (A) Number of calls issued by refugees before December 2017 vs after March 2017 for all antennas. The full line shows a model $\rho_f = a \cdot \rho_i^b$ where ρ_f and ρ_i are the final and initial densities, respectively. The linear model of the log-transformed variables has root mean squared error $RMSE = 0.25$ and coefficient of determination $R^2 = 0.95$. (B) Percentage change in number of calls between March and December 2017 vs the number of outgoing calls in March 2017. The two quantities are negatively correlated with Spearman coefficient $S = -0.46$, $p < 0.01$.

References

1. No Lost Generation. http://nolostgeneration.org
2. Turkey: Education Sector Dashboard Q4 January-December 2017. http:data2.unhcr.org/en/documents/download/63145
3. Turkish Statistical Institute. http://www.turkstat.gov.tr
4. Areas, G.A.: Gadm database of global administrative areas, version 2.0. URL: http://www. gadm. org [accessed 201–03–24] (2012)
5. Barber, C.B., Dobkin, D.P., Huhdanpaa, H.: The quickhull algorithm for convex hulls. ACM Transactions on Mathematical Software (TOMS) **22**(4), 469–483 (1996)
6. the children, S.: Humanitarian situation analysis of syrian under temporary protection in turkey. https://resourcecentre.savethechildren.net/library/humanitarian-situation-analysis-syrian-under-temporary-protection-turkey (2017)
7. Deville, P., Linard, C., Martin, S., Gilbert, M., Stevens, F.R., Gaughan, A.E., Blondel, V.D., Tatem, A.J.: Dynamic population mapping using mobile phone data. Proceedings of the National Academy of Sciences **111**(45), 15888–15893 (2014)
8. Okabe, A., Boots, B., Sugihara, K., Chiu, S.N.: Spatial tessellations: concepts and applications of Voronoi diagrams, vol. 501. John Wiley & Sons (2009)
9. Report, H.R.: *"When I Picture My Future, I See Nothing"* barriers to education for syrian refugee children in turkey. http://www.hrw.org/report/2015/11/08/when-i-picture-my-future-i-see-nothing/barriers-education-syrian-refugee-children (2015)
10. Salah, A.A., Pentland, A., Lepri, B., Letouzé, E., Vinck, P., de Montjoye, Y.A., Dong, X., Dağdelen, Ö.: Data for Refugees: The D4R Challenge on Mobility of Syrian Refugees in Turkey. ArXiv e-prints (Jul 2018)

11. Salvatier, J., Wiecki, T.V., Fonnesbeck, C.: Probabilistic programming in python using pymc3. PeerJ PrePrints **4**, e1686 (2016)
12. Tastan, C.Celik, Z.: The education of syrian children in turkey: Challenges and recommendations. (2017)
13. UNHCR: Syrian regional refugee response. https://data2.unhcr.org/en/situations/syria/location/113

Developing Integration Policy for Refugees through Mobile Phone Data Analysis: A Study on Türk Telekom Customers

Ibrahim Zincir[1], Tohid Ahmed Rana[2], Ayselin Yıldız*[3] and Dilaver Arıkan Açar[4]

[1] Yaşar University, Department of Computer Engineering, Izmir, Turkey
[2] Yaşar University, Department of Computer Programming, Izmir, Turkey
[3] Yaşar University, Department of International Relations, Izmir, Turkey
[4] Yaşar University, Department of International Relations, Izmir, Turkey
*ayselin.yildiz@yasar.edu.tr

Abstract. Managing migration and developing effective integration policies require reliable, updated information and comparable statistical data. However, traditional data sources such as public censuses, poor official statistics or geographically limited field studies do not provide safe and updated migration data on a highly dynamic phenomenon. Mobile phone data as a new and alternative source offers higher quality, comparable, up-to-date and better data that can ensure increased support for statistical systems of policy makers. This paper aims to bring relevant insight for policy makers in Turkey and possibly abroad for better integration of refugees by analyzing the database based on anonymized mobile Call Detail Record of phone calls and SMS messages of Türk Telekom customers. The results show that mobile phone data is able to identify approximate location of the refugees, their mobility trends and activities over time along with their general communication patterns. The algorithms used in this study are able to differentiate each individual caller and their location information based on the communication patterns derived from the data. The results provide policy makers a handy alternative data source to get information and verify their statistics related to refugees and develop better integration policies by optimizing limited welfare resources effectively.

Keywords: Data Analysis, Social Integration, Refugees, Mobile Phone Data.

1 Introduction

Turkey, located on the Eastern Mediterranean migration route[1], has become a major "transit" and "destination" country for many migrants, asylum-seekers and refugees in the recent years (İçduygu 2013, Kaiser and Kaya 2015). As İçduygu and Sert (2009) notes, Turkey's location between the politically and economically unstable East and

[1] Eastern Mediterranean migration route begins in Asia, Central Asia, Middle East or the Horn of Africa and ends in Cyprus, Greece or Bulgaria via Turkey both by sea and by land.

the prosperous West makes it both an obvious transit route and an attractive destination in itself for migrants (İçduygu and Sert 2009). Although this is not a new fact, following the migration crisis of 2015, when over a million irregular migrants and refugees have reached Europe mostly from Syria, Africa and South Asia (IOM, 2015), Turkey has started to attract more attention from the international society in terms of its migration policies. European Border and Coast Guard Agency, also known as FRONTEX, reported 885,386 migrants of the nearly one million migrants had reached the European Union (EU) via the Eastern Mediterranean route in 2015 (FRONTEX, 2016). However, beyond 2015 migration crisis, Turkey has always been considered as a country where migration "is" and "will" be a significant issue of concern. As İçduygu (2005) notes, as long as the ongoing political turmoil and violence persist in conflicted neighbouring areas, people will continue to leave their homelands to search for prosperity, security and protection from persecution. Thus, Turkey is always exposed to new waves of especially irregular migration and asylum seekers. These movements have implications and challenges not only for social, economic and political dynamics in Turkey but also for the neighboring regions.

Currently, Turkey hosts the world's largest number of refugees; more than 3.5 million Syrians and additionally 346,800 refugees and asylum-seekers of various nationalities (DGMM, 2018a, UNHCR 2018). Beyond being a transit country, as a destination country where millions of refugees are residing, Turkey develops its own integration policy, which is currently lacking, to better serve not only the migrants and refugees but also to keep the solidarity and social acceptance among the host society. In this context, policy makers and practitioners require to benefit from relevant and actual data in order to develop well targeted and efficient policies on; a) managing migration and refugee movements through/to Turkey, b) develop an inclusive and comprehensive integration policy for the settled migrants.

Accordingly, new data sources such as mobile phone calls, internet searches, or interaction on social media provide the policy makers with more timely and evidence based data on migration. As it was also widely agreed at UN High-level Dialogue on Migration and Development (2013) and also at the Report of the High-Level Panel of Eminent Persons on the Post-2015 Development Agenda (2013), policy makers need better data on migration and development. Lack of data often leads to inefficient public policies on migration management and integration due to miscalculations about the scale and impact of migration. The problem is not only the lack of robust migration data but also how to confirm and update the data that comes from poor official statistics. Especially in the migration field, dynamic people movements have to be quickly reflected in the statistics in order to develop effective policies.

This report aims to provide an insight for policy makers to demonstrate how mobile phone data can be utilized to facilitate the development of better migration policies in Turkey. It focuses on the integration aspect of Syrian refugees in Turkey and specifically prioritizes to contribute the themes on "integration" and to some extent on "unemployment". The report uses the large dataset which is composed of anonymous

mobile phone data and made available for the Data for Refugees (D4R) Challenge (Salah et al.). D4R Challenge is initiated by Türk Telekom in partnership with the Turkish Academic and Research Council (TÜBİTAK) and Boğaziçi University.

The first section of the report provides brief information on the current nature of refugees in Turkey and the main challenges of integration. The second part explains the main problem and specific aim of this research. The third part explains the methodology and data analysis. The final part evaluates the research findings and concludes with some suggestions for policy makers. It also provides insights for further academic research that can be improved with additional data utilization.

2 Refugees in Turkey and Challenges of Integration

In the last decades, mobility of people through and towards Turkey have become more diversified in terms of the changing dynamics in types, flows and destinations. In this context, Turkey emerges as a major significance and also a concern in terms of its migration policies that are highly influenced by the high levels of irregular migration and asylum flows through its territory. Currently, Syrians constitute the highest number of refugees in Turkey and their socio-economic integration refers to one of the primary challenging cases of migration management in Turkey. Every progressive step and ad-hoc practices on Syrians' integration contributes to Turkey's wider integration policy, which is currently evolving.

After the civil war in Syria that has erupted in 2011, Turkey followed an open-border policy and provided protection to all destitute Syrians[2]. As of 2018, among 3.5 million registered Syrians, only 6 percent of whom are sheltered in 25 camps in 10 provinces of Turkey while 3.3 million Syrians are living dispersed around various locations throughout the country (DGMM 2018b). In addition to Syrians, 346,800 refugees and asylum-seekers from other countries are provided international protection in Turkey (UNHCR, 2018). These include people from various countries such as Iraq, Afghanistan, Iran, Somalia, Pakistan, Yemen, Eritrea and Palestine. Syrians are registered under "temporary protection status" and granted free access to education, health services, social aid and right to work. These rights and benefits have been extended to everyone from all nationalities granted international protection in Turkey. In this context, Turkey has been generous in providing humanitarian protection and aid

[2] Due to keeping its geographical limitation to 1951 Geneva Convention, Turkey does not grant refugee status for the asylum applicants from non-European countries of origin. According to Regulation No. 29153 on Temporary Protection (2014 Regulation), which was enacted in line with Article 91 of the Law Number 6458 on Foreigners and International Protection, Syrians in Turkey are offered a group-based "temporary protection". The regulation applies to all Syrian nationals, refugees as well as stateless persons from Syria seeking international protection, including those without identification documents. It ensures *non-refoulment,* and grants right to legal stay as well as free basic health care, education and social assistance for registered Syrians.

for not only Syrians but all refugees, even when the numbers have become overwhelming.

One of the biggest challenges of migration management in Turkey is the lack of overarching official "integration" policy. Turkey currently develops its integration policy based on its ad-hoc humanitarian practices in particular that are developed for Syrians. The temporary stay of Syrians in Turkey is prolonged and a high number of them started to permanently settle in Turkey. Most studies report that the majority of Syrian refugees are inclined to stay in Turkey for the long term, since even if the conflict in Syria ends, peace and stability will not be restored immediately (Yıldız and Uzgören 2016, Kirişci 2014, Erdoğan 2015). Actually, even before the Syrian case, Turkey has been always exposed to new waves of refugees, which would certainly bear additional socio-economic challenges. Thus, a sustainable integration policy including in all the best practices being implemented in the fields of education, health, employment and social integration is an urgent necessity for Turkey. In doing so, the existing legislative loopholes and administrative shortcomings could be well determined and resolved. Despite Turkey's long-lasting experience on migration movements, the institutionalization of the policy area is highly dependent on the provision and regular tracking of "fast changing" and "hardly obtained" accurate statistics due to the very mobile nature of the issue.

3 Main Problem and Aim of the Research

This paper aims to demonstrate how data mining can be utilized to develop solutions and offer benefits for social integration of refugees in Turkey. Accordingly, by using different datasets, it attempts to provide an analysis based on mobile phone usage of randomly selected in total 20 anonymous Türk Telekom customers who are composed of 10 refugees and 10 non-refugees (Salah et al.).

Within the limits of the available data, firstly, the analysis focuses on mainly two features; namely "location" and "timing" of the calls. Upon these two features, by employing basic statistics to the datasets modified by the project team, it tries to map out the accurate residence, mobility story and the possible employment status of the refugees. This new information provides a unique comparable input for the statistical records of policy makers of mainly two institutions; Turkish Ministry of Interior Directorate General for Migration Management (DGMM) and Turkish Ministry of Labour, Social Services and Family.

Secondly, through using five algorithms (Decision Tree, Ripper, K-Nearest Neighbours, Hoeffding Tree and Naive Bayes) and implementing them on the sample group, the analysis figures out both refugee and non-refugee customers' behaviour of mobile phone usage. Depending upon the users' tracked habit of mobile phone usage, the algorithms proved to be able to not only differentiate refugee and non-refugee users, but also to identify and confirm the individual caller behaviour and to find out

the exact ID of the user with more than 90 per cent success rate. This would also help to get better result from a noisy data which identifies the "refugee users" as a heterogeneous group composed of refugees, foreigners with residence/work permit, international/exchange students, and the ones registered with temporary status.

The main problem that motivates this research stems from the fact that a significant number of Syrians are eventually mobile in Turkey and they do not live in the cities and districts that they are first registered. As a result, there are shortcomings of current migration statistics and there exists an information gap concerning the provincial residency of these mobile refugees, which hinders the policy makers' planning, decision making, efficient allocation of available resources and managing welfare services. For example, most of the Syrians who had been first registered at the border cities just after they entered Turkey, moved to different cities in search for employment and better living conditions. This creates challenges for the public institutions in terms of managing their resources and offering sustainable and adequate public services for the refugees. According to the 2014 Regulation, Syrians can only benefit from public services such as education, access to free health services and social aids in the cities that they are registered (with few exceptions on specific criteria). Therefore, public institutions, civil society organizations and local authorities develop their policies and manage the allocation of resources on the basis of number of registered Syrians in their district. As long as Syrians do not transfer their registrations from one city to another, which is not a common and easy case due to some administrative difficulties, the domestic movement of Syrians creates unexpected burden and aggregation on some public institutions in specific locations. The update of statistics is possible through Syrians' own initiatives to register themselves or transfer their registrations. Consequently, the statistics of DGMM on the exact number of Syrians by province (DGMM, 2018) could have an alternative accurate comparable data to confirm their numbers through such a mobile data analysis with to some extent negligible deviation. In addition to this, despite the fact that DGMM does not share the official number of refugees by district but only by province/city, the mobile data analysis provides a general estimation about the location of refugees for academics and researchers. As a result, based on the problem of determining and confirming the real number of refugees in a specific location, the data mining analysis of the mobile phone calls might facilitate the development of public policies and allocation of welfare resources for the better social integration Syrians in Turkey.

As the second problem, the number of employed Syrians in the labour market of Turkey can only be estimated. This is mainly because Syrians are mostly working as unregistered, as part of informal economy. The Regulation on Provision of Work Permits for People under Temporary Protection, introduced in January 2016, allows Syrians registered under temporary protection to get work permits. The number of issued work permits is reported as 20,968 in 2017 (Turkish Ministry of Labour, Social Services and Family, 2018). The appraised efforts of Turkey to ensure working rights for refugees, the ones under international protection and Syrians under temporary protection is noteworthy which also reflects an inclusive approach of Turkey's evolv-

ing integration policy. Nevertheless, the total number of issued work permits for Syrians has remained rather low with regards to the fact that there are more than 2 million Syrians in Turkey who are at working age. As of September 2018, the number of Syrians who are at the working age (15-64) is 2.051,186 (DGMM, 2018). However it is estimated that almost 800,000 Syrians are working in Turkey as part of informal economy mainly in the sectors of agriculture, textile, construction and services.

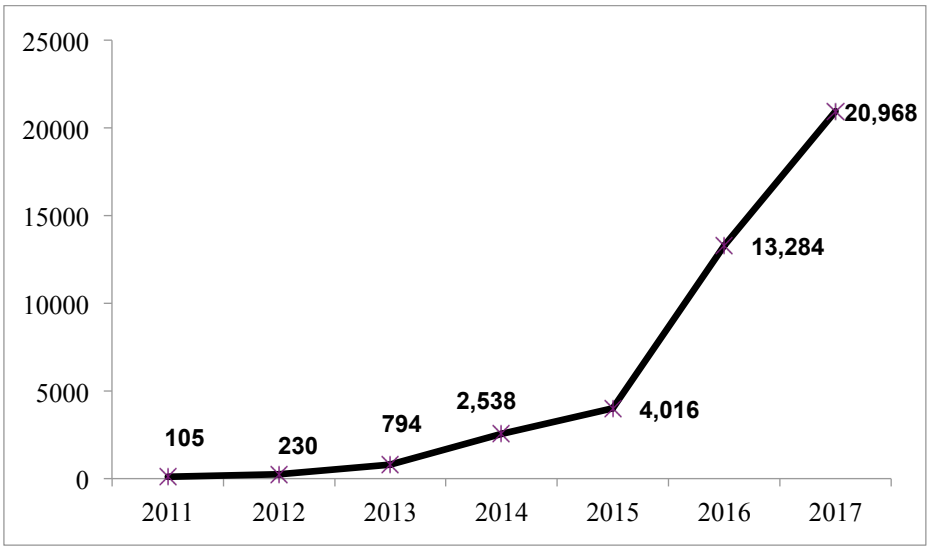


Fig. 1. Number of work permits issued for Syrians in Turkey.
Source: Data compiled for the study from Turkish Ministry of Labour, Social Services and Family Statistics, 2018.

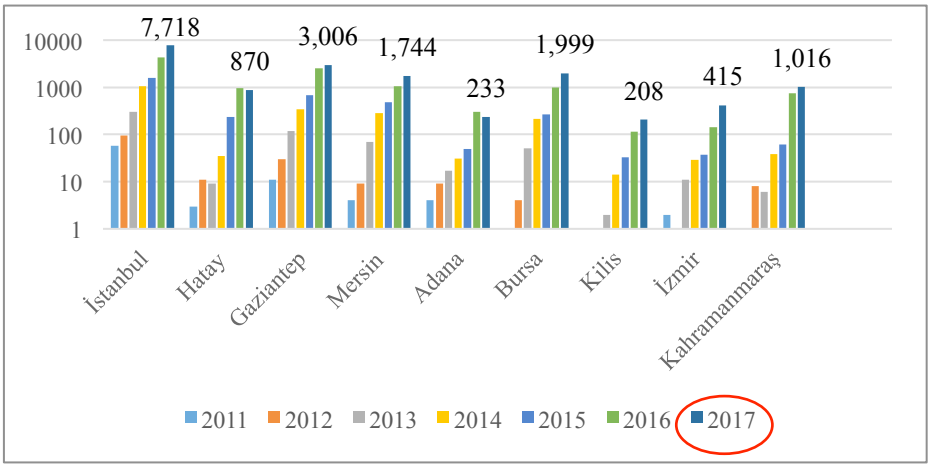


Fig. 2. The number of work permits granted to Syrian nationals by top-ten cities in Turkey. Source: Data compiled for the study from Turkish Ministry of Labour, Social Services and Family Statistics, 2018.

Conducting an analysis on the call hours and the locations of the mobile users can provide a rough idea about where the users spent their day and night times. Although not necessarily to be considered as the working place, the analysis can provide a general hint for the possible status of the refugee, whether unemployed or employed, spending specific hours of the day in a different location, or move back-and-forward from one city to another as seasonal worker. It also provides a picture on patterns of changing conditions, expectations and way of life of refugees. If the gender information of the user is also provided, the data might offer a broader interpretation of the employment status. For example, in Torbalı district of Izmir, it is a crucial problem that many Syrians are working as undocumented labours in agriculture and they are subject to labour exploitation (Reidy, 2016). They live in informal encampment and they are mostly coming to Torbalı from border cities in search for seasonal work. The mobile phone data can at this point provide a rather accurate estimation about counting the uncountable, unregistered Syrian workers.

In this context, this paper aims to provide some insights for the policy makers to help decision-making and policy planning concerning refugee mobility and optimizing the public services for their better social integration. In doing so, it offers a comprehensive analysis based on understanding the communication activities of users and interactions among;

a) Users' mobile phone usage habits,
b) Population structures (distribution of refugees) in specific geographic locations,
c) Timing of the calls of the selected population.

4 Methodology and Data Analysis

This analysis utilizes the large-scale mobile Call Detail Record (CDR) database, which is composed of the anonymized mobile CDR of phone calls and SMS messages of Türk Telekom customers. The dataset is created from 992,457 customers of Türk Telekom, of which 184.949 are tagged as "refugees", and 807,508 as Turkish citizens (Salah et al. p. 3). Keeping in mind the limitation that, it is not certainly possible to address a particular CDR belongs to a refugee or not, based on the patterns of aggregated records (Salah et al.), for each modified dataset, the analysis used 20 anonymous user data which is composed of 10 refugee and 10 non-refugee data. The users are randomly selected from Dataset 2 and Dataset 3.

The analysis is conducted through four steps over the given data in Dataset 2 and Dataset 3;

1) Data normalization,
2) Modifying 6 datasets by merging and associating different data in Database 2 (incoming and outgoing voice call activity, incoming and outgoing SMS activity) and Database 3 (incoming and outgoing voice call activity),
3) Implementing machine learning algorithms,
4) Evaluating the results.

The research findings are concluded by employing two different analytical approaches. First, "basic statistics" is used and the data of users' locations and time stamps of their calls are analyzed in order to understand the communication activities of the refugees. Accordingly, the data is modified to understand the mobile phone usage characteristics of the refugees. First, the time stamp value of the data is taken and then from this value three new columns are generated;

1) The value of day of the month (1,2,3,…,30,31)
2) The value of month of the year (1,2,3,…,11,12)
3) The value of hourly intervals (0,1,2,3,…,22,23)

As another new input which improves datasets, the time period for one day is grouped/split into 3 intervals; 00:00-8:00; 08:00-19:00; 19:00-24:00.

The findings are interpreted and some conclusions have been drawn through a statistical process. Additionally, WEKA data mining software is used to analyse the datasets. The results give an idea about the "time bounded mobility activity" and "time-bounded location" information about the refugees. For this purpose, both Dataset 2 (includes voice and SMS data of a particular group of users who are observed for a period of 2 weeks) and Dataset 3 are used.

In the second approach, "machine learning algorithms" (Decision Tree, Ripper, K-Nearest Neighbours, Hoeffding Tree and Naive Bayes) are applied over the modified datasets, which provide an analysis on the mobile data usage habits/patterns of the given users and then identify the caller IDs and district IDs of the users. The district ID identification is only possible for Dataset 3.

The analysis is conducted by a team of four researchers; two from computer engineering and two from international relations field. Working in collaboration under the UNESCO Chair on International Migration at Yasar University, the research employs a multidisciplinary approach, which is highly lacking in migration studies. In this respect, the expertise of both computer engineers and migration experts are merged and pooled in to provide feasible solutions to the problems of migration management by using available but idle large-scale data.

4.1 Limitations of the Data

The dataset defines the caller "refugee" with reference to the caller's identity card number starting with 98 or 99. In Turkey, not only the ones with international protection status or refugees but all foreigners, regular migrants are provided with these official ID cards by the Turkish government and they are given ID numbers starting with 98 and 99. Thus, the datasets include such biases of a heterogeneous group, which requires different integration policy approaches with different priorities and needs. For example, according to DGMM, as of September 2018, 716,494 residence permits have been issued for foreigners in Turkey (DGMM 2018c). Roughly, this means foreigners who are legally residing, working or studying in Turkey constitutes 19% of the group addressed as "refugees" in the given dataset. Therefore, beyond the fact that mobile market share of Türk Telekom is only 24,7% in Turkey across all operators (Bilgi Teknolojileri ve İletişim Kurumu 2017) the main limitation exists with the representativeness of the data where "refugee" customers are defined in a heterogeneous group of foreigners. As a result this report does not intend to be representative, rather it contributes to the integration policies for refugees through sampling and aspires to present how data mining can be useful for policy making.

Another limitation is since the ID numbers of the mobile phones of the users in the given databases are not identical, the analysis does not able the researchers to connect databases. In Dataset 2, concerning both incoming/outgoing voice and SMS activity the callers are not the same users. This hinders to figure out the users' behaviour of mobile phone communication activities by using enriched data.

4.2 Data Normalization and Modifying Datasets

First, from the original dataset, which is composed of three different callee prefixes, "refugee", "non-refugee" and "unknown", the "unknown" users have been removed. The analysis is based on six new sub-samples generated by the project team; four of them are created from Dataset 2 and two of them from Dataset 3. After sub-sampling, duplications have been removed and both datasets have been modified for the analysis.

In Dataset 2, which is composed of "incoming voice call, outgoing voice call, incoming SMS and outgoing SMS) are modified by adding the city and district data to the dataset. The features in these subsets are "CALLER_ID, SITE_ID, DAY_OF_CALL, MONTH_OF_CALL, HOUR_OF_CALL". The four new datasets are prepared by including the city, district data (acquired from BTS_ID).

In Dataset 3, the data of "incoming voice call" and "outgoing voice call" are modified again by adding city, base station and time stamp data. In total these six datasets are modified for the analysis by removing the duplications. The features in these subsets are "CALLER_ID, ID, CITY_ID, DAY_OF_CALL, MONTH_OF_CALL, HOUR_OF_CALL".

The finalized datasets inherited the referred features to extract the classification models are as follows;

Dataset 2: who made the call (CALLER_ID), the base station where the mobile device is connected during the respective connection (SITE_ID), day of the week the respective connection is established (DAY_OF_CALL), month of the year the respective connection is established (MONTH_OF_CALL), hour of the day the respective connection is established (HOUR_OF_CALL)

Dataset 3: who made the call (CALLER_ID), location of the prefecture where respective connection is established (ID), location of the city where respective connection is established (CITY_ID), day of the week the respective connection is established (DAY_OF_CALL), month of the year the respective connection is established (MONTH_OF_CALL), hour of the day the respective connection is established (HOUR_OF_CALL)

5 Evaluation

5.1 Volume of the Communication Activity of Refugees

One of the findings revealed through WEKA data mining tool obviously demonstrates that, the total volume of communication activity of refugees is significantly lower than the non-refugee mobile phone users. The data reveals that while non-refugee users have in total 223,360 communication activities, this remains at only 15,599 activities for refugees.

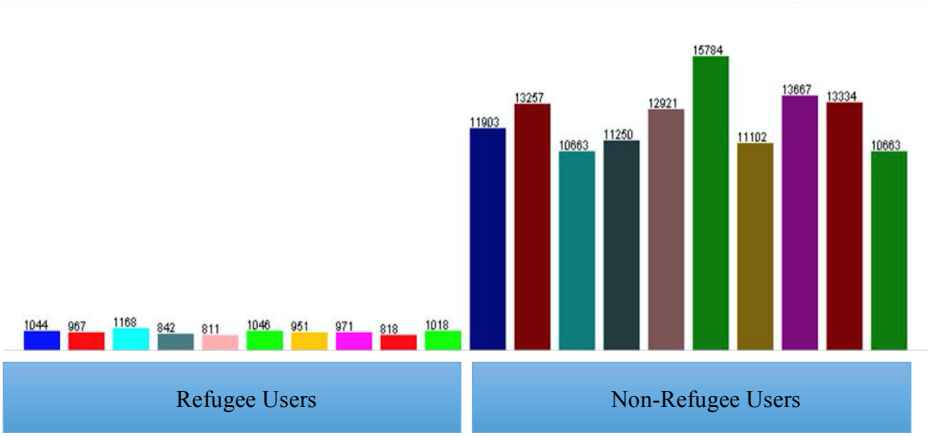


Fig. 3. Outgoing Voice Call (Dataset 3).

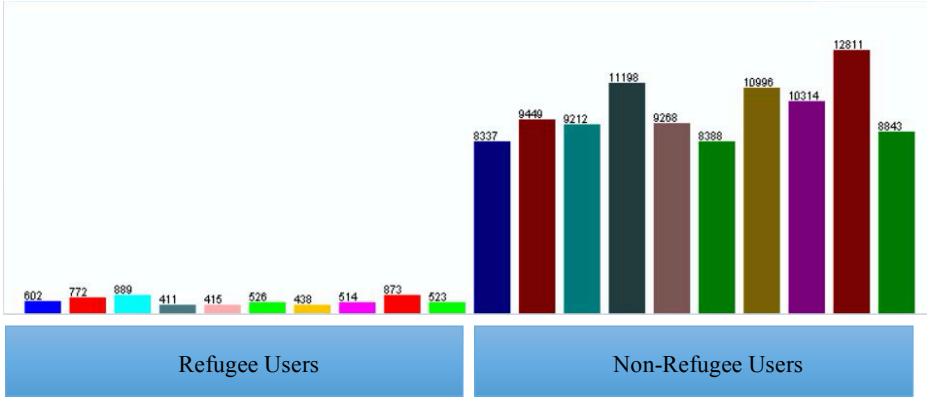


Fig. 4. Incoming Voice Call (Dataset 3).

If further data is made available in terms of the status of called user as well whether he/she is also refugee or not, the analysis might provide insights on social communication network among refugee and host population. Then, the data is able to identify the social integration of refugees through their social communication networks and the interactions with the host society.

Dataset 2 is analysed in terms of SMS activity and it demonstrated that while refugees had 4,703 SMS activity (2,442 SMS in and 4,703 SMS out), non-refugee users had 14,871 SMS activity in total (7,892 SMS in and 19,574 SMS out). This means 24% of all SMS activities belong refugee users.

Table 1. Refugee Users SMS in (Dataset 2).

Count of HOUR_OF_SMS_IN	Column Labels																							
Day hours	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	Grand Total
ADANA	10	8	6	3								5	2	7	4	14	5	3	8	6	6	11	13	7	118
ANKARA	15	8						1	8	12	13	11	7	17	14	10	18	19	26	22	34	29	33	29	326
ANTALYA	3	3	2	4	3								2	2	2		2	1	3	1	1	5	4	6	44
GAZIANTEP	62	39	12					2	4	14	34	79	103	94	108	116	124	111	102	81	76	80	90	69	1400
ICEL	25	17	8	9	3	4	10	30	10	13	28	22	44	23	18	32	27	36	22	27	19	36	36	55	554
Grand Total	115	75	28	16	6	4	10	33	22	39	75	117	158	143	146	172	176	170	161	137	136	161	176	166	2442

Table 2. Non-refugee Users SMS in (Dataset 2).

Count of HOUR_OF_SMS_IN	Column Labels																							
Day hours	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	Grand Total
BALIKESIR								10	1	2	1	1	5												20
BURSA	244	146	64	37	14	12	31	86	94	114	123	119	166	215	313	343	353	354	282	402	423	457	589	447	5428
ISTANBUL	38	26	16	2	2				2	16	29	26	19	22	37	54	40	76	49	51	35	47	38	49	674
IZMIR	137	105	89	87	56	38	53	56	37	36	28	43	49	54	64	72	77	84	86	110	99	96	104	97	1757
MANISA						5	1	2					1	4											13
Grand Total	419	277	169	126	72	55	85	154	134	168	181	189	240	295	414	469	470	514	417	563	557	600	731	593	7892

Table 3. Refugee Users SMS out (Dataset 2).

Count of HOUR_OF_SMS_OUT	Column Labels																					
Day hours	0	1	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	Grand Total
ANKARA	15	8			1	8	13	14	11	7	19	15	10	18	20	26	23	33	29	32	28	330
ANTALYA						2	18	45	60	44	46	45	51	53	57	32	32	31	33	30	16	595
GAZIANTEP	33	9		1	4	9	17	31	63	69	73	81	70	81	50	58	46	50	50	56	39	890
ICEL	7		6	23	28	6	11	9	13	36	13	11	19	27	48	27	31	12	36	36	47	446
Grand Total	55	17	6	24	33	25	59	99	147	156	151	152	150	179	175	143	132	126	148	154	130	2261

Table 4. Non-refugee Users SMS out (Dataset 2).

Count of HOUR_OF_SMS_OUT	Column Labels																						
Day hours	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	Grand Total
ADANA	12					1	5	8	18	36	22	14	12	14	32	24	30	64	44	56	62	46	70	46	616
AFYON																					6	13	11		30
BILECIK																		2	7						9
BURSA	168	90	36	20	4	3	35	27	16	46	50	63	99	148	213	159	241	228	152	260	225	393	459	275	3410
ESKISEHIR																			1						1
GAZIANTEP	20	9		1		4	3	9	33	69	66	79	79	49	78	61	77	67	76	56	101	88	86	82	1193
HATAY	2	1	2				2				2	3	2	7	6	4	4	3	3	3	3	2	3	3	55
IZMIR	80	56	22	22	18	14	38	24	32	28	4	20	8	22	18	26	50	74	78	104	110	62	98	80	1088
KAYSERI	2					2	2	4	9	7	14	8	20	13	16	18	11	5	11	7	3	4	3	6	165
KILIS														1			3	3	2	2	14	2			27
KMARAS												5	4												9
KONYA	1																						6	2	9
KUTAHYA																			3	8	4				15
MANISA													2	2											4
OSMANIYE							3																		3
SIVAS	9	2			1	6	19	26	24	28	7	17	12	12	14	22	32	19	18	19	15	18	13	12	345
Grand Total	294	158	60	43	23	30	107	98	132	214	165	209	238	268	377	314	448	465	395	515	543	628	749	506	6979

When the voice call and SMS activity ratios are compared it could be observed that refuges are more intended to use SMS activity than voice calling as seen in Table 5 below.

Table 5. Preferred communication activity among refugee and non-refugee customers.

Communication Activity	Refugees	Non-refugees
Voice call	7%	93%
SMS	24%	76%

5.2 Interpreting Communication and Mobility Activity Based on Location and Time

The basic statistics analysis demonstrates the communication activity of refugees by examining their daily mobile phone usage habits (over 24 hours) which gives an idea about some patterns. It provides location information which helps to understand the mobility activity of users. The data on location and time (space-time behaviour) allows the policy makers to identify more precisely where and when the refugees undertake their daily activities. Moreover, it presents an accurate geographic information about the residency of the refugees. The data also demonstrates the daily mobility activity on location basis. The daily mobility can be traced over day and night time period. Within the limitations of the data, for this research the information is limited

to 2 weeks, however if the data is extended with the access to one-year period, the whole trajectory of a refugee can be better traced. It might provide information for mobility story and change of location of refugees.

Table 6 demonstrates the mobile phone usage of a selected refugee customer between 1-30 January 2017. The table shows how many times the refugee called someone/somewhere, from where these calls are done and during which hours of the day these calls have taken place. It can be interpreted that the selected refugee high probably lives in Oğuzeli district of Gaziantep referring to 430 calls made from Oğuzeli. The highest number of calls made during night time and early in the morning (20:00-00:00 and 00:00-07:00) have taken place in Oğuzeli. Table 5, as a complementary data, also supports the fact that the refugee most probably lives in Oğuzeli as 202 calls have been received in Oğuzeli with the highest at night time as well.

Table 6. Number of calls out of a refugee customer in one month (Dataset 3).

Count of HOUR_OF_CALL	Column Labels																	
Day hours	0	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	Grand Total
GAZIANTEP	**6**	**16**	**32**	**31**	**14**	**19**	**29**	**20**	**19**	**24**	**14**	**44**	**49**	**67**	**56**	**75**	**31**	**33**	**579**
ISLAHIYE				3		1													4
KARKAMIS				1				1					2						4
NURDAGI			1				3												4
OGUZELI	6	12	17	21	10	12	15	5	11	5	6	31	35	57	52	71	31	33	430
SAHINBEY		3	13	6	2	1		4	1	10	2	3	4	6	4	2			61
SEHITKAMIL		1	1		2	5	11	10	7	9	6	10	8	4		2			76
HATAY				**2**	**1**	**1**													**4**
ANTAKYA					1														1
HASSA				2		1													3
KILIS			**11**	**43**	**45**	**60**	**24**	**41**	**47**	**45**	**44**	**44**	**24**	**25**	**5**	**2**	**1**		**461**
ELBEYLI			4	26	36	46	20	33	44	41	34	31	9	5					329
POLATELI			1	3			1	1					2	1					9
SEHIRMERKEZI-KILIS			6	14	9	14	3	7	3	4	10	13	13	19	5	2	1		123
Grand Total	**6**	**16**	**43**	**76**	**60**	**80**	**53**	**61**	**66**	**69**	**58**	**88**	**73**	**92**	**61**	**77**	**32**	**33**	**1044**

Table 7. Number of received calls of a refugee customer in one month (Dataset 3).

Count of HOUR_OF_CALL	Column Labels																		
Day Hours	0	1	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	Grand Total
GAZIANTEP	**3**	**1**	**7**	**18**	**19**	**10**	**14**	**22**	**20**	**20**	**14**	**24**	**12**	**23**	**22**	**29**	**37**	**26**	**17**	**338**
ISLAHIYE				1			3													4
KARKAMIS									1											1
NURDAGI				2				3												5
OGUZELI	3	1	6	5	8	7	6	7	5	7	2	10	9	19	19	17	28	26	17	202
SAHINBEY				8	8				5	2	4		1	1	2	9	9			49
SEHITKAMIL			1	2	3	3	5	12	9	11	8	14	2	3	1	3				77
HATAY					**1**	**6**	**3**													**10**
ANTAKYA						6	1													7
HASSA					1		1													2
KIRIKHAN							1													1
KILIS				**7**	**26**	**26**	**34**	**31**	**18**	**15**	**23**	**23**	**20**	**16**	**12**	**3**				**254**
ELBEYLI					3	19	25	27	16	13	19	21	14	3	3					163
POLATELI				1	1								1	1		1				5
SEHIRMERKEZ I-KILIS				6	22	7	9	4	2	2	4	2	5	12	9	2				86
Grand Total	**3**	**1**	**7**	**25**	**46**	**42**	**51**	**53**	**38**	**35**	**37**	**47**	**32**	**39**	**34**	**32**	**37**	**26**	**17**	**602**

When the communication activity of this selected refugee is traced over according to the each day of January 2017, the mobility trajectory of him/her becomes much clear.

The following map demonstrates the mobility of the selected refugee based on the number and order of his/her usage of mobile phone. The mobility story is interpreted by taking into account the location and timing of the calls done and received.

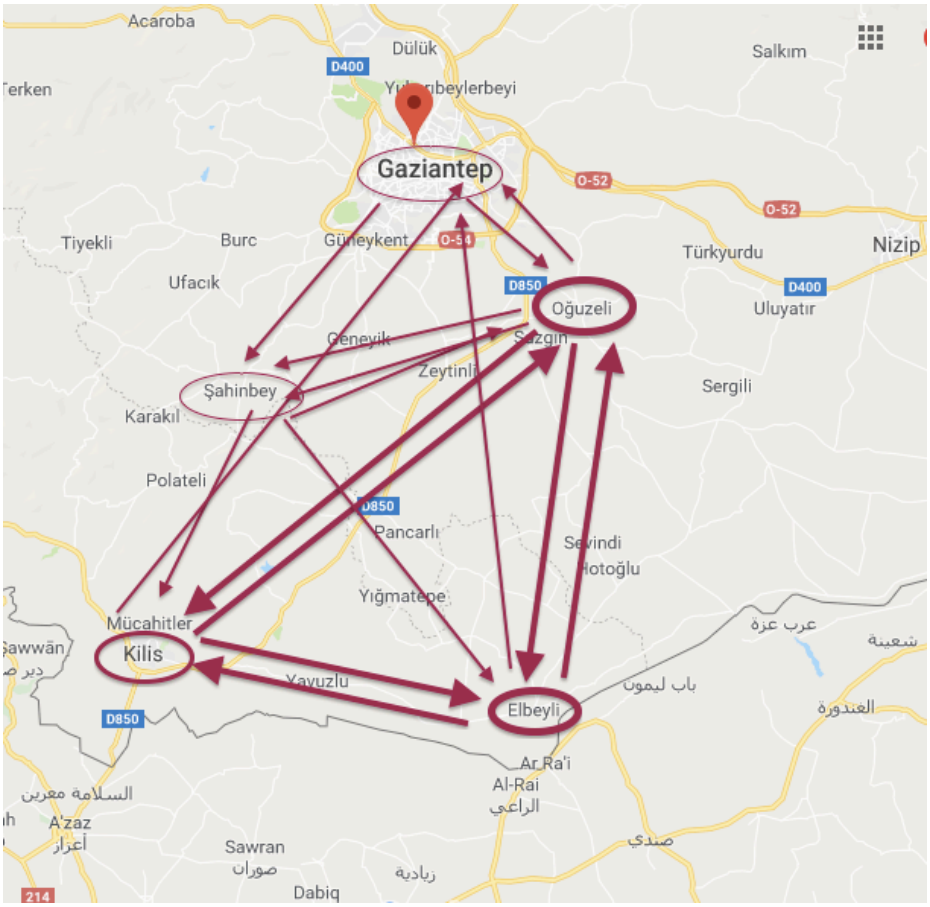


Fig. 5. Original project data sample implemented on Google map.

The mobility trajectory of the refugee on hourly and daily basis gives the idea that the refugee resides in Oğuzeli and he/she spent day time (working hours slot) in Elbeyli. It seems the refugee travels regularly to Elbeyli mostly over Kilis city center and sometimes over Şahinbey. The refugee also had some visits back and forward between Gaziantep/ Şehitkamil and Oğuzeli. Although not certain, it is probable that the refugee might be working in Elbeyli and residing in Oğuzeli according to his/her phone calls (out/in) during specific times of the day over one month.

This method can be applied to all refugee users by merging and interpreting the location and time information acquired from data analysis. By also using GIS tools, further studies are possible to map out the residency, possible employment location and the route of mobility if exists. These constitute very crucial information to clarify the accurate statistics by considering the existing informality of mobility activities in order to plan better integration policies.

5.3 Identifying the "Caller" and the "Location" by Using Machine Learning Algorithms

Based on the calling habits of the people with reference to the location and timing of the communication activities, some machine learning algorithms are successfully able to identify the caller or the district of the caller. While conducting the data mining over 6 datasets, some of the widely used algorithms are implemented and five of them achieved the best results with high success rates. These algorithms are Decision Tree, Ripper, K-Nearest Neighbours (KNN), Hoeffding Tree and Naive Bayes.

The final results obtained are presented in the tables below. In order to decide the most effective approach to understand data, Total Accuracy (TA), True Positive (TP), True Negative (TN), ROC Area and Precision (PR) rates are taken into consideration to analyse the data.

In modified Dataset 2 and Dataset 3, the algorithms are implemented to identify the callers (CALLER_ID) by examining their autonomous patterns depending upon location of activity and hourly usage. The following algorithms achieved more than 90% accuracy rate in identifying the CALLER_ID.

Table 8. Call in Caller ID (Dataset 2)

D2CALL_IN - CALLER_ID	Decision Tree	Ripper	KNN	Hoeffding Tree	Naive Bayes
TA	0,96944	0,993284	0,9665	0,725531	0,987669
TP	0,97	0,993	0,966	0,726	0,988
PR	0,971	0,994	0,967	0,726	0,988

Table 9. Call out Caller ID (Dataset 2)

D2CALL_OUT - CALLER_ID	Decision Tree	Ripper	KNN	Hoeffding Tree	Naive Bayes
TA	0,857796	0,90625	0,920187	0,53811	0,900479
TP	0,858	0,906	0,92	0,538	0,9
PR	0,862	0,897	0,92	0,545	0,9

Table 10. SMS in Caller ID (Dataset 2)

D2SMS_IN - CALLER_ID	Decision Tree	Ripper	KNN	Hoeffding Tree	Naive Bayes
TA	0,92636	0,956454	0,971163	0,736307	0,98684
TP	0,926	0,956	0,971	0,736	0,987
PR	0,932	0,957	0,971	0,758	0,987

Table 11. SMS out Caller ID (Dataset 2)

SMS_OUT - CALLER_ID	Decision Tree	Ripper	KNN	Hoeffding Tree	Naive Bayes
TA	0,937987	0,902056	0,966991	0,74632	0,979437
TP	0,938	0,902	0,967	0,746	0,979
PR	0,943	0,904	0,968	0,747	0,98

Table 12. Call in Caller ID (Dataset 3)

D3CALL_IN - CALLER_ID	Decision Tree	Ripper	KNN	Hoeffding Tree	Naive Bayes
TA	0,971378	0,968257	0,971149	0,961414	0,963685
TP	0,971	0,968	0,971	0,961	0,964
PR	0,971	0,968	0,971	0,962	0,964

Table 13. Call out Caller ID (Dataset 3)

D3CALL_OUT - CALLER_ID	Decision Tree	Ripper	KNN	Hoeffding Tree	Naive Bayes
TA	0,9287	0,9151	0,940334	0,904375	0,902698
TP	0,929	0,902	0,94	0,904	0,903
PR	0,93	0,91	0,94	0,908	0,903

In modified Dataset 3, the algorithms are able to find out the district (BTS_ID**)** of the caller through analysing the communication pattern of the user based on user and time of activity.

Table 14. Call in BTS ID (Dataset 3)

D3CALL_IN	Decision Tree	Ripper	KNN	Hoeffding Tree	Naive Bayes
TA	0,88073	0,87062	0,896659	0,842173	0,842468
TP	0,881	0,875	0,897	0,842	0,842
PR	0,851	0,868	0,89	0,806	0,809

Table 15. Call out BTS ID (Dataset 3)

D3CALL_OUT	Decision Tree	Ripper	KNN	Hoeffding Tree	Naive Bayes
TA	0,849612	0,835421	0,884923	0,786645	0,790475
TP	0,85	0,821	0,885	0,787	0,79
PR	0,848	0,831	0,881	0,758	0,765

This analysis provides a unique contribution to the policy makers of migration in terms of identifying the population structures base on each individual caller characteristics and also the location. The algorithms point out the fact that every caller has specific characteristics/behavior of mobile phone usage and some patterns/commonalities can be drawn for specific groups of population. Beyond differentiating refugee groups among a heterogeneous group of foreigners with divergent needs for integration and different mobile phone usage habits, the algorithms proved to identify each individual caller and their locations. Accordingly, algorithms help the policy makers to verify their statistics in terms of eliminating discrepancies by matching the number of population and their exact location. Moreover, it presents some insights about the communication behavior of refugee groups which can foster their integration by helping the policy makers in terms of offering new ways of communication to reach, guide and support their target group with specially designed tools for mobile phones.

6 Concluding Remarks

Turkey is not only a transit but also a destination country for many migrants, asylum-seekers and refugees who are searching for a better life and humanitarian protection. Therefore, Turkey is always exposed to new waves of migration movements since it is part of a migration system that is spread over a large geographical area including Europe, Middle East and Africa. Referring to its long history of hosting many migrants and providing protection to many refugee populations, Turkey's integration policy has been developing in recent years. While hosting more than 3.5 million refu-

gees in Turkey in the last 7 years, Turkey once more proved not only its ability to manage such a high number of mass movement, but also the hospitality and tolerance among its host population. However, as the temporary stay of Syrians has been prolonged, there are risks of increasing unrest and decreasing social acceptance among the host society which is strongly linked with the need for a well-established, inclusive and comprehensive integration policy.

Accordingly, not only to cater migrants and refugees but also to keep the peaceful co-existence and social acceptance among the host society, integration policies should manage the allocation of resources properly. Needs and challenges in education, health, housing, employment and social provisions should be carefully determined and planned while policies are being developed. In that respect, accurate information and statistics play significant role since migration involves many informalities concerning the movement of people, change of location and also engagement in labor market. Policy makers who are in dire need of accurate statistics are dependent to their official initiatives, refugees' own engagements with the authorities or some small-scale field work to confirm and update their information and statistics. However, new data sources such as mobile phone data provides a unique contribution to verify and also to provide new data for policy makers.

This research which utilizes the mobile phone data provided by Türk Telekom (Salah et al.), provides some insights for policy makers for better integration policies. Although the study is not representative due to some limitation of data and time, the sample studied demonstrates a methodology that can be utilized for large group of refugee populations as well. The research findings and some policy suggestions drawn from the analysis can be listed as follows:

1. Mobile phone data clearly helps the policy makers to understand how many refugees are actually living in specific locations. This is a crucial new information because as DGMM has provincial/city level directorates that keep the number of refugees in their area of responsibility (exceptions are Antalya, Hatay, Istanbul and Şanlıurfa) the analysis of mobile phone data could not only provide at city level but also at district level information about the actual number of refugees spread in a city. This is also an important information for local authorities such as municipalities. The data provides a unique new input to verify the official statistics and keep track of records of refugees in terms of their movements, residence, work and public service preferences.

2. The detailed analysis of mobile phone data modified according to time and location demonstrates the mobility story of refugees and their main places of visit. The traced information can give an idea about where the refugee is actually living, whether he/she moved to another place, and even whether he/she might be working depending upon the repeating or regular mobile phone usage hours and location information. Interpretation on employment status is a very important input because it is known that more than 80 % of Syrians who are in the labor market are working as unregistered.

The information about the age and gender distribution of this group is also not existing. Thus, the mobile phone data occurs as a trustable source of information to clarify the actual situation in informal labor market which constitutes the very important aspect of integration policies and livelihood practices. If the data is extended with age and gender information the methodology provides a better picture on the socio-demographic mapping of the refugee population and also explains better their mobility trajectory. At this point, data on internet usage (as it is widely accepted that refugees use mostly internet) and could be useful to understand the needs of refugees for better integration policies.

3. Although not representative but the data analysis drawn from the sample reveals that in comparison to non-refugee population, refugees tend to use SMS communication activity. The volume of their voice calls is significantly lower than the non-refugee callers. The reasons need to be explored in a further study but some of the reasons might be related with refugees' low level of income, limited social network in Turkey and low level of daily activity that requires or prioritizes communication.

4. Machine learning algorithms of Decision Tree, Ripper, K-Nearest Neighbours, Hoeffding Tree and Naive Bayes proved to differentiate the individual caller ID and also location depending upon the mobile phone usage pattern of the refugees and non-refugees. This finding is useful to determine the actual living location of refugees and help to verify statistics related to residences and mobility of refugees. Moreover, it ensures information on communication patterns of refugees which provides the policy makers some hints about the ways of effective communication to support and reach out to refugees with better policies.

Based on the conclusions drawn from this data analysis which worked on a sample taken from the given dataset, the study demonstrates that mobile phone data can be used to determine the exact location, mobility trajectory and patterns of communication of large refugee populations. Moreover, a deeper analysis developed by interrelating location and time of the calls presents some insights to interpret possible employment status of the refugee population. Considering the limitations with given data and limited time, the analysis does not come up with ambitious and grandiose policy suggestions but it provides a new methodology which clearly demonstrates how mobile phone data can be utilized for verifying, contributing and updating the actual official statistics on refugees in terms of identification of their location, time-bounded mobility activity and mobile phone communication patterns.

Finally, it is worth mentioning that further relevant research in the field of migration and refugee studies could be possible and feasible if the anonymous data shared is extended with gender and age info of the caller, information on the called number (refugee or not, individual or institution etc.) and duration of the call. Needless to say, these new features would make it more possible for the policy makers to remain relevant and develop well-targeted efficient integration policies.

References

1. Bilgi Teknolojileri ve İletişim Kurumu [Information and Communications Technologies Authority-Turkey], Üç Aylık Pazar Verileri Raporu, 2017 Yılı 1. Çeyrek. Sektörel Araştırma ve Strateji Geliştirme Dairesi Başkanlığı, (June 2017).
2. DGMM 2018a (Turkish Ministry of Interior Directorate General of Migration Management), [International Protection] http://www.goc.gov.tr/icerik6/uluslararasi-koruma_363_378_4712_icerik, last accessed 2018/08/05.
3. DGMM 2018b (Turkish Ministry of Interior Directorate General of Migration Management), [Temporary Protection] http://www.goc.gov.tr/icerik3/gecici-koruma_363_378_4713, last accessed 2018/08/05.
4. DGMM 2018c (Turkish Ministry of Interior Directorate General of Migration Management), [Residence Permits] http://www.goc.gov.tr/icerik3/ikamet-izinleri_363_378_4709, last accessed 2018/08/05.
5. Erdoğan, M.: Türkiye'deki Suriyeliler Toplumsal Kabul ve Uyum [Syrians in Turkey: Social Acceptance and Integration]. İstanbul Bilgi Üniversitesi Yayınları, Istanbul (2015).
6. Frontex [European Agency for the Management of Operational Cooperation at the External Borders of the Member States of the European Union] Risk Analysis Report 2016 (2016). doi:10.2819/416783, last accessed 2018/08/10.
7. International Organization for Migration (IOM): Irregular migrant, refugee arrivals in Europe top one million in 2015: IOM. [Press Release] (22 December 2015), https://www.iom.int/news/irregular-migrant-refugee-arrivals-europe-top-one-million-2015-iom, last accessed 2018/08/14.
8. İçduygu, A.: The politics of international migratory regimes: Transit migration flows in Turkey. International Social Science Journal 52 (165), 357–367 (2000).
9. İçduygu, A., Sert, D.: Country profile Turkey. Focus Migration 5, (April 2009).
10. İçduygu, A.: Turkey and international migration 2012–13. Migration Research Center at Koç University Report, Istanbul (November 2013).
11. Kaiser, B., Kaya, A.: Transformation of migration and asylum policies in Turkey. In: Güney A., Tekin, A. (eds.) The Europeanization of Turkish public policies. Routledge, New York (2015).
12. Kirişçi, K.: Syrian refugees and Turkey's challenges: Going beyond hospitality. Brookings, Washington, D.C. (2014). https://www.brookings.edu/wp-content/uploads/2016/06/Syrian-Refugees-and-Turkeys-Challenges-May-14-2014.pdf, last accessed 2018/08/14.
13. Regulation on Provision of Work Permits for People under Temporary Protection [Geçici Koruma Sağlanan yabancıların Çalışma İzinlerine Dair Yönetmelik], T.C. Resmi Gazete [The Turkish Official Gazette] 2016/8375, 11 January 2016, http://www.resmigazete.gov.tr/eskiler/2016/01/20160115-23.pdf, last accessed 2018/08/10.
14. Reidy, Eric.:Torbali: Third Stop on Mediterranean Route. News Deeply, (April 22, 2016). https://www.newsdeeply.com/refugees/articles/2016/04/22/torbali-third-stop-on-mediterranean-route.
15. Salah, A.A., Pentland, A., Lepri, B., Letouzé, E., Vinck, P., de Montjoye, Y.A., Dong, X. and Dağdelen, Ö., 2018. Data for Refugees: The D4R Challenge on Mobility of Syrian Refugees in Turkey. arXiv preprint arXiv:1807.00523.
16. The Report of the High-Level Panel of Eminent Persons on the Post-2015 Development Agenda: A New Global Partnership: Eradicate Poverty and Transform Economies through Sustainable Development (2013),

https://www.un.org/sg/sites/www.un.org.sg/files/files/HLP_P2015_Report.pdf, last accessed 2018/08/10.
17. Turkish Ministry of Labour, Social Services and Family Statistics 2018, http://cibs.csgb.gov.tr/, last accessed 2018/08/06.
18. UN High-level Dialogue on International Migration and Development (HLD), United Nations General Assembly Background Paper 3, http://www.un.org/en/ga/68/meetings/migration/pdf/HLD%20RT3%20background%20paper_final%20version%209.9.2013.pdf, last accessed 2018/08/14.
19. UNHCR Turkey: Key Facts and Figures June 2018, http://reporting.unhcr.org/sites/default/files/UNHCR%20Turkey%20Facts%20%26%20Figures%20-%20June%202018.pdf, last accessed 2018/08/10.

Measuring and mitigating behavioural segregation as an optimisation problem: the case of Syrian refugees in Turkey

Daniel Rhoads[1 [0000−0002−9037−1758]],
Javier Borge-Holthoefer[1 [0000−0001−9036−8463]], and
Albert Solé-Ribalta[1 [0000−0002−2953−5338]]

Internet Interdisciplinary Institute (IN3), Universitat Oberta de Catalunya (UOC)
Rambla del Poblenou, 156, Barcelona, Spain
{drhoads,jborgeh,asolerib}@uoc.edu

Abstract. Turkey hosts the largest population of Syrian refugees of any country in the world. As options to return home or settle in other countries remain limited, long-term integration of the refugee population into Turkish society is a major policy objective. Using a large dataset of mobile phone records provided by Turkeys largest mobile phone service operator, Trk Telekom, in the frame of the Data 4 Refugees project, we define, analyse and optimise inter-group integration as it relates to the communication patterns of two segregated populations: refugees living in Turkey and the local Turkish population. To achieve this, working from the records of call and SMS origins and destinations between and among both populations, we develop an extensible, statistically-solid, and reliable framework to measure the differences between the communication patterns of two groups. Utilising this new framework, we identify the districts of the Istanbul province where the variation between the ways the two populations communicate is largest. Finally, in order to show the potential of our framework, we provide and estimate the costs of some recommendations on how to target public and private investment to incentivise refugees to live outside of established refugee enclaves, increasing inter-group contact and integration.

Keywords: Mobile phone data · Segregation · Refugee Integration · Residential Mixing .

1 Introduction

Turkey hosts the largest population of Syrian refugees of any country in the world [31]. As the Syrian Civil War continues without a foreseeable end, and with policies in force to prevent refugee out-migration to the European Union, most analysts agree that policy action directed at Syrian refugee residents in Turkey should be based on the assumption that they will remain in the country for the long-term. With this in mind, policies have to become targeted to develop self-sufficiency among the refugee population, and to move them towards integration within the host community [53,13].

The analysis of segregated communities, due to its important implications for the lives of citizens and for social cohesion, has held the attention of policy-makers and academics in the field of social and urban sciences for some time. Segregation can have many dimensions [37] and comes with many faces: spatial [5], economic [27], occupational [20], gender [14,10,36], religious [48] or ethnic[39] segregation are just a few examples.

One of the first steps towards the understanding of segregation processes is the definition of measures that allow for its direct detection and quantification [19,38,49,9,23,35], or the quantification of similar concepts such as homophily [42,21,4]. Its identification permits the comprehension of the mechanisms that cause segregation to emerge as a phenomenon. Considered both from the theoretical field [25,45,43] and from current experimental analysis [8,32], segregation can develop as the result of one of the extremes of two opposing forces. On the one side, we find homophily, where similar members (of whatever category) are more affine to be connected. And, on the other side, cultural dissemination [6], driven by self-organized, social and political efforts which aim at maintaining communities at an equilibrium level of diversity, in a controlled way. When these two forces lose equilibria, either we reach segregation or complete cultural homogenisation. Considering homophily intrinsic to the human being only social enfold and public or private investment can incentivise refugees to leave enclaves and become more integrated in the local community by choice.

Long-term acceptance of Syrian refugees within the host community in Turkey has been identified as an important challenge as the refugees make semi-permanent lives in Turkey [22]. Even though many of the processes of integration are dependent on self-organisation, there is still an important role for policy action in promoting more integrated communities. Juzwiak et al. [28] identify several factors of immigrant integration that could be divided into three tiers, each one building off of the ones below. First, immigrants must have the basic rights to work and live, as well as a command of the local language. Next, there must be communication and interaction between the immigrant community and the local community. Finally, the interactions between the two communities should be positive. That is, both communities must be able to benefit from each other. The first tier factors are concerned basically with political and legislative acts, which need to be addressed by governmental institutions [2]. Bypassing this tier, which is out of the scope of the project and the provided data, we focus on the next important step, engendering communication and interaction between the Turkish local and Syrian refugee community.

Here, we define, analyse and optimise integration as it relates to communication patterns among different groups. Specifically, we are interested in studying the variation in communication patterns between two different populations in Turkey, Turkish locals and Syrian refugees. It is our claim that, in order to integrate separate communities, it is not enough to simply bridge the gap between several individual characteristics (e.g. social, economic and occupational) and spatial distributions, but that the communities must also be similar in the way they interact with each other. In this regard, the main assumption we rely on

is that, if two communities are equally distributed across a territory (that is, if there is no spatial segregation) and their calling requirements are equivalent, their communication patterns should be indistinguishable. This approach to behavioural segregation (as opposed to purely residential, spatial segregation) is paralleled in studies of the mobility patterns of segregated groups (*activity-space* segregation) [56], and in other work focusing on levels of contact between groups [11].

If we were to frame our measure of integration through communication patterns with respect to the classical dimensions of segregation, it would most likely be related to evenness [38]. Our approach, as we will describe in detail, is based on analysing the origin and destination of calls and SMS between the different segregated communities. From this analysis, we identify the districts of the Istanbul province that present the largest variations between the ways the two communities communicate. Finally, in order to show the potential of our framework, we provide and estimate the costs of some recommendations on how to target public and private investment to incentivise refugees to live outside of established refugee enclaves and clusters, increasing inter-group contact and integration.

2 Quantifying existing levels of spatial segregation

The development and presence of enclaves is a common phenomenon within immigrant communities. An immigrant enclave is an expression of spatial segregation as defined by the Dissimilarity Index [38] or by Louf et al. [35], who quantify segregation in terms of the deviation from the random distribution of populations in an area. As expected, refugees are not distributed equally across space in Turkey. Figure 1A plots the ratio of refugee to local population in Istanbul's 39 districts. Refugees are over-represented in districts with ratios above the horizontal red average line (representing the ratio of refugee-local of the Istanbul province), and under-represented in districts with ratios below the same line. Additionally, we provide the value of a well-known measure of segregation: the *Index of Dissimilarity*. The obtained value is not small, but nor is it as large as one might expect in a very segregated community [24] (e.g. values between 0.50 and 0.6 were found for the geographic segregation of the black and white population in the U.S. in 2000). Panel B and D of the same figure plot the distribution of the Turkish and refugee populations, respectively, in map form. The maps confirm that most of refugee enclaves are concentrated in the West-Center part of the province. On the contrary, we observe that the East part of the province seems practically inaccessible to refugees.

From another perspective, social integration can be also framed in terms of cost-benefit analysis [15]. In this conceptualisation, language acquisition, distance from family, and exposure to unfamiliar cultures would be considered costs, and are difficult to quantify in economical terms. However, that is not the case of the housing costs. Aside from the characteristics of individual houses, this cost reflects a variety of factors including access to services, employment, and

city resources [46,47,34]. Panel C of Fig. 1 provides complementary information to the spatial segregation analysis and shows that indeed Syrian refugees tend to live in cheaper, and so less favorable, neighbourhoods. This implies that some rent-reduction incentives might be effective in getting refugees to relocate out of enclaves.

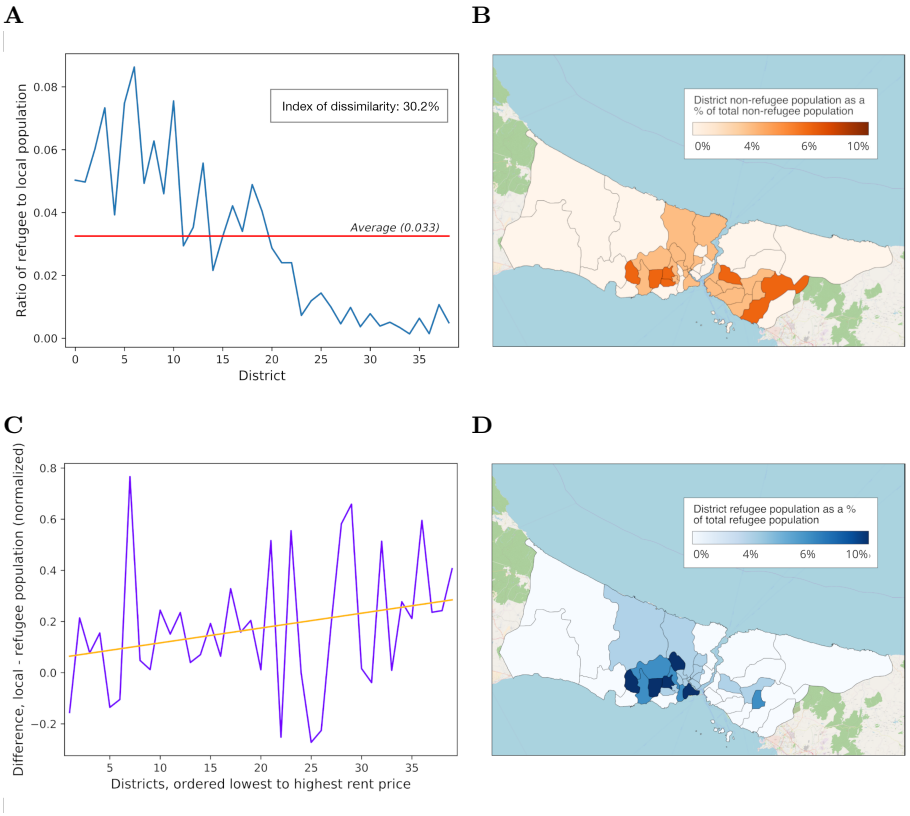


Fig. 1: Segregation analysis on the Istanbul province. Panel A illustrates the ratio of refugees to the local Turkish population across the 39 districts of the Istanbul province. The horizontal red line indicates the average ratio for the entire city. Deviations from this line indicate district-level segregation. Panel C shows the relationship between rent price paid for residence between Turkish and Syrian refugees. Data has been obtained from www.endeska.com. Panel B and D show, illustrated in map form, the distribution of the local Turkish and refugees populations, respectively, across the Istanbul province.

This analysis shown in this section presents an initial picture of the extent to which Turkish and refugee citizens are segregated spatially within Istanbul

province. The unexpectedly moderate results of the *Index of Dissimilarity* open up the possibility of exploring other possible measures of segregation, such as one sensitive to behavioural differences. The development and implementation of such a measure of behavioural segregation, through the analysis and comparison of group communication patterns, will be the subject of the rest of the paper.

3 Measuring behavioral segregation through communication pattern analysis

Segregation is usually, with a few exceptions [23,55], assessed in terms of the local demographic or socio-economic characteristics of each area of interest. However, segregation does not only regard the physical or spatial distribution of communities around an area, but also the relative level of harmonisation of local and cultural behaviours between groups [3,28]. Keeping in mind that the analysis of behavioural and cultural adoption is not easily quantifiable, here we develop a framework based on phone mobile phone use data to assess the extent to which communities differ in their behaviour and cultural habits [12].

3.1 Communication Network generation

The analysis we propose relies on the communication patterns between various collectives of people in terms of relational data, which in the following sections will be referred to as the Communication Network (CN). The CN will be represented, as is usual, as an adjacency matrix [41], O, where its entries o_{ij} correspond to the number of calls and SMS originated at location i and with destination j.

We built the Communication Network with the datasets provided for the Data 4 Refugees project [44,50]. The data is structured into 3 datasets, DS1, DS2 and DS3. The data in both DS1 and DS2 involves cellphone calls and SMS (which we will call 'communications' for convenience) as well as the identity of the cell antennas handling the communications. DS3 is made up of communication records, including the individual who made the call and the location of the call origin, for a pool of selected callers. To construct the CN, we made use of Datasets 1 (DS1) and 2 (DS2), as our analysis was concerned with communication between groups. DS3 was discarded.

As noted by the Data 4 Refugees project organisers, the indicator of refugee versus non-refugee for communication originators and receivers is imperfect. First, due possibly to simple human error, some communications registered as originating from refugees may have originated from non-refugees and vice-versa. Second, while the project is focused on Syrian refugees, who make up a significant portion of the refugee population in Turkey, the refugee indicator does not discriminate based on nationality. Thus, we refer to "Turkish callers" and "refugee callers", but not "Syrian callers".

To construct the required Communication Network, we require an estimation of the total communications made between each pair of cell antennas, divided

into four separate population sub-groups: communications originated by Turkish citizens and received by Turkish citizens (TT), communications originated by locals and received by refugees (TR), communications originated by refugees and received by refugees (RR) and communications originated by refugees and received by Turkish (RT). We achieved this through a combination of the information of DS1 and DS2.

DS1 consists of aggregate communication counts between unique cell phone antennas (site-to-site traffic) on an hourly basis, indicating, for each hour, the number of communications made, and the number of those communications which originated from refugees. Thus, while the proportion of communications originating from refugees and locals respectively was known, the proportion of each population receiving the communications was not. Hourly communication records were aggregated in order to calculate the total refugee-originated and local-originated communication between each pair of antennas.

DS2 is made up of individual communication records between unique originators and unique receivers, both identified according to their population (Turkish or refugee). For locational purposes, only the antenna originating the communications was available. From this information, we calculated, for each antenna, the proportion of total refugee-originated communications directed at locals, and the proportion directed at refugees; the same information was calculated for Turkish-originated communications from each antenna as well.

Thus, for each antenna, the proportion RR and RT sum to one, and the same holds for the proportion TT, TR. Next, we used these proportions to the aggregated information from DS1 to estimate the total RR, RT, TT, and TR communication counts for each unique pair of antennas.

Finally, for convenience (and noise reduction purposes) the antenna-to-antenna data was aggregated into district-to-district data. We considered districts a better unit of measurement, as they have more administrative meaning, and the aggregation of large amounts of antenna data lessens the risk of uneven geographical distribution of antennas skewing the interpretation of the data. Turkey is divided into 81 large administrative provinces, which are further subdivided into smaller districts, of which there are 923 in total. Our analysis focused on Istanbul, a province of Turkey made up of 39 districts.

It is important to note that communications originating from and received by the same district are also represented in our Communication Networks. That eventually will be addressed as O^{TT}, O^{RR} and O^{RT}. We omit the O^{TR} network since it is not required for our conducted analysis.

3.2 Aggregate communication pattern analysis: province scale

Once the Communication Network has been generated, we are ready to analyse the communication patterns of both collectives. We start with a macro-analysis of the average call destination probability of the area of study (Istanbul), independent of the individual district. This gives us an initial overview of how distinct both communication patterns are. A visual analysis of the results, see Fig. 2, suffices to show that both distributions have a similar shape, and so there

is not much difference between the communication habits of both collectives on average. However, detailed comparison at district level evidences a different situation. See 3. Panel A shows that while there are districts where the difference is small, in many others it is much larger than in the aggregated analysis. Panel B shows the differences in the distributions for three handpicked districts. The differences are visually evident.

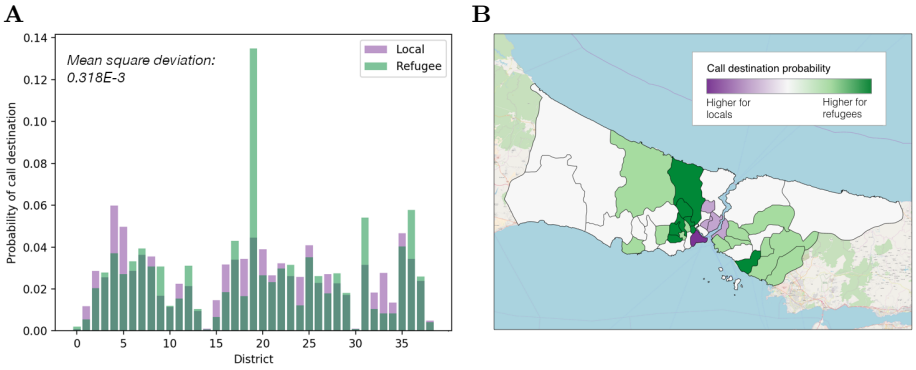


Fig. 2: Aggregated analysis of communication patterns of local Turks and refugees. Panel A shows the aggregated distribution of call destination (i.e. independent of district). Panel B presents the same data on an Istanbul district map.

The difference between the results obtained from the aggregated analysis and the initial local analysis might be indicative of the Simpson's Paradox [54] in the different Communication Networks. Within the aggregated whole of the province, each district has different proportions of refugee and local populations; additionally, social and economic factors vary by district and population. These considerations indicate the advantage of a local-scale analysis of the Communication Network in the characterisation of the differences between local and refugees communication patterns. In the next section, we address the formal structure of this local-scale Communication Network analysis, which forms the basis of the rest of the work.

3.3 Fine-grained communication pattern analysis: district scale

Given the Communication Networks across the different communities of interest, Turkish-Turkish, Refugee-Refugee and Refugee-Turkish, respectively O^{TT}, O^{RR} and O^{RT} (see Fig. 5A) we define our behavioural segregation measure in terms of the χ^2 test for homogeneity between the various outgoing communication distributions. Formally, the extent to which the two frequency counts are drawn from the same distribution is measured statistically by the p-value. In

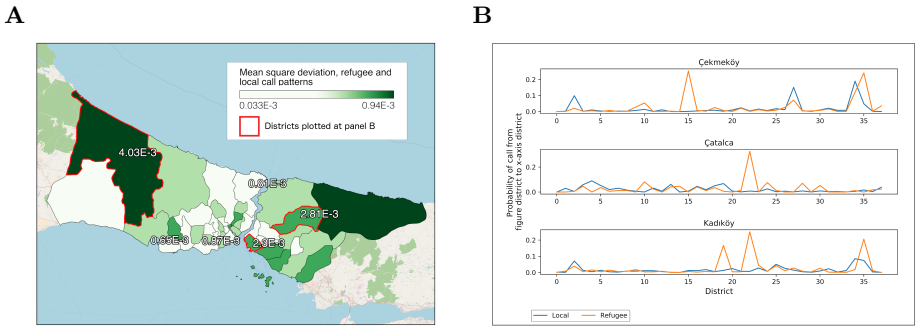


Fig. 3: District level analysis of the difference between communication patterns of the Turkish and refugee populations. Panel A illustrates the mean squared deviation of the distribution of probabilities of communication (call and SMS) destination originating from each district (e.g., the chances that a call from district i will be directed at district j and not a third district k). That is, for district i this is obtained as $\text{MSD}_i = \frac{1}{39}\sum_j (o_{ij}^{TT} - o_{ij}^{S\circ})^2$, where $o_{ij}^{S\circ} = o_{ij}^{SS} + o_{ij}^{ST}, \forall ij$. Panel B plots the distributions of the districts where the communication patterns for both communities is largest.

our case, the frequency counts correspond to the calls originating from district i and directed to each of the other districts j (represented as vector o_i) for both the Turkish population, $\mathbf{o}_i^{TT}$, and the refugee population, $\mathbf{o}_i^{RR}$. Thus,

$$\text{p-value}\left(\mathbf{o}_i^{TT}, \mathbf{o}_i^{RR}\right), \tag{1}$$

will assess statistically if both frequency counts an be indistinguishable or not. If the results of the test inform us that both samples come from a different distribution (the two call register samples are significantly different), we conclude that there is segregation in the area in terms of communication. If the test does not allow us to reject the null hypothesis (H_o: both samples come from the same distribution) we cannot conclude segregation exists in that area.

Conducting the χ^2 test for each district of the Istanbul province, we obtain an initial view of how segregation is distributed in the province. Figure 4 shows the results of comparing the communication patterns of the Turkish-Turkish network to the Refugee-Refugee network. We observe that segregation in terms of communication is spread over the entire area, and that there are very few regions that we could refer to as not segregated. We would like to highlight the difference between Fig. 3 and 4. Figure 3 aims to provide a visual guide of the different communication patterns which is not possible through χ^2 test analysis due to the considerable differences of both distributions. In turn, Fig. 4 shows the result of the segregation analysis in terms of p-values. We only provide the results considering the Turkish-Turkish vs. Refugee-Refugee datasets since in the others all of the districts presented communications patterns that were significantly different.

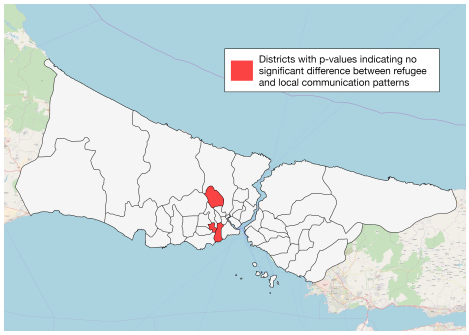


Fig. 4: Analysis of the difference in communication patterns between Turkish-Turkish vs Refugee-Refugee datasets in terms of p-values of the χ^2 homogeneity test on the communication destination counts. Red areas correspond to districts where p-value > 0.01. That is, areas where we cannot reject the null-hypothesis.

4 Mitigating behavioural segregation through residential mixing policies

Politicians [4,48], urban planers [39] and scholars [33,20] have been debating the solutions to segregation and concentration of poverty in Europe and North America since the 70's. One of the primary mechanisms developed, along with some criticisms [33], is residential and social mixing [40,5]. Mixing polices aim to promote the mobility of segregated communities to other neighbourhoods with low segregation. The final objective is to obtain neighbourhoods with heterogeneous populations given the required volumes of citizens of each community. Our work builds from such policies and aims to estimate the specific volumes of citizens that would need to move from their current district, as well as the districts they would need to move to, in order to improve segregation as measured by variations in Communication Networks.

4.1 Minimising segregation: residential mixing as an optimisation problem

As discussed, house or residential mixing aims at promoting the mobility of segregated communities into other less segregated neighbourhoods. Framing this concept within our definition of behavioural segregation, the problem can be rephrased as obtaining a mobility matrix $\mathbf{M}$, where each entry m_{ji} stands for the fraction of refugees living in district i that are required to be reallocated in district j, that maximise[1] the p-value of the χ^2 homogeneity test. Our interpretation of the problem is very similar to the definition of the Duncan Dissimilarity

[1] Recall that we reject the H_o when the p-value is less than our significance level, so maximising the p-value has the effect of forbidding the rejection of the H_o. That is, we cannot say there is segregation in the area.

Index [19], which is usually interpreted as the percentage of the minority population that would need to relocate in order to perfectly integrate the residential distributions in a region. The estimation of the best M can be formally described as an optimisation problem with the following variables. See Fig. 5, where the matrix of original communication records $\mathbf{O}$ is shown in column A; the mobility matrix $\mathbf{M}$, column B; and the resulting matrix of communication records $\hat{\mathbf{O}}$ after the mobility matrix have been applied, column C. Column B shows, in addition, the effect the mobility matrix has over the original call records. Note that moving 50% of the citizens from region A_1 to region A_2 implies that the same fraction of communications originated at A_1 are now originated at A_2 with equivalent destinations. The final non-linear optimisation problem corresponds to

$$\text{maximize} \quad \sum_i \text{p-value}\left(\mathbf{o}_i^{TT}, \hat{\mathbf{o}}_i^{RR}\right) \tag{2}$$

$$\text{s.t.} \quad \sum_i m_{ji} = 1 \ \forall i \tag{3}$$

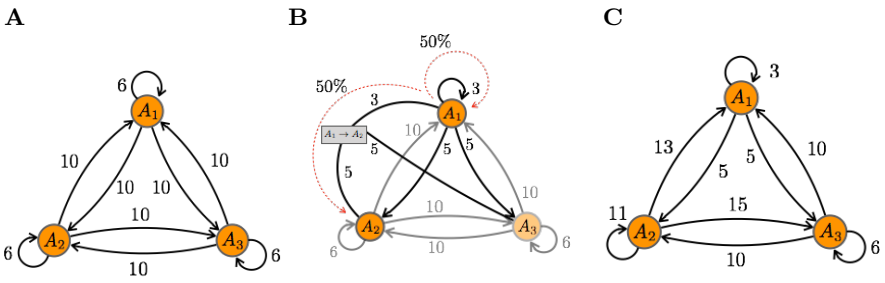
$$\sum_j \hat{\mathbf{o}}_{ij} \leq f_i \ \forall i \tag{4}$$

$$0 \leq m_{ji} \leq 1 \tag{5}$$

$$\text{where} \quad \hat{\mathbf{O}}^{RR} = \mathbf{M}\mathbf{O}^{RR} \tag{6}$$

where each m_{ji} is an unknown to be obtained. Restriction in Eq. 3 guarantees that the total number of communications is maintained. That is, in the mobility matrix, the sum from each origin to all the destinations must equal the total number of communications observed in the call record matrix $\mathbf{O}$. The restriction in Eq. 4 ensures that no district has more than f_i refugees. This restriction is important since the definition of enclaves has to do with the fraction of immigrants in a region with respect to its total population. In our case f_i is obtained such that the fraction of refugees living in a district never exceeds 10% of the total population. The restriction in Eq. 5 simply ensure that the different m_{ji} correspond to fractions.

The high non-linearity of the problem does not allow us to obtain satisfactory results optimising directly the previous optimisation problem. The fundamental complication was due to the very low p-values obtained with the initial call densities, $\mathbf{o}_i^{TT}$ and $\mathbf{o}_i^{RR}$. From that values we were unable to find good initialisation for unknowns m_{ji} that are close enough to a satisfactory mobility matrix solution. Instead, we developed a two steps procedure based on two similar optimisation problems. At the first step we modified the objective function (with equivalent restrictions) to find the mobility matrix that minimise the mean squared difference between vectors $\mathbf{o}_i^{TT}$ and $\hat{\mathbf{o}}_i^{RR}$. At the second step, using as initialisation vector the mobility matrix outcome of the previous optimisation, we minimised the sum of the χ^2 value for the different vectors $\mathbf{o}_i^{TT}$ and $\hat{\mathbf{o}}_i^{RR}$. The solution to the optimisation problem have been obtained by using the MatLab R2017a engine. We have used the *fmincon* function configured to use the Interior-Point



$$\mathbf{O} = \begin{bmatrix} 6 & 10 & 10 \\ 10 & 6 & 10 \\ 10 & 10 & 6 \end{bmatrix} \quad \mathbf{M} = \begin{bmatrix} 0.5 & 0 & 0 \\ 0.5 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad \hat{\mathbf{O}} = \mathbf{MO} = \begin{bmatrix} 3 & 5 & 5 \\ 13 & 11 & 15 \\ 10 & 10 & 6 \end{bmatrix}$$

Fig. 5: Description of the variables and structures of the non-linear optimisation problem. See main text for a description of the example.

algorithm. The source code as well as the aggregated data for the experiments and plots can be downloaded from *http://cosin3.rdi.uoc.edu/data.html*.

This double step process, similar to the original objective function in 2, gives very satisfactory results as Fig. 6 shows. Panel A shows the results mitigating segregation considering the Refugee-Refugee network. We observe that, parting from an initial situation in which more than 90% of the districts showed segregation, after the proposed mobility we reach a situation where only about 40% of the districts show segregation. When considering Refugee-Turkish communications, the results are also impressive (see 6B). In the initial situation 100% of the districts indicated segregation. However, after promoting the mobility, segregation remained significant in only 60% of the districts.

In order to provide a referenced view of the spatial segregation of the two populations after optimising segregation in terms of communication, we recomputed the *Index of dissimilarity* for the resulting Turkish and refugee population distributions. The results indicate that the proposed mobility for both networks (RR and RT) would have no effect on the initial spatial segregation. The new value of the index in both cases was 29%.

4.2 Economic incentives towards integration

The proposed optimisation problem in Eq. (2) provides us with information about the volume of communications that need to be shifted from one district to another. The density of communications originating from an area is known to be related to the population density of the area [17,18,30] as Appendix Fig. 1, drawn from the project data, shows. We then are able to use outgoing call

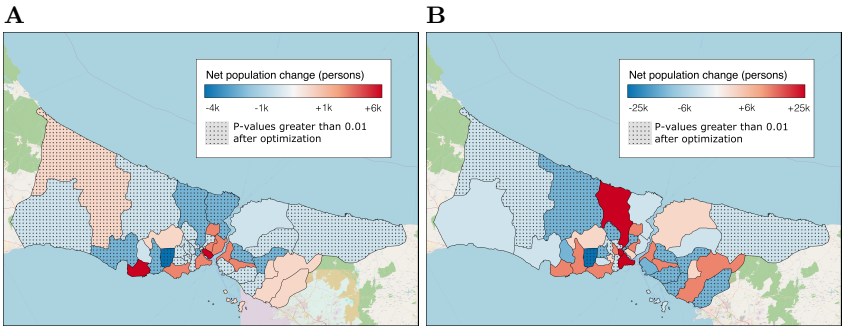


Fig. 6: Maps of the results of the optimisation, for the Refugee-Refugee and Refugee-Turkish networks respectively. Districts are coloured in a gradient according to their relative change in population. Districts where Turkish and Refugee communication patterns were harmonised (p-value from χ^2 test indicating no significant difference) have a dotted pattern

volume as a proxy for the amount of citizens for whom we need to incentivise mobility.

While higher rent prices are negatively related to both refugee and total population, the relationship is more evident for the refugee population (see Fig. 1C). This could be an opportunity for public and private actors interested in increasing host-refugee integration in Turkey to adjust the cost-benefit analysis of refugee location choice by subsidising rent in targeted areas of the city, thereby encouraging refugees to live away from enclaves and making inter-group contact more frequent.

In support of the viability of using rental subsidies as a way to incentivise refugee location choice, we examined the overall change in rent payments that would occur under the new population distribution considering rental markets for the 2017 period [1]. Performing the optimisation considering the RR communication, a total of 54948 refugees are required to be relocated. The resulting net increase in monthly rent cost is 9.021.009₺ (1.166.188€), which corresponds to 164₺ (21€) per household (of 4 members)/month. Performed considering the RT communication network, the optimisation resulted in a relocation of 399987 refugees. This corresponds to a net rent increase of 52.430.233₺ (6.777.903€), or 131₺ (17€) per household per month.

As can be seen in Fig. 7A and C, the changes in rent payment approximate a normal distribution with large variance, meaning that, under the adjusted population distribution, some refugees would considerably increase their savings on rent, and others would pay a quite larger price. The overall tendency, though, is a positive increase in the rent costs. The distribution of these changes in rent cost over the districts of Istanbul at the level of the individual is provided in Appendix Fig. 2. These figures provide an individual (refugee) point of view in terms of the increase or reduction in cost of living. Panels B and D of Fig. 7, on

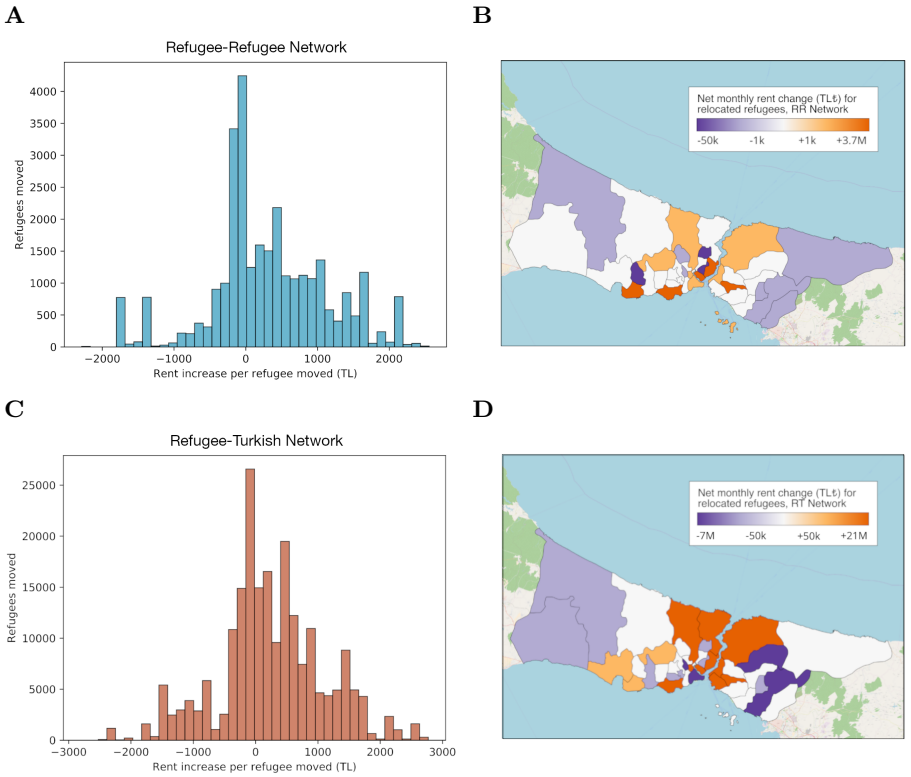


Fig. 7: The histograms at left illustrate the distribution of monthly rent changes after optimisation, for the Refugee-Refugee and Refugee-Turkish networks respectively. The height of the vertical bars indicates the number of relocated refugees whose rent payments increased or decreased by the value indicated on the horizontal axis. Subsequently, the maps at right indicate, for each respective network, the total monthly rent change per district. That is, the product, for each origin district j and destination district i, of the difference in rent cost between j and i and the number of refugees moved m to the destination district. Thus, formally, the reported rent change per district i is obtained as $v_i = \sum_{\forall j} m_{ij}(c_i - c_j)$

the other hand, provide a governmental or organisational perspective. The maps indicate the total investment that would be required in each district, in order to fully offset the increased rent payments of refugees. As we can see, the subsidies would be larger at the districts near the Bosphorus Strait. Surprisingly, these largest subsides are not regularly distributed among adjacent districts, and they correspond to the densest areas of the province.

5 Discussion

The method we propose in this paper allows for the quantification and potential mitigation of refugee segregation within a geographical area. The method goes beyond the spatial dimension usually considered and accounts for behavioural aspects of the different communities. From the combined analysis of communication data, the first conclusion we extract is that refugees and Turks do not communicate in the same way. There is a clear difference in their communications patterns obtaining a minimum of 36 out of 39 districts in the Istanbul province where their communication pattern is significantly different (p-values lower than 0.01). Two plausible reasons for that are the existence of strong cultural differences and the existence of enclaves. All the same, both reasons point to the segregation of refugees in specific areas. The basic assumption upon we rely is that the Turkish and Syrian refugees would improve integration if their communication patterns were indistinguishable. That is, their patterns of communication had reached a kind of natural equilibrium, considering the geo- and socio-economical situation of the city. We hypothesise that acting on the different collectives to merge the different communication patterns into a single one would bring the Turkish and Syrian refugee population closer to the stable social situation, allowing for more inter-communication and integration between the two communities.

Without a doubt, the reported results compose a very idealised situation and disregard known important aspects for integration, such as Syrian an Turkish cultural differences, which undoubtedly should be preserved and respected. We do not have any knowledge of the long term effects that achieving similar communication habits might have on these cultural aspects. Additionally, while we considered current rent conditions, the dynamics of urban politics in Istanbul did not enter into our analysis. Several authors have explored the complex relationship between the Turkish state, local government, real estate businesses, and residents in the context of the trend of "urban transformation" in the city [51,26]. This ongoing process of development and redevelopment, along with the construction of new specific city services such as the new city's airport [29], will clearly have an impact on any project that considers housing and rent. A carful study is needed to take these complex factors into account. Further research could also include other variables that we have not addressed here. For example, while our analysis is aggregated and anonymised, a similar procedure could be carried out with data tracking individuals over a period of time, to draw related but distinct conclusions. Additionally, "quality" of communications could be taken into consideration. Here, SMS and phone calls are given the same value. Perhaps even call duration could provide some measure of communication quality.

While we admit that the unaddressed aspects should without question be considered in subsequent works, the estimations given here can be of practical use in several ways. First, the developed procedure provides estimations on the amount of integration that can be obtained by using social and residential mixing strategies. Secondly, we provide quantitative and well-founded recommendations

about the volumes and destinations of the Syrian refugees that should be encouraged to relocate. This recommendation can be seen as a good starting point for governments and NGO's to analyse the situation, target their campaigns and optimise the economic investment in the area. Lastly, the optimisation framework proposed here can be easily complemented with other interesting parameters. In this work, we applied only one restriction to mobility, the one limitting the proportion of refugees per district. That said, mobility can be easily restricted in other ways; for instance, assuming the availability of the data, a restriction could be applied using employment data or labor demand in each district in order to achieve more socially accurate results.

Refugees choice of residential location influences their integration with the local community [16,7]. Policy-driven incentives such as rent subsidies could facilitate those who, for example, might choose to move away from an ethnic enclave if rent prices outside were lower. We have estimated that average rent paid among the relocating refugee population would not rise by more than by 21€per family per month. This is a barrier that could be too high for refugees who already have difficulties. However, it is also a barrier that governmental and NGO policy could reduce. Governments and NGOs have a range of options available to them to incentivise locational choice which are out of the scope of the scientific work presented here. However, several well-known approaches to the problem exist. For example, a simple program could involve a differential rent subsidy, or "voucher" [52], based on the relative rent price in target districts –moving to a more expensive district would lead to a higher subsidy.

As said, it is not our objective to influence particular policies, but instead to provide methods to quantify and give indication of what should be expected from house mixing policies. An optimal integration of the refugee and host population should probably be considered an organic process, as the meaning of integration here is connections between people, and connections are made voluntarily and maintained only by individual choice. In the event that governments and non-governmental entities decide to take a hands-off approach to integration policy, the proposed framework can be useful for analysing how the situation evolves and provide early warnings of recessive or problematic conditions.

As a final word, the authors of this work expect that any policy developed from our described findings will be aimed at helping refugees make positive integration choices for themselves.

References

1. Istanbul rental apartments price statistics, `https://www.endeksa.com/analiz/istanbul/endeks/kiralik/daire`
2. Ahmet, İ.: Syrian refugees in turkey-the long road ahead. Migration Policy Institute, Washington (2015)
3. Algan, Y., Bisin, A., Manning, A., Verdier, T.: Cultural integration of immigrants in Europe. Oxford University Press (2013)
4. Andersson, R., Bråmå, Å.: Selective migration in swedish distressed neighbourhoods: can area-based urban policies counteract segregation processes? Housing studies **19**(4), 517–539 (2004)

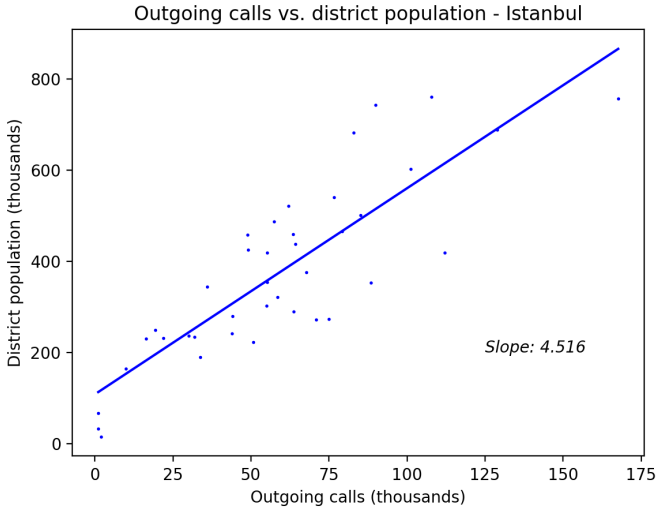
5. Atkinson, R., Kintrea, K.: Disentangling area effects: evidence from deprived and non-deprived neighbourhoods. Urban studies **38**(12), 2277–2298 (2001)
6. Axelrod, R.: The dissemination of culture: A model with local convergence and global polarization. Journal of conflict resolution **41**(2), 203–226 (1997)
7. Bauer, T., Epstein, G.S., Gang, I.N.: Enclaves, language, and the location choice of migrants. Journal of Population Economics **18**(4), 649–662 (2005)
8. Berry, J.W.: Acculturation: Living successfully in two cultures. International journal of intercultural relations **29**(6), 697–712 (2005)
9. Blackburn, R.M., Jarman, J., Siltanen, J.: The analysis of occupational gender segregation over time and place: considerations of measurement and some new evidence. Work, Employment and Society **7**(3), 335–362 (1993)
10. Blair, S.L., Lichter, D.T.: Measuring the division of household labor: Gender segregation of housework among american couples. Journal of family issues **12**(1), 91–113 (1991)
11. Blumenstock, J., Fratamico, L.: Social and spatial ethnic segregation: a framework for analyzing segregation with large-scale spatial network data. In: Proceedings of the 4th Annual Symposium on Computing for Development. p. 11. ACM (2013)
12. Canton, N.: Cell phone culture: How cultural differences affect mobile use. CNN online **28** (2012)
13. Chabkoun, M.: Can refugees return to syria, as many want them to? (2017)
14. Charles, M., Grusky, D.B.: Occupational ghettos: The worldwide segregation of women and men, vol. 71. LIT Verlag Münster (2005)
15. Danzer, A.M.: Economic benefits of facilitating the integration of immigrants. CESifo DICE Report **9**(4), 14–19 (2011)
16. Danzer, A.M., Yaman, F.: Do ethnic enclaves impede immigrants' integration? evidence from a quasi-experimental social-interaction approach. Review of International Economics **21**(2), 311–325 (2013)
17. Deville, P., Linard, C., Martin, S., Gilbert, M., Stevens, F.R., Gaughan, A.E., Blondel, V.D., Tatem, A.J.: Dynamic population mapping using mobile phone data. Proceedings of the National Academy of Sciences **111**(45), 15888–15893 (2014)
18. Douglass, R.W., Meyer, D.A., Ram, M., Rideout, D., Song, D.: High resolution population estimates from telecommunications data. EPJ Data Science **4**(1), 4 (2015)
19. Duncan, O.D., Duncan, B.: A methodological analysis of segregation indexes. American sociological review **20**(2), 210–217 (1955)
20. Duncan, O.D., Duncan, B.: Residential distribution and occupational stratification. American journal of sociology **60**(5), 493–503 (1955)
21. Easley, D., Kleinberg, J.: Networks, crowds, and markets: Reasoning about a highly connected world. Cambridge University Press (2010)
22. Erdogan, M.M.: Urban Refugees from" detachment" to" harmonization": Syrian Refugees and Process Management of Municipalities: the Case of Istanbul. Marmara Belediyeler Birligi (2017)
23. Farber, S., O'Kelly, M., Miller, H.J., Neutens, T.: Measuring segregation using patterns of daily travel behavior: A social interaction based model of exposure. Journal of transport geography **49**, 26–38 (2015)
24. Frey, W.H., Myers, D.: Racial segregation in us metropolitan areas and cities, 1990–2000: Patterns, trends, and explanations. Population studies center research report (05-573) (2005)
25. Gracia-Lázaro, C., Lafuerza, L.F., Floría, L.M., Moreno, Y.: Residential segregation and cultural dissemination: An axelrod-schelling model. Physical Review E **80**(4), 046123 (2009)

26. Gündoğan, A.Z.: Divergent responses to urban transformation projects in turkey: common sense and state affinity in community mobilization. Urban Geography pp. 1–25 (2018)
27. Jargowsky, P.A.: Take the money and run: Economic segregation in us metropolitan areas. American sociological review pp. 984–998 (1996)
28. Juzwiak, T., McGregor, E., Siegel, M.: Migrant and refugee integration in global cities: The role of cities and businesses. The Hague Process on Refugees and Migration and UNU-MERIT & its School of Governance (2014)
29. KAHRAMAN, C., Alkan, G.: Istanbuls third airport in terms of transportation geography: Geopolitics, regional and economic effects. PEOPLE: International Journal of Social Sciences **3**(3) (2018)
30. Khodabandelou, G., Gauthier, V., El-Yacoubi, M., Fiore, M.: Population estimation from mobile network traffic metadata. In: World of Wireless, Mobile and Multimedia Networks (WoWMoM), 2016 IEEE 17th International Symposium on A. pp. 1–9. IEEE (2016)
31. Kirici, K., Brandt, J., Erdoan, M.M.: Syrian refugees in turkey: Beyond the numbers (Jun 2018), https://www.brookings.edu/blog/order-from-chaos/2018/06/19/syrian-refugees-in-turkey-beyond-the-numbers/
32. Koopmans, R.: Trade-offs between equality and difference: Immigrant integration, multiculturalism and the welfare state in cross-national perspective. Journal of ethnic and migration studies **36**(1), 1–26 (2010)
33. Lees, L.: Gentrification and social mixing: towards an inclusive urban renaissance? Urban Studies **45**(12), 2449–2470 (2008)
34. Limsombunchai, V.: House price prediction: hedonic price model vs. artificial neural network. In: New Zealand Agricultural and Resource Economics Society Conference. pp. 25–26 (2004)
35. Louf, R., Barthelemy, M.: Patterns of residential segregation. PloS one **11**(6), e0157476 (2016)
36. Maccoby, E.E., Jacklin, C.N.: Gender segregation in childhood. In: Advances in child development and behavior, vol. 20, pp. 239–287. Elsevier (1987)
37. Massey, D.S.: Reflections on the dimensions of segregation. Social Forces **91**(1), 39–43 (2012)
38. Massey, D.S., Denton, N.A.: The dimensions of residential segregation. Social forces **67**(2), 281–315 (1988)
39. Musterd, S.: Social and ethnic segregation in europe: levels, causes, and effects. Journal of urban affairs **27**(3), 331–348 (2005)
40. Musterd, S., Andersson, R.: Housing mix, social mix, and social opportunities. Urban affairs review **40**(6), 761–790 (2005)
41. Newman, M.: Networks: An Introduction. Oxford University Press (2010)
42. Newman, M.E.: Mixing patterns in networks. Physical Review E **67**(2), 026126 (2003)
43. Ódor, G.: Self-organizing, two-temperature ising model describing human segregation. International journal of modern physics C **19**(03), 393–398 (2008)
44. Salah, A.A., Pentland, A., Lepri, B., Letouzé, E., Vinck, P., de Montjoye, Y.A., Dong, X., Dağdelen, Ö.: Data for refugees: The d4r challenge on mobility of syrian refugees in turkey. arXiv preprint arXiv:1807.00523 (2018)
45. Schelling, T.C.: Dynamic models of segregation. Journal of mathematical sociology **1**(2), 143–186 (1971)
46. Selim, H.: Determinants of house prices in turkey: Hedonic regression versus artificial neural network. Expert Systems with Applications **36**(2), 2843–2852 (2009)

47. Selim, S.: Determinants of house prices in turkey: A hedonic regression model. Doğuş Üniversitesi Dergisi **9**(1), 65–76 (2011)
48. Smith, A.: Religious segregation and the emergence of integrated schools in northern ireland. Oxford Review of Education **27**(4), 559–575 (2001)
49. Somers, R.H.: A new asymmetric measure of association for ordinal variables. American sociological review pp. 799–811 (1962)
50. Telekom, T.: Data for refugees turkey: D4r. d4r.turktelekom.com.tr/ (2018)
51. Torus, B., Yönet, N.A.: Urban transformation in istanbul
52. Varady, D.P., Walker, C.C.: Housing vouchers and residential mobility. Journal of Planning literature **18**(1), 17–30 (2003)
53. Vignal, L.: Perspectives on the return of syrian refugees. Forced Migration Review (57), 69–71 (2018)
54. Wagner, C.H.: Simpson's paradox in real life. The American Statistician **36**(1), 46–48 (1982)
55. Wang, Q., Phillips, N.E., Small, M.L., Sampson, R.J.: Urban mobility and neighborhood isolation in america's 50 largest cities. Proceedings of the National Academy of Sciences (2018). https://doi.org/10.1073/pnas.1802537115, http://www.pnas.org/content/early/2018/07/03/1802537115
56. Wong, D.W., Shaw, S.L.: Measuring segregation: An activity space approach. Journal of geographical systems **13**(2), 127–145 (2011)

A Appendix

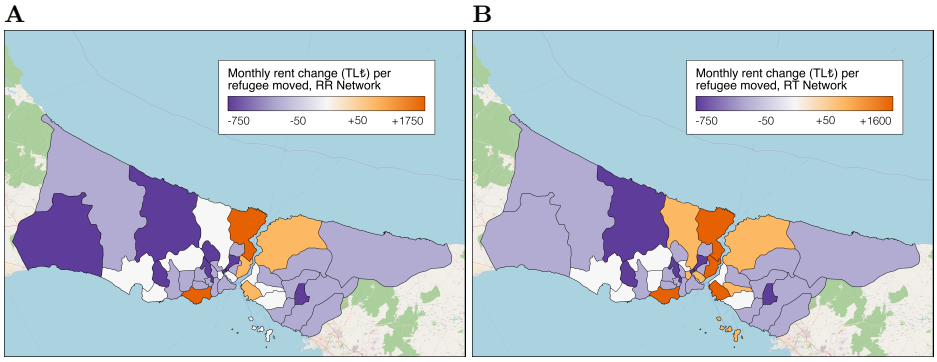
A.1 Additional Figures



Appendix Figure 1: Observed relationship between the amount of communications originated in each district of Istanbul and the population living in the district. We confirm a linear relationship as expected.

Id	District Name	Id	District Name	Id	District Name
1	Adalar	14	Bykekmece	27	Maltepe
2	Arnavutky	15	atalca	28	Pendik
3	Ataehir	16	ekmeky	29	Sancaktepe
4	Avclar	17	Esenler	30	Saryer
5	Baclar	18	Esenyurt	31	ile
6	Bahelievler	19	Eyp	32	ili
7	Bakrky	20	Fatih	33	Sultanbeyli
8	Baakehir	21	Gaziosmanpaa	34	Sultangazi
9	Bayrampaa	22	Gngren	35	Tuzla
10	Beikta	23	Kadky	36	mraniye
11	Beykoz	24	Kathane	37	skdar
12	Beylikdz	25	Kartal	38	Zeytinburnu
13	Beyolu	26	Kkekmece	39	Silivri

Appendix Table 1: Table with id's and district names



Appendix Figure 2: Maps, for each corresponding network, of the average monthly rent change per refugee arriving in each district. These maps complement the distributions of Fig. 7.

Refugee Mobility: Evidence from Phone Data in Turkey

Michel Beine[1], Luisito Bertinelli[1], Rana Comertpay[1], Anastasia Litina[2], Jean-Francois Maystadt[3,4], and Benteng Zou[1]

[1] University of Luxembourg, Avenue de la Faïencerie 162A, L-1511, Luxembourg
michel.beine@uni.lu
luisito.bertinelli@uni.lu
rana.comertpay@uni.lu
benteng.zou@uni.lu
[2] University of Ioannina, University Campus, 455 00 Ioannina, Greece
alitina@cc.uoi.gr
[3] University of Antwerp, Prinsstraat 13, 2000 Antwerpen, Belgium
[4] Lancaster University, Baggily, Lancaster LA1 4YW, UK
j.maystadt@lancaster.ac.uk

Abstract. Our research report employs the D4R data and combines it to several other sources to study one of the multiple aspects of integration of refugees, namely the mobility of refugees across provinces in Turkey. In particular, we employ a standard gravity model to empirically estimate a series of determinants of refugee movements. These include the standard determinants such as province characteristics, distances across provinces, levels of income, network effects as well as some refugee-specific determinants such as the presence of refugee camps and the intensity of phone call interaction among refugees. Importantly, we explore the effect of certain categories of news events, notably protests, violence and asylum grants. Considering news as an indicator of policy implemented at the provincial level we gain a better understanding as to how policy can facilitate refugee mobility and thus enhance integration. To benchmark our findings, we estimate the same model for the mobility of individuals with a non-refugee status.

Keywords: Social Integration, Refugee Mobility, Gravity Model of Migration, Poisson Pseudo-Maximum Likelihood

1 Introduction

Predicting human mobility in complex emergencies is essential for both operational and methodological reasons. For the former, understanding how displaced population make their mobility decisions may help relief operations to better target those in need of assistance. For the latter, researchers investigating the consequences of forced displacement on hosting areas have either assumed that forcibly displaced people have little agency, considering their location as quasi-random [Godøy, 2017; Grönqvist et al., 2012] or have overlooked the dynamic nature of such location decision. Anecdotal evidence suggests that for instance, refugees may move multiple times in their country of asylum [Bose, 2013; Bose, 2014].

In our study, we aim at better understanding the mobility of refugees within Turkey, which we consider as being a potential measure of social integration. Under the terms of the 1951 Refugee Convention, refugees are entitled to enjoy the freedom of movement. The determinants of refugee movements however are scarcely studied in the literature so far. To this end, we exploit spatially explicit call detail records provided by Salah et al. (2018) within the Data for Refugees Turkey ("D4R") challenge. More specifically, we look at the location of 100,000 randomly selected mobile transactions (50,000 refugees and 50,000 non-refugees) recorded by cell towers to define likely

decisions to move between provinces. Reconstructing bilateral migration flows at the monthly and at the province level, we can then apply a gravity model to understand the main determinants of refugee movement across provinces – in contrast to the mobility of non-refugees in Turkey. Our empirical findings suggest that distance, levels of GDP, network effects and policy likely influence refugees' mobility decisions. In particular, we highlight the importance of policy-related events and changes on movements across the country.

To benchmark our results we use the same model to study the mobility of the non-refugee population in our sample of individuals. We find that while the standard gravity model determinants also matter for the non-refugee population mobility, the policy-related determinants have a differential effect in magnitude.

Comparing the two groups of individuals, the data suggests that the non-refugee population moves not only more frequently but also longer distances than the refugee population. In the light of the fact that refugees are mostly free to move within Turkey (in some provinces restrictions may be in place but they are not strictly implemented), we can thus infer that the imperfect integration of refugees in the society (economic, market and/or social integration) is the main reason for reduced mobility. As convergence of the mobility of the two groups is thus a desirable outcome and a good proxy of the level of integration, it is essential to study some of the drivers of refugee mobility in order to be able to develop policies that can further facilitate and encourage it.

Beyond the relevance of our study for relief operations, our contribution is threefold. First, we contribute to an emerging literature exploiting mobile phone data to characterize human mobility in emergency situations. To the best of our knowledge, existent studies have focused on mobility responses to natural disasters (e.g. Blumenstock et al., 2016) and predicting disease propagation (e.g. Wesolowski et al., 2012).[5] Our analysis relates bilateral migration flows among refugees (and non-refugees to benchmark our results) who live in provinces in Turkey to push and pull factors.

Second, the economic literature has a long tradition of exploiting the gravity model to model migration decisions (Ravenstein, 1985, 1989). Despite its simplicity, the gravity model has shown impressive predictive power, making it an essential input for forecasting exercises between and within countries (Crozet, 2004; Mayda, 2010; Garcia et al., 2015; Beine et al., 2016)]. However, to the best of our knowledge, none of these studies have applied the gravity model to refugee migration. Combined with highly disaggregated data, its simplicity and predictive power may make it an interesting tool for emergency operations.

Third, aggregating the individual data at the province level allows us to further enrich the mobile phone data with a number of additional resources. We have two types of additional data. First, we construct a series of socioeconomic characteristics of provinces (notably using satellite data) which we can easily combine with our proxies for integration. Second, we construct a number of indices related to the incidence of news that could directly or indirectly concern the refugee population. In particular we have indices for the following types of news: leadership change, boycotts, violent protests, economic aid, humanitarian aid and asylum grants. We can then use the gravity model to study whether the implementation of some policies or the incidence of events (as captured and disseminated by the news) can affect refugee and non-refugee mobility.

The policy implications of our research are direct as we can trace mobility reactions to particular policy measures and assign a positive or a negative sign to it. Our research agenda aspires to expand the measures of integration used and further study the reaction of integration measures to policy decisions as well as to the incidence of various events.

[5] This strand of the literature builds upon advancements over the last two decades on the use of new technologies such as remote sensing, geographical information systems, and global positioning systems to study mobility patterns in non-emergency contexts (Deville et al., 2014).

The structure of the report is the following. Section 2 describes the empirical strategy used in this report. Under section 3, sub-sections 3.1 and 3.2 respectively present the data used and some descriptive statistics to help understand better the sample of our study. Section 4 provides the empirical results for our main research question. Section 5 conducts some robustness tests on the estimation of mobility determinants from section 4. Section 6 provides the implications of our results for policy and derive some recommendations. Section 7 concludes. The corresponding tables and figures are presented in section 8 at the end of this present report.

2 Methodology

Our empirical strategy is based on the utility maximization approach, proposed by Roy (1951), and further extended by Grogger and Hanson (2011) and Beine et al. (2011). The model has been frequently applied to migration (Beine et al., 2016). It is based on agents' decision to migrate in order to maximize their well-being, and leads to the pseudo-gravity framework, which can be readily estimated. The model predicts that migration flows can be expressed as follows:[6]

$$M_{odt} = \frac{e^{ln(y_{dt})+s_{dt}-\phi_{odt}}}{\Sigma_k e^{ln(y_{kt})+s_{kt}-\phi_{okt}}} \tag{1}$$

where M_{odt} is the expected migration flows between location o and location d at time t; ϕ_{odt} represents the accessibility of location d for potential migrants (i.e. migration costs); y_{dt} represents the attractiveness of location d in terms of utility (e.g. expected wage, ...); s_{dt}, the ability of location o to send migrants (e.g. public expenditures, ...); and k stands for all locations, other than o, i.e. potential destinations. The parameters of the model can then be estimated after a logarithmic transformation. In our application, we are estimating these parameters over a relatively short time-frame, spanning January to December 2017. We can derive the following specification:

$$ln(M_{odt}) = \beta' ln\frac{y_{dt}}{y_{ot}} + \gamma' s_{dt} - \delta' s_{ot} - \epsilon\phi_{odt} + \varepsilon_{odt}. \tag{2}$$

where $\varepsilon_{odt} = \phi_o + \psi_d + \pi_t + \varepsilon_{odt}$ can measure origin, destination and time fixed effects, and an independent and identically distributed (iid) error term. M_{odt} represents the bilateral mobility flow of refugees between provinces o and d at month t. $\phi_(odt)$ will take account of the cost of moving to d for potential candidates to mobility in province o. y_{dt} and y_{ot} capture both pull and push factors in provinces d resp. o, proxied by province level income. Month fixed effects are introduced to correct for seasonality in migration patterns.

Finally, given the large number of zeros in the bilateral migration flows, relying on standard estimation techniques (e.g. OLS) would likely lead to inconsistent coefficient estimates. As widely adopted in the literature (Beine et al., 2011), we call upon Poisson regression models that relies on pseudo maximum likelihood estimates (Santos Silva and Tenreyro, 2006; Santos Silva and Tenreyro, 2011).

3 Data and descriptive statistics

We first describe our data in sub-section 3.1. Second, we provide some descriptive statistics in sub-section 3.2 that will help visualize and therefore better explore our sample.

[6] Beine et al. (2016) detailed the derivation of the random utility maximization model of migration providing micro-foundations to the empirical specification of the gravity model.

3.1 The Data

In the above specification, the construction of our dependent variable is of key importance. In order to make the most of the D4R dataset, we aggregate our data at the province level. This allows us not only to use any of the three available datasets at this level, but also to combine it with any other data that can be collected or constructed at the same level. In the current project, we proxy integration with a measure of mobility. This is measured by a migration rate, which is of the form *Migration Rate_'r'_'i'* where *'r'* refers to the refugee (i.e. R) or non-refugee (i.e. NR) status of the observation, and *'i'* corresponds to the minimum number of calls generated from a given province to characterize the latter as the residence location (i.e. frequency filter of *'i'* calls, in ou case, we set *'i'*=10).[7] It is worthwhile noticing that we have restricted our analysis only to calls occurring during weekends, thereby increasing the likelihood to focus on location of residence rather than workplace.

We employ two main sets of determinants of mobility. First, we use the standard gravity model controls, i.e., variables that relate to the attractiveness (resp. repulsiveness) of district d (resp. o) for prospective refugees, the so-called pull (resp. push) factors.

In the absence of systematic data at the province-monthly income level for Turkish provinces we proxy for the economic attractiveness using night-light density at province and month level. We obtain data on province-level night-lights in Turkey from National Oceanic and Atmospheric Administration ("NOAA")'s National Centers for Environmental Information ("NCEI"). NOAA provides users with public access to geographical data and information. The use of satellite data in order to proxy economic activity at fine units for which systematic data are not available, is nowadays a standard practice in economics (see e.g., Henderson et al. (2011, 2012)).

Networks at the destination province also play a key role in reducing migration and assimilation costs (Beine et al., 2011). Networks are an essential pull factor as they are likely to provide information or financial support to newcomers [Munshi, 2003; Beaman, 2012]. In the absence of official measures for refugee networks at the province-monthly level, we proxy for such existing networks, using the number of calls from refugees. We construct this variable by using the dataset 1 provided by Salah et al. (2018) where we compute the total number of calls from refugees per province.

Proximity between pairs of provinces is measured using geodesic distances. It captures practical difficulties of moving across provinces. Last, we construct a binary variable indicating the presence or the absence of a refugee camp.

The second set of variables that we construct is aimed to capture policy related issues. Our source dataset is the Global Database of Events, Language and Tone (hereinafter referred as to "GDELT"). GDELT captures world-wide news media over 30 years, in over 100 languages and is updated daily. The provided database consists of over a quarter billion georeferenced event records in over 300 categories. The platform is open for research and analysis. It provides news for a large number of events.

The variables we have relied upon are as follows: rally for leadership change; boycotts; violent protest; economic aid; humanitarian aid and asylum grants. In GDELT, an event is given an id *GlobalEventID* and there exists a variable *EventBaseCode* which shows to which category this particular event belonged to. CAMEO (Conflict and Mediation Event Observations) event codes are defined in a three-level taxonomy. This enables us to aggregate events at various resolutions of specificity. In our study we also aggregate the data at the province level that allows us to merge it with the existing dataset.

[7] Under section 5, we construct a stricter mobility measure, i.e., we replicate our analysis with a frequency filter of 20 calls.

The aid variables indicate news that either humanitarian and/or economic aid is provided in province d (resp. o) at month m. Aid is crucial as it eliminates or at least partially alleviates financial concerns. Political factors have also been found to matter in other contexts. Researchers from various disciplines have been interested in measuring the impact of national policies on asylum seekers' health (Steele et al., 2002; Mills, 2012; Ziersch et al., 2017). A study from Greyling (2016) finds that government assistance, culture, economic factors, crime, refugee status, reasons for leaving the home countries, time spent and number of people staying in a house in the host country are all policies that affect asylum seekers in South Africa. We therefore augment the specification with variables capturing political factors such as boycotts, rally for leadership change and protest against the local authorities. Last, the news for asylum grants are directly linked with policy considerations that have a direct impact on the decisions of refugees and their ability to integrate and to move freely around the country.

3.2 Descriptive Statistics

Our sample is composed of 64,800 bilateral observations for which we have information about all variables in our baseline specification (See Table 1).[8]

According to our mobility measure, bilateral movements of refugees between provinces is limited, and amounts to 0.1%, i.e. on average, one refugee out of 1000 changes province from one month to another.

As we have already mentioned, we use the level of night-light density to infer income in different provinces (see Figure 1 for a Map in August 2017). Based on our study sample, every province in Turkey is shown to have at least one district for which the night-light density is the highest. (≈ 49.80). This generally corresponds to the city center of the provincial capital. Almost half of the provinces in Turkey have at least one district with a night-light density level of 0. The mean night-lights density equals ≈ 1.82, and its distribution is skewed to the right.

The shortest distance corresponds to the distance between the provinces *Gaziantep* and *Kilis* while the longest distance is between *Hakkari* and *Edirne*. The mean distance is $\approx$ 574km and approximately corresponds to the distance between *Mus* and *Yozgat*.

The total number of calls in our sample is almost 10 times higher than the number of refugee calls.[9] *Istanbul* is the province with the largest amount of calls, while *Bayburt* receives the lowest amount of calls.

In October 2017, *Ankara* was the province in which most events related to economic and humanitarian aid took place and also in which most events related to violent protests occurred in May 2017. Events related to asylum grants also concerned mainly *Ankara* during the same month, whereas most events related to rallies for leadership change took place in *Istanbul* in April 2017.

There are 10 provinces in which refugee camps are present. These provinces are *Adana*, *Adiyaman*, *Gaziantep*, *Hatay*, *Kahramanmaras*, *Kilis*, *Malatya*, *Mardin*, *Osmaniye* and *Sanliurfa*.

Finally, tables 2 and 3 offer a comparison between the mobility of refugees and non-refugees in our sample based on the frequency of their moves and the distance they traveled. Interestingly, according to our mobility measure, non-refugees move more often and further than refugees. Approximately 94% of refugees did not move with $\approx$ 4% of them moving once and $\approx$ 2% twice.

[8] The number of observations results from pairing each province with another province, given the bilateral nature of mobility. We do so for every month of the year 2017. We miss information on night-light for the month of June.

[9] It is worthwhile keeping in mind that the number of calls is based on the universe of calls from dataset 1 from Salah et al. (2018), whereas our mobility variable is computed using the sample provided from the third dataset.

While also a large amount ($\approx 87\%$) of non-refugees in our sample did not move, the remaining $\approx$ 12% moved at least once. Table 3 shows that while non-refugees traveled ≈ 112 km on average, non-refugees only traveled ≈ 37 km. Among the sub-sample of non-refugees who moved, ≈ 840 km were traveled in comparison with ≈ 615 km for refugees. This result implies that while most refugees did not move, those who did, according to our mobility measure, did travel quite a long distance.

4 Empirical Findings

This section presents our main empirical results.

Columns (1) to (6) of Table 4 explore the determinants of refugee movements. Each column corresponds to a new specification with the addition of events which respectively relate to rallies for leadership change, boycotts, violent protests, economic aid, humanitarian aid and asylum grants. Each event is considered both at the origin and destination provinces. All columns include a month fixed effect, a dummy variable for the presence of refugee camps at both the origin and destination districts. All these regressions also include level of night-lights at origin and destination, distance, number of calls and number of refugee calls (at origin and destination) as explanatory variables. Following the underlying pseudo-gravity model in a double log form, we use a logarithmic transform of these variables. We report robust standard errors to ensure the accuracy of inference.

As shown from the coefficients of night-lights at origin in columns (1) to (6), (low) income acts as a push factor for refugees in all specifications (even though it is marginally insignificant at standard levels in two cases). This result is quite standard in the migration literature, where people tend to leave low income places, to join higher income destinations.[10] However, in the present case, we are unable to highlight this latter feature, i.e. even though refugees seem to leave above all low income provinces, our results do not systematically indicate that higher income regions are preferred destinations. Put differently, refugees might not be able to reach wealthier regions although they tend to leave poorer ones, which deviates from the standard pull factor story, where higher incomes are usually considered as an important motivation for the choice of destination by migrants.

Proximity between provinces has an expected negative impact on movements and is significant at the 1% confidence level. This is a standard result in the literature on population movement. Also number of calls has an expected positive impact on mobility, which is again significant at the 1% confidence level. However a very interesting result emerges here; the number of refugee calls positively influences mobility with a significance of 5% (even 1% in specification column (2)). This result is in accordance with the literature around migration networks; migrants tend to move to regions where other migrants have already settled.

The second half of the explanatory variables in Table 4 focus on event data, as described in 3.1. As observed from Column (1), refugees tend to leave provinces with an ongoing rally for leadership change. Perhaps this captures political instability and a pre-election rhetoric that might be directed against the presence of refugees. Interestingly though it does not act as a pull factor. Column (2) illustrates that higher incidence of boycotts is associated with lower mobility. Provinces with higher number of boycott-related news could be more active on the political and humanitarian front and this may encourage immigrants to settle. As in Column (1), higher incidence of boycotts at the destination though does not attract immigrants. Violent protest news (Column 3) also do not confer any statistically significant effect on refugee mobility.

[10] it is also reassuring as to the fact that light density is a good proxy for income per capita at the province level.

Column (4) shows that refugees tend to move to provinces with more economic aid and leave as economic aid decreases. Humanitarian aid (Column 5) does not have any effect on the decision to move. Last, Column (6) shows that grants of asylum generate mobility towards these regions and results are significant at the 5% level.

Finally, the presence of a refugee camp in the origin or destination province does not attract or distract refugees from these places.

5 Robustness

This section presents results from the robustness tests we have conducted. Table 5 is subdivided in two panels, displaying only the results of our main variables of interest while Table 6 displays the results for all variables of the baseline specification in which dependent variable becomes the mobility of non-refugees.[11]

In Table 5 and under panel 1, we test for the robustness of the frequency filter that has been adopted to include observations in our dependent variable. In particular, we take districts for which at least 20 calls have been reported as opposed to 10 calls in our main specification. Although the significance of our results is overall reduced, interpretations remain qualitatively the same; in particular (i) rallies for leadership change at the origin are no longer significant, however the coefficient becomes significant for rallies at the destination; (ii) results on protest variables are mixed. In this specification they matter when they take place at the destination; (iii) economic aid at origin reduces the number of leavers, whereas it increases the number of new comers at destination; (iv) the same is true for (destination) provinces where more asylum has been granted.

Panel 2 of Table 5 reports results when adding origin and destination fixed effects. This allows us to take account of time invariant determinants specific to departure resp. arrival provinces, such as geographic feature, climatic factors, but also political institutions and demographic factors which can be hypothesized to be stable across our period of observation (i.e. January to December 2017). Results of these regressions do not deviate substantially from previous outcomes, i.e.rallies for leadership and boycotts, economic aid and asylum grants (and the absence thereof) remain important factors for refugees' decision to move and where to go.

In Table 6, we rely on the same specification as in Table 4, but change the sample, from refugees to non-refugees. The purpose of this is to serve as a counter-factual, i.e. do determinants of mobility of refugees differ from the rest of the population? Furthermore, mobility of refugees and the determinants thereof, is part of a component of social integration, and should therefore be compared to the reference group of non-refugees. We find that low income at origin plays a strong repulsive role for non-refugees, which is similar to our results on refugees, but the magnitude is stronger here. As expected, distance is negatively linked to migration. The refugee network (measured as the number of refugee calls) turns out to be insignificant, which is reassuring in terms of relevance of our measures. Also of noteworthinesses is the strong repellent effect that the presence of refugee camp at the province level have on potential non-refugees. Interestingly, the coefficients on event variables provide analogous results compared to Table 4, as to the direction of the coefficients. The same type of policy/events has the same (qualitatively) impact on the decision to move. However what changes is the magnitude of the effect which is systematically higher for non-refugees. This formalizes the summary statistics that indicate that refugee mobility is more constraint compared to non-refugee mobility.

[11] Complete result tables are available from the authors of this report upon request.

6 Policy Implications and Recommendations

Our report attempts a preliminary exploration of one dimension of integration, namely mobility of refugees within Turkish provinces. Applying standard econometric techniques on novel data compiled from several sources we can summarize our findings, for the sake of policy making, as follows: (i) Non-refugees move further and more frequently compared to refugees. Thus any policy measures should be directed to convergence of refugee and non-refugee movement. (ii) The standard determinants of mobility for immigrants also apply for refugees, i.e., income of the origin/source province, distance between provinces and network effects.

However access to rich provinces is potentially restricted for refugees. However, as a matter of fact, the very presence of refugees can become an engine of growth for the poor provinces. Besides, access to all provinces should be targeted as a policy measure, in order to assure equal mobility opportunities for everyone. (iii) Ensuring political stability in a province is an essential attracting factor for refugee population. (iv) Economic aid also facilitates the mobility of refugees and eliminates some of the mobility constraints. Economic aid can also attract refugees to particular provinces, thus it is an effective means of intervening in refugee mobility. (v) Asylum grants also matter a lot when it comes to mobility.

Last, but not least what our data hints to is that not only the implementation of particular policies affects the mobility of refugee population but also the proper dissemination of news may as well act as an incentive to this direction.

7 Concluding Remarks

This report is an attempt to provide an empirical estimation of the determinants of one measure of integration, i.e., for refugee movements in Turkey using phone data. We apply the standard gravity model to a novel dataset on phone calls conducted by refugees and we combine the data to several other datasets at the province level. We find that the standard gravity determinants apply, such as distance, origin and source income as well as networks. Furthermore, policy interventions that are facilitated with political stability, asylum granting and economic aid also matter thus suggesting that there is ample room for policy making.

We benchmark our analysis with a non-refugee sample to illustrate that any policy should be targeted to the convergence between refugee and non-refugee movement (non-refugees move further and more frequently). The same determinants apply in both samples, however the impact of each of these determinants is stronger for non-refugees. Thus any policy should be targeting at mitigating any factor that cause such differences.

8 Appendix

The following pages present tables and figures, which we refer to in the main text.

Table 1: Definitions and summary statistics of the variables from baseline specification

	Definition	Obs.	Mean	Std.	Min.	Max.
Dependent Variable						
Mobility of Refugees	This measure has been built by generating an index *migr_rate_'R'_'10'_'WE'_'In'* from dataset 3[a] where *'R'* indicates calls from refugees, *'10'* corresponds to the frequency filter, e.g. the minimum number of calls generated from a given district to characterize the latter as a destination, *'WE'* includes weekend calls and finally, *'In'* stands for ingoing calls.	64,800	0.001	0.02	0	1
Explanatory Variables						
Night-Lights	Monthly cloud free composites (excludes data impacted by stray light, lightning, lunar illumination, and cloud-cover) in geotiff format and collected across the globe at 750-meter resolution.	64,800	1.82	4.57	0	49.80
Distance[b]	This variable measures the geodesic distances, i.e. the length of the shortest curve between two points along the surface of a mathematical model of the earth, based on the coordinates of the centroids of each province.	64,800	574	322.62	38.71	1,559.53
Number of Calls	Constructed with dataset 1[c] where the total number of calls per antennae is obtained. Relying on GIS software, we link these antennae coordinates to turkish provinces.	64,800	416,922.8	1,654,334	2,585	2,13e+07
Number of Refugee Calls	Constructed with dataset 1[d] where the number of refugee calls per antennae is obtained. Coordinates of the districts in which these antennae are obtained using GIS software. These districts are then linked to their respective provinces.	64,800	50,105.81	199,618.8	0	2,653,137
Events[e]:						
Rally for Leadership Change	Dissent collectively, gather, or rally demanding leadership change	64,800	0.02	0.39	0	11
Boycotts	Refuse to work or cooperate until demands for political, social, economic, or other rights are met	64,800	0.16	1.03	0	13
Violent Protests	Protest forcefully, in a potentially destructive manner	64,800	0.28	1.86	0	27
Economic Aid	Extend, provide monetary aid and financial guarantees, grants, gifts and credit	64,800	1.16	5.30	0	60
Humanitarian Aid	Extend, provide humanitarian aid, mainly in the form of emergency assistance	64,800	0.76	3.59	0	42
Asylum Grants	Provide, grant asylum to persons	64,800	0.37	1.94	0	27
Presence of Refugee Camps[f]	Binary variable indicating the presence or the absence of a refugee camp in a province.	64,800	0.12	0.33	0	1

[a] See Salah et al. (2018)
[b] The centroid coordinates are based on the WGS 1984 datum and we rely on Vincenty (1975) equations to calculate distances.
[c] See Salah et al. (2018)
[d] See Salah et al. (2018)
[e] See Gerner et al. (2002).
[f] Data retrieved from https://data.humdata.org/dataset/syria-refugee-sites

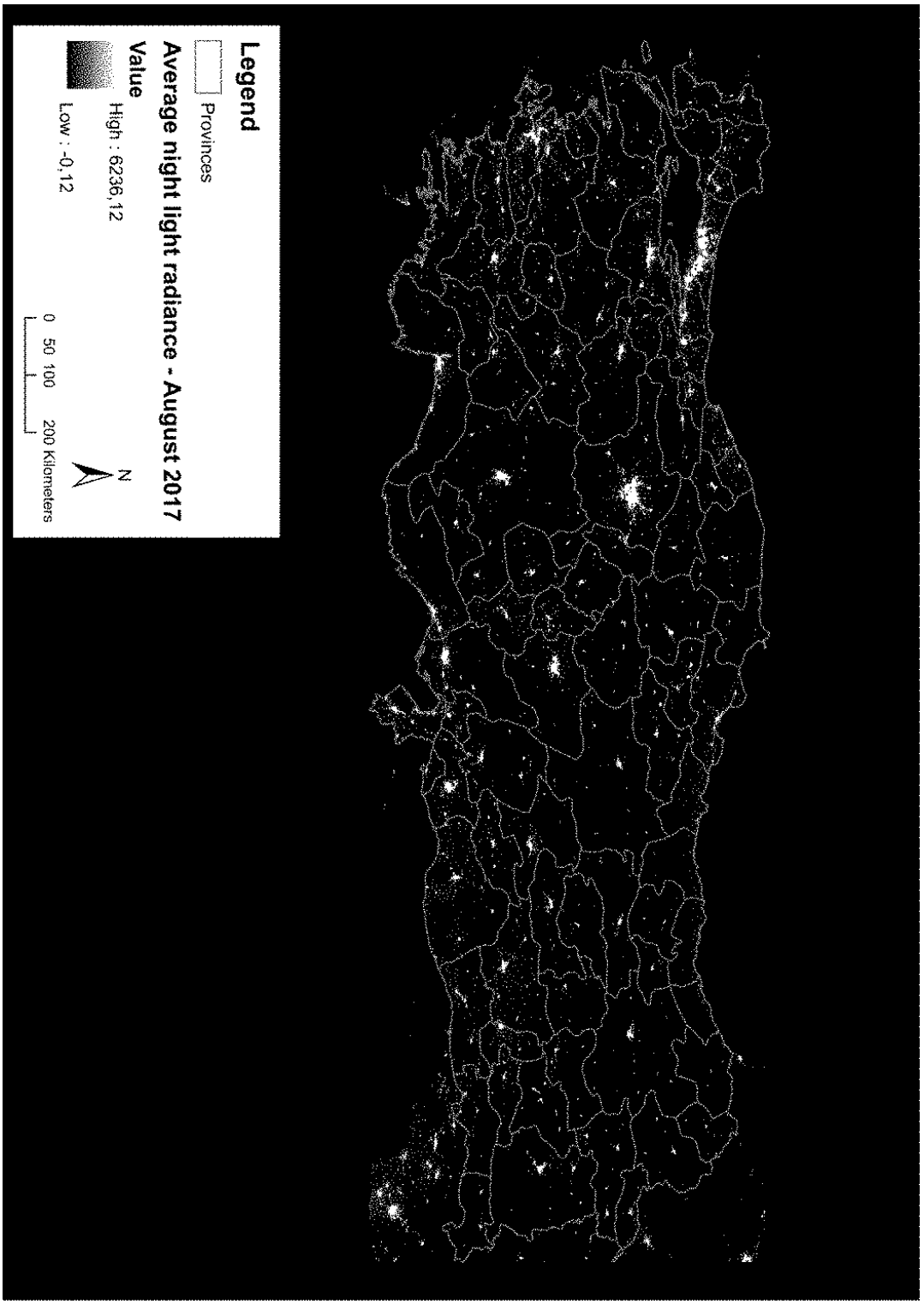


Fig. 1: Night-Lights in Turkey (August 2017; province borders)

Table 2: Mobility of Refugees and Non-Refugees: Frequency of their moves

	Refugee Mobility with a Frequency Filter of 10 calls		**Non-Refugee Mobility with a Frequency Filter of 10 calls**	
	Number of Refugees	Percent	Number of Non-Refugees	Percent
Moves:				
0	17,797	93.93	13,766	86.65
1	796	4.2	872	5.49
2	294	1.55	950	5.98
3	42	0.22	176	1.11
4	14	0.07	96	0.6
5	4	0.02	18	0.11
6	0	0	5	0.03
7	0	0	4	0.03

Notes: (i) The variables *Refugee Mobility with a Frequency Filter of 10 calls* and *Non-Refugee Mobility with a Frequency Filter of 10 calls* are of the form *Migration Rate_'r'_'i'* where *'r'* refers to the refugee (i.e. R) and non-refugee (i.e. NR) status of the observation and *'i'* corresponds to the minimum number of calls generated from a given province to characterize the latter as the residence location (i.e. frequency filter of 10 calls); (ii) While in our analysis the observations are provided for each pair of districts, here we provide descriptive statistics at the individual level.

Table 3: Mobility of Refugees and Non-Refugees: Distance Traveled

	Distance Traveled	
	Refugees	Non-Refugees
Average Distance (km)	37.3	112.1
Average Distance for Movers (km)	614.9	839.7

Table 4: Determinants of Refugee Movements in Turkey

	Dep. Var: Mobility of Refugees					
	(1)	(2)	(3)	(4)	(5)	(6)
Events:	Rallies for Leadership Change	Boycotts	Violent Protests	Economic Aid	Humanitarian Aid	Asylum Grants
Log Night-Lights at Origin	-0.231*	-0.161	-0.288**	-0.213	-0.234*	-0.212*
	(0.125)	(0.137)	(0.122)	(0.132)	(0.126)	(0.126)
Log Night-Lights at Destination	-0.0727	-0.0822	-0.0985	-0.0947	-0.0843	-0.0954
	(0.0842)	(0.0868)	(0.0887)	(0.0907)	(0.0870)	(0.0878)
Log Distance	-0.950***	-0.932***	-0.951***	-0.943***	-0.945***	-0.938***
	(0.112)	(0.114)	(0.113)	(0.114)	(0.113)	(0.113)
Log Number of Calls	0.329**	0.316**	0.329**	0.330**	0.329**	0.338**
	(0.141)	(0.146)	(0.142)	(0.144)	(0.143)	(0.145)
Log Number of Refugee Calls	0.429***	0.356**	0.419***	0.398**	0.401**	0.367**
	(0.153)	(0.155)	(0.154)	(0.156)	(0.158)	(0.155)
Events at Origin	0.146**	-0.0759***	-0.0315	-0.0238***	-0.0162	-0.0367**
	(0.0710)	(0.0281)	(0.0194)	(0.00744)	(0.0133)	(0.0153)
Events at Destination	0.0570	-0.0381	0.0174	0.0169***	0.0132	0.0347**
	(0.0710)	(0.0322)	(0.0185)	(0.00639)	(0.0106)	(0.0146)
Refugee Camps at Origin	-0.359	-0.347	-0.344	-0.351	-0.353	-0.344
	(0.219)	(0.218)	(0.219)	(0.218)	(0.219)	(0.218)
Refugee Camps at Destination	-0.372*	-0.145	-0.333*	-0.272	-0.290	-0.215
	(0.198)	(0.219)	(0.197)	(0.208)	(0.211)	(0.208)
Constant	-9.965***	-9.427***	-9.845***	-9.872***	-9.796***	-9.610***
	(1.105)	(1.140)	(1.116)	(1.128)	(1.131)	(1.130)
Observations	64,800	64,800	64,800	64,800	64,800	64,800
R-squared	0.043	0.044	0.043	0.042	0.043	0.044

Summary: This table presents the estimates from our baseline specification (See Section 2) and establishes some determinants of refugee movements within Turkey. Our analysis controls for month fixed effects and includes dummy variables to control for the presence of *refugee camps at the origin* and *refugee camps at destination*. Columns (1) to (6) show the coefficients or our empirical specification 2. Each column corresponds to a new specification with the addition of events which respectively relate to rallies for leadership change, boycotts, violent protests, economic aid, humanitarian aid and asylum grants. Each event is considered both at the origin and destination districts.

Notes: (i) Our dependent variable is measured by a migration rate, which is of the form *Migration Rate_'r'_'i'* where *'r'* refers to the refugee (resp. non-refugee) status of the observation, and *'i'* corresponds to the minimum number of calls generated from a given province to characterize the latter as the residence location (i.e. frequency filter of 10 calls); (ii) Robust standard errors are reported in parentheses; (iii) *** denotes statistical significance at the 1 percent level ($p < 0.01$), ** at the 5 percent level ($p < 0.05$), and * at the 10 percent level ($p < 0.10$), all for two-sided hypothesis tests.

Table 5: Results from Robustness Tests for the Mobility of Refugees

	Dep. Var: Mobility of Refugees					
	(1)	(2)	(3)	(4)	(5)	(6)
Events:	Rallies for Leadership Change	Boycotts	Violent Protests	Economic Aid	Humanitarian Aid	Asylum Grants
Panel 1	**Mobility of Refugees with a frequency filter of 20 calls**					
Events at Origin	0.0650	-0.0647	-0.0150	-0.0223**	-0.0281	-0.0151
	(0.0687)	(0.0466)	(0.0301)	(0.0104)	(0.0188)	(0.0196)
Events at Destination	0.186**	0.0350	0.0696***	0.0161*	0.0143	0.0330*
	(0.0745)	(0.0293)	(0.0207)	(0.00898)	(0.0133)	(0.0198)
Observations	64,800	64,800	64,800	64,800	64,800	64,800
R-squared	0.030	0.029	0.036	0.029	0.032	0.030
Panel 2	**Mobility of Refugees with fixed effects**					
Events at Origin	0.148**	-0.0878**	-0.0131	-0.0540***	-0.00698	-0.0377
	(0.0672)	(0.0417)	(0.0260)	(0.0173)	(0.0174)	(0.0329)
Events at Destination	0.0342	-0.158**	-0.0247	0.0119	0.00840	0.0429**
	(0.0508)	(0.0641)	(0.0265)	(0.0135)	(0.0186)	(0.0195)
Observations	52,510	52,510	52,510	52,510	52,510	52,510
R-squared	0.086	0.087	0.083	0.084	0.087	0.087

Summary: This table displays results of a number of robustness tests. Under panel 1, the mobility of refugees is measured by a frequency filter of 20 calls instead of 10. Panel 2 corresponds to the case in which fixed effects are introduced in our gravity model equation 2.

Notes: (i) Our dependent variables are measured by a migration rate, which is of the form *Migration Rate_'r'_'i'* where *'r'* refers to the refugee (i.e. R) status of the observation; (ii) Robust standard errors are reported in parentheses; (iii) *** denotes statistical significance at the 1 percent level ($p < 0.01$), ** at the 5 percent level ($p < 0.05$), and * at the 10 percent level ($p < 0.10$), all for two-sided hypothesis tests.

Table 6: Results from Robustness Tests for the Mobility of Non-Refugees

	Dep. Var: Mobility of Non-Refugees					
	(1)	(2)	(3)	(4)	(5)	(6)
Events:	Rallies for Leadership Change	Boycotts	Violent Protests	Economic Aid	Humanitarian Aid	Asylum Grants
Log Night-Lights at Origin	-0.500***	-0.419***	-0.629***	-0.471***	-0.461***	-0.465***
	(0.133)	(0.145)	(0.125)	(0.145)	(0.138)	(0.140)
Log Night-Lights at Destination	0.0453	0.0538	0.0788	0.0475	0.0516	0.0351
	(0.0527)	(0.0531)	(0.0544)	(0.0534)	(0.0526)	(0.0533)
Log Distance	-0.754***	-0.746***	-0.767***	-0.753***	-0.754***	-0.747***
	(0.0822)	(0.0813)	(0.0825)	(0.0820)	(0.0812)	(0.0819)
Log Number of Calls	0.910***	0.906***	0.901***	0.907***	0.898***	0.930***
	(0.106)	(0.106)	(0.105)	(0.106)	(0.105)	(0.107)
Log Number of Refugee Calls	0.0288	0.0302	0.0586	0.0446	0.0621	0.000289
	(0.114)	(0.115)	(0.112)	(0.113)	(0.111)	(0.113)
Events at Origin	0.364***	-0.0823**	0.00942	-0.0301***	-0.0114	-0.0326
	(0.0724)	(0.0334)	(0.0183)	(0.00863)	(0.0114)	(0.0227)
Events at Destination	-0.0548*	-0.00713	0.0144	0.00430	0.00427	0.0225**
	(0.0309)	(0.0270)	(0.0131)	(0.00453)	(0.00661)	(0.0100)
Refugee Camps at Origin	-0.114	-0.123	-0.0901	-0.119	-0.125	-0.117
	(0.141)	(0.142)	(0.141)	(0.141)	(0.141)	(0.141)
Refugee Camps at Destination	-0.509***	-0.505***	-0.558***	-0.535***	-0.565***	-0.478***
	(0.124)	(0.130)	(0.120)	(0.121)	(0.120)	(0.125)
Constant	-13.56***	-13.65***	-13.56***	-13.73***	-13.71***	-13.56***
	(0.771)	(0.775)	(0.755)	(0.751)	(0.738)	(0.762)
Observations	64,800	64,800	64,800	64,800	64,800	64,800
R-squared	0.151	0.154	0.150	0.152	0.152	0.154

Summary: This table corresponds to the mobility of non-refugees in Turkey and is the counter-factual in our analysis. Our analysis controls for month fixed effects and includes dummy variables to control for the presence of *refugee camps at the origin* and *refugee camps at destination*. Columns (1) to (6) show the coefficients or our empirical specification 2. Each column corresponds to a new specification with the addition of events which respectively relate to rallies for leadership change, boycotts, violent protests, economic aid, humanitarian aid and asylum grants. Each event is considered both at the origin and destination districts.

Notes: (i) Our dependent variable is measured by a migration rate, which is of the form *Migration Rate_'r'_'i'* where *'r'* refers to the non-refugee (resp. refugee) status of the observation, and *'i'* corresponds to the minimum number of calls generated from a given province to characterize the latter as the residence location (i.e. frequency filter of 10 calls); (ii) Robust standard errors are reported in parentheses; (iii) *** denotes statistical significance at the 1 percent level ($p < 0.01$), ** at the 5 percent level ($p < 0.05$), and * at the 10 percent level ($p < 0.10$), all for two-sided hypothesis tests.

Bibliography

Beaman, A. L. (2012). Social Networks and the Dynamics of Labour Market Outcomes: Evidence from Refugees Resettled in the U.S. *Review of Economic Studies 79*, 128–161.

Beine, M., S. Bertoli, and J. F.-H. Moraga (2016). A practitioners' guide to gravity models of international migration. *The World Economy 39*(4), 496–512.

Beine, M., F. Docquier, and C. Ozden (2011). Diasporas. *Journal of Development Economics 95*, 30–41.

Blumenstock, J., N. Eagle, and M. Fafchamps (2016). Airtime transfers and mobile communications: Evidence in the aftermath of natural disasters. *Journal of Development Economics 120*, 157–181.

Bose, P. S. (2013). Building sustainable communities: Immigrants and mobility in Vermont. *Research in Transportation Business & Management 7*, 81–90.

Bose, P. S. (2014). Refugees in Vermont: Mobility and acculturation in a new immigrant destination. *Journal of Transport Geography 36*, 151–159.

Crozet, M. (2004). Do migrants follow market potentials? an estimation of a new economic geography model. *Journal of Economic geography 4*(1), 439–458.

Deville, P., C. Linard, S. Martin, M. Gilbert, F. R. Stevens, A. E. Gaughan, V. D. Blondel, and A. J. Tatem (2014). Dynamic population mapping using mobile phone data. *Proceedings of the National Academy of Sciences 111*(45), 15888–15893.

Garcia, A. J., D. K. Pindoliay, K. K. Lopiano, and A. J. Tatemzz (2015). Modeling internal migration flows in sub-Saharan Africa using census microdata. *Migration Studies 3*(1), 89–110.

Gerner, D. J., R. Abu-Jabr, P. A. Schrodt, and Ömür Yilmaz (2002). Conflict and Mediation Event Observations (CAMEO): A New Event Data Framework for the Analysis of Foreign Policy Interactions. In *of Foreign Policy Interactions." Paper presented at the International Studies Association.*

Godøy, A. (2017). Local labor markets and earnings of refugee immigrants. *Empirical Economics 52*(1), 31–58.

Greyling, T. (2016). The expected well-being of urban refugees and asylum-seekers in Johannesburg. South African. *South African Journal of Economic and Management Sciences (SAJEMS) 19*, 232–248.

Grönqvist, H., P. Johansson, and S. Niknami (2012). Income inequality and health: lessons from a refugee residential assignment program. Working Paper Series 2012:11, IFAU - Institute for Evaluation of Labour Market and Education Policy.

Grogger, J. and G. H. Hanson (2011, May). Income maximization and the selection and sorting of international migrants. *Journal of Development Economics 95*(1), 42–57.

Henderson, J. V., A. Storeygard, and D. N. Weil (2012). Measuring Economic Growth from Outer Space. *American Economic Review 102*(2), 994–1028.

Henderson, V., A. Storeygard, and D. Weil (2011). A Bright Idea for Measuring Economic Growth. *American Economic Review 101*, 194–99.

Mayda, A. M. (2010). International migration: A panel data analysis of the determinants of bilateralflows. *Journal of Population Economics 23*(4), 1249–1274.

Mills, K. (2012). Under the radar: Impact of policies of localism on substance misuse services for refugee and asylum seeking communities. *International Social Work 55*, 662–674.

Munshi, K. (2003). Networks in the Modern Economy: Mexican Migrants in the U. S. Labor Market. *The Quarterly Journal of Economics 118*, 549–599.

Ravenstein, E. G. (1985). The laws of migration. *Journal of the Royal Statistical Society 48*(2), 167–235.

Ravenstein, E. G. (1989). The laws of migration – second paper. *Journal of the Royal Statistical Society 52*(2), 241–305.

Roy, A. D. (1951). Some Thoughts On The Distribution Of Earnings. *Oxford Economic Papers 3*(2), 135–146.

Salah, A., A. Pentland, B. Lepri, E. Letouzé, P. Vinck, Y.-A. de Montjoye, X. Dong, and z. Dağdelen (2018). Data for Refugees: The D4R Challenge on Mobility of Syrian Refugees in Turkey. arXiv preprint arXiv:1807.00523.

Santos Silva, J. and S. Tenreyro (2006). The Log of Gravity. *The Review of Economics and Statistics 88*(4), 641–658.

Santos Silva, J. and S. Tenreyro (2011). Further simulation evidence on the performance of the Poisson pseudo-maximum likelihood estimator. *Economics Letters 112*(2), 220–222.

Steele, L., L. Lemieux-Charles, J. P. Clark, and R. H. Glazier (2002). The Impact of Policy Changes on the Health of Recent Immigrants and Refugees in the Inner City: A Qualitative Study of Service Providers' Perspectives. *Canadian Journal of Public Health / Revue Canadienne de Santé Publique 93*(2), 118–122.

Vincenty, T. (1975). Direct and Inverse Solutions of Geodesics on the Ellipsoid with Application of Nested Equations. *Survey Review 23*, 88–93.

Wesolowski, A., N. Eagle, A. J. Tatem, D. L. Smith, A. M. Noor, R. W. Snow, and C. O. Buckee (2012). Quantifying the impact of human mobility on malaria. *Science 338*, 267–270.

Ziersch, A., M. Walsh, C. Due, and E. Duivesteyn (2017). Exploring the Relationship between Housing and Health for Refugees and Asylum Seekers in South Australia: A Qualitative Study. *International Journal of Environmental Research and Public Health 14*(9), 1036.

Mobility and Calling Behavior to Assess the Integration of Syrian Refugees in Turkey

Antonio Luca Alfeo[1,2] ✉, Mario G. C. A. Cimino[1]
, Bruno Lepri[3], and Gigliola Vaglini[1]

[1] University of Pisa, largo Lucio Lazzarino 1, Pisa, 56126, Italy
[2] University of Florence, via di Santa Marta, 3, Florence, 50121, Italy
[3] Bruno Kessler Foundation, Trento, via S. Croce, 77, 38123, Italy
luca.alfeo@ing.unipi.it mario.cimino@unipi.it lepri@fbk.eu
gigliola.vaglini@unipi.it

Abstract. By absorbing more than 3.4 millions Syrians in the last years, Turkey has shown a remarkable resilience. But host community hostility toward these newcomers is rising. To address this issue the formulation of effective integration policies is needed. However, the design and implementation of an effective integration policy demands tools aimed at (i) understanding the integration of refugees despite the complexity and the width of this phenomenon; and (ii) evaluating the effectiveness of the policy. In this context, great benefits can be provided by replacing the well-known paper-and-pencil survey with big-data driven measures. In this work, we propose a set of metrics aimed at providing insights and assessing the integration of Syrians refugees, by analyzing a large call details record (CDR) dataset. Specifically, we aim at assessing the integration of refugees, by exploiting the similarity between refugees and locals in terms of calling behavior and mobility, considering different spatial and temporal features. Together with the already known methods for data analysis, in this work we use a novel computational approach to analyze users' mobility: computational stigmergy, a bio-inspired scalar and temporal aggregation of samples. Computational stigmergy associates each sample to a virtual pheromone deposit (mark) defined in a multi-dimensional space and characterized by evaporation over time. Marks in spatiotemporal proximity are aggregated into functional structures called trail. The stigmergic trail summarizes the spatiotemporal dynamics in data and allows to compute the similarity between them. Our analysis employs a real-world CDR including calls from refugees and locals in Turkey throughout 2017.

Keywords: Social Integration · Safety and Security · Unemployment · Turkey · Refugee · Computational Stigmergy.

1 The Challenge

In the context of Syrian refugee crisis, Turkey is both an *effective* and *affected* country [1]. Indeed, it provides protection and facilities to more than three million refugees; but, on the other hand, an increasing hostility is emerging in the

local Turkish communities, due to the magnitude and the duration of the humanitarian crisis [2]. In order to prevent the growing of societal tensions over Syrian refugees, there is the need to formulate effective long-term integration policies [3] [4]. However, the formulation of an effective policy demands tools aimed at evaluating and understanding the integration of refugees despite the complexity and the width of this phenomenon. In this context, great benefits can be provided by complementing the paper-and-pencil surveys, the interviews, and the focus groups with a data-driven approach [9].

An interesting approach is to use data mining techniques to analyze the aggregated behavior of users, finding a number of groups based on behavioral similarity [37]. This approach can reveal interesting social phenomena occurring among refugees and locals [38].

One source of data that offers great potential for this kind of analysis are information captured from mobile phones [14], which have been used to analyze many effects of the migratory phenomena, i.e., the ones on political elections [8], job markets [7] or on the spread of epidemics [5].

In this work we analyze the Call Detail Records (CDR) datasets provided within the D4R data challenge [10] with the aim of unfolding which conditions can contribute to the integration of refugees. Moreover, we aim at providing some data-driven indicators of the integration of Syrian refugees in Turkey, in order to allow policy makers at evaluating the effectiveness of the strategies aimed at fostering the integration of refugees. In order to provide the reader with few insights about how each D4R datasets is used in our analysis, we briefly present each one of them:

- The *Antenna Traffic Dataset (ATD)* contains one year site-to-site traffic on an hourly basis. Each site is an antenna with known GPS location. Specifically, for each antenna we have a timestamp, the outgoing antenna, the incoming antenna, the total number of calls, the total number of refugees' calls, the total calls duration, and the total refugees' calls duration.
- The *Fine Grain Mobility Dataset (FGMD)* contains the antenna identifiers used by a group of randomly chosen users for a period of 2 weeks, for a total amount of 26 periods during 2017. This data has been anonymized by replacing the user number with a random ID, which prefix means refugee (i.e., 1), non-refugee (i.e., 2), and unknown (i.e., 3). Specifically, in this dataset we can find, for each call, the caller id, the timestamp, the callee prefix, and the antenna id.
- The *Coarse Grain Mobility Dataset (GGMD)* contains the calls details for a unique group of users throughout the whole year, but with a more coarse spatial coordinates, i.e. the district. In this case there is not any reference to the callee, while the caller can be identified as refugee or local according to the prefix of his/her caller id (as in the case of FGMD).

These datasets allow for different applications in terms of analysis width, number of individuals, spatial accuracy and time duration. To list few examples: (i) the information contained in the ATD can be used to describe the spatial distribution

of refugees calls, since it takes into account the whole population of refugees even if as a whole group; (ii) the information contained in the FGMD can be used to define trajectories of refugees and locals, or to assess refugee to refugee if this has more interactions (calls) with locals or other refugees, even if on a bi-weekly base; and (iii) the information contained in the CGMD can be used for long-term analysis (several months or year-round), and district-wise trajectories, or analyzing the difference in call patterns during this period.

In the following sections we present the analysis of these data. Specifically, in Section 2 we describe our approach and the metrics we aim to exploit. In Section 3 the experimental setup is depicted, and the results obtained are presented in section 4. Finally, we draw the conclusions of this study in section 5.

2 Method

In order to assess the integration of refugees, it is essential to establish metrics able to capture this phenomenon. These metrics should consider both on short (daily) and long (bi-weekly or monthly) term mobility and calling behavior of refugees and locals. Indeed, many works in the literature [29] highlight the improvement obtained by including individuals mobility and behavior in the model, with respect to pure statistical one. It follows the list of the metrics we propose for our analysis:

- *Residential Inclusion by District (RI)*: we can assume that most of the calls during the night and early morning hours come from people's homes. Indeed, based on this assumption many works in the field of the CDR analysis infer the location of an individual's home as the place from which he/she mostly call between 8 pm and 8 am [25]. Thus, by observing the percentage of calls made by refugees (via the ATD dataset) between 8 pm and 8 am per antenna $a \in d$ is possible to assess the coexistence of resident locals and refugees in a given the districts d and a given month m. This metric is defined between 0 (no resident refugees' in the district) and 1 (only resident refugees' in the district).

$$RI_{d,m} = \frac{|calls_{a,m}(R)|_{a \in d}}{|calls_{BS,m}(R) + calls_{a,m}(L)|_{a \in d}} \tag{1}$$

- *District Attractiveness (DA)*: A district is considered attractive if the flow of people who move to it is on average higher than the flow of people who move from there in a given month (i.e. the people netflux). As for the assumptions used in the RI metrics, a person resides in a given district and month if that district is the most recurrent location from which he/she makes calls between 8 pm and 8 am. Specifically, given $residentRefugee_{d,m} = \{R | r : home_r(m) = d\}$ i.e. the set of the refugees who live in the district d during the month m, the District Attractiveness (computed via the CGMD dataset) is defined as:

$$DA_{d,m} = |residentRefugee_{d,m+1}| - |residentRefugee_{d,m}| \tag{2}$$

– *Refugee's Interaction Level (IL)*: it is defined as the percentage of phone calls toward locals made by a given refugee in a given period, (computed via the FGMD dataset). It represents how much the refugee is socially connected to the locals [11], i.e. 0 means no calls toward locals and 1 means only calls toward locals. Each level is defined as a range of 20% within this scale. In this, as in many studies in this field [24], we consider the IL a solid metric for measuring individual integration.

$$IL_r = \frac{|calls_{r \to L}|}{|calls_{r \to L}| + |calls_{r \to R}|} \tag{3}$$

– *Refugee's Calling Regularity (CR)*: let us consider the time series of the call frequency (i.e. the calling pattern) made by each individual. Specifically, we build the calls pattern as the number of phone calls made by a person in a given hour of the day during a period of time. We normalize this amount with the average number of calls per hour in order to be comparable despite the different amount of calls made by each person. The period of time taken into account can be a month if the Calling Regularity is computed with the CGMD dataset, bi-weekly if it is used the FGMD dataset. In general the calling pattern may be due to several factors, e.g. daily routines, habits, or working schedule.

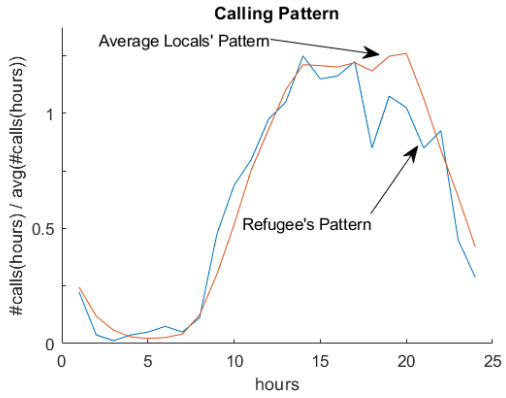


Fig. 1. Calling patterns representation.

Even if it is not possible to determine which component has a predominant role in generating a specific calling pattern, we can assume that similar routines will most likely generate similar calling patterns. Moreover, routines similarity is often linked to integration [31] [32]. Thus, the similarity between the calling patterns of locals and refugees, may be a proxy of integration. As an example, each refugee that is employed is supposed to have a calling pattern (thus, a daily routine) similar to the average calling pattern of

the locals, since they are mostly employed [15]. In this context, the more a refugee's calling pattern CP_r is similar to the average local's calling pattern LCP the more it is considered regular. The similarity between two calling patterns is computed (Eq. 4) as their cosine similarity [26]. This metric is defined between 0 (completely different calling pattern w.r.t. locals) and 1 (identical calling pattern w.r.t. locals).

$$CR_r = \frac{CP_r \cdot LCP}{||CP_r|| \cdot ||LCP||} \tag{4}$$

– *Refugee's Mobility Similarity (MS)*: by collecting the locations of each call (via FGMD dataset) occurred during the day we can build the daily trajectories of an users' mobility. The similarity of the trajectories of refugees T_r and locals T_l implies the sharing of some urban space at the same time and may affect (or be affected by) the integration of the refugees [12] [13]. The computation of this similarity is based on the principle of *stigmergy*. Stigmergy is a self-organization mechanism used in social insect colonies [27]. Basically, individuals in the colony affect each other behavior by marking a shared environment with pheromones when a specific condition occurs (e.g. the presence of food). The pheromone marks aggregate with each other in the trail if they are subsequently deposited in proximity to each other, otherwise they evaporate and eventually disappear. Thus, the resulting pheromone trail steers the whole colony toward the region in which the condition above (e.g. the discovery of food) occurs consistently.
This pheromone-like aggregation mechanism can be employed in the context of data processing, providing self-organization of data [28] while unfolding their consistent spatio-temporal dynamics [22]. By exploiting *computational stigmergy*, each sample of the trajectory is transformed in a digital pheromone deposits (i.e. mark) and released in a three-dimensional virtual environment in correspondence of each sample coordinate and time of appearance. Marks are defined by a truncated cone with a given width. Marks aggregate in the *stigmergic trail*, which is characterized by evaporation (i.e. temporal decay δ). The evaporation may be counteracted if marks are frequently released in proximity to each other, due to their aggregation, whereas isolated mark progressively evaporates and disappear. Eq. 5 describes the trail at time instant i.

$$T_i = (T_{i-1} - \delta) + Mark_i \tag{5}$$

Since only consistent spatio-temporal dynamics in data generate a stable pheromone trail, the trail itself can be considered as a summarization of these dynamics [21]. By matching trails, we provide a general similarity measure for spatiotemporal trajectories. The similarity between trails is obtained by using the Jaccard similarity [30] [22], i.e. the ratio between the volume of the intersection and the union of the stigmergic trails (Fig.2).
The similarity of the spatiotemporal trajectories of refugees T_R and locals T_L (Eq. 6) is defined between 0 (completely different trajectories) and 1 (identical trajectories).

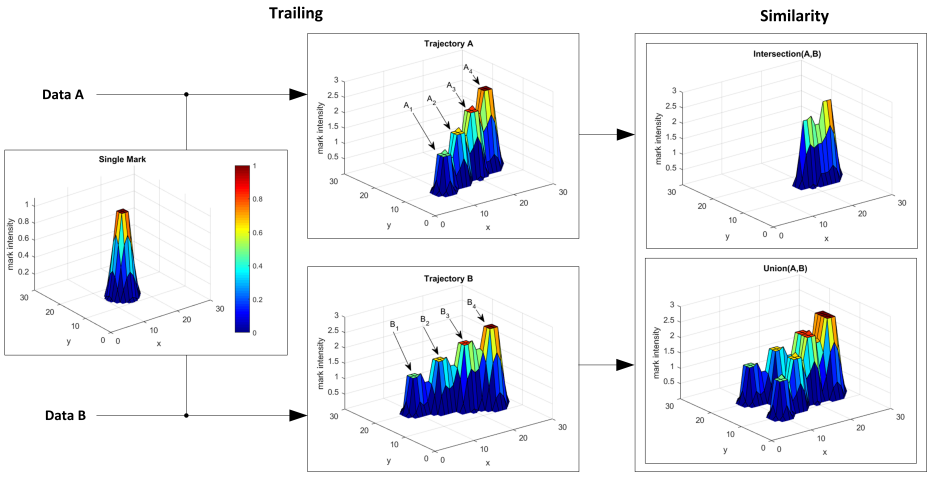


Fig. 2. Phases of the mobility similarity computation. We represent the trails obtained from the deposit of 4 consecutive samples (A_1, A_2, A_3, A_4 and B_1, B_2, B_3, B_4) of the trajectories (A and B), their intersection and their union, which are used to compute their similarity

$$MS_{R,L} = \frac{|T_R \cap T_L|}{|T_R \cup T_L|} \tag{6}$$

3 Experimental Setup

Since our investigation includes an analysis of mobility, call behavior, and district characterization it is necessary to focus our research in areas that ensure (i) an high calling activity made by refugees. Indeed, in order to have representative behavioral models we have to avoid areas characterized by sparse data; and (ii) a good spatial resolution, which means an high density of antennas, since the granularity of the trajectories will be determined by this; in fact, with few antennas in the area under investigation, all trajectories will be roughly similar; and (iii) high number and diversification of districts per area; indeed, the district-based metrics can explain the settlement choice of each refugee. This effect is especially noticeable in the presence of many different districts close to each other since this allows refugees to move from one district to another according to their socio-economic integration level and its change in time. Therefore, our first survey aim at finding the areas with these characteristics. Thus, we analyze the density of antennas (Fig. 3) and the total amount of calls (in seconds) made by refugees (Fig. 4) with a spatial discretization of 10 km per squared areas over the whole Turkey by exploiting the ATD dataset.

As shown by our results, the cities of Istanbul, Ankara and Izmir are the most promising areas to conduct our analysis since they have the larger density

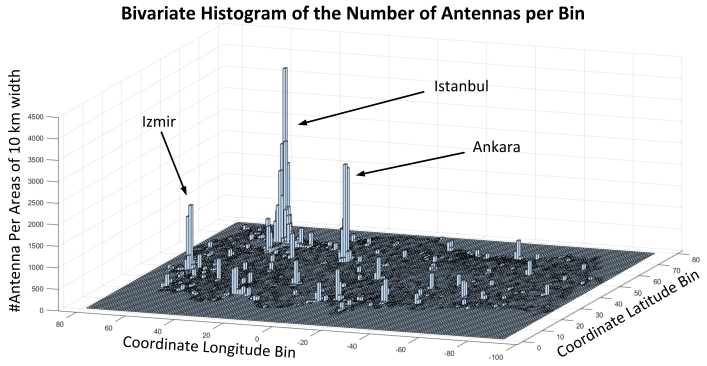


Fig. 3. Number of Antennas per squared area of 10x10 km.

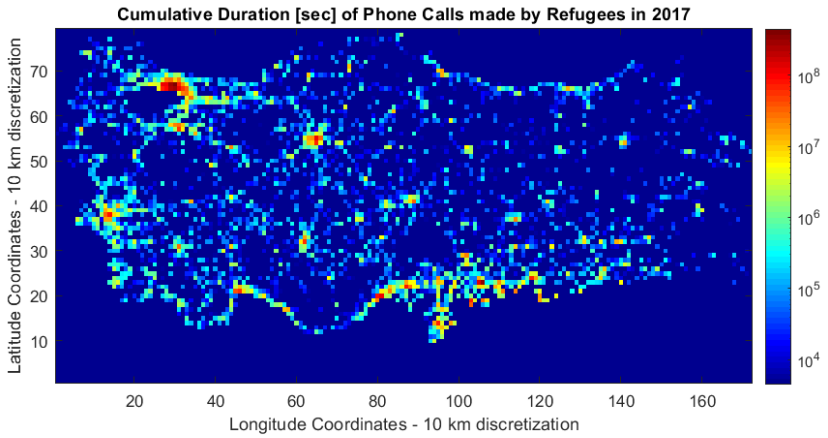


Fig. 4. Total amount of calls (duration) per squared area of 10x10 km. The metropolitan areas of Istanbul, Izmir and Ankara are the areas with the largest amount of calls.

of antennas and the larger calling activity made by refugees. This result is also comforted by other external data sources [10] [23]. Due to these reasons, those 3 cities are ideal areas to analyze both mobility and interaction with the locals. In addition, Istanbul's metropolitan area alone consists of 69 districts [10] with a variety of different characteristics (e.g., different housing costs or job opportunities). For this reason, our analyses on districts will be focused on Istanbul.

4 Results and Discussion

In this section, we describe the process and discuss the findings of our analysis.

4.1 Calling regularity and Interaction with locals

In order to verify if the Calling Regularity can be actually used as an integration proxy, we analyze the relationship between the Interaction Level and the calling regularity of each refugee in Istanbul. In order to have a reliable model of the calling pattern, we select the refugees with an average amount of calls per day equal or greater to 2. We compute the Pearson correlation coefficient between the average of the Calling Regularity of the refugees and their Interaction Level. We repeat this procedure for each period in the FGMD dataset, focusing our analysis on Istanbul.

With an average correlation of 0.795, Interaction Level and the Calling Regularity are strongly and positively correlated, providing us with the insight that refugees that exhibit greater interaction with locals have also daily routines which are similar to them. This result comforts the findings of other studies in the field [31] [32] and allows us to promote this measure as a metric for integration that should be taken into consideration by policy makers, at least in the case of Syrian refugees in Turkey.

4.2 Districts, Inclusion and Calling Regularity

Given the possibility of using Calling Regularity as an integration metric, we try to use it to obtain more insights at the district level. Specifically, we analyze the relationship between the District Attractiveness, the Residential Inclusion and the Calling Regularity of the refugees in each district of Istanbul according to the cost of living in the district itself. As an indicator of the cost of living per district, we consider the average rent cost per square meter in each district during 2016 (the data owner is an online housing website that would like to stay anonymous).

Firstly, we assess the impact of the presence of refugees on the attractiveness of a district. In order to do so, we compute the correlation between each district's yearly (i.e. averaged over 2017) Residential Inclusion (RI) and the District Attractiveness (DA). Figure 5 shows the correlation matrix obtained with the yearly RI and DA per district.

With a correlation coefficient equal to 0.494 and a p-value of 0.0016 we can consider RI and DA significantly and positively correlated. This means that **refugees are more likely to move and stay in districts with a greater the number of refugees**. Pushing the investigation on a more fine level (i.e. Monthly-wise) we focus on the relation between the RI of a given district and month and the average CR of the refugees living in that district during that month. In order to work with representative calling patterns the CR is computed with refugees having at least 100 calls in the CGMD dataset and settled in Istanbul for at least half of the whole year. We compute the correlation coefficients between the RI of each district and month and the average CR of the refugees living in that district during that month.

With an average correlation of 0.548, the RI in a given month and district appear to be positively correlated with the CR of the refugees living in that district

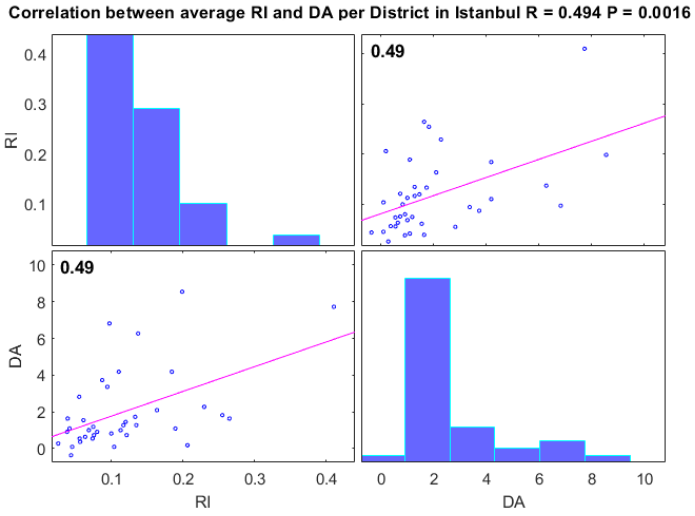


Fig. 5. Correlation matrix obtained with the yearly RI and DA per district. On the diagonal the distribution of the average RI and DA respectively, whereas the others are the bivariate scatter plots with a fitted line.

during that month. In other words, the districts with the highest Residential Inclusion of refugees are also the districts where the routine of the refugees is more similar to the locals. This might suggests that a minimum number of refugees per area is required for the dynamics of integration to be triggered, as suggested in [33]. To understand the order of magnitude of the amounts we are talking about we show the distribution of RI by month and district (Figure 6).

In Figure 6 it is evident that many districts have a low RI, thus depicting a scenario of minor coexistence of refugees and locals in most of the districts. Moreover, in the few districts (and months) with higher RI, the RI value never exceeds 50%. Thus, the more evenly distributed the residents (locals and refugees in an area) are, the greater the similarity between the routines of locals and refugees.

Finally, we include the cost of living in a certain district in the analysis. In order to study the relationship between DA and the cost of living, we computed the correlation coefficient between the average cost of living and the DA of each district. With a coefficient equal to -0.23, we can conclude that DA and the cost of living are loosely and inversely correlated (i.e. the cheaper is the cost of living the more attractive is a district).

The results in the last section, show that the Calling Regularity can be considered a proxy for social integration. However, as already specified in Section 2, the Calling Regularity may be even linked to the employment of a refugee. Unfortunately, it is not possible to verify directly this implication due to the lack of details about refugee employment, since they are often employed in the informal sector [16] [17]. Yet, it is possible to study this implication according

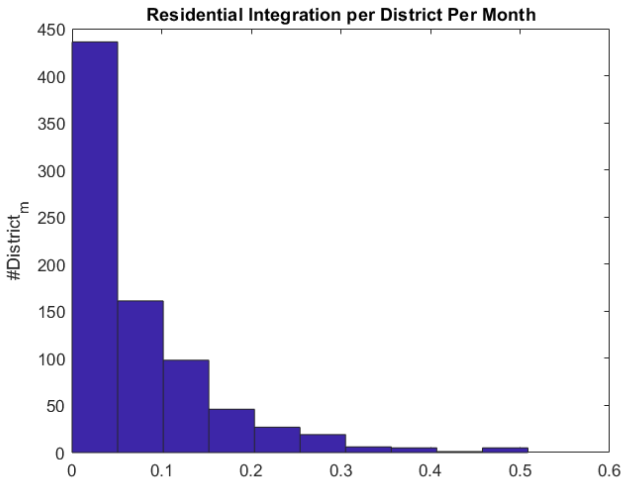


Fig. 6. Histogram of the distribution of RI per each district in Istanbul and each month.

to the social and economic characteristics of the location where refugees live. Indeed, depending on their economic well-being and the level of integration, migrants and refugees may choose different settlement solutions [19] [18]. For example, only an individual who has enough economical resources (e.g. who has some kind of job) can afford to live in an area that offers better opportunities. On the other hand, those who are not integrated and/or not working often find themselves socially isolated from the locals and relegated to poor neighborhoods.

In order to provide additional insights into this, we exploit the average rent cost per square meter in a given district in the year 2016 as an indicator of the cost of living for that specific district. Specifically, we compute the correlation between the cost of living in a given district and the average CR of refugees living in that district (Fig. 7).

With a correlation coefficient equal to 0.5 and a p-value equal to 0.003, the average CR of refugees living in the district exhibits a significant and positive correlation with the cost of living in that district. This means that, although it may be influenced by some factors not detectable by the data under analysis, the CR is a proxy for the daily routine similarity and for the economic capacity of refugees (i.e. the ability to meet a certain cost of living), thus it is a tool able to capture both necessary conditions occurring with the employment of refugees [39]. For this reason, this metric should be taken into account when dealing with the problem of integration of refugees because having a job is one of the first promoter of refugees' integration [34], but is also hard to analyze it since the refugees' employment often happen in the informal sector [17].

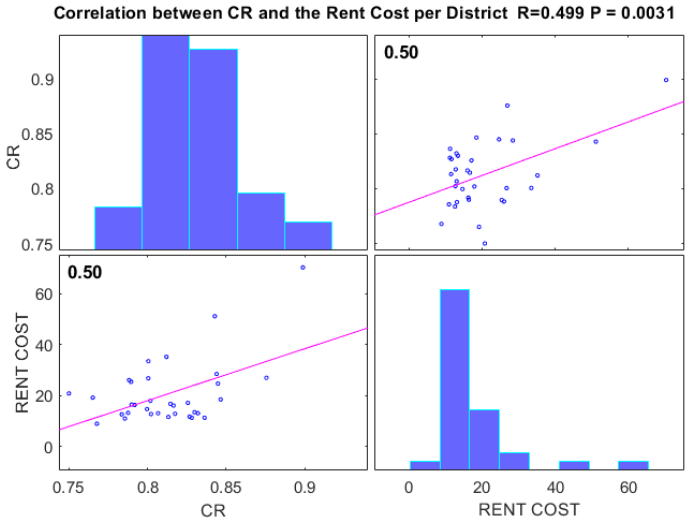


Fig. 7. Correlation matrix obtained with the average CR in a district and the cost of living per district. On the diagonal the distribution of the average CR and cost of living respectively, whereas the others are the bivariate scatter plots obtained with these variables together with the fitted line.

4.3 Mobility and Interaction with locals

Another fundamental driver of integration can be the sharing of urban spaces with the locals [36]. However, its positive contribution in the integration dynamics it is not obvious. Indeed, it can allow the progressive integration in the social structure of the hosting city. However, on the other hand the shared urban areas may not be easily defined and perceived as a safe space [35] thus leading to the occurrence of social friction in those areas.

In order to understand the contribution of sharing the same urban space with the locals, we analyze the relationship between the Mobility Similarity and the Interaction Level on a daily bases. Specifically, we create the cumulative trajectories of the group of refugees with a given Interaction Level, i.e. the stigmergic trails obtained with all the samples of the people in that group. Then, we compute the Mobility Similarity with the cumulative trajectories obtained with an equally sized group of locals. Regarding the size of these groups, it is worth highlight that the Mobility Similarity measure is sensitive to the number of users employed in the creation of the cumulative trajectories, i.e. the more the users the higher the likely to have more similar cumulative trajectories. In addition to this, the size of the groups with a given Interaction Level varies significantly according to it (Fig. 8).

Thus, in order to have a fair comparison between the similarities computed with different groups, we set the size of each group as the minimum size among all the groups. Finally, we collect the Pearson correlation coefficients between the

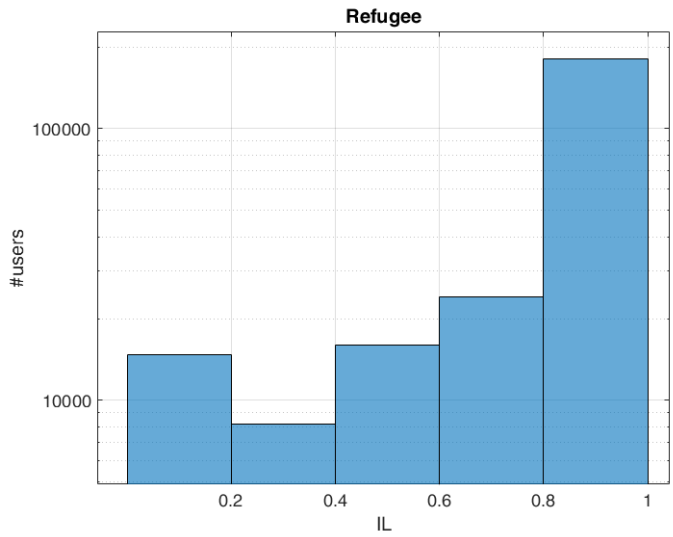


Fig. 8. Number of refugees in Istanbul according to their Interaction Level (FGMD). The smallest group is made up of 8184 people. The figure is in Log scale.

Interaction Level of each group and the resulting Mobility Similarity. We repeat this procedure multiple times by randomly subsampling the people for each group larger than the smallest one. In Fig. 9 we present the distribution of the obtained correlation coefficients by means of boxplots. It is evident that in the 3 cities analyzed the Mobility Similarity is strongly correlated with the interaction level. Indeed, the 95% confidence interval of the correlation coefficients results as 0.91 ± 0.01 in Istambul, 0.83 ± 0.06 in Ankara, and 0.92 ± 0.04 in Izmir. On the basis of the obtained results, it is possible to claim that the more the refugees have interactions with locals, the more they share urban spaces with the locals. This allows us to say that sharing of urban spaces is a positive factor in the dynamics of integration of refugees. Thus, the policies designed to improve refugees' integration should take into account Mobility Similarity to assess their impact.

4.4 Integration and Social Friction

Since we have seen how Mobility Similarity and Interaction Level are able to capture the integration of refugees, we now attempt to use them to study the effects of the events that are certainly caused or can cause the disruption of refugees' integration: the occurrence of social frictions. In order to look for the features that characterize a social friction, it is necessary to start with few examples of publicly known social frictions. Specifically, we collect a set of such events and we compare the Mobility Similarity and Interaction Level in 2 weeks before and after each event. We have found a number of occurrence of such events by

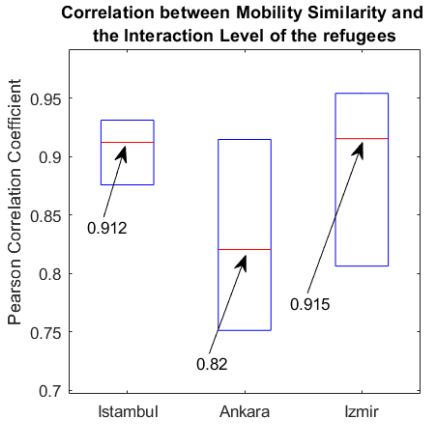


Fig. 9. Boxplot of the correlation coefficients between the Mobility Similarity and the Interaction Level, over multiple trials with subsampling. The cases of Istanbul, Ankara and Izmir.

searching for them over the internet [43] [44] and exploiting a publicly available news collector, i.e. the GDELT Project [40]. The GDELT Project monitors the world's broadcast, print, and web news from all over the world and makes it possible to query them according to locations, subjects involved, and emotions. By querying for events involving refugees in Turkey, we were able to obtain a pool of potential events that we checked manually to select only the ones related to actual social frictions and police interventions. The final pool of events taken into consideration is displayed in Table 1.

Date	**Location**	**Source**
March 6	Izmir	[41]
April 12	Istanbul	[42]
May 15	Istanbul	[43]
May 16	Istanbul	[44]

Table 1. Dates and locations of the social friction events taken into account.

Once these events have been identified, we study the impact of these social friction by calculating the Mobility Similarity (with repeated trials according to the methodology described in the last section) and the percentage of calls made toward the locals, according to the Interaction Level of the refugees. These measures will be derived with data from different periods (FGMD dataset). In order to make them comparable and highlight the fluctuations with respect their average across different periods, a normalization with (i.e. divided by) their average per period is performed. Finally, we present the ratio between MS and

the percentage of calls in the two weeks before and after each event. If this ratio is greater than 1, it indicates that after the event, the integration measure taken into consideration has decreased. As an example, in the Figure 10 we show the results obtained with the event of May 16 in Istanbul.

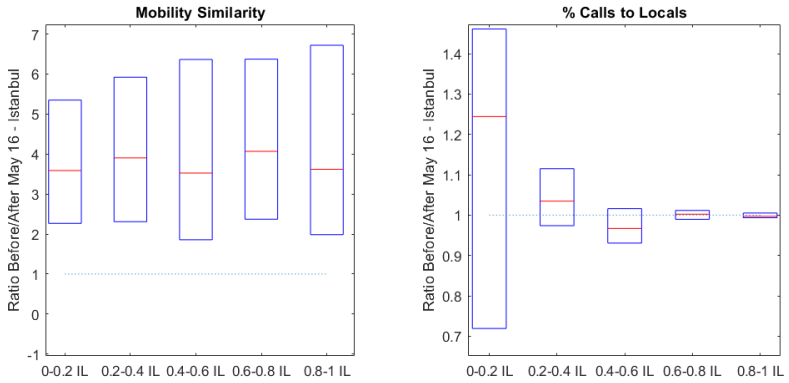


Fig. 10. Mobility Similarity (left) and the percentage of calls made toward refugees (right): ratio between the values two weeks before and two weeks after the 16th of May in Istanbul. A ratio greater than 1 indicates that, after the event, the integration measure taken into consideration has decreased. The ratios are separated for different ILs of the group of refugees.

It is apparent that the social friction affects the behavior of the refugees by reducing the amount of shared urban space with the locals (i.e. lowering the Mobility Similarity after the event). Moreover, in terms of calls made toward locals, the social friction event affects the group of refugees with lower level of interaction with locals way more than the more integrated groups. Indeed, on average, they exhibit a lower percentage of calls made toward locals and a greater variability. Moreover, this trend is confirmed on every event we are taking into account, as shown in the the aggregate results in Figure 11. Indeed, the quartiles of the percentage of calls made toward locals are arranged as [0.55, 1.05, 1.41] with the refugees with the lower Interaction Level, whereas are [0.98, 0.99, 1] with the refugees with the greater Interaction Level. Here, even the MS results more affected in the group of refugees with lower Interaction Level, who tend to be more segregated after the social friction event. Indeed, the median of the distribution of the ratios obtained with the Mobility Similarity with the lower and greater Interaction Level are respectively 5.31 and 3, which means that the Mobility Similarity of the refugees with lower Interaction Level decreased 77% more with respect to the refugees with greater Interaction Level. Based on the obtained results, the proposed metrics result able to capture the effect of a social friction and should be taken into account when addressing application such as attempting to identify or measure the impact of social friction events.

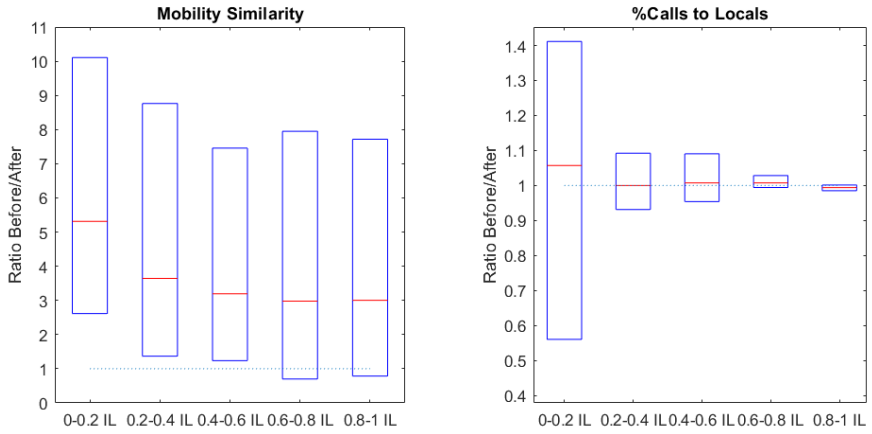


Fig. 11. Mobility Similarity (left) and the percentage of calls made toward refugees (right): ratio between the values two weeks before and two weeks after each social friction. A ratio greater than 1 indicates that, after the event, the integration measure taken into consideration has decreased. The ratios are separated for different ILs of the group of refugees.

5 Conclusion

In this work, we have proposed a set of metrics to assess the integration of Syrian refugees in Turkey. Each refugee-related measure takes into account their behavioral and spatio-temporal patterns with different approaches. Specifically, (i) the daily mobility of refugees is analyzed by means of stigmergic trails, a biologically-inspired computational method that allows to compare spatio-temporal patterns (i.e. the spatio-temporal trajectories of refugees and locals), and (ii) the calling regularity, i.e. the similarity between the time series of the frequency of the daily calls made by refugees and locals. Thanks to these metrics and to few more districts-wise descriptive metrics, we have provided a number of insights about the integration of refugees in the main Turkish cities. Specifically, (i) both Mobility Similarity and Calling Regularity are positively and significantly correlated with the level of interaction between refugees and locals, and have proved to offer great potential as measures of the integration related phenomenon with different applications; (ii) the integration is fostered by the simultaneous presence of refugees and locals who reside in the same area in a fair amounts; (iii) the Calling Regularity is also a proxy for refugee's economic capacity, which can imply refugee's employment, and (iv) both Mobility Similarity and the amount of calls made toward the locals are affected by events such as social friction involving refugees; however, the behavior of less integrated refugees appears to be significantly more affected by this kind of events. Given the promising results obtained with these metrics, their application should be further explored on different scenarios. For example, by retrieving more data about other cities we can

gain more insights and employ a different spatial resolution for the geospatial analysis. Future work will focus on consolidating such experiments.

References

1. Keyman, Fuat, (2016), Turkey at the Heart of the Refugee and ISIL Crises: Can the Buffer State be a Solution?, International Law & Politics, Vol. 12, No. 1, pp: 5-12.
2. Ayla Albayrak, Tensions Rise Between Syrian Refugees, Turks, The Wall Street Journal, May 12, 2013.
3. Ager, A., Strang, A., O'may, F., Garner, P. (2002). Indicators of Integration: A Conceptual Analysis. Report to the Home Office Immigration Research and Statistics Service.
4. Carpi, E., Pnar enouz, H. (2018). Refugee Hospitality in Lebanon and Turkey. On Making The Other. International Migration.
5. A. M. Tompkins and N. McCreesh, Migration statistics relevant for malaria transmission in senegal derived from mobile phone data and used in an agent-based migration model, Geospatial health, vol. 11, no. 1s, 2016.
6. D. Gundogdu, O. D. Incel, A. A. Salah, and B. Lepri, Countrywide arrhythmia: emergency event detection using mobile phone data, EPJ Data Science, vol. 5, no. 1, p. 25, 2016.
7. S. Silm and R. Ahas, Ethnic differences in activity spaces: A study of out-of-home nonemployment activities with mobile phone data, Annals of the Association of American Geographers, vol. 104, no. 3, pp. 542559, 2014.
8. Altindag, Onur; Kaushal, Neeraj (2017) : Do Refugees Impact Voting Behavior in the Host Country? Evidence from Syrian Refugee Inflows in Turkey, IZA Discussion Papers, No. 10849, Institute of Labor Economics (IZA), Bonn
9. Hardy, K., Maurushat, A. (2017). Opening up government data for Big Data analysis and public benefit. Computer law security review, 33(1), 30-37.
10. Salah, A.A., Pentland, A., Lepri, B., Letouz, E., Vinck, P., de Montjoye, Y.A., Dong, X. and Dadelen, ., 2018. Data for Refugees: The D4R Challenge on Mobility of Syrian Refugees in Turkey. arXiv preprint arXiv:1807.00523.
11. Blumenstock, J., Fratamico, L. (2013, December). Social and spatial ethnic segregation: a framework for analyzing segregation with large-scale spatial network data. In Proceedings of the 4th Annual Symposium on Computing for Development (p. 11). ACM.
12. Licoppe, C., Diminescu, D., Smoreda, Z., Ziemlicki, C. (2008). Using mobile phone geolocalisation for sociogeographicalanalysis of coordination, urban mobilities, and social integration patterns. Tijdschrift voor economische en sociale geografie, 99(5), 584-601.
13. Hebbani, A., Colic-Peisker, V., Mackinnon, M. (2017). Know thy Neighbour: Residential Integration and Social Bridging among Refugee Settlers in Greater Brisbane. Journal of Refugee Studies.
14. Gundogdu, D., Incel, O. D., Salah, A. A., Lepri, B. (2016). Countrywide arrhythmia: emergency event detection using mobile phone data. EPJ Data Science, 5(1), 25.
15. Almaatouq, A., Prieto-Castrillo, F., Pentland, A. (2016, November). Mobile Communication Signatures of Unemployment. In International Conference on Social Informatics (pp. 407-418). Springer International Publishing.

16. Del Carpio, X. V., Wagner, M. C. (2015). The impact of Syrian refugees on the Turkish labor market.
17. Balkan, B., Tumen, S. (2016). Immigration and prices: quasi-experimental evidence from Syrian refugees in Turkey. Journal of Population Economics, 29(3), 657-686.
18. Fawaz, M. (2017). Planning and the refugee crisis: Informality as a framework of analysis and reflection. Planning Theory, 16(1), 99-115.
19. United Nations High Commissioner for Refugees (UNHCR) and UN-Habitat (2014) Housing, Land and Property Issues in Lebanon: Implications of the Syrian Refugee Crisis.
20. A.L. Alfeo, M.G.C.A. Cimino, A. Lazzeri, G. Vaglini, "Detecting urban road congestion via parametric adaptation of position-based stigmergy", Intelligent Decision Technologies, IOS Press, Vol. -, Issue -, Pages 1-28, 2017.
21. A.L. Alfeo, P. Barsocchi, M.G.C.A. Cimino, D. La Rosa, F. Palumbo, G. Vaglini, "Sleep behavior assessment via smartwatch and stigmergic receptive fields", Personal and Ubiquitous Computing, Springer, Vol. -, Issue -, Pages 1-17, 2017.
22. A.L. Alfeo, M.G.C.A. Cimino, S. Egidi, B. Lepri, A. Pentland, G. Vaglini, "Stigmergy-Based Modeling to Discover Urban Activity Patterns from Positioning Data" in Proc. SBP-BRiMS The International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation (SBP-BRiMS 2017), pp. 292-301, Washington, DC, USA, 5-8 July, 2017.
23. UNICEF Turkey Humanitarian Situation Report 7 MARCH 2017. Available at: `https://www.unicef.org/appeals/files/UNICEF_Turkey_Humanitarian_Situation_Report_March_2017.pdf`
24. Lifanova, A., Ngan, H. Y., Okunewitsch, A., Rahman, S., Guzmn, S., Desai, N., Yildirim, M. (2016). New Locals: Overcoming Integration Barriers with Mobile Informal and Gamified Learning. In Proceedings of the International Conference on Information Communication Technologies in Education (pp. 132-41).
25. Alexander, L., Jiang, S., Murga, M., Gonzlez, M. C. (2015). Origindestination trips by purpose and time of day inferred from mobile phone data. Transportation research part c: emerging technologies, 58, 240-250.
26. Dong, Y., Pinelli, F., Gkoufas, Y., Nabi, Z., Calabrese, F., Chawla, N. V. (2015, September). Inferring unusual crowd events from mobile phone call detail records. In Joint European conference on machine learning and knowledge discovery in databases (pp. 474-492). Springer, Cham.
27. Marsh, L., Onof, C.: Stigmergic epistemology, stigmergic cognition. Cognitive Systems Research, 9(1-2), 136–149, (2008).
28. Vernon, D., Metta, G., Sandini, G.: A Survey of Artificial Cognitive Systems: Implications for the Autonomous Development of Mental Capabilities in Computational Agents. IEEE Transactions on Evolutionary Computation, 11(2), 151–180, (2007).
29. Singh, V. K., Bozkaya, B., Pentland, A. (2015). Money walks: implicit mobility behavior and financial well-being. PloS one, 10(8), e0136628.
30. Niwattanakul S., Singthongchai J., Naenudorn E., Wanapu S.: Using of Jaccard coefficient for keywords similarity. In Proceedings of the International MultiConference of Engineers and Computer Scientists, 1, 6, (2013).
31. Clark, F. A. (2000). The concepts of habit and routine: A preliminary theoretical synthesis. The Occupational Therapy Journal of Research, 20(1suppl), 123S-137S.
32. Jansen, T., Chioncel, N., Dekkers, H. (2006). Social cohesion and integration: Learning active citizenship. British Journal of Sociology of Education, 27(02), 189-205.

33. McIsaac, E. (2003). Nation building through cities: A new deal for immigrant settlement in Canada. Caledon Institute of Social Policy.
34. Bakker, L., Cheung, S. Y., Phillimore, J. (2016). The asylumintegration paradox: Comparing asylum support systems and refugee integration in the Netherlands and the UK. International Migration, 54(4), 118-132.
35. Lyytinen, E. (2015). Refugees Conceptualizations of Protection Space: Geographical Scales of Urban Protection and HostRefugee Relations. Refugee Survey Quarterly, 34(2), 45-77.
36. Madanipour, A. (1998). Social exclusion and space. Social exclusion in European cities, 76.
37. Figueiredo, M. A. T., Jain, A. K. (2002). Unsupervised learning of finite mixture models. IEEE Transactions on pattern analysis and machine intelligence, 24(3), 381-396.
38. Hubl, F., Cvetojevic, S., Hochmair, H., Paulus, G. (2017). Analyzing refugee migration patterns using geo-tagged tweets. ISPRS International Journal of Geo-Information, 6(10), 302.
39. Llorente, A., Garcia-Herranz, M., Cebrian, M., Moro, E. (2015). Social media fingerprints of unemployment. PloS one, 10(5), e0128692
40. Available at: https://www.gdeltproject.org/
41. 130 migrants intercepted in Aegean heading for Chios. Available at: http://www.ansamed.info/ansamed/en/news/nations/greece/2017/03/06/130-migrants-intercepted-in-aegean-heading-for-chios_b60eec09-29a6-42a6-b846-3e747f71f000.html
42. Kurdish militants claim responsibility for Turkey tunnel attack. Available at: https://www.reuters.com/article/us-turkey-blast/kurdish-militants-claim-responsibility-for-turkey-tunnel-attack-idUSKBN17E0MW
43. One killed in brawl between locals and Afghan, Syrian migrants in Istanbul. Available at: http://www.hurriyetdailynews.com/one-killed-in-brawl-between-locals-and-afghan-syrian-migrants-in-istanbul-113111
44. Istanbul police evacuate refugees after clashes with locals. Available at: https://www.trtworld.com/turkey/istanbul-police-evacuate-300-refugees-after-clashes-with-locals-358027

Characterizing the Mobile Phone Use Patterns of Refugee Hosting Provinces in Turkey

Erika Frydenlund[1], Meltem Y. Şener [2], Ross Gore[1], Christine Boshuijzen-van Burken[3], Engin Bozdag[4], Christa de Kock[5]

[1] Old Dominion University, USA
[2] Istanbul Bilgi University, Turkey
[3] Eindhoven University of Technology, The Netherlands
[4] 4TU Centre for Ethics and Technology, The Netherlands
[5] Stellenbosch University, South Africa
efrydenl@odu.edu

Abstract. We use coarse-grained mobile phone data from a large Turkish mobile phone provider and cross-reference this data with social media data and a qualitatively composed violent events list to explore the integration of refugees in Turkey. The data provides grounds for fruitful future research. It suggests that border communities with the refugee sending country have much different communications patterns than non-border communities. Additionally, proximity to refugee camps may increase negative sentiment on social media towards refugees, which we suggest may be a proxy for understanding 'compassion fatigue.' These findings provide directions for future research on integration.

Keywords: Integration, Twitter Sentiment, Syrian Refugees.

1 Background

As a host country, Turkey has the largest number of refugees in the world, where the vast majority live outside of camps in urban and peri-urban environments [1]. Since most of the millions of refugees live in urban areas, the impact of forced migration on integration and neighborhood relations is critical for maintaining safety and order. From a data perspective, what do "good" neighborhood relations look like? We use mobile phone data provided by a large mobile phone company in Turkey, in combination with Twitter sentiment analysis, and a set of violent events that occurred in Turkey in 2017 to explore the relationship between refugee and non-refugee communication, negative sentiments expressed on social media, and geography (proximity to refugee camps and the Syrian border). In particular, we focused on violent events hypothesizing that an increase in communication by both refugees and non-refugees after an attack is a proxy for both groups expressing concern over the incident; in other words, an increase in communications by both groups could signal some form of integration. However, an increase in communication by the host population alone prior to an event could signal lack of integration, as citizens organized prior to the attack. The data reveal patterns

about cultural use of mobile phones in after violent events, as well as link geography and negative social media sentiment.

From early on in the process of hosting so many Syrian refugees, Turkey has made great strides in attempting to provide the necessary social support to promote integration, or at least self-sufficiency, as integration is generally a long-term process. Turkey, as a host government, supported an education strategy that allowed large numbers of Syrians the chance to continue their education [2,3]. These strategies were not without their challenges [4], but given the exceptional number of Syrians entering Turkey, any avenues for education can provide a means for integrating newcomers linguistically and culturally.

Host communities can often experience some difficulties in the wake of a large influx of a new population. As far back as 2013 and 2014, border provinces such as Hatay reportedly experienced increased tensions as ethnic balances shifted in the wake of the Syrians fleeing into Turkey [5]. Even in this early phase of migration, historic cross border relations between Turkey's border communities and Syria that should make integration easier became strained under the pressures of long-term refugee hosting of such vast numbers of refugees [6, 7]. This has included the demands placed on the economy such as lower wages, competition over jobs, and higher housing prices [8]. These economic tensions affect both host and newcomer populations, giving rise to negative and xenophobic sentiments that can affect long-term integration strategies [9].

Though Turkey has been generous to the Syrian refugees, it has also been receiving a large number of individual asylum seekers from elsewhere, such as Iran, Iraq, Somalia, and Afghanistan. Syrians are given special protection as guests under Turkish law, but the same does not apply to others seeking asylum in or transit through Turkey who must each apply on a case by case basis [10]. It should be noted that in our data, we cannot distinguish between countries of origin. We assume, based on the number of refugees from Syria versus other countries who are given permission and access to rights in Turkey that the majority of 'refugees' in our dataset are of Syrian origin. Still, Syrians suffer from uncertainty of their status and future in Turkey, and though many citizens still broadly feel compassion for the humanitarian case of receiving refugees, the lived reality is putting pressure on many cultural differences and infrastructure challenges [11]. Many of these challenges are ones that Turkey cannot face alone as it attempts to accommodate changing residency and citizenship demands as well as the extreme burden placed on infrastructure and social institutions [12].

A report from the International Crisis Group (ICG) in 2016 that was built upon qualitative interviews collected along the Turkish border with Syria suggested that the length of stay experienced by Syrians has inhibited integration. The report indicates that part of the struggle with integration is language, which many Syrians did not attempt to overcome as they thought their stay in Turkey would be short. Many Syrians they interviewed did not have Turkish friends and worried about increasing segregation by job type, neighborhood of settlement, and language and culture differences [13].

Given the sensitivity of mobile phone data and the ability to uniquely identify individuals, the D4R Challenge took great strides to ensure anonymity and safe use of the data. As a consequence, it is very difficult to generalize about individuals' behaviors

or integration strategies. Thus, we must cross-reference the mobile phone data with other sources of data in order to draw some conclusions. We found that it was difficult to draw any conclusions about integration per se, so we concentrated our efforts on what could be proxy measures of integration or dis-integration. To accomplish this, we focused on violent attacks—where the refugees and non-refugees' responses are a proxy for (dis)integration—as they have enormous negative consequences in terms of refugee integration. Violent events involving refugees demonstrate to the newcomers that they are not welcome and do not belong to the host society. To be able to formulate and implement any long-term integration program, the host society should first give refugees the message that they are welcome in their new country, including providing mechanisms for decreasing tensions between the host society and refugees. This research uses big data to understand whether mobile phone data can tell us something about how differently refugees and non-refugees respond via SMS and calls after a violent event that involved Syrian refugees, the largest refugee community hosted in Turkey. The premise of this study is that we may be able to understand how mobile phone usage correlates with other quantitative and qualitative data that characterizes certain locations in Turkey based on integration levels.

2 Methodology

The study relies upon triangulation of mobile phone data, social media data from Twitter, and locations of refugee camps in reference to qualitatively relevant violent events that occurred in Turkey in 2017. The following section provides a summary of the study approach.

2.1 Ethical Considerations and Privacy

Given the sensitive nature of mobile phone data, Turk Telecom highly anonymized the data and requested signed statements from all researchers to comply with the privacy policy before the research could be started. Privacy concerns were extremely important to our research team; therefore, in addition to the measures put in place by Turk Telecom, our research team set up a separate directive of principles by which to abide during the course of the study. The dataset provided by Turk Telecom (TTG) is anonymized and as a result it does not relate to an identified or identifiable natural person and the individuals are no longer identifiable. Per the data challenge, the only purpose of data processing under this project is scientific research which shall be proportionate to the aim pursued, respect the essence of the right to data protection, and provide for suitable and specific measures to safeguard the fundamental rights and the interests of the data subject. The team protocol put in place included sharing data internally only on a need-to-know basis. We accomplished this by ensuring that the data scientists could access the data and relied on them to provide aggregate measures and visualizations to advance the research agenda. The data was never shared with any third parties, was securely stored, and we used the data only for the intended purposes

of the challenge. All researchers agreed to these principles before we embarked on the study.

While big data presents an opportunity to see social phenomena in new ways, we are aware of the limitations of doing research with big data for crisis situations. Some researchers warn that this type of analysis can lead to surveillance and control, particularly among displaced populations [14, 15], and we took this challenge to heart. In addition to surveillance, big data presents a number of scientific challenges including 'noisiness' of data and selection bias (particularly of displaced and other marginalized populations) of who can own a mobile device or who participates on social media [15]. We cannot control what others do with the data, but we chose to pursue hypotheses that we felt best reflected the humanitarian spirit of the data challenge. The research presented here was guided by the ideal that big data may be able to reveal policies that are effective for integration or future avenues of research that assist refugees and host communities in coping with cohabitation.

2.2 Phone Data

We exclusively used the coarse-grained mobility dataset which tracked anonymous individuals continuously throughout the course of 2017. This dataset represented 100,000 users (half refugee, half non-refugee identified based on their registration card needed to subscribe to a mobile phone line). This dataset included the province name, but not the exact location within the province, of the calls and texts that were recorded [16].

2.3 Social Media Data

An increasing number of high-stakes decisions are now made based upon predictive models of publicly available sentiment data. The most common source for such predictions is currently Twitter due to its public nature [17], large userbase [18], and high accessibility by both academic and industry researchers [19]. Tweets, which are short text-based messages created by users of Twitter and posted to their online profiles, cover a wide variety of topics depending upon user preferences, and this diversity allows for the investigation of human behavioral patterns across many disciplines [20].

Sentiment is generally viewed as a desirable predictor in this type of modeling because of its simplicity of computation, its applicability regardless of text type, and its ability to reduce any length of text into a single numerical summary value. Specifically, one common challenge in the statistical modeling of text is a "p>N" scenario in which the number of words/variables exceeds the sample size. Various types of dimension reduction, which restructure large groups of variables into their hypothetically latent causes, are used to increase statistical power to detect meaningful relationships [21]. Sentiment analysis is arguably the most popular type of dimension reduction; it can be used to reduce any quantity of words to a single value representing a continuum ranging from highly positive to highly negative tone [22].

Given the sensitivity of linking Twitter data to the phone data, we took extra precautions to store the data and report our findings. Access to the raw data was strictly

limited to our team during the project period. The raw data set where the Twitter data and phone data are connected will be deleted after the project, leaving only the aggregate data, analyses, and measures that inform our findings and report. Our collection used Twitter's streaming API which provides low latency access to Twitter's global stream of Tweet data. The procedure adhered to Twitter's terms of use/service.

We collected geo-tagged tweets with locations inside Turkey for the year 2017 that referenced the following hashtags or phrases to cross-reference with our mobile phone data:

- #ÜlkemdeSuriyeliİstemiyorum
- #ulkemdesuriyeliistemiyorum
- #suriyelilersınırdışıedilsin
- #suriyelilerseçmendeğildir
- #SuriyelileriGeriGonderin
- "Suriyeli istemiyorum diye ırkçıysam ırkçının kralıyım bundan da gurur duyarım"
- #IDon'tWantSyriansInMyCountry
- #IDontWantSyriansInMyCountry
- #DeportSyrians

Note that these are all negative hashtags and phrases. Positive hashtags and phrases were more difficult to locate and are the subject of current ongoing investigation.

There are issues that must be addressed with how well a geo-tagged Twitter data set can represent the sentiment of a population. Only 15% of online adults regularly use Twitter, and 18–29 year-olds and minorities tend to be more highly represented on Twitter than in the general population. Furthermore, on Twitter, 95% of users never geo-tag a single tweet and only ~ 1% of users geo-tag the majority of the tweets they post. Also, the extent to which the individual 'tweeter' is represented in our Twitter corpus is biased. Very passive users (< 50 tweets per year) and very active users (> 1000 tweets per year) geo-tag a smaller percentage of tweets than moderate users (50–1000 tweets per year) [23]. Ultimately, these limitations mean that the Twitter data set which informed our study is a non-uniform subsample of statements made by a non-representative portion of the Turkish population.

2.4 Location Data

In order to get a sense of how close people were living to formal refugee settlements, as opposed to urban integration, we identified the location of refugee camps using data from the Humanitarian Data Exchange by UNHCR.[1] A map of these locations is shown in **Fig. 1**.

[1] https://data.humdata.org/dataset/turkey-refugee-camps

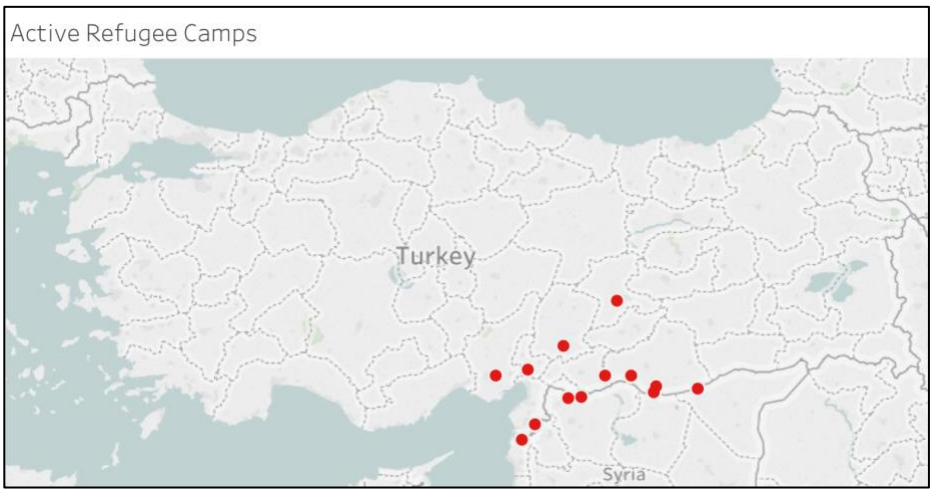


Fig. 1. Active Refugee Camp Locations in Turkey

2.5 Event Data

We compiled a list of 16 violent events related to refugees in Turkey during the study year of 2017. These events were gathered from various international and Turkish news sources and qualitatively determined to be relevant to the study of refugee integration in Turkey. For each event, we identified the city in which the violent event towards refugees took place. These events are located on the map (**Fig. 2**) with the descriptions in the table below (**Table 1**).

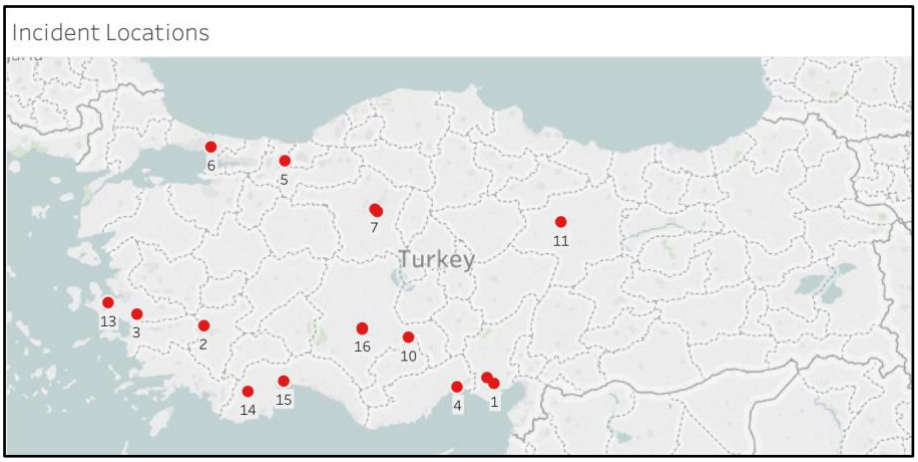


Fig. 2. Violent Incident Locations in 2017

Table 1. Violent Incidents by Date with Description

ID	Date	Description
1	27 February	Some local residents drove to an area where Syrians were living and lit their tents on fire and fired rifles in Adana Yuregir.[2]
2	18 March	After a fight between a group of Syrian teenagers and residents, 500 people came together and threw stones at the houses of Syrians in Denizli Saraykoy.[3]
3	5-6 April	Because of a rumor that a Syrian person hit a child, 300 people got organized and went into the neighborhood where Syrians live in Izmir Torbali District Pamukyazi. 500 Syrians had to leave the neighborhood.[4]
4	15 May	A group attacked the houses and workplaces of Syrian families in Mersin Akdeniz District Sevket Sumer Neighborhood.[5]
5	16 June	A fight occurred between Syrian workers and Turkish residents in Sakarya Hendek.[6]

[2] http://www.milliyet.com.tr/adana-da-suriyeliler-ile-mahalle-sakinleri-adana-yerelhaber-1870768/

[3] http://www.cumhuriyet.com.tr/haber/turkiye/702265/Saraykoy_de_gergin_gece__Suriyeliler_3_kisiyi_dovdu.html

[4] https://www.yeniakit.com.tr/haber/izmir-valiliginden-suriyeli-aciklamasi-299416.html

[5] http://www.hurriyet.com.tr/gundem/sosyal-medyada-orgutlendiler-dun-gece-saldirdilar-40459388

[6] https://www.cnnturk.com/turkiye/sakaryada-suriyeli-gerginligi-buyuyor-yaralilar-var?page=1

ID	Date	Description
6	1 July	After a fight between Syrian and local teenagers, a group of locals got together in the evening and raided a Syrian family's house in Istanbul Uskudar Yavuzturk Neighborhood.[7]
7	3 July	Attempt to lynch Syrians in Ankara Demetevler because of rumors on social media.[8]
8	13 July	A group of people who were wearing masks raided the workplaces of Syrians in Adana Seyhan Mirzacelebi.[9]
10[10]	20 August	After a fight between a group of Turkish residents and two Syrian men, a group of people who don't want Syrians in their neighborhood organized on social media and circled a building where Syrians were staying in Konya Karapinar.[11]
11	24 August	A mob convened in front of a house where Syrians were living in Sivas Istiklal Neighborhood. They wanted the Syrians to vacate the house.[12]
12	7 September	Two people on a motorcycle fired rifles on a street where Syrians are living in Konya Karatay Kerimdede Neighborhood.[13]
13	21 September	In Urfa tension arose between Syrians who were registered at the adult education center and the parents of the children at a primary school sharing the space. Later in the day parents came to the school garden to protest the Syrians.[14]
14	22 September	After a fight between a group of Syrian teenagers and the owner of a bakery, a group of 600 people attempted to attack Syrians in Antalya Elmali.[15]
15	8 October	During the street wedding of a Syrian couple, residents of that street attacked the Syrians in Antalya Kepez Mehmet Akif Neighborhood. One Syrian died, and another was wounded.[16]

[7] https://www.evrensel.net/haber/325186/uskudarda-suriyeli-ailenin-evine-saldiri

[8] http://www.hurriyet.com.tr/gundem/ankarada-suriyelilerle-vatandaslar-arasinda-gerginlik-40508566

[9] http://www.cumhuriyet.com.tr/haber/turkiye/94361/Satirli_ve_maskeli_grup_Suriyelilerin_isyerlerini_basti_.html

[10] Note: Incident number 9 was removed from the original dataset

[11] http://www.hurriyet.com.tr/karapinarda-taciz-kavgasi-1-suriyeli-oldu-1-40556700

[12] http://www.hurriyet.com.tr/sivasta-suriyeli-gerginligi-40559616

[13] https://www.evrensel.net/haber/331758/konyada-suriyelilere-ates-acildi-2si-suriyeli-4-yarali

[14] https://www.evrensel.net/haber/333065/urfada-bir-okulda-suriyeli-siginmacilarla-gerilim-yasandi

[15] https://m.sondakika.com/haber/haber-antalya-elmali-da-suriyeli-gerginligi-10057266/

[16] http://www.milliyet.com.tr/dugun-sonrasi-laf-atma-kavgasinda-1-antalya-yerelhaber-2324578/

ID	Date	Description
16	3 November	200 people who did not want Syrians in their neighborhood attacked the houses and workplaces of Syrians in Konya Karatay District Sems Tebrizi neighborhood.[17]

For each city in Turkey we tracked the average percent increase in the percentage of calls and texts from refugees and non-refugees following each violent incident (including the day of the event and two following days). Note a negative number reflects a percent decrease in calls and texts. The average refugee and non-refugee call/text number in the city is determined by looking at number of calls and texts sent by refugees / non-refugees in the city over the course of the year. We also computed the average for each city across all violent events.

3 Findings

Table 2 summarizes the findings of the data analysis. Note, all are statistically significant correlations (p=0.00) with weak+ to strong- effect sizes (0.25 – 0.80):

Table 2. Overall Data Analysis Findings

Finding	Effect Size
Calls and texts correlate closely with one another.	Strong
Changes in refugee communication and non-refugee communication within a city occurring before and after a specific violent event exhibit similar patterns that correlate with one another.	Strong
There is a lot of variation across all the cities in terms of refugee and non-refugee communication occurring at the time of a specific violent event.	N/A
Cities closer to where the event occur experience a larger increase in communication related to a violent event. This is true for refugees and non-refugees.	Moderate
Given a city and an event, there is more of an increase in refugee communication (calls and texts) than non-refugee communication. Statistically significant difference between the two populations.	N/A
Negative twitter sentiment correlates with increases in refugee and non-refugee communications.	Moderate
Negative twitter sentiment tracks correlates with proximity to violent events.	Moderate

Violent events tend to occur more often in places where negative sentiment is expressed on Twitter. This is only one avenue of social media, expression, but there appears to be some relationship. This is not surprising as many of the events in the

[17] http://www.hurriyet.com.tr/gundem/tehlikeli-gerginlik-200-kisiyle-saldirdilar-40633451

violent incidents we covered were mobilized via social media such as Facebook. This indicates that there may be a way to monitor the risk of violent events between refugees and citizens through social media, but more research would be needed to know what type of error rate to expect from this type of monitoring. Refugees and non-refugees alike experience an increase in communication—both phone calls and SMS text messages—when they are close to a violent event in the dataset. Refugees tend to experience more of an increase in communication over non-refugees when faced with violent incidents related to Syrian refugees. This difference in communications patterns may be an indicator against integration. Members of the host society appear to be less interested in a violent incident against Syrians, where refugees are very affected and possibly transmit that concern through increased phone and SMS communications.

When looking at the distance of the area to the nearest refugee camp, we can see from **Fig. 3** that places very close to a refugee camp (in dark red) also tend to have a high number of calls by non-refugees (x-axis) and high frequency of negative sentiment tweets against refugees (y-axis). One thing to notice in the graph is that locations associated with both negative sentiment towards refugees on social media as well as a relatively high call volume around violent incidents tend to occur near refugee camps. This may coincide with intergroup contact theory in the integration literature that suggests more frequent contact leads to more integration [24]. Perhaps refugees who are somewhat isolated by living in camps do not interact as frequently with citizens as those who live in urban and peri-urban areas, leading to more negative sentiment. It should be noted here that most of the violent events we used did not occur near refugee camps (see **Fig. 1** and **Fig. 2**).

Since the vast majority of refugees do not live in camps, but rather in urban or peri-urban locations [1], it is not surprising that the events generally occurred around large cities rather than specifically near camps. Variation in the negative sentiment expressed on Twitter in Turkey correlates with a large number of non-refugees being somewhat close to a refugee camp. It seems that variation on negative Twitter sentiments is explained more by the proximity of the camp than the number of refugees in a city. This may suggest that urban and peri-urban based settlements, even though they generate occasional violent incidents, help to disrupt growing negative sentiments. In fact, as we can see near refugee camps around the world, large and prolonged settlements of refugees organized by governments and operated by NGOs can breed significant local hostilities. This gives some empirical evidence to support the growing practice of promoting self-settlement in urban areas, not just for economic reasons, but also to decrease hostile feelings towards refugees. While this does not speak to integration directly, it does indicate that the foundation of facilitating harmonious co-habitation by host and migrant populations is to promote self-settlement and urban/peri-urban solutions over camp-based solutions.

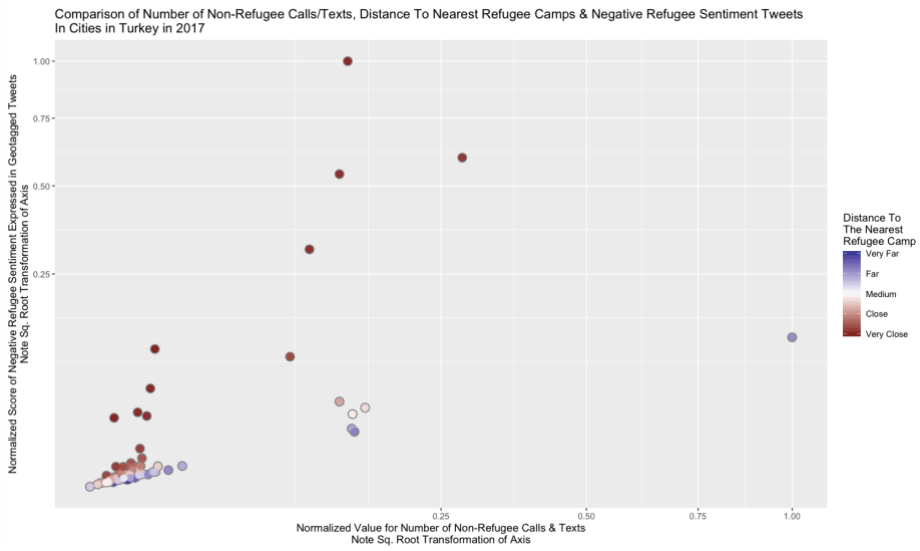


Fig. 3. Number of Non-Refugee Communication and Distance to Nearest Refugee Camp as related to Negative Refugee Sentiment Tweets 2017

This idea about proximity to refugee locations led us to a question about geographical dispersion of refugee populations and the possible differences between communication patterns of refugees and non-refugees. The 2018 ICG report on metropolitan areas makes a baseline assumption that "the potential for anti-refugee violence is highest in the metropolitan areas of Istanbul, Ankara, and Izmir where communities see Syrians as culturally different and resent their competition for low-wage jobs or customers, especially within the informal economy" [25]. The report contends that border communities are much more integrated because they have long done business across interstate lines and have cultural and linguistic ties that may not exist in other areas.

These findings are somewhat in contrast to what we found regarding negative Twitter sentiment and proximity to refugee camps—the vast majority of which are along the Syria/Turkey border. In our data, border provinces are more likely to express negative sentiments towards refugees on Twitter. These communities are both close to the border and home to Turkey's refugee camps. These findings of negative sentiment are in keeping with the tensions between border inhabitants and Syrians who found themselves staying in Turkey far longer than they expected indicated by qualitative fieldwork conducted by ICG in 2016 [12]. It is not possible here to disaggregate the effects of those two factors from the negative Twitter sentiment. In other words, we cannot tell from the data available here whether negative sentiment is caused by proximity to the border, to the refugee camps, neither, or both. Comparing to **Fig. 2**, however, we can see that the violent incidents that occurred involving Syrian refugees corresponds roughly to the same locations that experience the most negative sentiment on Twitter. Note, those cities not included in **Fig. 4** registered very low (near or at zero)

negative Twitter sentiment with regards to the hashtags and phrases we used for this study. Future assessments should explore alternative hashtags that may be unique to particular regions or cities in Turkey.

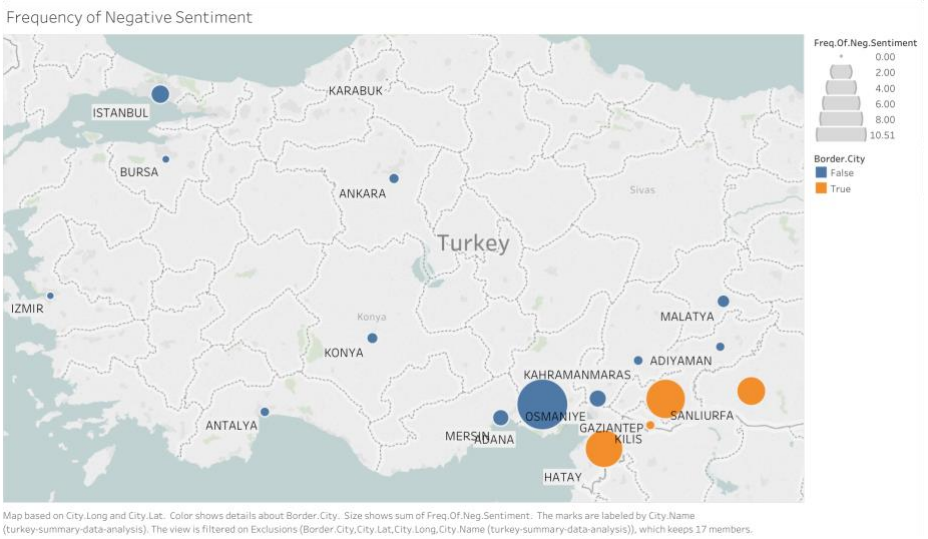


Fig. 4. Frequency of Negative Sentiment (marker size indicates relative frequency) and proximity to border (orange indicates border province)

When comparing border provinces (shown in orange in **Fig. 4**) and non-border provinces, there is a statistically significant difference in the volume of calls conducted by refugees and non-refugees in the dataset. In total call volume, border communities experience a much higher call volume than non-border communities after a violent event involving Syrians in Turkey. These two phenomena are likely related, though we do not know the content of the calls and thus find correlation between the two events (incident and high call volume). Along the border, there is a smaller difference in the call volumes between refugees and non-refugees ($p = 0.03978$), meaning, from the dataset, border communities have similar communications behavior between refugees and non-refugees after a violent event. These relationships do not hold for texts. This pattern could possibly indicate that both refugees and non-refugees care somewhat equally about the effects of a violent event involving refugees. What we cannot tell, however, is the tone or intention of the calls. It is not possible to know whether these calls were mobilizing xenophobic or integrative sentiment.

These conflicting findings, in comparison with the ICG report, require future investigation. While similar call patterns may support the hypothesis that border communities are more integrated—at least in as much as they culturally rely on phone calls to communicate after violent events—the increased negative sentiment towards refugees near camps and near border communities indicates some other underlying explanations. Further research is needed to pull apart the factors that may be influencing negative social media sentiment and the difference in communication patterns between

border communities and non-border communities and refugees and non-refugees. More must also be done to understand why the same relationships do not hold for texts. This finding suggests that perhaps texts and calls serve different sociological functions.

These findings point broadly toward areas that require deeper investigation. Our research suggests two particular findings. First, the assumption that border communities are more integrated than cities due to their historical connection with neighboring countries may not hold true when the number of refugees has increased so dramatically and remained for so long. Further research is needed to understand the dynamics between camp-based refugees in provinces that neighbor the refugees' country of origin and the 'compassion fatigue' described in so many refugee-related studies, including the ICG reports [12, 25]. Second, it appears there is a different culture for communication between border communities and non-border communities. In border communities, the similarity between refugees and non-refugees may indicate some amount of integration—at least in the cultural use of mobile phones. Our data showed that the difference in communication patterns between refugees and non-refugees becomes increasingly less pronounced the closer the province is to a refugee camp. More research is needed to understand how and why people use phones, particularly in situations of forced displacement and prolonged hosting of refugee communities. The data for this study cannot provide more insight into these cultural communication differences.

4 Policy Implications

This research opened up more questions than provided definitive answers that can inform policy decisions. We limit our policy recommendations to suggestions for future research. First, border cities appear to be able to inform our understanding of integration. This may be for both the positive and the negative, where 'compassion fatigue' erodes what was once a path to integration, and time and proximity lend themselves to inevitable co-evolution of cultural practices such as communication. Our data cannot provide specific solutions or policies regarding this aspect of integration, but it does point to the notion that more work must be done to understand these nuances of border life in the hopes that their experiences can inform integration practices throughout the host country.

Second, the study found that there are statistically significant differences in the call volumes of refugees and non-refugees after a violent event that involves Syrians in Turkey. While we cannot gauge the tone or intent of these communications, there is clearly communication about the event happening. This phenomena supports the notion that more mechanisms must be put in place to defuse tensions after a violent event in order to limit the erosion of progress to local integration strategies.

Third, our research suggests that fewer negative sentiments are expressed farther away from refugee camps, despite claims that urban centers are more likely to experience refugee violence. More research must be done to understand why this is. Our preliminary work seems to support the growing body of refugee literature that indicates urban and peri-urban settlements are not only good for economic growth and

opportunity, but also perhaps in dissolving some of the negative sentiment towards refugees. Again, our findings are based on data that are not representative of the entirety of Turkey or any one host community's experiences. The data suggests, however, that this may be an area that is worth investing in more research in the future, not just for the benefit of integration in Turkey, but also for extrapolating Turkey's urban host policies to other host states around the world.

References

1. UNHCR. 3RP 2017 Progress Report. (2017). https://data2.unhcr.org/en/documents/download/60340
2. AFAD (Disaster and Emergency Management Presidency) Survey on Syrian Refugees in Turkey. (2013). https://data2.unhcr.org/en/documents/download/40157
3. Aras, B. and S. Yasun. The Educational Opportunities and Challenges of Syrian Refugee Students in Turkey: Temporary Education Centers and Beyond. (2016). IPC Mercator Policy Brief: http://research.sabanciuniv.edu/29697/1/syrianrefugees.pdf
4. Kirisci, K. and E. Ferris. Not Likely to Go Home: Syrian Refugees and the Challenges to Turkey- and the International Community. Turkey Project Policy Paper, Number 7. (2015). https://www.brookings.edu/wp-content/uploads/2016/06/Turkey-Policy-Paper-web.pdf
5. Çağaptay, S. The Impact of Syria's Refugees on Southern Turkey, Policy Focus 130, Washington Institute for Near East Policy. (2014). Washington DC. http://www.washingtoninstitute.org/policy-analysis/view/the-impact-of-syrias-refugees-on-southern-turkey
6. Dinçer, O.B., et al. Turkey and Syrian Refugees: The Limits of Hospitality. (2013). Washington DC and Ankara: Brookings Institution and International Strategic Research Organization: https://www.brookings.edu/research/turkey-and-syrian-refugees-the-limits-of-hospitality/
7. Kanat, K. B. and K Ustun. Turkey's Syrian Refugees: Toward Integration. (2015). SETA Report: http://file.setav.org/Files/Pdf/20150428153844_turkey%E2%80%99s-syrian-refugees-pdf.pdf
8. Tumen, S. The Economic Impact of Syrian Refugees on Host Countries: Quasi-Experimental Evidence from Turkey. American Economic Review 106(5): 456-460. (2016). https://pubs.aeaweb.org/doi/pdfplus/10.1257/aer.p20161065
9. Yildiz, A. and E. Uzgoren. Limits to Temporary Protection: Non-Camp Syrian Refugees in Izmir, Turkey. Southeast European and Black Sea Studies 16(2): 195-211. (2016). https://www.tandfonline.com/doi/full/10.1080/14683857.2016.1165492?src=recsys
10. Corabayir, M. The Evolving Approach to Refugee Protection in Turkey: Assessing the Practical and Political Needs. (2016). Transatlantic Council on Migration, A Project of the Migration Policy Institute: https://www.migrationpolicy.org/research/evolving-approach-refugee-protection-turkey-assessing-practical-and-political-needs
11. Erdoğan, M. *Türkiye'deki Suriyeliler: Toplumsal Kabul ve Uyum*, İstanbul: İstanbul Bilgi Üniversitesi Yayınları. (Syrians in Turkey: Social Acceptance and Integration Research). (2015). https://data2.unhcr.org/en/documents/download/46184
12. International Crisis Group. Turkey's Refugee Crisis: The Politics of Permanence. (2016). Europe report no. 241: https://d2071andvip0wj.cloudfront.net/241-turkey-s-refugee-crisis-the-politics-of-permanence_0.pdf

13. İçduygu, A. Syrian Refugees in Turkey: The Long Road Ahead. Report. (2015). Migration Policy Institute. https://www.migrationpolicy.org/research/syrian-refugees-turkey-long-road-ahead
14. Fonio, C., & Boersma, K. Big data, surveillance and crisis management. In Big Data, Surveillance and Crisis Management (pp. 15-30). (2017). Routledge.
15. IOM. Bulletin, Issue No. 5. (February 2018). ISSN 2523-5060.
16. Salah, A.A., Pentland, A., Lepri, B., Letouzé, E., Vinck, P., de Montjoye, Y.A., Dong, X. and Dağdelen, Ö. (2018). Data for Refugees: The D4R Challenge on Mobility of Syrian Refugees in Turkey. arXiv preprint arXiv:1807.00523.
17. Public streams: Twitter developers. 6 February 2018 <https://dev.twitter.com/streaming/public>.
18. Second quarter letter to twitter shareholders. 6 February 2018. <http://goo.gl/ihmm46>.
19. Kwak, H, et al. What is twitter, a social network or a news media? Proceedings of the 19th International Conference on World Wide Web. ACM, (2010): 591–600.
20. Go, A, R Bhayani and L Huang. Twitter sentiment classification using distant supervision. Technical. (2009). Stanford.
21. Guyon, I and A Elisseeff. An introduction to variable and feature selection. Journal of Machine Learning Research 3(3). (2003): 1157–1182.
22. Pang, B and L Lee. Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval 2(2). (2008): 1–135.
23. Smith A, Brenner J. Twitter use 2012. Pew Internet & American Life Project. 2012.
24. Pettigrew, T. F. (1998). Intergroup Contact Theory. *Annual Review of Psychology*, *49*(1), 65-85. doi:10.1146/annurev.psych.49.1.65
25. International Crisis Group. 2018. *Turkey's Syrian Refugees: Defusing Metropolitan Tensions*. Europe Report no. 248: https://d2071andvip0wj.cloudfront.net/248-turkey-s-syrian-refugees.pdf

Improve Education Opportunities for Better Integration of Syrian Refugees in Turkey

Marco Mamei[1], Seyit Mümin Cilasun[2], Marco Lippi[1], Francesca Pancotto[3], and Semih Tümen[4]

[1] DISMI, University of Modena and Reggio Emilia, Italy
marco.mamei@unimore.it, marco.lippi@unimore.it
[2] Central Bank of the Republic of Turkey , Turkey
Seyit.Cilasun@tcmb.gov.tr
[3] DCE, University of Modena and Reggio Emilia, Italy
francesca.pancotto@unimore.it
[4] Department of Economics, TED University, Turkey
semihtumen@gmail.com

Abstract. The integration of Syrians' refugees in Turkish society is crucial for the long-term well-being of both populations. Education and schooling is one of the most important elements to integrate Syrians' children and prevent a "*lost generation*". In this project we investigate two main aspects related to refugees' education: "How to improve Syrians' access to schooling?" and "What is the impact of Syrians' schooling on Turkish society?" The analysis presented in the paper provides quantitative elements to analyze and optimize education resources with respect to refugees' and natives' needs, supporting the claim that education plays a key role in improving integration in the society.

Keywords: education, social integration

1 Introduction

Turkey is the largest refugee hosting country worldwide with 3.5M registered Syrian refugees. Most of these refugees will not return to Syria anytime soon. According to a recent research documenting the extent of economic destruction in Syria, more than 50% of Syrian economy got destroyed, which means that economic returns to going back to Syria are almost zero [4]. Considering that Syrians will continue to live in Turkey for many years, and that a significant portion will not turn back to their country, their integration in the Turkish society is crucial.

As more than one third of the Syrian population residing in Turkey is composed of school-age children, schooling and education have a major role in such an integration process. Schooling has a fundamental role to normalize the lives of children, to support their interactions with peers (Syrians and Turkish) and to provide them the skills for their future professions. When children are left out education, they are prevented from escaping poverty and become more vulnerable

to ghettoization and radicalization. Education of Syrian children is fundamental for both those who will stay in Turkey and for those who will eventually return to Syria. The former will contribute economically, socially and culturally to Turkish society, the ones who return to Syria will rebuild their country with the education they received in Turkey. "*In this sense, education is an important soft power strategy.*"[7]. However, for the perspective of any hosting society, mass refugees inflow represents a huge challenge. Refugees generate competition for several resources and many public services get congested. This congestion creates conflicts and one of the major conflict area is "education". Host countries have limited education resources (schools and teachers), which makes policy making and resource re-allocation a major issue.

One of the main assets to address these complex challenges is to have up-to-date fine-grained information about refugees and their activities. Mobile phone data (i.e., Call Detail Records (CDRs) [16]) allows to track activities of refugees and natives at a fine grained scale, therefore it is a natural response to this need for information. Mobile phone data has been used to estimate the socioeconomic status of territories [11, 14] and individuals [2], to analyze the dynamics of cities [8], to model the spreading of diseases [20], and to predict crime levels [3].

In this project we will use mobile phone data to analyze some of the challenges to integrate Syrian refugees on the education system in Turkey and the impact to the Turkish society. Educational institutions do not simply transmit human capital, they also pass on social capital in the form of social rules and norms. So, investments in education will not only work toward increasing Syrians' school enrollment rate, they will also impact their integration in the society. Our goal is to propose some new findings and recommendations along this dimension:

1. We analyze the distribution of refugees across the country and their possible impact on education facilities. We analyze logistic obstructions to schooling of Syrians' children, and propose and evaluate an optimization mechanism to identify areas where new education resources are required considering both refugees and natives needs.
2. We analyze the relationship between education and social integration. We analyze the impact of Syrian refugees schooling to social integration (measured via numbers of calls between natives and Syrians). We analyze the impact of Syrian refugees on natives education choices and their impact on Turkish economic development.

Overall, our purpose is to develop evidence-based research with outputs directly implementable by policy makers.

2 Background on Syrian Refugees Education

Providing accurate figures about Syrian refugees' education is a challenging task. On the one hand, the number of refugees is constantly increasing (in 2017 alone it grew by 20%), On the other hand, policies and applications about Syrians education are changing frequently. Because of this, analyses and reports on Syrians

education in Turkey quickly lose their currency (this is a strong reason for the need of innovative information sources, like mobile phone data). Nevertheless, as a general background, we report some data from existing documents [5–7].

There are about 3.5M registered Syrian refugees in Turkey. About 8% of Syrian refugees live in one of the 23 Temporary Sheltering Centers (TSCs) across the country. The remaining 92% live in cities. Among them, more than 1M are school-age children. Thanks to the efforts of Turkey and of the international community, the number of Syrian refugees' regularly attending primary education rose from about 25% in 2016 [5], 55% in 2017 [7], and 63% in 2018[5]. Despite this notable increase, the situation is far from solved. Prior to the conflict, the primary school enrollment rate in Syria was 99% and lower secondary school enrollment was 82%, with high gender parity.

Syrian refugees' education could be attained in public schools (about 65,000 schools) and in Temporary Education Centers (TECs): private schools typically run by Syrian charities and offering courses in Arabic and and intensive Turkish language courses (about 400 TECs – serving about 230K students). However, in 2017, in an attempt to better integrate Syrian kids in the education system in public schools, the Ministry of National Education (MoNE) has decided to start the closing TECs. Therefore, while in 2016, 80% of Syrians preferred to study at TECs, Since the beginning of the 2016-2017 academic year, MoNE has put a halt to new registration to preschools as well as to the 1st, 5th and 9th grade at the TECs (Turkish school system is termed *4+4+4*: 4 years primary education, first level, 4 years primary education, second level, and 4 years secondary education). The newcomers to these grades are now directed to public schools (provided a sufficient Turkish language skills). In 2017, about 50% of schooled Syrians went to TECs and 50% to public schools.

2.1 Congestion of Education Resources

Turkey faces many issues in taking care of the education of Syrian children. First of all, cities like Sanliurfa, Gaziantep and Istanbul – where most of the Syrians live – had already issues regarding educational infrastructure. The number of students per teacher and per classroom in these cities are above the average of Turkey. In addition, 19% of the primary schools in Turkey have double-shift education[6] and 46% of students study at these schools. 65% of primary school students in Gaziantep, Adana and Bursa and 55-64% of students in cities like Istanbul, Ankara, Izmir, Sanliurfa, Mersin and Osmaniye are educated at schools with the double shift system. Cities with high population of Syrians are at the same time the cities where schools have many troubles and where double-shift education is a common phenomenon. The current plan to close down TECs

[5] http://www.hurriyetdailynews.com/63-percent-of-syrian-children-in-turkey-attend-school-education-minister-132276

[6] Many school buildings have been arranged for double-shift sessions in neighborhoods heavily populated by Syrians. Those schools provide education to Turkish students in the first shift and Syrians in the second shift starting at 14:30.

(although well justified to improve integration) will worsen the situation. For this reason, physical capacities of public schools that Syrians will be directed to in the case of closures of TECs must be assessed carefully and eventually new resources should be directed to congested areas. As stated in the Introduction, addressing correctly this congestion problem is fundamental to improve integration and acceptance of Syrians in Turkish society.

2.2 Obstructions and Mobility

Although important results have been already achieved, an important challenge faced by Turkey is to increase Syrians' school enrollment rate. While the two most prominent obstructions to schooling are economic issues (children work to support their family), and language barrier (e.g., inability to understand courses, feeling of alienation in the school environment, concern about forgetting Arabic), in this section, we focus on another important obstruction that is directly related and measurable with mobile phone data: transportation and mobility.

An important obstructions in Syrian children's access to education is about the distance between the homes of the kids and schools. Syrians usually live in the poorest neighborhoods and near the industrial zones where even the locals have issues accessing education and urban life. The fact that some schools start at later hours force children being on the road after the dark. As the majority of the Syrian students do not have the economic means to buy service from private transportation companies, this causes some families to avoid sending their children (daughters especially) to school. Another factor in keeping children out of school is the high mobility rates in the lives of Syrians. It was observed that the majority of unschooled children in Istanbul and Ankara just migrated to these cities from southern provinces. As Syrian families continuously move between the cities and relocate frequently, absenteeism becomes an inevitable consequence for the children.

In general, these logistic obstructions must be taken into account, together with the above congestion problem, when planning new education resources.

2.3 Ongoing Activities

There are several studies and ongoing activities to improve Syrians education in Turkey and reduce congestion on existing resources. Among them, the PICTES Project (Promoting Integration of Syrian Children to Turkish Education System) is one of the largest initiatives. It is a 500M Euro project between Turkey, EU, Kreditanstalt für Wiederaufbau and the World Bank to support integration and education activities [6]. The project is implemented in 23 provinces where the population of Syrians under temporary protection is the largest. The project includes school constructions (about 200M Euros), transportation services (about 30M), and many other activities (e.g., teaching Turkish, Arabic language training, remedial/catch-up courses, educational materials).

Providing reliable and timely data on Syrian refugees' locations and activities is critical to support planning and decision making in these projects.

3 Materials and Methods

The D4R initiative provides a unique dataset to analyze the behavior of Syrian refugees at a fine grained scale [16]. In this section we describe the main information extracted from this data and used in our work.

- Dataset 1 (Antenna Traffic) provides site-to-site traffic (aggregated CDRs) on an hourly basis. Following [18, 17], we extracted several features associated to this data ranging from simple aggregations to graph-based metrics describing the connectivity graph across multiple regions. In general we use this data as a proxy for Syrian refugees location and activities.
- Dataset 2 (Fine Grained Mobility) provides detailed (BTS-level) information about calls and sms made by a random sample of active users (individual CDRs). We use this dataset for two main purposes: on the one hand it allows to precisely identify the home (primary) location of the sample of users [13] and to analyze those locations in relation with schools around. On the other hand, this data allows to measure for each user the number of calls/sms placed to other Syrians and/or to other Turkish people. Aggregating the number of calls between Syrian-Syrian, Syrian-Turkish, Turkish-Turkish we can get information on the social integration in the random sample.
- Dataset 3 (Coarse Grained Mobility) provides coarse (District-level) information about calls and sms made by a random sample of active users (individual CDRs). We used this data to analyze large-scale mobility behavior and re-locations of Syrian refugees across the country.

In addition to D4R data, we got access to the following data sources:

- **Population and School Enrollment.** Data from Turksat and UNHCR about Syrians and natives distribution across the country. Data is divided also by gender and age-group. In particular we obtained fine grained location data about the education choices of Turkish natives. For each district we have the number of male and female students attending primary, secondary, high school, university, master and doctorate schools. Critically, we did not get access the same information about Syrians refugees. However, we obtained from [5] data about province-level distribution of the number of Syrians frequenting public schools and temporary education centers (TECs) in 2016.
- **Education Resources.** Data from Ministry of National Education (MoNE) about the distribution of public schools by district and type (pre-schools, primary schools and high schools). Data about province-level distribution of TECs in 2016. We extracted this data from [5]. Data from Open Street Map (OSM) containing the exact locations of a sample of about 7,000 schools (1034 in Ankara, 642 in Kayseri, 621 in Istanbul, 547 Konya, etc.)
- **Economic Indicators.** Data about province-level per-capita GDP from 2004 to 2014 from Turkstat. Data about district-level per-capita GDP from 2014 to 2017 proxied by the analysis of satellite night-lights from NOAA VIIRS/DNB dataset[7].

[7] https://explorer.earthengine.google.com/#search/tag:lights

Data analysis has been conducted with standard statistical techniques (summary statistics and OLS) favoring interpretability of the results. In the following section, we describe data analysis and experiments in detail.

4 Research Questions and Experiments

In this project we investigate a number of important aspects related to refugees education: refugees' education access, logistic obstructions to schooling, congestions in access to education resources, the impact on natives education choices, the relationship between education and social integration and well-being. We organize the section in a preliminary analysis on the representativeness of the data and in two main research questions:

0. **Does mobile phone data reflect the distribution of refugees across the country?**
1. **How to improve Syrians' access to schooling?**
   - What are the main logistic obstructions for Syrian refugees' schooling?
   - Can we identify areas in which education resources are particularly congested?
   - Can we use this information to prioritize the construction of new education resources?
2. **What is the impact of Syrians' schooling on Turkish society?**
   - Does schooling and education have an impact in social integration?
   - Does refugees influx have an impact on natives education choices?
   - Does integration and education have an impact on economic well-being?

Data-driven research addressing these questions can provide critical information to policy makers to enact interventions.

4.0 Does mobile phone data reflects the distribution of refugees and natives across the country?

This introductory question is intended to understand the representativeness of the data and to analyze its main biases. As described in [16], D4R dataset is collected from 992K Türk Telecom customers, of which 185K are tagged as refugees and 807K as Turkish citizens. D4R data contains also the province-level distribution of such Türk Telecom customers. The Turkish citizens have been sampled mainly for the cities with high-presence of registered refugees, so they do not reflect the distribution of citizens across the country. Figure 1(a-b) shows correlation, at the province level (NUTS-3 regions), of sampled Türk Telecom data (refugees and natives) and estimates from official statistics considering March 2017 (the number of Syrian refugees in Turkey in 2017 grow by 20%, so the monthly analysis can provide more accurate figures). While there is a good correlation between refugees' distributions, it is possible to see the sample bias for

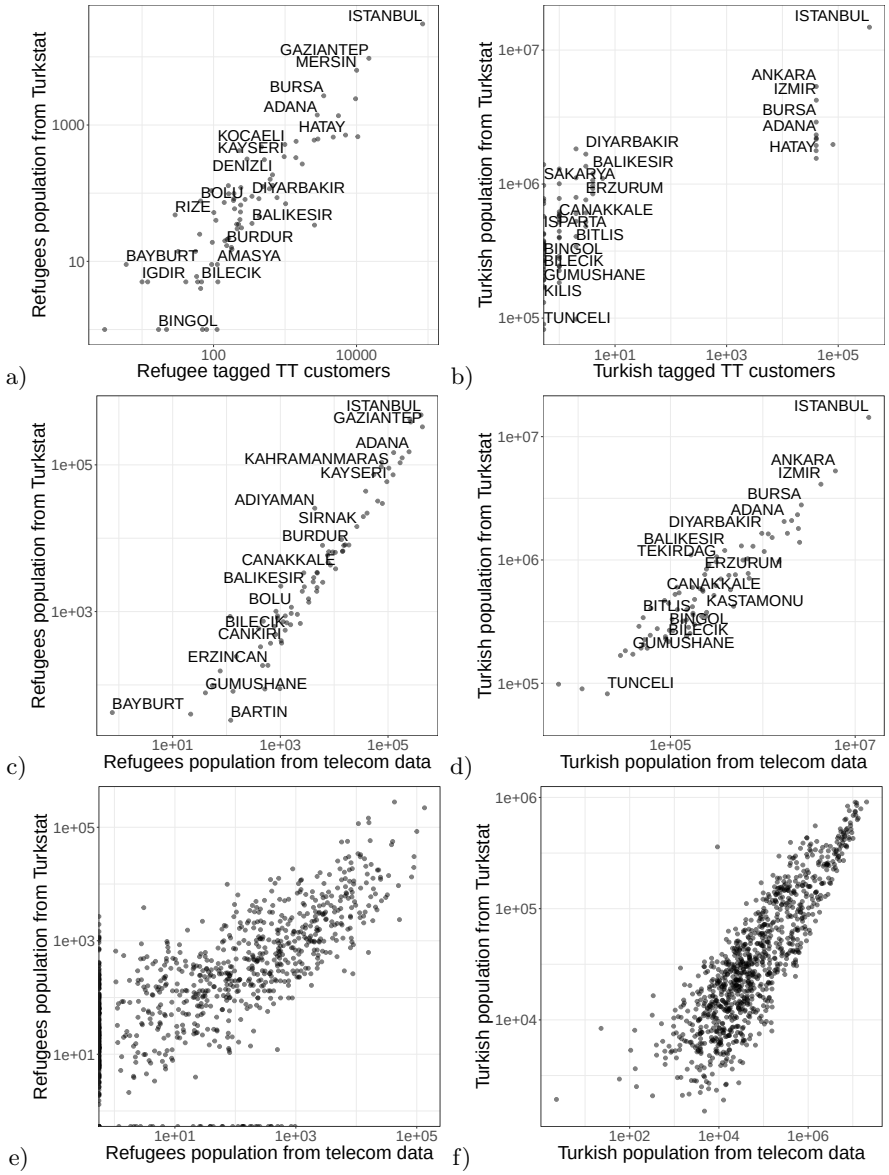


Fig. 1. a-b) Province-level correlation between Türk Telecom refugees and natives in D4R dataset and from Turkstat. **c-d)** Province-level correlation between refugees/natives population from telecom data and from Turkstat. **e-f)** District-level correlation between refugees/natives population from telecom data and from Turkstat.

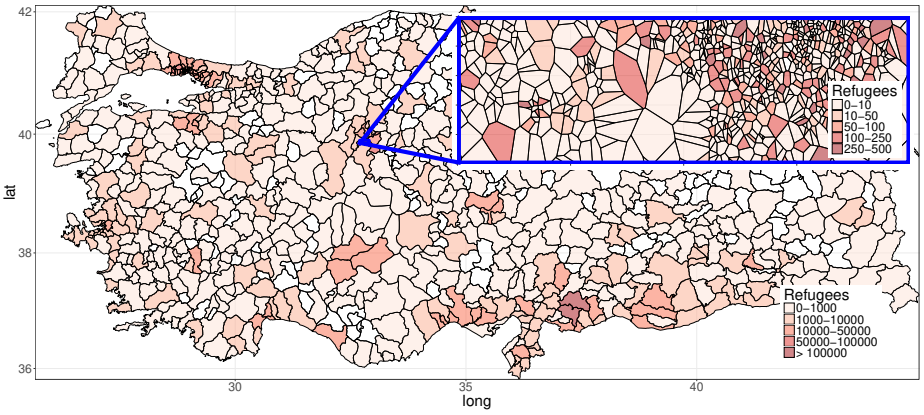


Fig. 2. Refugess' density across the country. The inset shows a BTS resolution map obtained by a Voronoi tessellation of the telecom antennas. In the city centers, Voronoi regions are typically smaller than 500 m^2.

citizens, exhibiting a bi-modal distribution with low-sampled and high-sampled provinces.

From this data, we computed the Türk Telecom province-level market-share associated with the sampled users: for each province, we divided the number of sampled Türk Telecom customers (refugees and natives) by the corresponding Turkstat data. Of course, this ratio is much lower than the true market share as it reflects only the sampled users.

To estimate people presence on the basis of mobile phone data, we analyzed monthly-aggregate outgoing call volume. Outgoing call volume has been used in previous work as a proxy for people presence [9, 10]. For each province, monthly-aggregate outgoing call volume has been scaled according to (i.e., divided by) the computed province-level market-share. This process should compensate biases in the Türk Telecom customers sampling process. We fitted a linear regression between aggregate outgoing call volume (scaled by province-level market-share) and Turkstat data and used regressed value as population estimate from telecom data. In the case of Turkish population we fit two regressions: one for high-sampled provinces (right cluster in Figure 1(b)) and one for low-sampled provinces.

Figure 1(c-d) shows correlation, at the province (NUTS-3 regions) level, between refugees/turkish estimate from telecom data and from Turkstat ($\rho = 0.94$ for refugees, $\rho = 0.98$ for turkish). Figure 1(e-f) shows correlation, at the district (LAU regions) level, between refugees/turkish estimate from telecom data and from Turkstat ($\rho = 0.68$ for refugees, $\rho = 0.88$ for turkish).

It is important to remark that the not perfect correlation, especially between refugees from from telecom data and from Turkstat at the district level, is not

necessarily a limit of telecom data. It might be the case that Turkstat data do not accurately reflects the presence of refugees at the district level, the use of telecom data in this domain is exactly to improve over existing information.

It is also interesting to notice that the correction for (sample-based) market-share allows to efficiently correct biases in the sampling process. Nevertheless, the small sample of natives for some provinces might still bias further analysis. Therefore, in the following, we often use mobile phone data to analyze refugees' distribution while we revert to official statistics for natives. On the basis of these results, we can create fine grained maps of areas frequented by refugees. Considering a Voronoi tessellation of cell network's antennas (BTS), we can create maps at resolution much finer than districts - LAU regions. In the city centers, Voronoi regions are typically smaller than 500 m^2. See Figure 2.

Discussion. Overall, mobile phone data, once corrected for sample bias, well reflects the distribution of refugees and natives across the country. We speculate that refugees' distribution at the district level computed from mobile phone data can be more accurate than official statistics and that the not perfect correlation between telecom and Turkstat data can be ascribed to this fact.

4.1 How to improve Syrians' access to schooling?

What are the main logistic obstructions for Syrian refugees' schooling? A number of reports and surveys [5–7] indicates that the sheer distance to school is an important obstruction for Syrian refugees' schooling. We performed an experimental analysis to understand which are the most disadvantaged areas with this regard.

We downloaded from Open Street Map (OSM) the location of 7K school buildings in Turkey. From D4R Dataset-2 we find the location where each refugee "lives", by computing the average location of his/her calls and sms.

For each district, we then computed the average distance between refugees whose main location is in the district and the closest school. More specifically, given N refugees $R_{1...N}$ living at coordinates (r_{ix}, r_{iy}) and M schools $S_{1...M}$ at coordinates (s_{ix}, s_{iy}), we first compute, for each refugee R_i, the closest school: $S_{R_i} = argmin_j\ dist((r_{ix}, r_{iy}), (s_{jx}, s_{jy}))$. Then:

$$avg.\ dist.\ to\ school = \frac{1}{N}\sum_{i=1}^{N} dist((r_{ix}, r_{iy}), (s_{R_ix}, s_{R_iy}))$$

In Figure 3 we present results for different provinces. The maps show the average distance between Syrian refugees "living" in a given district and closest schools. It is important to remark that, on the one hand, this estimate represents a best-case scenario where a student can attend the closest school available. We did not have information about the capacity in terms of number of seats for each school. On the other hand, the OSM database has a limited coverage and not all the schools are mapped. Nevertheless these maps can indicate disadvantaged

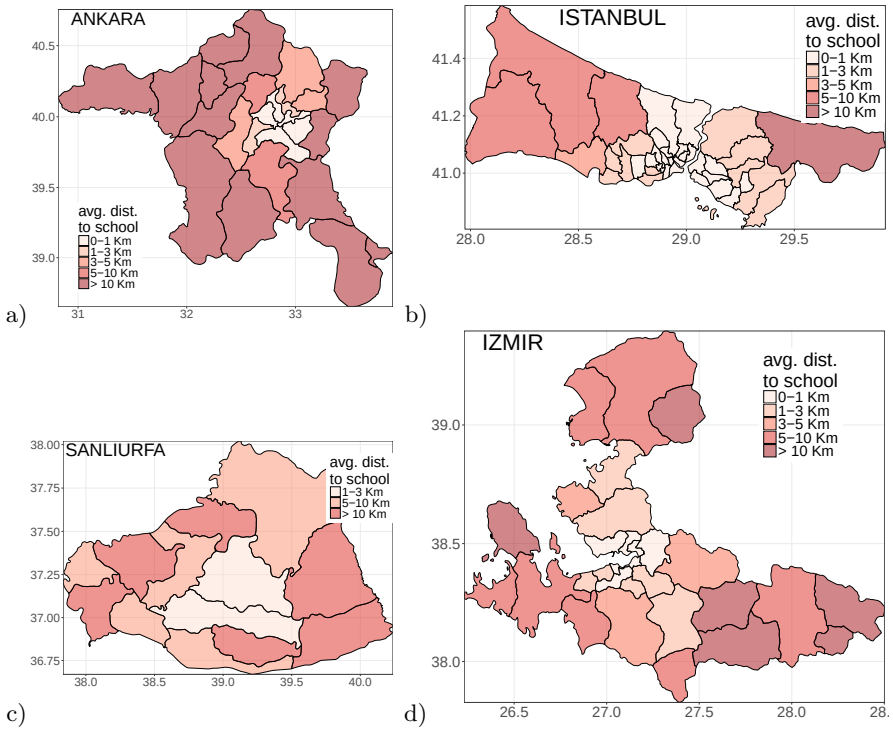


Fig. 3. Average distance between Syrian refugees "living" in a given district and closest schools. **a)** Ankara province (1034 schools in OSM db). **b)** Istanbul province (621 schools in OSM db). **c)** Sanliurfa province (122 schools in OSM db). **d)** Izmir province (431 schools in OSM db).

areas with respect to schools' reachability. In our future work, we plan to refine this analysis by comparing the male and female Syrian student populations, as this obstruction, which is linked to the perception of "safety" getting to and from school, is particularly limiting girls participation in education.

Another obstruction to schooling described in [5–7] is high mobility rates in the lives of Syrians. For example, the majority of unschooled children in Istanbul and Ankara just migrated to these cities from southern provinces. To understand Syrians' refugee mobility we focused on Dataset 3 considering the whole observation period (1 year). We computed three mobility indicators:

– Radius of gyration, that is a synthetic and easy-to-compute parameter describing the spatial extent of user traces. It is defined as the deviation of user positions from the corresponding centroid position. It is given by: $r_g = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(p_i - p_{centroid})^2}$ where $p_i = (x_i, y_i)$ represents the i^{th} posi-

tion recorded for the user and $p_{centriod} = (r_x, r_y)$ is the center of mass of the user's recorded displacements obtained by $p_{centroid} = \frac{1}{n}\sum_{i=1}^{n}(p_i)$
- The distribution of the number of unique provinces and districts visited by refugees
- The distribution of the number of unique "home" provinces and districts of refugees. For each month, we computed the "home" province and district of each refugee as the one where they spend more time. This is different from the above distribution as a person can have a stable primary "home" province and district while visiting and commuting across many. A change in "home" provinces and districts indicates that the person has probably relocated to the new area.

The distribution of the radius of gyration reveals that the vast majority of Syrian refugees live in a well defined area (80th percentile = 30 Km). However, there is a long tail of refugees spanning larger distances (95 percentile = 250 Km). Figure 4a) shows the distribution of all visited provinces and districts: 77% of Syrian visit more than one province, 99% of Syrian visit more than one district. Figure 4b) shows the distribution of home provinces and districts: 28% of Syrians has more than one home province. 54% of Syrians has more than one home district. Under our assumptions and according to D4R data, this means that about 1/3 of Syrians has changed residency province in 2017.

Discussion. Mobile phone data allows to quantify logistic obstructions to schooling. It is possible to see that for all the provinces there are disadvantaged districts where the average distance to the closest school is large (5Km or even more). In the lack of affordable and safe transportation and considering the late hours at school (due to double shift system) for a large number of Syrian refugees, this kind of distances represent an important obstruction to schooling. In addition, a sizable number of Syrian refugees move in a wide area (5% move further than 250 Km). Moreover, it seems that about 1/3 of Syrians has changed residency province in 2017. Indeed mobility and relocations among provinces can be a obstruction for school planning for Syrian children.

Can we identify areas in which education resources are congested and prioritize new education resources accordingly? For the perspective of any hosting society, mass refugees inflow represents a huge challenge. Refugees generate competition for several resources there included education. Identifying congested areas and allocating education resources there is a valuable process to ease conflict between natives and refugees, and to support refugees education and integration. Moreover, given that the distance to school is an important obstruction to schooling, a natural remedy is to build or potentiate schools to improve reachability by refugees. There are indeed initiatives in this direction, for example the PICTES Project allocates 200 million Euros to school construction.

A concrete use of mobile phone data is to use fine-grained data about refugees' locations to support decision makers on where to create new education facilities

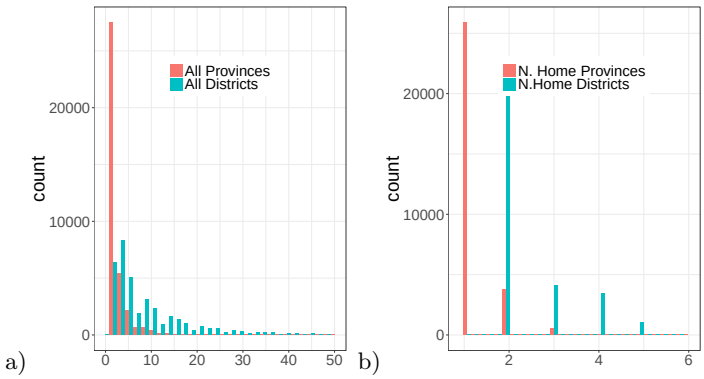


Fig. 4. a) Distribution of all visited provinces and districts. **b)** Distribution of home provinces and districts

taking into account both natives' and refugees' needs. We envisioned a simple procedure that, taking into account both natives and Syrian refugees' needs, identifies provinces and districts where the education resources are most congested. Then, focusing on a specific district, it identifies areas where to build new schools or potentiate existing ones with the goal of minimizing travel distance for refugees. We are perfectly aware that the process of creating schools and allocating resources is a complex operation involving many factors (political, economic) outside our data. We do not think that our procedure *per-se* can be a valid decision-support tool. However it could enrich existing planning solutions.

We can compute a simple congestion metric for education resources at the province/district level as the ratio between the number of children (both Turkish and Syrians) and the number of schools in the area. Unfortunately we do not have data about school capacity and number of teachers that would notably improve the estimate. We obtained number of Turkish children and number of schools directly from Turkstat and MoNE. The number of Syrian children has been obtained from mobile phone data by scaling the total number of Syrians by the ratio of children across Syrians' population (44% from UNHCR[8]).

The results are depicted in Figure 5(a-b) at the province-level, and in Figure 5(c) at the district-level (focusing on the Ankara province).

Given a budget of X schools to build, we can simply allocate X_i schools to province P_i proportionally to the education-resources' congestion C_i in that province. Table 1 shows the result of an experiment to allocate 75 new schools across the country. The column "New School" is the number of schools to build in each province according to the above proportional criteria. For comparison we obtained from [6] data about the construction of 75 new schools by the PICTES project (that we consider a sort of optimal allocation). The two columns have Spearman rank correlation of 0.8, p-value ≈ 0.

[8] https://data2.unhcr.org/en/situations/syria/location/113

The same mechanism can be applied at the district level. For example focusing on Ankara, our approach would allocate the 3 schools to build in the district of Altindag. Interestingly enough, also the 6 schools PICTES is building in Ankara are all in the Altindag district [9].

Finally we can try to identify the exact location where to build a school by minimizing distance Syrians' refugees have to travel to reach the closest school (this to tackle one of the main logistic obstruction discussed before). More specifically, a school S to be build in district D is set at coordinates (s_x, s_y) so that the distance between refugees $R_{1....N}$ living at coordinates (r_{ix}, r_{iy}) is minimized:

$$(s_x, s_y) = argmin_{x,y} \frac{1}{N} \sum_{i=1}^{N} dist(x, y, r_{ix}, r_{iy})$$

An exemplary result of this allocation are in Figure 5(c), where Syrian refugees' main locations are represented by red circles, existing schools by green squares, and the location of new schools to build by blue triangles.

Discussion. Mobile phone data allows to precisely localize refugees' whereabouts across the country. This – together with natives' population statistics – allows to better estimate the *demand* for education resources and to compute fine-grained congestion maps to pinpoint the most stressed areas. This kind of information is an important asset to optimize education resources. Of course, congestion maps could be notably improved via better information about school type, their capacity and the number of available teachers. For example, there is a recent discussion in Turkey on the lack of enough demand for certain types of schools (e.g., Imam Hatip High Schools – religious schools), so some of the results could be reverted considering those schools. In general data-driven allocation results of (education) resources, like the ones presented, can provide useful information and guidelines to policy makers [1].

4.2 What is the impact of Syrians' schooling on Turkish society?

In addition to the above issues about refugees access to education, it is very important to analyze the impact of refugees influx in Turkish society and how education impact social integration of Syrians' refugees. As Syrians' access to education is constantly improving, the focus should gradually shift to the wider perspective of integration through education.

Does schooling and education have an impact in social integration? To address this question, we take into account the total number of calls and sms that are exchanged between natives and refugees (both incoming and outgoing). This information can be extracted from Dataset 2 where information about both

[9] http://pictes.meb.gov.tr/www/saha-ziyaretleri-ankarada-devam-ediyor/icerik/44

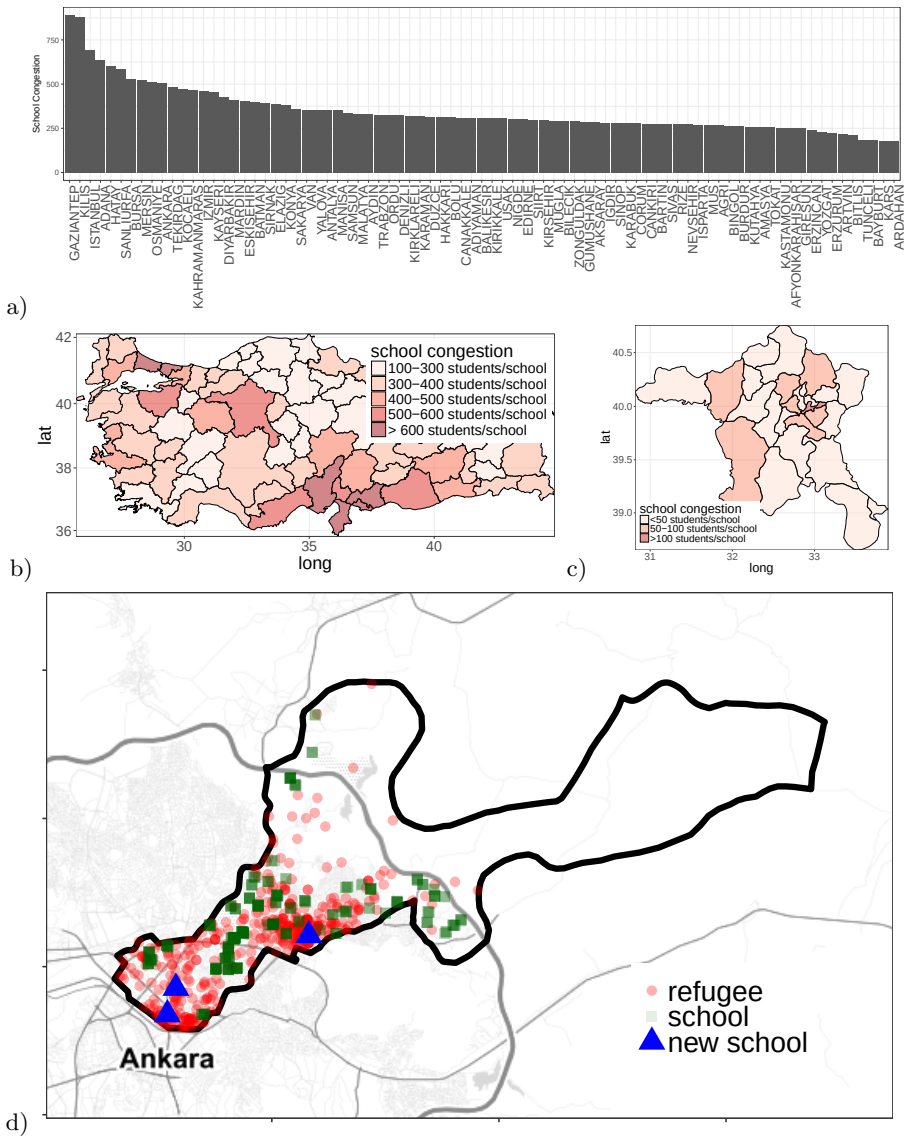
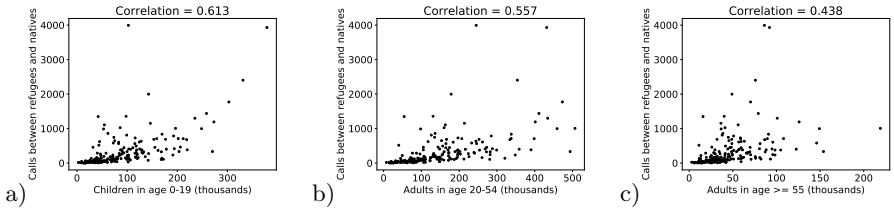


Fig. 5. a-b) school congestion at province level. **c)** school congestion at district level district. **d)** Refugees and schools distribution and where to build new schools

Table 1. New school allocation by our approach and PICTES project

Province	New Schools	PICTES
Gaziantep	7.00	5.00
Kilis	7.00	6.00
Istanbul	7.00	6.00
Adana	4.00	6.00
Hatay	4.00	5.00
Sanliurfa	4.00	6.00
Bursa	3.00	6.00
Mersin	3.00	5.00
Tekirdag	3.00	0.00
Ankara	3.00	6.00
Osmaniye	3.00	1.00
Diyarbakir	2.00	2.00
Eskisehir	2.00	0.00
Izmir	2.00	4.00
Kayseri	2.00	1.00
Kahramanmaras	2.00	5.00
Mardin	2.00	5.00



a) b) c)

Fig. 6. Correlation between the total number of children (a), adults (b) and seniors (c), and the total number of calls between natives and refugees. Data are for a single month (one point is one district), but identical trends hold for different months, as well as for sms rather than calls.

communication end-points are provided. We named those variables $x2y_{calls}$ and $x2y_{sms}$. Such quantities are highly indicative of the level of integration between the two parts of the population. We compute Pearson's correlation between the total number of people in different age intervals (we obtained this data from Turkstat), and quantities $x2y_{calls}$ and $x2y_{sms}$ aggregated at the district level. In particular, we distinguish between children (age 0-19), adults (age 20-54) and seniors (age $\geq$ 55) and assume that these numbers reflect both natives *and* Syrians distribution. An example of the obtained results is shown in Figure 6: clearly, for all age categories, the correlation is positive (the higher the number of people, the higher the number of calls) but for the category of children it results to be significantly higher ($p < 0.01$ according to a Wilcoxon paired test computed over months). Although our analysis do not support causal claims, this fact suggests that children are an important vehicle for integration between Syrians

and natives. In the context of education, results might suggest that building new schools (and addressing the congestion problem described above) can be an appropriate means to attract more children, and hence to promote integration. In addition, we related $x2y_{calls}$ with the total number of Syrians present in TECs or public schools, respectively. Integration through public education appears to be very effective: a correlation coefficient equal to 0.3 in the case of TECs grows up to 0.7 when considering Syrians educated in public schools (see Figure 7).

Discussion. Education has an important role in supporting the social integration of Syrians' refugees. On the one hand, the presence of children is positively correlated with Turkish-to-Syrian interactions (x2y). Therefore, building new schools can be an appropriate means to attract more children, and hence to promote integration in the area. On the other hand, this data shows that education at public schools over temporary education centers (TECs) improve integration, in agreement with recent policies by the Turkish Ministry of National Education.

What is the impact of Syrian refugees on natives education choices? Another important aspect to consider is the impact of refugees influx on the education choices of natives. The existence of Syrian refugees as a competitor in the labor market also affects the long-term educational behavior of native children. Given that refugees are mostly employed informally in low-skill jobs, returns to low-skill employment will be low in Turkey in the near future. This observation may generate a substantial structural change in the schooling decisions of youth in Turkey. A recent article [19] ascribe to Syrian influx a 4% increase in high-school enrollment of Turkish youths, especially from lower parental backgrounds. So, the Syrian refugee crisis affects the educational outcomes of both the natives and refugees, which will have important consequences on the long-term human capital capacity of both groups.

To address this question we analyzed the correlation between the Syrian refugees' percentage and the fraction of Turkish natives who attained a certain

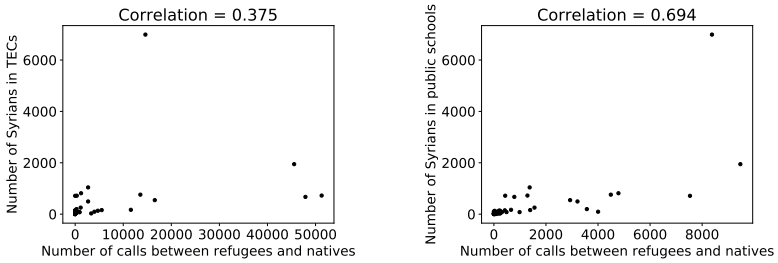


Fig. 7. Correlation between the total number of Syrians in TECs (a) or public schools (b) with respect to the total number of calls between natives and refugees.

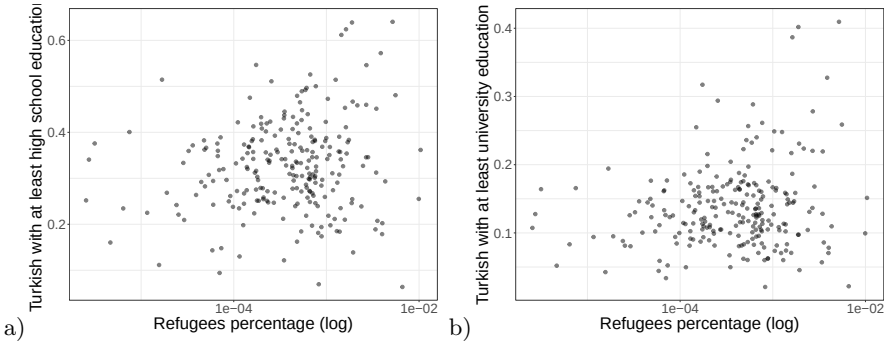


Fig. 8. Correlation between refugees percentage and natives' schooling level. **a)** high-school. **b)** univeristy

level of schooling (high-school, university). Results are in Figure 8. High school $\rho = 0.14$, p-value $= 0.001$. University $\rho = 0.19$, p-value $= 0.001$. We can see a positive correlation between schooling level and refugees' presence.

Discussion. Schooling level of Turkish youth seems positively correlated with the presence of refugees. As returns to low-skill employment diminishes, Turkish youths, especially males from lower parental backgrounds, try to improve their education in order to shoot for better jobs. Despite the general appreciation for higher education, this trend can be at the root of two main drawbacks. On the hand, this "surplus" of students tends to enroll in low-quality high schools. Therefore, the country will have a higher stock of "low-ability" youths with general high school education [19]. This may exacerbate the youth unemployment problem in the near future as the country fails to create enough jobs for higher educated individuals. On the other hand, the low-skill market becomes dominated-by / dependent-on low-skill refugees. This again creates conflicts in the society: firms become dependent on Syrians to keep labor cost down, natives are concerned by competition on the job market. This story totally goes against "a healthy integration process". For these reasons, policies intended to educate refugees, give them work permit, place them into sectors with labor shortages based on a long-term job interventions programme, minimize fierce competition for low-pay/low-quality jobs become even more important.

Does integration and education have an impact on economic well-being? As a final analysis, we tried to understand the relation between refugees' education and integration with economic well-being.

We estimate a regression with the level of (per capita) GDP in 2014 (last available official information) as dependent variable to assess the correlation between economic situation in Turkish provinces and the conditions of interactions between natives and refugees. We use as independent variables in the regression:

Table 2. Syrians integration and economic activity

	Dependent variable:		
	GDP Official statistics	GDP nightlight	
	(2014)	(2018)	(2017)
	(1)	(2)	(3)
sync1	0.801^{***}	2.855	4.942
	(0.208)	(8.493)	(8.358)
scale(prov_university)	-	1.954^{***}	2.199^{***}
	-	(0.491)	(0.484)
scale(SYRIANS.AT.PUBLIC.SCHOOLS)	0.074^{***}	-0.883^{***}	-0.974^{***}
	(0.015)	(0.271)	(0.267)
scale(Refugee.percentage)	-0.088^{***}	-0.346	-0.313
	(0.010)	(0.208)	(0.204)
prov_x2y	0.113^{***}	0.004^{***}	0.004^{***}
	(0.014)	(0.001)	(0.001)
Constant	9.057^{***}	0.894	0.639
	(0.024)	(0.917)	(0.902)
Observations	966	75	75
R^2	0.325	0.914	0.927
Adjusted R^2	0.323	0.908	0.922
Residual Std. Error	(df = 961)0.290	(df = 69) 1.593	(df = 69) 1.568
F Statistic	(df = 4; 961) 115.840^{***}	(df = 5; 69) 147.339^{***}	(df = 5; 69) 175.718^{***}
Note:	*p<0.1; **p<0.05; ***p<0.01		

- Synchronization of calls: it is a variable, computed from Dataset 1, indicating the tendency of a region to have activities in sync with other regions. Formally, for each couple of regions i and j we compute the average daily Mutual Information (MI) between the calls outgoing or incoming in that regions. In [14], this variable has been considered as an indicator of social interaction and bridging social capital. Here we focus on the synchronization between calls made by natives and refugees as a signal of integration.
- Refugee percentage
- Sum of communications between Turkish and Syrians (variable $x2y$): indicates the interaction (in terms of telephone calls) of natives and refugees
- Number of Syrians at public Schools

We find that our synchronization metric is positively related to GDP. Similarly, the value of communication interaction existing between natives and refugees ($x2y$) is positively correlated to economic activity (see Table 2). We find also a negative relationship with the percentage of refugees, which perhaps may indicate that areas where refugees have been located are actually those with lower economic activity, while on the contrary there is a positive effect of both the number of Syrians accessing public schools which might suggest that a policy of integration implemented via a shared public education could be a positive tool to improve the economic situation.

However, it is important to emphasize that a number of confounding variables can bias these results. For example, there are some economically advantaged areas (such as Izmir, Bursa, and parts of Istanbul) that never had TECs (or relatively few). There enrolment in Turkish Public Schools by Syrians was always relatively higher, while enrolment in TECs is dominant in the economically disadvantaged south-east.

Provided that national official statistics on economic activity report data only up to 2014, we run a robustness check on our results using data on night light (NOAA VIIRS/DNB dataset)in 2017 and 2018 as a proxy for economic activity [12]. Data in this case are processed at province level. We use as independent variables the same as in the previous model but also including the percentage of University students at province level (variable *prov_university*).

We find that the synchronization variable is no longer significant while the variable measuring the interaction in calls between natives and refugees is positive and significant, in line with previous result of integration supporting economic activity. Also, the presence of higher level of university students is positively related to stronger economic activity while the opposite is true for the presence of Syrians at public schools. We interpret this change of sign as a potential indicator of the presence of shadow economy (see [15]) which is not captured in the official statistics data while in turn is present in the rilevation made using night light: more Syrians in public schools identify probably areas where shadow economy run by immigrants is lower.

Discussion. Higher integration between natives and refugees seems positively related to stronger economic activity. This effect may be explained through the channels of higher levels of trust and the presence of norms of civic cooperation in areas where higher integration has been achieved and/or the coexistence between natives and refugees is more integrated.

Despite these encouraging results, we fully acknowledge that no causal claims can made from this analysis and further work is needed to control for the many confounding variables present.

5 Conclusion

Investments in education will not only work toward increasing Syrians' school enrollment rate, they will also impact their integration in Turkish society. The analysis presented in the paper provides quantitative elements supporting policies to reduce conflicts over education resources, improve Syrians education and integration and contribute to the well-being of Turkish and Syrian society.

References

1. Bansak, K., Ferwerda, J., Hainmueller, J., Dillon, A., Hangartner, D., Lawrence, D., Weinstein, J.: Improving refugee integration through data-driven algorithmic assignment. Science **359 6373**, 325–329 (2018)

2. Blumenstock, J., Cadamuro, G., On, R.: Predicting poverty and wealth from mobile phone metadata. Science **350**(6264) (2015)
3. Bogomolov, A., Lepri, B., Staiano, J., Letouzé, E., Oliver, N., Pianesi, F., Pentland, A.: Moves on the street: Classifying crime hotspots using aggregated anonymized data on people dynamics. Big Data **3**(3), 148–158 (2015)
4. Ceylan, E.S., Tumen, S.: Measuring economic destruction in Syria from outer space. https://theforum.erf.org.eg/2018/06/19/measuring-economic-destruction-syria-outer-space/ (2018)
5. Coşkun, İ., Emin, M.N.: A Road Map For The Educaion of Syrians In Turkey: Opportunities and Challenges. SETA Publications 60 (2016)
6. Coşkun, İ., Ökten, C.E., Dama, N., Barkçin, M., Zahed, S., Fouda, M., Toklucu, D., Özsarp, H.: Breaking Down Barriers Getting Syrians Children Into Schools in Turkey. SETA Publications 90 (2017)
7. Coşkun, T., Zafer, C.: The Education of Syrian Children in Turkey: Challenges and Recommendations. Egitim-Bir-Sen Stratejik Arastirmalar Merekezi (2017)
8. De Nadai, M., Staiano, J., Larcher, R., Sebe, N., Quercia, D., Lepri, B.: The death and life of great italian cities: A mobile phone data perspective. In: Proceedings of the 25th International Conference on World Wide Web. pp. 413–423. ACM (2016)
9. Deville, P., Linard, C., Martin, S., Gilbert, M., Stevens, F.R., Gaughan, A.E., Blondel, V.D., Tatem, A.J.: Dynamic population mapping using mobile phone data. Proceedings of the National Academy of Sciences of the United States of America **111 45**, 15888–93 (2014)
10. Douglass, R., Meyer, D., Ram, M., Rideout, D., Song, D.: High resolution population estimates from telecommunications data. EPJ Data Science **4 4**, 1–13 (2015)
11. Eagle, N., Macy, M., Claxton, R.: Network diversity and economic development. Science **328**(5981), 1029–1031 (2010)
12. Henderson, J., Storeygard, A., Weil, D.: Measuring economic growth from outer space. American Economic Review **102**(2), 994–1028 (2012)
13. M. Mamei, M. Colonna, M.G.: Automatic identification of relevant places from cellular network data. Pervasive and Mobile Computing Journal **31**, 147–158 (2016)
14. Mamei, M., Pancotto, F., De Nadai, M., Lepri, B., Vescovi, M., Zambonelli, F., Pentland, A.: Is social capital associated with synchronization in human communication? an analysis of italian call records and measures of civic engagement. EPJ Data Science **7**(1), 25 (2018)
15. Medina, L., Schneider, F.: Shadow economies around the world: What did we learn over the last 20 years? (2018)
16. Salah, A.A., Pentland, A., Lepri, B., Letouzé, E., Vinck, P., de Montjoye, Y.A., Dong, X., Özge Dağdelen: Data for Refugees: The D4R Challenge on Mobility of Syrian Refugees in Turkey. arXiv:1807.00523 (2018)
17. Schmid, T., Bruckschen, F., Salvati, N., Zbiranski, T.: Constructing sociodemographic indicators for national statistical institutes by using mobile phone data: estimating literacy rates in senegal. Journal of the Royal Statistical Society, Series A **180 4**, 1163–1190 (2017)
18. Smith-Clarke, C., Mashhadi, A., Capra, L.: Poverty on the cheap: Estimating poverty maps using aggregated mobile communication networks. In: ACM Conference on Human Factors in Computing Systems. Toronto, Canada (2014)
19. Tumen, S.: The impact of low-skill refugees on youth education. In: Conference on the Impacts of Refugees in Hosting Economies. Los Angeles (CA), USA (2018)
20. Wesolowski, A., Eagle, N., Tatem, A.J., Smith, D.L., Noor, A.M., Snow, R.W., Buckee, C.O.: Quantifying the impact of human mobility on malaria. Science **338**(6104), 267–270 (2012)

Towards an Understanding of Refugee Segregation, Isolation, Homophily and Ultimately Integration in Turkey Using Call Detail Records

Jeremy Boy[1], David Pastor[2], Marguerite Nyhan[1], Daniel Macguire[3], Rebeca Moreno Jimenez[3], and Miguel Luengo-Oroz[1]

[1] UN Global Pulse
firstname@unglobalpulse.org
[2] Technical University Madrid
[3] UNHCR Innovation

Abstract. Integration is a complex and gradual legal, economic, social, and cultural process; it burdens both the receiving society, and the settling population; and its outcomes are difficult to measure. Integration is also one of the main durable solutions recognized by the UN High Commissioner for Refugees. In this report, we propose a series of methodologies based on Call Detail Records (CDRs) that measure key indicators related to integration, and that can help monitor the impacts of integration programmes and policies. We also create a series of interactive visual tools that aim to support decision making in this scenario, allowing to point out interesting regions related to persons of concern. We develop a framework for measuring segregation, isolation, and homophily using population estimates reconstructed from CDRs. We also study the evolution of communication patterns and mobility traces of refugees. Finally, we corroborate these measures with records of refugee registration, and data from a large-scale cash-transfer programme.

Keywords: CDRs, Turk Telekom, D4R, Refugees, Turkey, Integration, Segregation, Homophily, Mobility.

1 Introduction

Refugee (R) integration is generally understood as a multidimensional long-term and non-linear process that is influenced by the institutional environment of the receiving society, and the personal capacities of the settling population [1]. It can also be considered as the access potential of the settling population to rightful welfare services, as detailed in the UN High Commissioner for Refugees' (UNHCR) Convention and Protocol Relating to the Status of Refugees [2]. Social and spatial segregation are key challenges to the successful implementation of integration programmes.

In this report, we develop a framework for understanding the segregation, isolation, and homophily of Rs in Turkey using Call Detial Record (CDR) data

provided by Turk Telekom [4]. These data are separated into three sets [5]. Dataset 1 ($D1$) provides antenna (or *cell tower*) traffic for the whole year of 2017 on an hourly basis. Dataset 2 ($D2$) provides fine-grained mobility information for sample groups of Turk Telekom customers, regenerated every two weeks throughout 2017. Dataset 3 ($D3$) provides coarse grained mobility information for a consistent subset of customers throughout the whole year of 2017. In all three sets, data are labeled *refugee* (R) or *citizen* (C), according to the status of the customer—see [5] for details. We first use our framework with population estimates derived from $D2$. We then study the evolution of communication patterns and mobility traces of Rs, using $D1$, $D2$, and $D3$ to better understand their potential integration, and the effect of specific integration programmes. Ultimately, the goal of this work is to develop a series of analytic tools that humanitarian stakeholders can use to better assess and monitor integration, in Turkey and elsewhere.

2 Towards Integration

The refugee situation in Turkey is of great concern to UNHCR and its partners. It is estimated that 3.5 million Rs reside in Turkey (as of late 2017) [6][7], and that 90% of them live outside of camps. Integration is therefore one of the main durable solutions recognized by UNHCR. However, measuring integration is particularly challenging. The Migration Policy Group has developed an Integration Evaluation Tool for UNHCR [3], which includes four main types of indicators: 1) policy indicators, 2) administrative indicators, 3) financial inputs, and 4) outcomes. Here, we focus on 4). UNHCR also uses indicators of social integration categorized into legal, social, economic, and political dimensions [3][33]. Here, we focus specifically on *residential segregation*, which we break into *segregation*, *isolation*, and *homophily* because these have well-established measures [35]; and *participation in local activities/groups*, which we measure through *communication patterns*, and *mobility traces*.

2.1 Indicators

Segregation Segregation is generally defined as an imposed restriction on the interaction between people considered different [34]. It arises both from spatial conditions that force people into isolation, and from social dynamics that lead people to interact more with others they are similar to. Most measures of segregation focus on spatial conditions. For example, the *residential segregation* indicator used by UNHCR [33] is defined as "the degree to which refugees live in a [geographic unit] with many or few other refugees, generally and from [their] home country of origin." This strongly relates to measures of *dissimilarity* (or *evenness*), which consider the spatial distribution of different groups among units in an area. Segregation is low when majority and minority populations are evenly distributed within that area. More specifically, dissimilarity measures the percentage of the population that would have to relocate for each unit to have the same group distribution as the overall area [19].

Isolation Isolation refers to the extent to which distinct groups share common residential areas [35]. Like dissimilarity, it focuses on the spatial separation of people. In fact, isolation is often considered a co-dimension of segregation. However, isolation relates to the perception, or the experience of segregation. Measures of isolation (or of *exposure*) indicate the probability that a member of one group will meet with a member of another group [19]. Isolation is low when the probability of exposure is high.

Homophily Homophily is the pervasive tendency people have to associate with others who share similar traits [21]. It has been documented broadly across characteristics like age, race, gender, religion, and profession [36]. According to Carrington, it can be influenced by: 1) the composition of the population; 2) personal preferences; and 3) social structures (i.e., opportunity structures that favor homophilous associations.) [37]. Interestingly, homophily has already been studied using CDRs [17]. We build on this work, using homophily as an initial measure of the social segregation of Rs.

Participation According to the *participation in local activities/groups* indicator used by UNHCR [33], participation is an aspect of civic involvement. The description of the indicator stresses the importance of looking at group memberships that encompass the wider local population, which help expand Rs' social networks. Previous work on data-driven insights for humanitarian action has shown the potential of CDRs and other big data sources to estimate the social networks, mobility, and socio-economic profiles of groups of people [9][10][11][12][13][14][15]. It has also shown that CDRs can support monitoring and decision making processes [16]. In this light, participation could be measured through a temporal analysis of the evolution of Rs' communication patterns—e.g., through measuring the variations in the entropy of networks. Further indicators derived from the mobility of cell phone users could provide geographical information on R concentrations, and how these are distributed throughout the country. These indicators could then typically help measure Rs relative difficulty to access welfare services.

3 Results

3.1 Segregation, Isolation, and Homophily

Segregation We first explored the possible segregation of Rs within districts, using population estimates derived from the high resolution data in the voice call subset of $D2$ ($D2_v$). Fig. 1 shows the *dissimilarity* of districts, i.e., the uneven distribution of Rs and Cs at each cell tower in a district. The sparsity of the map reflects a lack of sufficient data to reconstruct population estimates for all areas. Interestingly, we see that levels of dissimilarity are generally low across the country, indicating a rather even distribution of R and C callers within each district. Nevertheless, several locations in the South East, near the Syrian

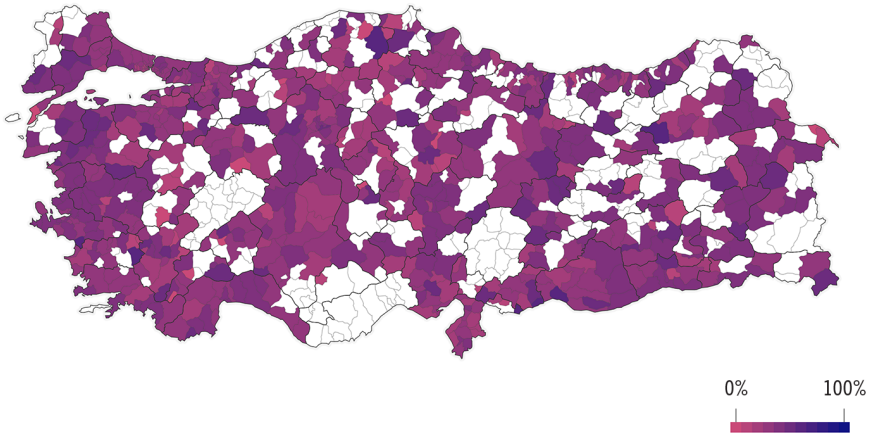


Fig. 1. Dissimilarity at district level. Magenta (0%): the refugee (R) and citizen (C) populations are evenly distributed between cell tower coverage areas within the district (i.e., within the voronoi polygons of each tower). Blue (100%): the R population is completely segregated within the district. The gradient between indicates the percentage of the population (R or R) that would need to be resettled to achieve an even distribution of population by type.

border show relatively high levels of R segregation. For example, Rs in districts like Elbeyli (Kilis), Karkamis (Gaziantep), or Suruc (Sanliurfa) seem to live in areas where there are low concentrations of Cs[4]. Indeed, 60% of the population in these districts would have to be relocated to reach evenness. Note that these high levels of segregation may be due to the attested presence of R camps [29].

Isolation We then looked at the possible isolation of Rs, specifically in the more segregated areas identified in Fig. 1. Fig. 2 shows the *exposure* of Rs to Cs within the different estimated areas of reach of cell towers in a district. All three districts mentioned above show low levels of exposure, suggesting that Rs are not only segregated in those areas, but that they might also experience the sociological effects of that segregation, since they likely do not often interact with Cs [19]. This may be due to the proximity with the Syrian border, and to the aforementioned presence of camps. Further, other districts show up more prominently as areas of R isolation on this map than they do on the map of segregation, like Arguvan (Malatya), Divrigi (Sivas), or Cal (Denizli). Nevertheless, both maps show overall common trends.

[4] Technically speaking, according to the way we determine the location of callers, a more precise formulation would be: Rs in these districts *radiate* from areas where there are low concentrations of Cs (see Sec. 4).

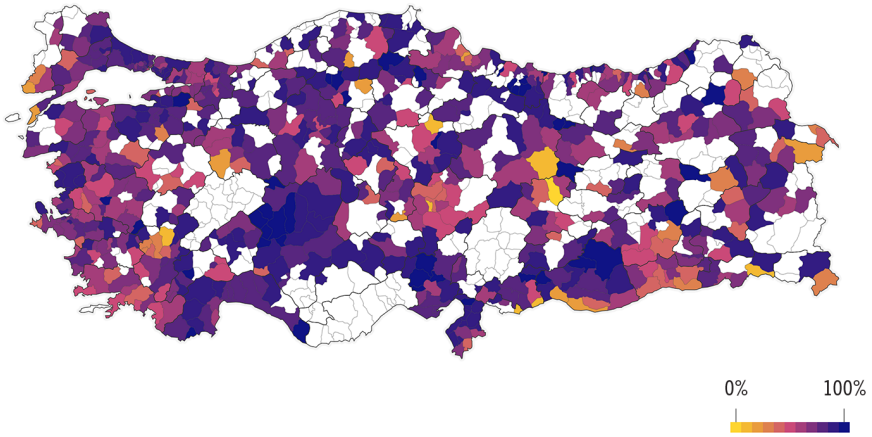
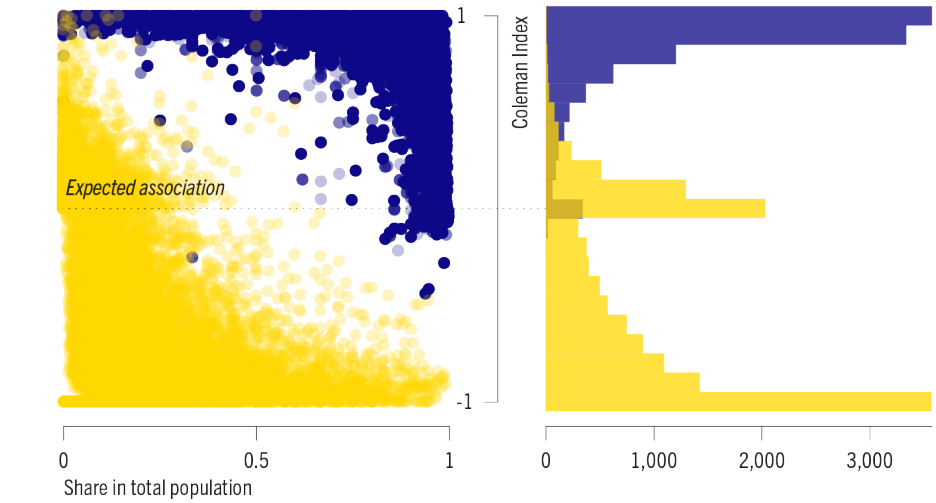


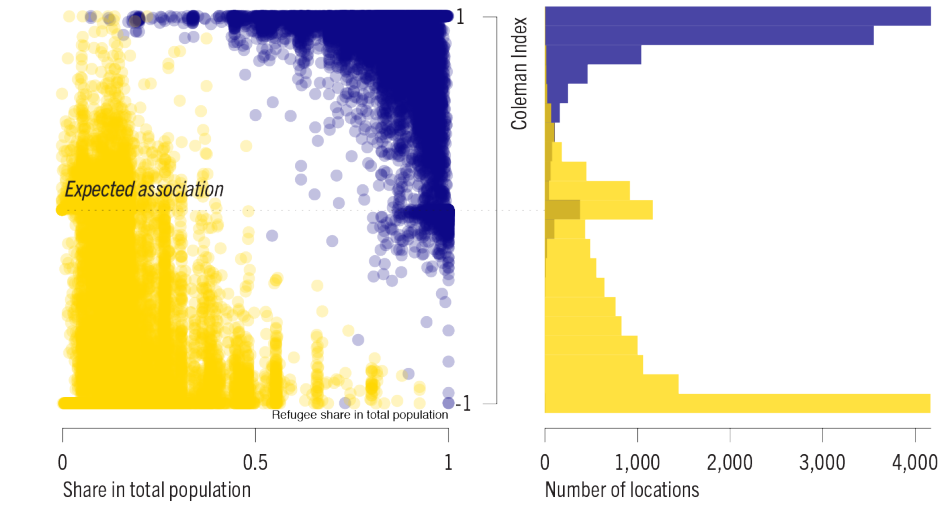
Fig. 2. Exposure at district level. Gold (0%): the probability that a refugee (R) would interact with a citizen (C) within a two-week time frame is 0. Blue (100%): the probability of interaction is 1.

Homophily Continuing, we explored the *homophily* of Rs and Cs at the cell tower level. Following the description in [17], we calculated the average homophily of individuals from both populations, using two different approaches to estimating the number of individuals with which a caller may be in close physical proximity. We refer to these approaches as the *simple count* and *buffered area* approaches (these are detailed in Section 4.1).

Fig. 3 shows the average homophily of R (Gold) and C (Blue) callers at each cell tower for which we had sufficient data, over the share of Rs / Cs in the population in close proximity to those callers. Chart [A] uses the *simple count* population estimate, while chart [B] uses the *buffered area* estimate. Three things stand out. First, whatever the approach to estimating population, Rs interact primarily with Cs in most locations (high *heterophily*), and Cs interact mainly with other Cs (high *homophily*). Second, it is clear that the more the population share of Rs is high, the more their tendency to interact with Cs is high. This is particularly interesting, as it goes against the basic assumption that people are generally more inclined to interact with others that are like them [21]. In addition, while there is no easily modeled trend for homophily over population share here, it seems this information could be useful for separating Rs from Cs in a classification algorithm. Third, the comparison of both charts shows how sensitive the measure of homophily is to population estimates—especially around the expected rate of interaction (or *association*) for Rs (index value of 0).

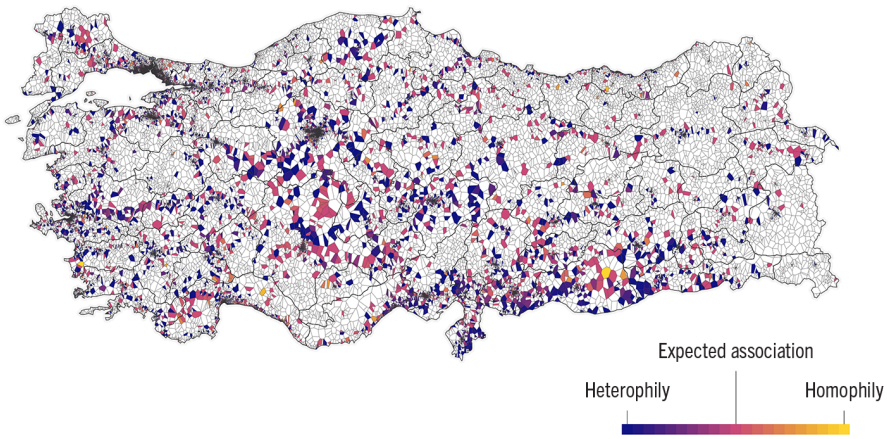


[A]

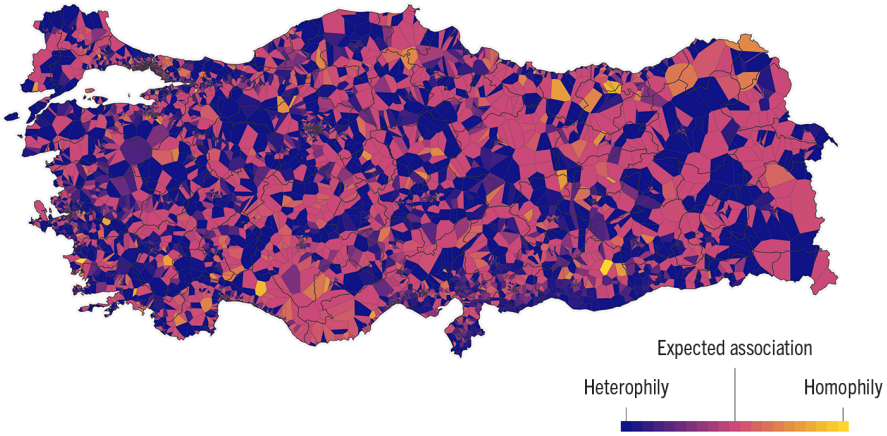


[B]

Fig. 3. Homophily at cell tower level of refugees (R) and citizens (C) based on two different approaches to estimating the caller population within reach of a cell tower. The scatterplots show the average homophily for each cell tower over the population share of Rs / Cs. The barcharts show the distributions of the index. [A]: the population is simply estimated by counting the number of individuals at each cell tower (*simple count* approach). [B]: the population is estimated by adding the number of individuals at cells towers within a 22 km radius of each cell tower (loosely following the implementation description in [17]—*buffered area* approach). Gold: homophily of the R population. Blue: homophily of the C population.



[A]



[B]

Fig. 4. Homophily at cell tower level, based on the *buffered area* population estimate. [A]: the voronoi polygons for each tower in the base station location dataset are represented to show the overall sparsity of the data. [B] the voronoi polygons are recalculated based on the subset of towers for which the index can be calculated. [B] is less precise than [A], but is easier to read. Blue (-1): refugees (Rs) interact most with citizens (Cs) (high heterophily). Magenta (0): Rs interact with Cs at an expected rate, considering their relative share in the total caller population. Gold (1): Rs interact mostly with themselves (high homophily).

Fig. 4 further shows homophily at the cell tower level with more or less resolution[5]. Interestingly, we see in map [B] that areas near the Syrian border (including the Elbeyli, Karkamis, and Suruc districts) show high levels of R heterophily. Note this observation holds whatever the population estimation approach. We hypothesize that similar to the meeting bias identified by Currarini et al [20], there may be *dependency bias* at play in the way Rs interact with Cs. Rs are likely to have to (at least) call Cs for e.g., work, housing, or community activities. On the contrary, Cs likely do not depend on Rs for such basic necessities. That said, these results might also imply that current integration programmes are successful in providing opportunity structures that favor heterophilous contact. However, we do stress that the results presented in this section only account for Turk Telekom customers. Further modeling of the population estimates using CDR data is required to draw any definitive conclusion on segregation, isolation, and homophily.

3.2 Participation through Communication and Mobility

Communication patterns We explored the development of Rs' communication networks throghout the year by analyzing the evolution of the volume of R communications at each cell tower. We used information from $D1$, both for voice calls ($D1_v$) and SMS ($D1_s$). We differentiated three levels of R activity at a cell tower: *low* ($\leq 33\%$), *medium* ($> 33\%, \leq 66\%$), and *high* ($> 66\%$).

Fig. 5 shows cell towers across Turkey labeled according to refugee activity in $D1_v$ for the months of February and November. Interestingly, most locations with high levels of R activity match the areas in which high numbers of Syrian Rs are registered [23]. However, some sites in the East (Batman area) and North are not accounted for by UNHCR data. These unmapped locations with high levels of R activity might be useful for identifying new concentrations of Rs. Further, we see that many sites with high levels of R activity are located in large population areas, like Istanbul. These could be considered for a deeper investigation of segregation in cities.

Fig. 6 gives an overview of communication patterns at each cell tower in $D1_v$. It shows the aggregated connectivity between the three types of R-activity-labeled locations for each day of the year. We see that the number of locations with medium and high levels of R activity increases between August–November. Locations with medium levels of R activity tend to increase their connectivity with sites with low levels of R activity, while locations with high levels of R activity tend to increase their connectivity with other sites of the same type. These patterns seem temporally consistent with the increase in payouts of the Emergency Social Safety Net (ESSN) cash programme [22], and could be considered a proxy indicator for social integration, provided further validation data.

We conducted the same analysis with $D1_s$.We found a similar increase between August–October, mainly for locations with high levels of R activity, outgo-

[5] All maps and visualizations discussed here are available in high resolution at https://d4r-turktelekom.unglobalpulse.net/.

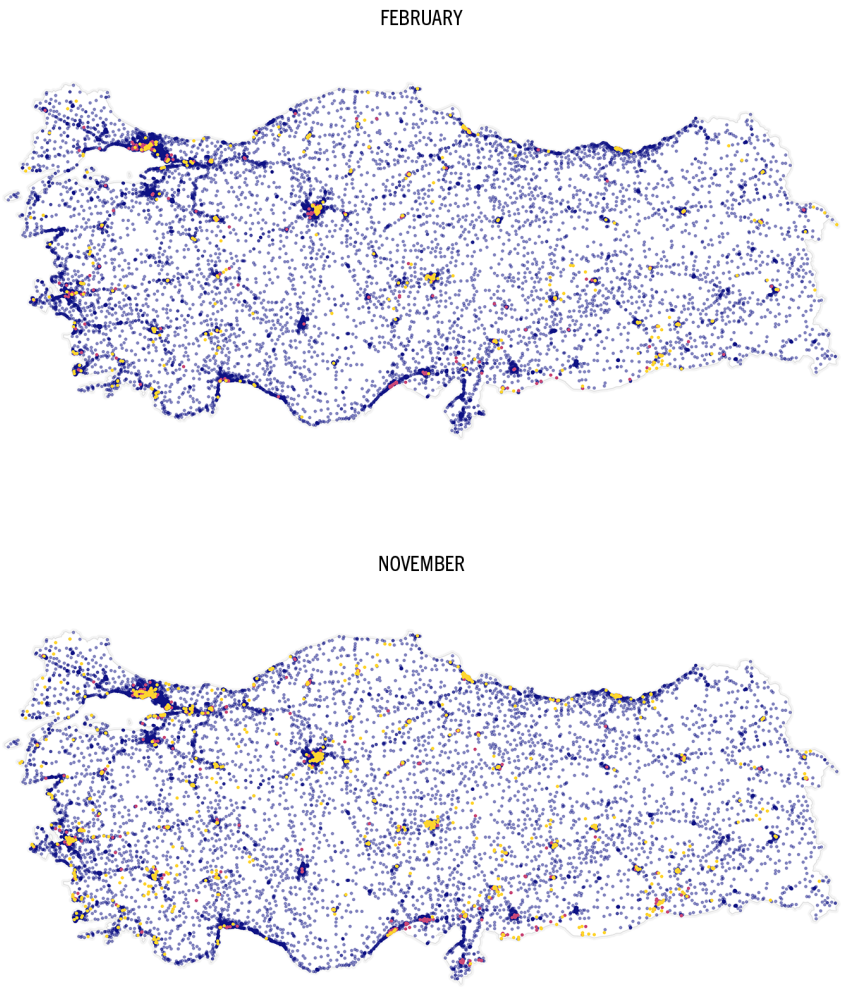


Fig. 5. Labelled sites for the voice subset of Dataset 1 in February (top) and November (bottom), where the largest number of refugee (R) spots were detected. Each dot represents a cell tower. Blue: less than a third of all calls involve Rs (low level of refugee activity). Magenta: between a third and two thirds of all calls involve Rs (medium level of refugee activity). Gold: more than two thirds of all calls involve Rs (high level of refugee activity).

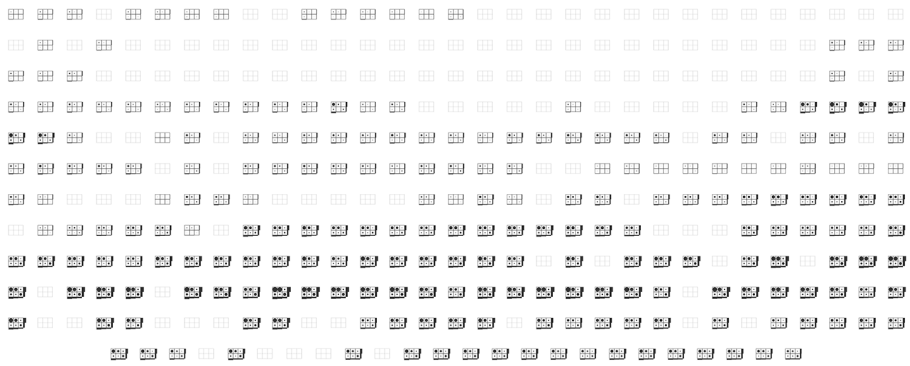


Fig. 6. Daily flux of outgoing (medium and high—rows in each matrix) and incoming (low, medium, and high—columns in each matrix) volumes of communication. Volume is encoded in the size of each circle. Grayed out plots indicate missing data.

ing to sites of the same type, and to sites with low levels of R activity. This suggests we can assume SMS is a relevant communication system for Rs in Turkey. Further, the increase among locations with high levels of R activity indicates that the sites with high concentrations of Rs tend to self-communicate.

Entropy We then computed the entropy of connected sites derived from $D1$. Fig. 7 focuses on the provinces of Mersin, Istanbul, and Gaziantep. We see an increase in entropy throughout the year for locations with medium levels of R activity. A peak is reached around September, which corresponds with the ESSN payouts timeline. Meanwhile, sites with low levels of R activity seem more stable.

Mobility hotspots Using $D2_v$ again, we explored Rs' mobility traces at the cell tower level within the two-week time frame imposed by the sampling of the data (see [5]). Fig. 8 shows where Rs travel to and from, and how much they move between locations. We immediately see that the majority of movement is short distance (large circles that indicate movement between clusters of towers located in close proximity), within urban areas, like Istanbul / Bursa (the biggest circle clusters towers in both cities), Mersin (Icel), Ankara, and Izmir. We refer to these locations as mobility *hotspots*.

Comparing this map with the ones in Fig. 5, we further see that some hotspots like Mersin do not show up clearly as areas with high levels of R activity—especially earlier in the year. The opposite is also true: areas with high levels of R activity show up in the East of the country, where there are no mobility hotspots. This concentration of R activity in areas with no mobility could indicate the presence of vulnerable groups that do not have sufficient resources to move. Conversely, mobility hotspots where levels of R activity are low could indicate short-term stays. Confirming this is important for understanding how Rs flow through the country.

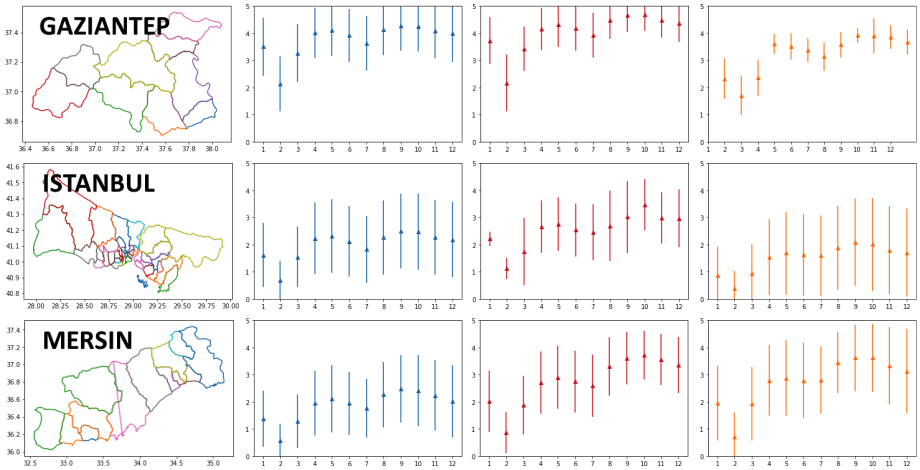


Fig. 7. Evolution of the entropy of sites connected for the cell towers (mean and std) computed form Dataset 1. Blue: entropy for low refugee (R) activity sites. Magenta: entropy for medium R activity sites. Gold: entropy for high R activity sites. y-axis: entropy. x-axis: months.

Accessibility Finally, we explored the relative accessibility (in terms of travel distance and time) of different districts by creating an interactive isochrone map from the lower resolution data in $D3$. Fig. 9 shows both the spatial and temporal distances between the Erdemli district (Mersin) and all other districts to and from which callers in $D3$ moved.We see that some districts, like Siverek (Sanliurfa), are temporally very close to Erdemli, even though they are spatially far apart. Conversely, Elbeyli (Kilis) is much further temporally than it is spatially. This can be used as a proxy for how easy it is to travel from one area to another. It could also be used for measuring the access capacity of Rs to (located) rightful welfare services, provided further disagregation of the data between Rs and Cs.

4 Methodology

The goal of this work is to develop a methodology and a series of interactive visual tools for improving different humanitarian organisations' staffs' understanding of the outcomes of integration programmes. Here, we present the methodology we deployed to obtain the results presented in the previous section.

4.1 Measuring Segregation, Isolation, and Homophily

We focused on the voice subset of $D2$ for measures of segregation, isolation, and homophily. This was intended to get the highest resolution possible. We initially inspected the composition of both the voice ($D2_v$) and sms ($D2_s$) subsets of $D2$, and found that only **0.1%** of unique individual identifiers in $D2_v$ and **0.27%** in

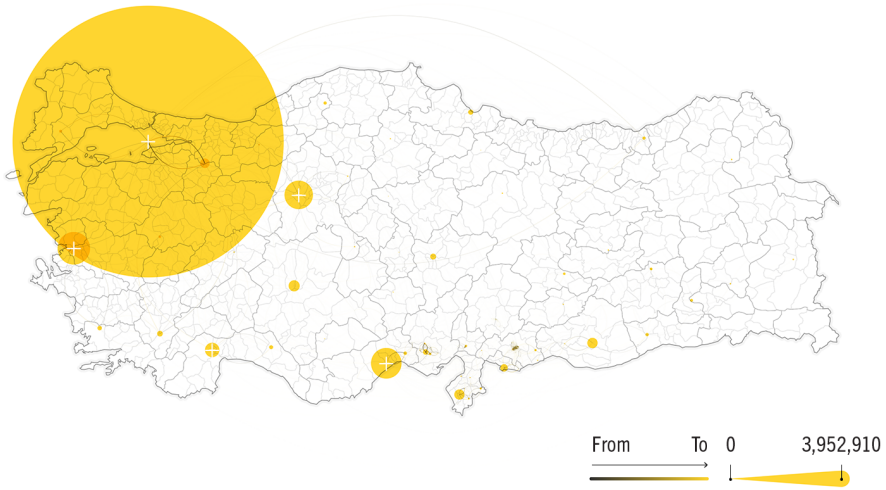


Fig. 8. Refugee mobility between cell towers derived from Dataset 2 for the whole year. For representation practicality, towers are aggregated into 300 clusters using a Kmeans algorithm [31]. Size encodes the volume of movement. Large gold circles indicate high volumes of movement within clustered towers. Crosshairs show the center points of large cluster. Arcs show movement between clusters. Direction is encoded using a gray (from) to gold (to) gradient.

$D2_s$ overlapped. Considering it unlikely that only such a minuscule proportion of the sampled population would both make calls and send SMS during the two-week sampling time frame (as opposed to only making calls or only sending SMS), we concluded this very small overlap was likely an artifact of the random *id* assignment, and that the records in $D2_v$ and $D2_s$ should not be associated.

Estimating Caller Populations We calculated the number-of-calls-at-a-cell-tower-weighted center of mass for all individuals' (*callers*—N = **3,298,947**) movements. We then determined which cell tower (among all those visited by a given individual) was closest to that center of mass, and assigned the caller to it. We also calculated the area of movement of each caller (the area of the polygon connecting all the cell towers that picked up one of their calls), as well as their radius of gyration (ROG), following the simplified procedure described in [8].

Fig. 10 shows a comparison of the bootstrapped mean areas of movement of Rs (N_R = **588,347**) and Cs (N_C = **2,710,600**). Fig. 11 shows a comparison of the bootstrapped mean ROG of Rs and Cs. While there is clear evidence in both figures that Cs cover more ground than Rs within the two-week time frame samples, the differences are quite small. For example, the bootstrapped mean ROG is **19 km** 95%CI[18.5, 19.5] for Rs, and **22.1 km** 95%CI[21.8, 22.4]. This is consistent with the idea that most movement is concentrated in hotspots, as

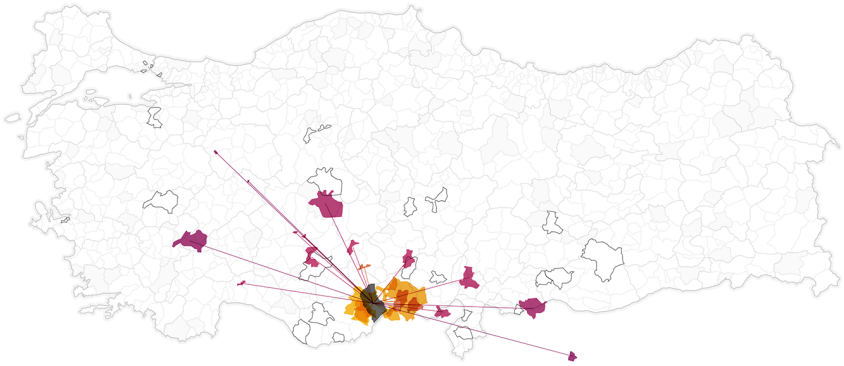


Fig. 9. Isochrone map of temporal distances between districts derived from Dataset 3. The node-link graph layout shows the temporal distances between the district of interest (here Erdemli, Mersin) and other districts for which data are available. It is optimized to relate maximum distances in space and time, as well as to preserve direction between the district of interest and the districts it is connected to. The color code goes from gold, to magenta, to blue to indicate short to far temporal distances. It is redundant with the spatial encoding of the graph.

seen in Fig. 8. In addition, it is noteworthy that the average distances traveled are only slightly above the mean district radius (≈**16 km**—this value is based on the average land area of districts in Turkey, considered as a disk). This suggest that on average both Rs and Cs in the caller population hardly travel outside the area of a district within a two-week time frame.

We then counted the population at cell tower level in two different ways. First, we simply counted the number of callers assigned to each tower. We refer to this approach as the *simple count* approach. Second, we estimated the number of callers any individual might encounter around the cell tower s/he was assigned to by summing up all the caller counts at cell towers located within the average

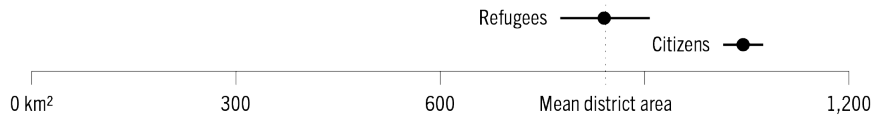


Fig. 10. Bootstrapped mean (10k replicates of random 10% samples of the data) area covered by refugees and citizens, with a reference to the mean district area in Turkey.

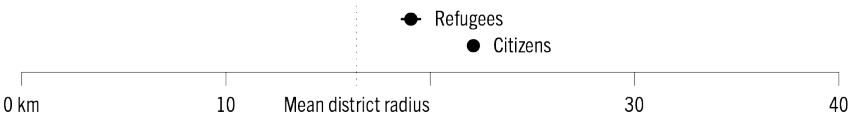


Fig. 11. Bootstrapped mean (10k replicates of random 10% samples of the data) radius of gyration for refugees and citizens, with a reference to the main district radius—calculated using the mean district area, and considering it as a disc.

ROG of the total caller population (≈**22 km**). We refer to this approach as the *buffered area* approach. This produced the two population estimates we compare in Fig. 3.

Calculating segregation Using the *simple count* population estimate, we first calculated the *index of dissimilarity* [27] at district level. This is defined as:

$$D = \frac{1}{2}\sum_{i=1}^{n}|\frac{c_i}{C_T} - \frac{r_i}{R_T}| \tag{1}$$

where n is the number of spatial units (cell tower coverage areas here); c_i is the number of Cs at cell tower i; C_T is the total number of Cs in the higher administrative unit (the district here); r_i is the number of Rs at cell tower i; and R_T is the total number of Rs in the district. Results are shown in Fig. 1. Note that the *index of dissimilarity* has several known limitations [19][30], but it is still widely used, and is recommended for measuring segregation [19].

Calculating isolation We then calculated the *index of exposure* [19], still using the *simple count* population estimate. There are several variations of this measure [32]. The one we worked with is defined as:

$$R_r c = \sum(\frac{n_{ir}}{Nr})(\frac{n_{ic}}{n_i}) \tag{2}$$

where n_{ir} is the number of Rs at cell tower i; n_{ic} is the number of Cs at cell tower i; N_r is the number of Rs in the district; and n_i is the total population at cell tower i. Results are shown in Fig. 2.

Calculating homophily Finally, we used both the *simple count* and *buffered area* population estimates to calculate the *Coleman index of homophily* [24]. Blumenstock & Fratamico offer some guidance on using the latter population estimation approach in what is essentially a calculation of Coleman's index using CDRs [17]. This index is defined for directed networks and focuses on the *outdegree* of nodes. It can be broken down into four distinct parts. The first part is simply the relative population share of a group of type l (here either Rs or Cs) within the total population. This measure is denoted $w_l = \frac{n_l}{n}$. The second

part is a basic measure of homophily defined by the ratio of outdegree links that connect individuals of a type l to others of that type, over the total number of outdegree links that connect individuals of type l to others of any type (here Rs + Cs). This measure is denoted $H_l = \frac{m_{ll}}{m_l}$. The third part is a comparison between H_l and what can be expected given a uniform random assortment of the population. The difference is called *excess homophily*, and is denoted $H_l - w_l$. The last part consists in normalizing the excess homophily by its maximal value $1 - w_l$, as groups that represent a very large share of the population (w_l) will never experience large excess homophily [21]. The index is then defined as:

$$C_l = \begin{cases} \frac{H_l - w_l}{1 - wl} & \text{if } H_l \geq w_l \\ \frac{H_l - w_l}{wl} & \text{if } H_l < w_l \end{cases} \tag{3}$$

To calculate H_l, we separately counted the number of calls made by Rs at each tower to other Rs (whatever their location) creating a $R2R$ dataset, and calls made by Cs to Cs creating a $C2C$ dataset. We then calculated w_l first using the *simple count* population estimate, then using the *buffered area* estimate. Results are shown in Figs. 3 and 4.

4.2 Measuring Participation through Communication and Mobility

Establishing communication patterns We labelled each cell tower according to the level of R activity in $D1$, using 33% and 66% thresholds (see Sec. 3.2). The result is shown in Fig. 5. We then created a daily communications graph (3x3), in which nodes corresponded to the labeled cell towers, and links represented the total duration of calls between two nodes throughout a day. We also created a monthly version of the graph, with cumulative volumes of calls as links. Results for $D1_v$ are shown in Fig. 6.

Visualizing mobility hotspots We identified travels as the relative displacement between two consecutive records in the CDRs of a same caller in $D2_v$. We separated out R and C travels, and aggregated them for each population according to their origin–destination, creating two directed graphs of all R and C movements. Nodes corresponded to cell tower locations, and links represented the total number of travels between two nodes throughout the year. We then clustered origin and destination cell tower locations into 300 clusters using a Kmeans algorithm [31] to make the graph more legible on a map. The result is shown in Fig. 8.

Visualizing accessibility We identified the travels made by all individuals in $D3$, using the same method as above. We aggregated travels according to their origin–destination, and filtered out the ones for which there were less than five occurences in the data. This resulted in a subset of **226,600** travels, connecting

129 sites. We then created a directed graph (129x129) of movements between locations, where nodes corresponded to districts, and links represented the lowest 10^{th} percentile of timestamp differences between two consecutive records used to identify the travel. The result is show in Fig. 9.

5 Discussion

We have presented a framework for understanding and quantifying the segregation, isolation, homophily, and ultimately integration of Rs in Turkey using CDR data. In doing so, we have explored the potential of the three datasets provided by Turk Telekom for the year 2017 [5].

$D1$ enabled us to analyze communications within and between R and C populations. We discovered concentrations of R populations through high levels of R activity. This can be helpful for targeting persons of concern who need support for integration. The consistency of our results with the information gathered about the ESSN cash programme is also promising for the design, monitoring, and assessment of humanitarian initiatives. Other descriptors of the communication graph dynamics, such as the entropy of links for each node, can be useful proxy indicators for integration when interpreted in the context of segregation, accessibility to resources, and participation in social activities. $D2$ enabled us to measure the segregation, exposure, and homophily of Rs at a high resolution, using population estimates reconstructed from the CDRs. Among other things, the *dissimilarity index* could be further investigated to plan population reallocation. The population estimates can also be useful for optimizing resource allocation, and for triggering early-warning systems if irregular situations are detected. Visualizing the mobility of Rs (Fig. 8) also helped us identify mobility hotspots. Finally, $D3$ enabled us to explore the relative accessibility of districts. It also allowed an initial assessment of internal migrations (not presented in this report due to space limitations).

UNHCR registration data, and data from the ESSN programme further provided useful references to humanitarian support in the country. However, more validation data is necessary for a deeper quantified understanding of R integration. Additional historic information would also be useful for establishing seasonal baselines, and to compensate for biases linked to seasonal patterns.

5.1 Limitations

Population Estimation One of the main issues in calculating segregation, exposure, and homophily is that these measures are extremely sensitive to population estimates. All results shown in Figs. 3 and 4 only account for caller populations. Further, Fig. 3 highlights how different ways of estimating the population at tower level can lead to important variations in the Coleman index.

Deville et al [25] have proposed a model for estimating populations based on CDRs. The model is created in two steps. First, the caller density (σ_c) is calculated at the tower level, using a voronoi tessellation to estimate the area

of coverage of each tower. Densities are then added for each administrative area at the level of interest (e.g., at district or province level), and are adjusted to take into account the area of coverage of the different cell towers within the boundaries of the administrative area, as well as the administrative area itself. Second, the resulting caller densities are compared to ground truth census data using the following equation:

$$\rho_c = \alpha \sigma_c^{\beta} \tag{4}$$

where ρ_c is the census-derived population density for administrative unit c; σ_c is the calculated number of callers for the corresponding administrative unit, established in the first step; α is a scale ratio; and β is the superlinear effect of ρ_c on σ_c. For practicality, this equation can be transformed into $log(\rho_c) = log(\alpha) + \beta log(\sigma_c)$. The values α and β can then be obtained through a standard weighted linear regression (see [25] for details). Results of this regression for our *simple count* population estimate at province level are shown in Fig. 12 for the total population [A], and the R population [B].

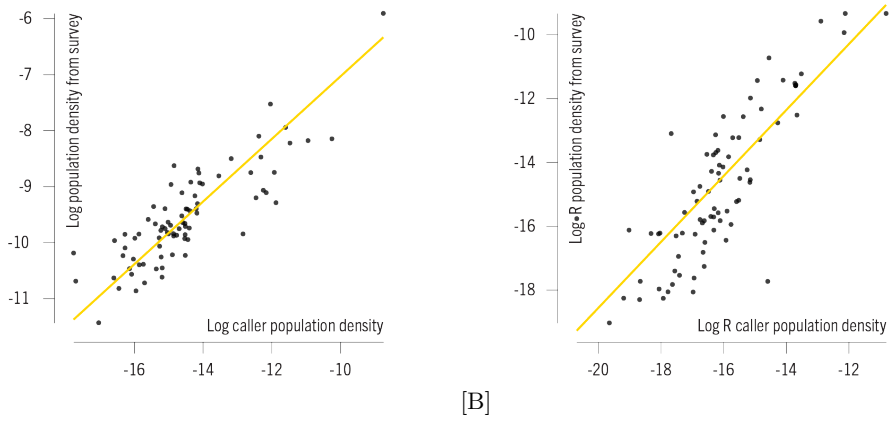


Fig. 12. Fit for the regression models of population densities from census (March 2017) over calculated caller population density (log–log scale). Each dot represents a province. The log densities are negative because the actual values are very small—they were calculated per square meter. [A] Total population density (Refugees + Citizens): R^2 = **0.87**, Slope = **0.56**, Intercept = **-1.45**, $p <$ **0.01**. [B] Refugee (R) density: R^2 = **0.77**, Slope = **1.03**, Intercept = **2.09**, p = **0.03**. Finally, Citizen (C) density (not shown here due to space limitations): R^2 = **0.84**, Slope = **0.54**, Intercept = **-1.66**, $p <$ **0.01**.

Eq. 4 has proven effective in high-income countries like France and Portugal [25], as well as in developing countries such as Senegal [26]. Fig. 12 seems to

indicate that the relationship also holds for Turkey—as attested by $R^2 > \mathbf{0.7}$ in both cases. However, while these initial results look promising, the inferred population estimates are far from perfect when compared to the census reference data. For example, the mean absolute percentage error (MAPE) for the total population estimates is **23.4%**. In addition, Pestre et al [26] have found that simply projecting the regression coefficients obtained at one administration level down to lower administrative levels can more than double the MAPE, which is an issue if we are to attempt to model the population at cell tower level based on coefficients obtained at the province level. We intend to continue exploring the modeling of population estimates by factoring in the known proportion of Turk Telekom customers within the general population (at province level), and by testing the model at lower administrative levels (e.g., at district level)—although ground truth information on R populations at these levels may be harder to get. We also intend to turn to $D3$ for modeling the population at district level, even though this means we will loose the cell tower resolution.

Sampling Effects in $D2$ We calculated the populations at tower level using $D2$ disregarding any possible effect of the two-week sampling method employed to create the dataset. One immediate pitfall of this approach is that it does not account for potential seasonality. The way people interact with each other may change over the course of the year. Further, this two-week sampling prevents from accounting for the yearly behavior of individuals. People (be they R or C) may move and encounter other populations, which may affect the segregation, exposure, and homophily indexes. We plan to explore the impact of these effects in the future.

6 Conclusion

Throughout this work, we have uncovered useful signals, created maps, and identified patterns that can help understand and monitor R segregation and integration dynamics at a high spatio-temporal resolution using CDRs. The report has shown that the potential of CDRs to improve the design of policies, to optimize resource distribution and allocation, and to monitor population is promising. This contributes to establishing a more comprehensive way of interconnecting humanitarian data, which will allow a more precise evaluation of integration programmes in the future.

As next steps, we intend to perform a multi-scale analysis of R mobility in Turkey, integrating $D2$ and $D3$ in a systematic way, and considering the effects of the aggregation in $D3$. Daily-based mobility could be an indicator of segregation depending on the geographical context. $D3$ could also reveal mobility as a coping strategy developed by Rs to facilitate their integration, or it could show structural migratory patterns of forcibly displaced populations throughout the country.

We further propose to use accessibility to public centers responsible for delivering basic welfare services (e.g., hospitals, schools, public administrations,

courts, etc.) as a proxy measure for the well-being outcome of integration. We embrace Cascetta et al's behavioral definition of accessibility, which describes it as "the expected number of opportunities 'available' for a subject to perform an activity, where 'available' means that the opportunity is perceived as a potential alternative to satisfy one's needs, and it can be reached given the spatio-temporal constraints of the individual's schedule" [38]. We will use a combination of the tools created using CDRs and other datasets to determine the accessibility of services to Rs residing in Turkey. We will then use these metrics to improve our understanding of integration from the 'personal capacities of the settling population' perspective—a human rights-based approach.

Finally, we will create retrospective simulations of real scenarios that involved planning and monitoring integration programmes to explore how data-driven decisions could have been made, or better informed by our framework and interactive visual tools, and we will use these insights to identify and measure the true operational value of mobile data in this context. These tools will be useful for humanitarian stakeholders who want to monitor humanitarian actions, and improve the distribution of financial aid.

References

1. Valtonen, K.: From the margin to the mainstream: Conceptualizing refugee settlement processes. In: Journal of refugee studies 17(1) (2004)
2. UN General Assembly: Convention Relating to the Status of Refugees. In: Treaty Series, 189 (1951)
3. http://www.refworld.org/pdfid/532164584.pdf
4. http://d4r.turktelekom.com.tr/presentation/data
5. Salah A. et al: Data for Refugees: The D4R Challenge on Mobility of Syrian Refugees in Turkey. (2018)
6. https://data.humdata.org/group/tur
7. https://goo.gl/FpAHeC
8. https://goo.gl/WBSF3m
9. Pastor-Escuredo D. et al: Flooding through the Lens of Mobile Phone Activity. In: 4th IEEE Global Humanitarian Technology Conference. San Jose, CA, USA (2014)
10. Mubangizi M., et al: Malaria surveillance with multiple data sources using Gaussian process models. In: Proceedings of the 1st International Conference on the Use of Mobile ICT in Africa. Stellenbosch, South Africa (2014)
11. Zufiria P., et al: Mobility profiles and calendars for food security and livelihoods analysis. In: Actes du Challenge D4D, Netmob (2015)
12. Nyhan M., et al: Exposure Track—The impact of mobile-device-based mobility patterns on quantifying population exposure to air pollution. Environmental Science Technology 50(17) (2016)
13. Nyhan M., et al: Predicting vehicular emissions in high spatial resolution using pervasively measured transportation data. Atmospheric Environment 140 (2016)
14. Pramestri ZA. et al: Estimating the Indicators on Education and Household Characteristics and Expenditure from Mobile Phone Data in Vanuatu. Netmob (2017)
15. UN Global Pulse and UNHCR: Social Media and Forced Displacement: Big Data Analytics and Machine-Learning. White Paper (2017)

16. Blondel V., Decuyper A., Krings G.: A survey of results on mobile phone datasets analysis. In: EPJ Data Science (2015)
17. Blumenstock J., Fratamico L.: Social and spatial ethnic segregation. In: Proceedings of the 4th Annual Symposium on Computing for Development, 11. Cape Town, South Africa (2013)
18. Alaa AM., Ahuja K., van der Schaar M.: Evolution of Social Networks: A Microfounded Model. arXiv preprint arXiv:1508.00205. (2015)
19. Weinberg D., Iceland J., Steinmetz E.: Measurement of Segregation by the U.S. Bureau of Census. In: Racial and Ethnic Residential Segregation in the United States: 1980-2000.
20. Currarini S., Jackson MO., Pin P.: An Economic Model of Friendship: Homophily, Minorities and Segregation. In: Econometrica 77(4) (2009)
21. Currarini S., Matheson J., Vega-Redondo F.: A simple model of homophily in social networks. In: European Economic Review. (2016)
22. https://data2.unhcr.org/en/documents/download/62568
23. https://data2.unhcr.org/en/documents/download/61354
24. Coleman j.: Relational analysis: The Study of Social Organizations with Survey Methods. In: Human Organization, 17(4) (1958)
25. Deville P. et al: Dynamic population mapping using mobile phone data. In: Proceedings of the National Academy of Sciences 111(45) (2014)
26. Pestre G., Letouze E., Zagheni E.: The ABCDE of big data: assessing biases in call-detail records for development estimates. In: Annual World Bank Conference on Development Economics (2016)
27. Duncan OD., Duncan B.: A methodological analysis of segregation indexes. In: American sociological review 20.2 (1955).
28. http://enceladus.isr.umich.edu/race/seg.html
29. http://www.refworld.org/pdfid/5a2fb20d4.pdf
30. Cortese CF., Frank RF., Cohen JK.: Further considerations on the methodological analysis of segregation indices. In: American sociological review (1976).
31. http://turfjs.org/docs/clustersKmeans
32. Bell, W.: A Probability Model for the Measurement of Ecological Segregation. In: Social Forces, 32(4) (1954)
33. The Expert Group on Refugee and Internally Displaced Persons Statistics: International Recommendations on Refugees Statistics. (2018)
34. Freeman LC.: Segregation in Social Networks. In: Sociological Methods Research, 6 (1978)
35. Massey D., Denton N.: The Dimensions of Residential Segregation. In: Social Forces, 67(2) (1988)
36. McPherson M., Smith-Lovin L., Cook JM.: Birds of a Feather: Homophily in Social Networks. In: Annual Review of Sociology, 27(1) (2001)
37. Carrington PJ.: Log-linear distance models of homophily in small groups. In: Methodological Innovations (2016)
38. Cascetta E., Cartenì A., Montanino M.: A behavioral model of accessibility based on the number of available opportunities. In: Journal of Transport Geography, 51 (2016)