Geografie en inwoneraantallen als verklarende factoren voor variatie in het Nederlandse dialectgebied

Wilbert Heeringa
Rijksuniversiteit Groningen

John Nerbonne
Rijksuniversiteit Groningen

Renée van Bezooijen
Radboud Universiteiten Nijmegen

Marco René Spruit
Meertens Instituut Amsterdam

Symposium kwantitatieve benaderingen in taal- en letterkundig onderzoek en elders in de geesteswetenschappen. Een kennismaking

Meertens Instituut Amsterdam
Donderdag 28 juni 2007

Overzicht

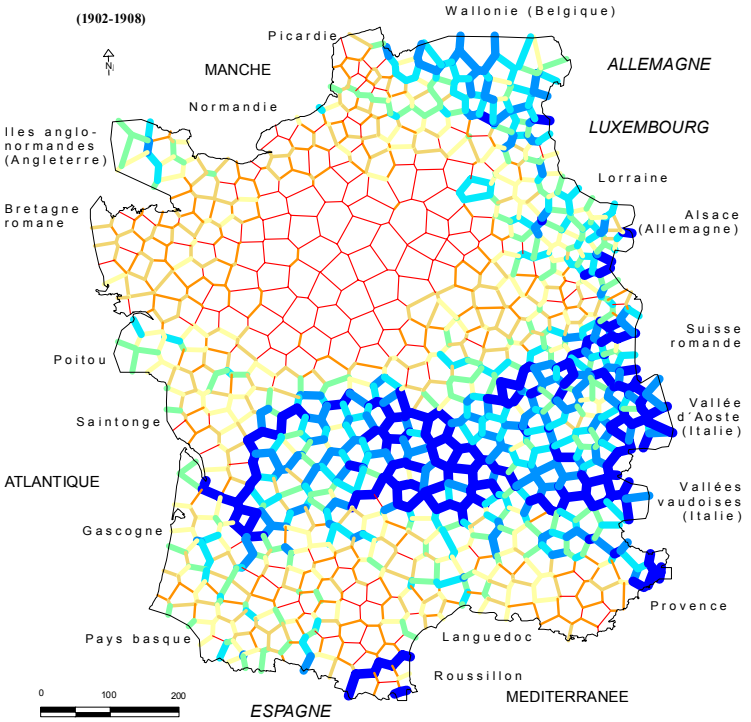
- Dialectometrie, Levenshtein-afstand
- Verklarende factoren in zwaartekrachtmodel
  - Geografie
  - Producten inwoneraantallen
- Geografie en producten inwoneraantallen
  - Zwaartekrachtmodel
  - Meervoudige regressieanalyse
- Conclusie

Dialectometrie

- **Dialectometrie**: de meting van het dialect.
- Term geïntroduceerd door Jean Séguy, directeur van de *Atlas linguistique de la Gascogne*.
- Afstand tussen twee naburige dialectplaatsen: het aantal items waarvoor de dialectplaatsen verschillend zijn, uitgedrukt in een percentage.
- Hans Goebl en Edgar Haimerl (Salzburg): vergelijkbare metingen, en geografische weergave van afstanden op een kaart.
- Cor en Geer Hoppenbrouwers: meet voor ieder dialect het aantal positief gemarkeerde features: het aantal geronde klinkers, het aantal stemhebbende medeklinkers, enz., op basis van fonetische transcripties.
- Voor ieder dialect 21 features, dus 21 frequenties (histogram).
- Maat van overeenkomst tussen twee dialecten: correlatie tussen de twee corresponderende reeksen van 21 features.

ALF

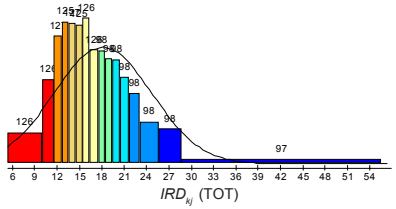
Série A: cartes 1-1421



Algorithme d'intervallisation
MEDMW 8-tuple

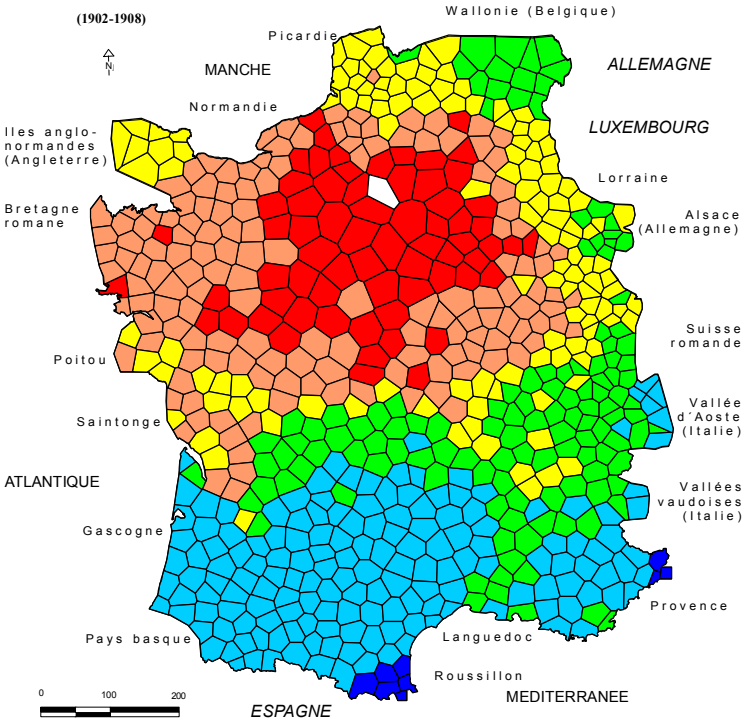
	de	à	points ALF
1	5,29	11,67	252
2	11,67	13,81	252
3	13,81	15,87	252
4	15,87	18,12	252
5	18,12	20,17	196
6	20,17	22,54	196
7	22,54	26,77	196
8	26,77	65,80	195
			Σ = 1792

Distribution de fréquence (distance)
MEDMW 16-tuple



ALF

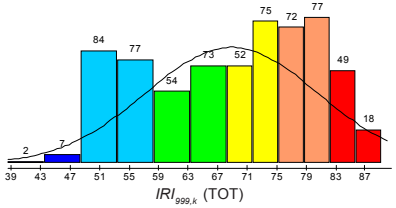
Série A: cartes 1-1421



Algorithme d'intervallisation
MINMWMAX 6-tuple

	de	à	points ALF
1	38,52	48,65	9
2	48,65	58,78	161
3	58,78	68,91	127
4	68,91	76,11	127
5	76,11	83,30	149
6	83,30	90,50	67
			Σ = 640

Distribution de fréquence (similarité)
MINMWMAX 12-tuple



Levenshtein-afstand

- Nadeel methoden Séguy, Goebl en Haimerl: twee items zijn òf gelijk òf ongelijk, geen gradualiteit.
- Nadeel methode gebr. Hoppenbrouwers: methode is niet gevoelig voor volgorde van de segmenten in een woord, bijv. [kəniːn] en [kniːnə] worden niet onderscheiden.
- In 1995 gebruikte Kessler de Levenshtein-afstand voor het meten van afstanden tussen Ierse dialecten.
- Geeft graduele afstanden, is gevoelig voor volgorde van de segmenten in woorden.
- Ook toegepast op Nederlandse dialecten (Nerbonne et al. 1996, Heeringa 2004), Sardische dialecten (Bolognesi & Heeringa 2002), Noorse dialecten (Gooskens & Heeringa 2004) en Duitse dialecten (Nerbonne & Siedle 2005).

Levenshtein-afstand

- Voorbeeld: *konijn* wordt uitgesproken als [kənɛːn] in het dialect van Amsterdam, en als [kniːnə] in het dialect van Zwollekerspel. Hoe veranderen we de ene variant in de andere?
- Dit kan op meerdere manieren. Het Levenshtein-algoritme kiest de operaties zodanig dat de totale kosten minimaal zijn:

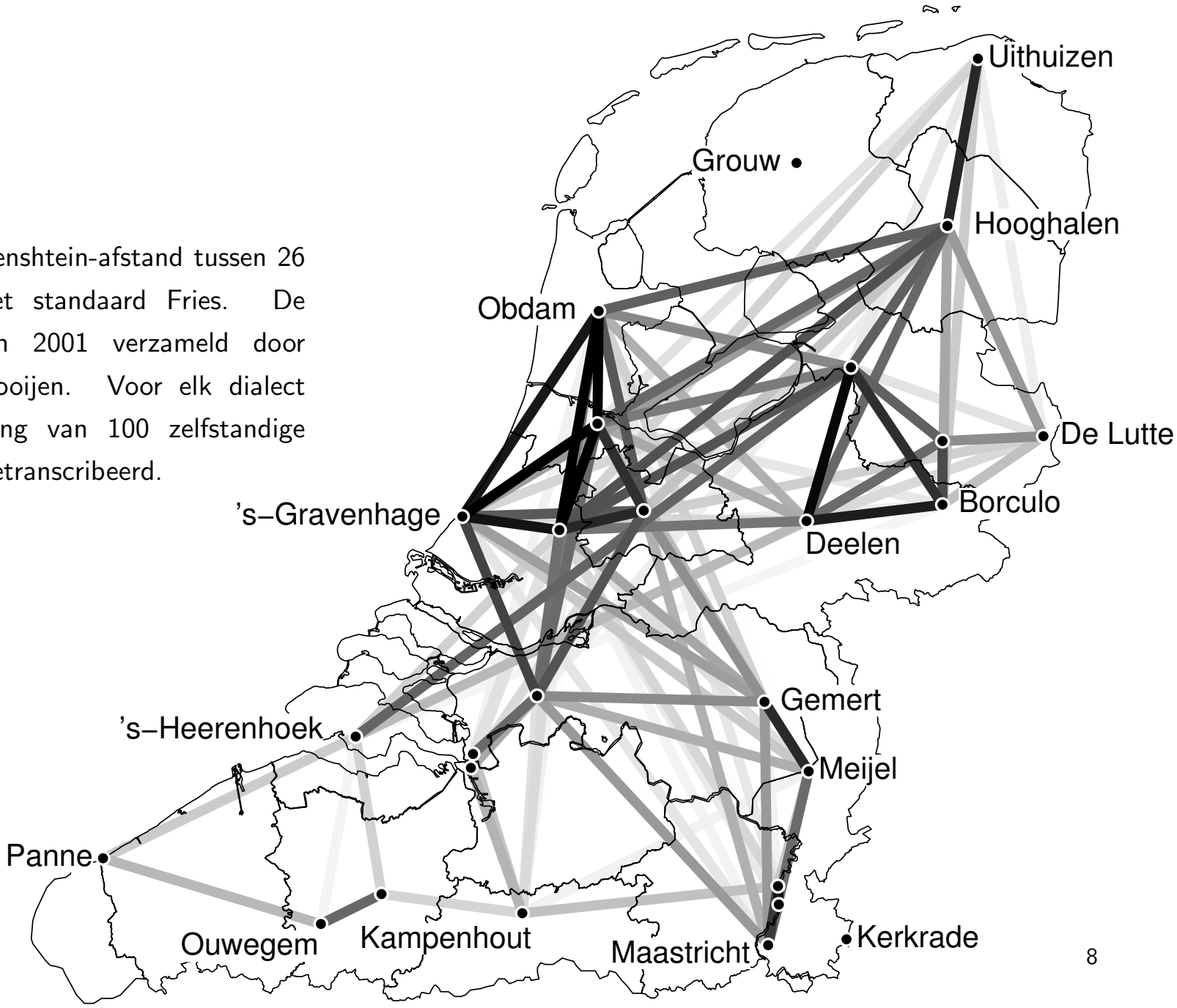
k	ə	n	ɛː	n		verwijder ə	1
k		n	ɛː	n		vervang ɛː door iː	1
k		n	iː	n		voeg toe ə	1
k		n	iː	n	ə		
							3
1	2	3	4	5	6		

- Deel de Levenshtein-afstand door de lengte van de oplijning: 3 / 6 = 0.5. Als percentage: 50%.

Levenshtein-afstand

- Afstand tussen twee dialecten: gemiddelde Levenshtein-afstand voor een reeks woordparen.
- Verfijning: gebruik graduele gewichten, namelijk de akoestische segmentafstanden.
- We staan alleen oplijningen toe waarin:
  - een klinker correspondeert met een klinker
  - een mederklinker correspondeert met een medeklinker
  - de [j] of [w] correspondeert met een klinker
  - de [i] of [u] correspondeert met een medeklinker
  - de schwa correspondeert met een sonorant

Gemiddelde Levenshtein-afstand tussen 26 dialecten en het standaard Fries. De data werden in 2001 verzameld door Renèe van Bezooijen. Voor elk dialect werd de vertaling van 100 zelfstandige naamwoorden getranscribeerd.

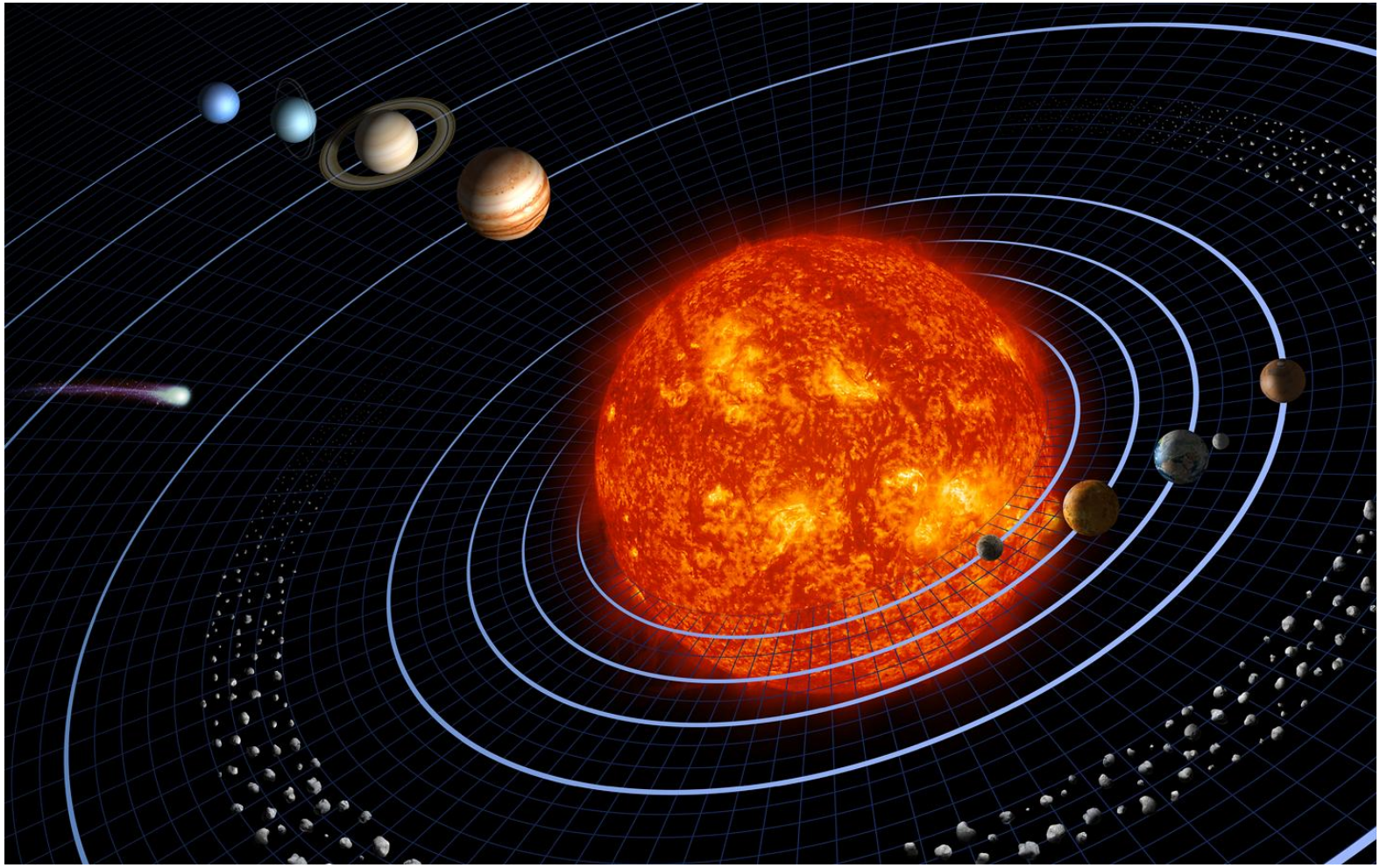


Zwaartekrachtmodel

- Zwaartekrachtmodel van Isaac Newton:

$$F = G\frac{m_1 \times m_2}{r^2}$$

F is de aantrekkingskracht
m_1 en m_2 zijn de gewichten van beide objecten
r is de afstand tussen beide objecten
G is de kracht tussen twee objecten van elk 1 kg op 1 m afstand van elkaar



Plaatje op `http://en.wikipedia.org/wiki/Gravitation`. De zwaartekracht houdt de planeten in hun baan rond de zon, en de maan in zijn baan rond de aarde.

Zwaartekrachtmodel

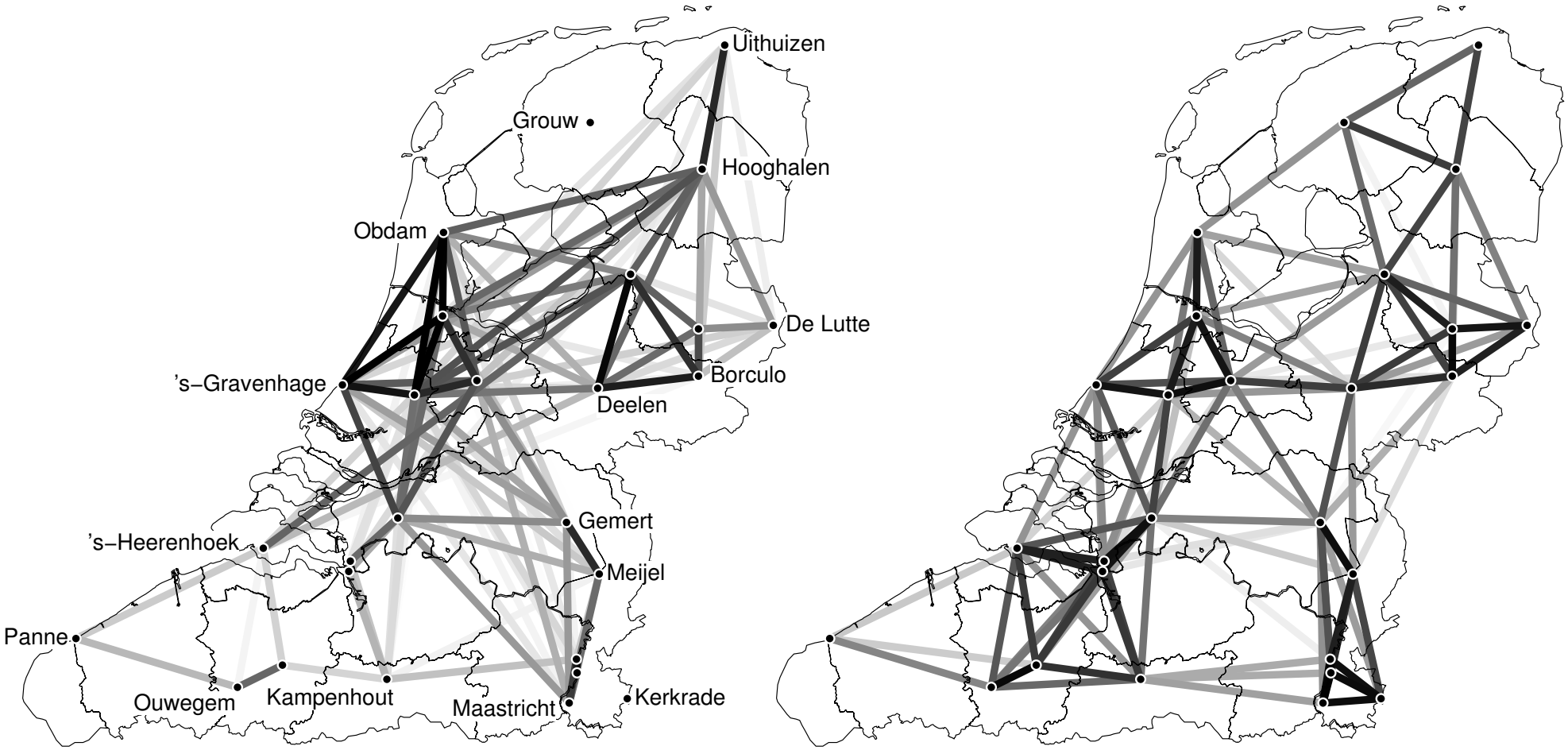
- Trudgill 1974, 1983 (ook Hinskens 1992, 1993) gebruikten het model als index voor taalkundige invloed tussen plaatsen:

$$F = G\frac{p_1 \times p_2}{r^2}$$

- Geografie: hoe dichter plaatsen bij elkaar liggen, hoe meer contact. Kans dat een inwoner gaat naar een punt op een denkbeeldige cirkel rond zijn of haar woonplaats: $1/r^2$.
- Inwoneraantallen: iedere inwoner uit de ene plaats kan in contact komen met iedere inwoner in de andere plaats: $p_1 \times p_2$
- G is de taalkundige verwantschap tussen beide plaatsen.

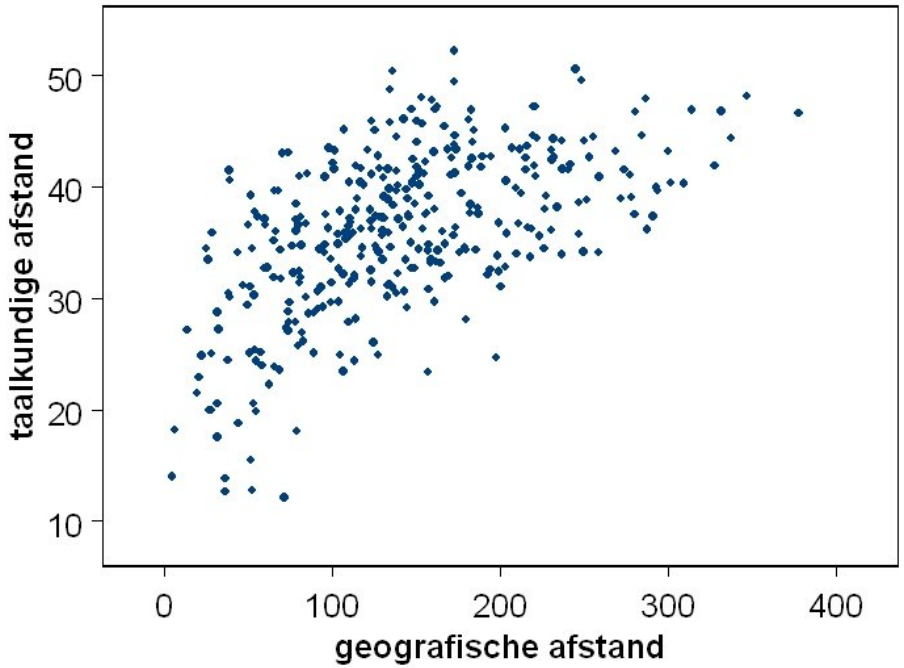
Zwaartekrachtmodel

- In navolging van Nerbonne & Heeringa 2006 gebruiken we het model als index van sociaal contact.
- G vervalt, taalkundige verwantschap is de inverse van taalkundige afstand, dat laatste willen we juist verklaren met het model.
- Hypothesen:
  - De taalkundige afstand D is recht evenredig met de kwadratische afstand: $D \propto r^2$.
  - De taalkundige afstand D is omgekeerd evenredig met de producten van de inwoneraantallen: $D \propto \frac{1}{p_1 p_2}$.
  - De taalkundige afstand D is omgekeerd evenredig met de mate van sociaal contact tussen twee plaatsen: $D \propto \frac{1}{F}$



Links: gemiddelde Levenshtein-afstanden, rechts: hemelsbrede geografische afstanden. Correlatie r=0.58.

Taalkundige afstanden vs. geografie



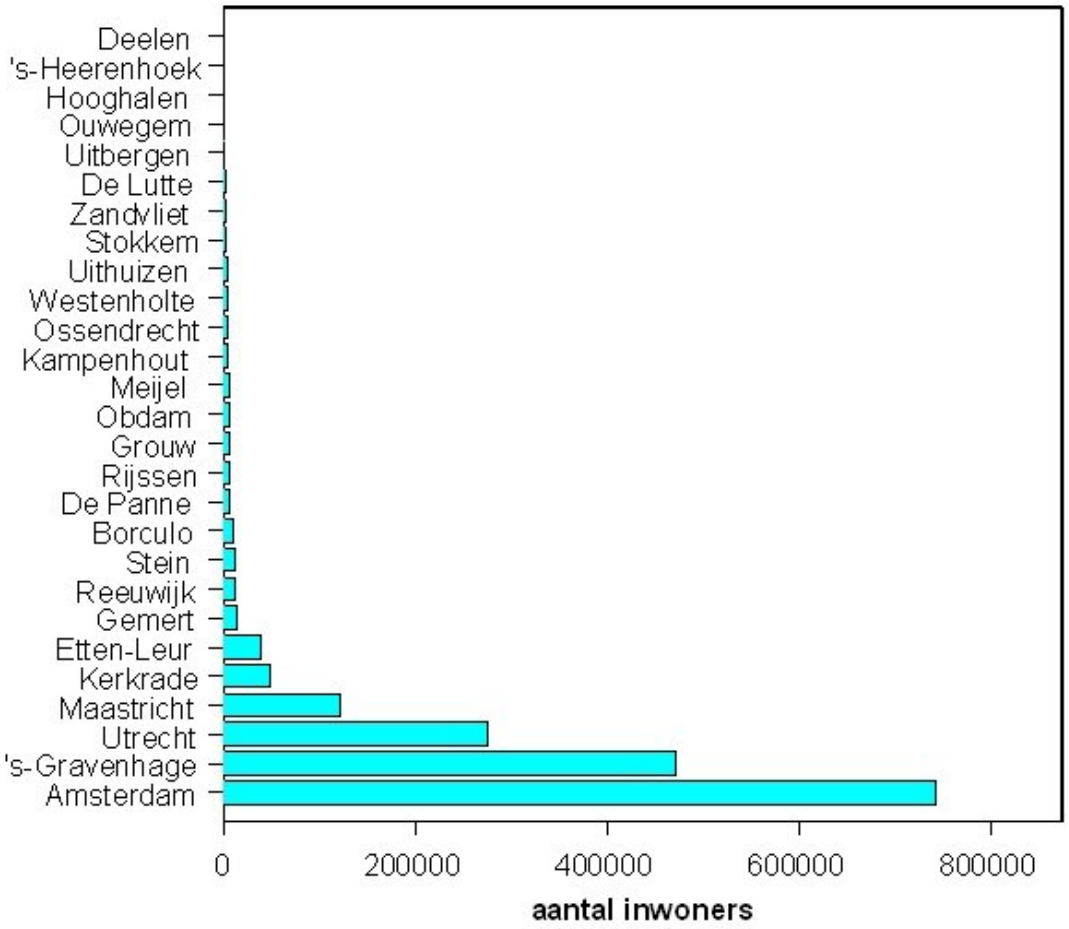
Er is een sterke corrrelatie (samenhang) tussen geografie en taalkundige afstand: r=0.58.

Taalkundige afstanden vs. geografie

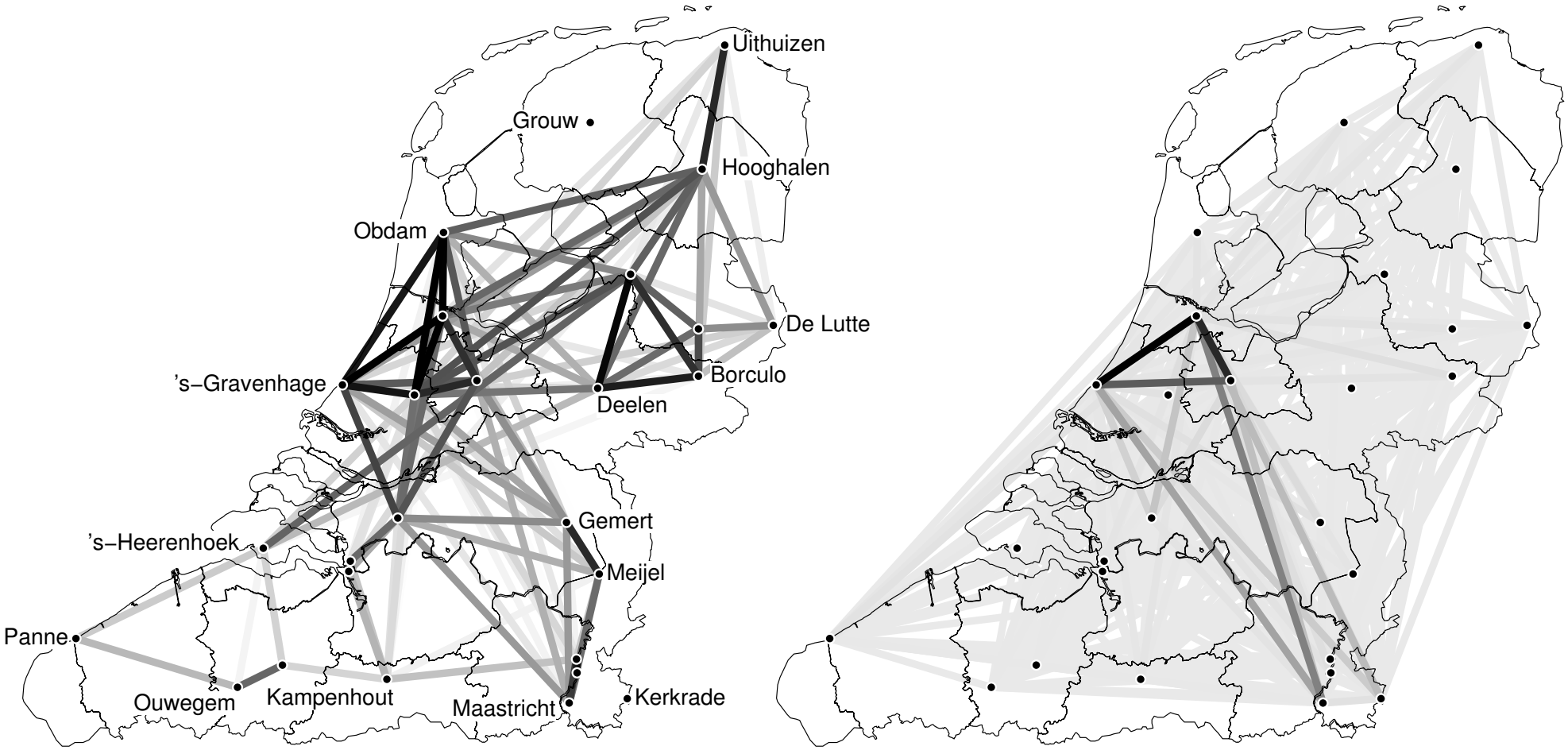
- In de zwaartekrachtformule is geografie kwadratisch: r^2.
- De grafiek suggereert eerder dat geografie een logaritmisch verloop heeft.
- Vier transformaties:

transformatie geografie	correlatie geografie	verklaarde variantie geografie
kwadratisch	0.49	24%
geen	0.58	33%
wortel	0.61	37%
logaritmisch	0.62	39%

- De logaritmische afstanden correleren significant beter dan de kwadratische afstanden.

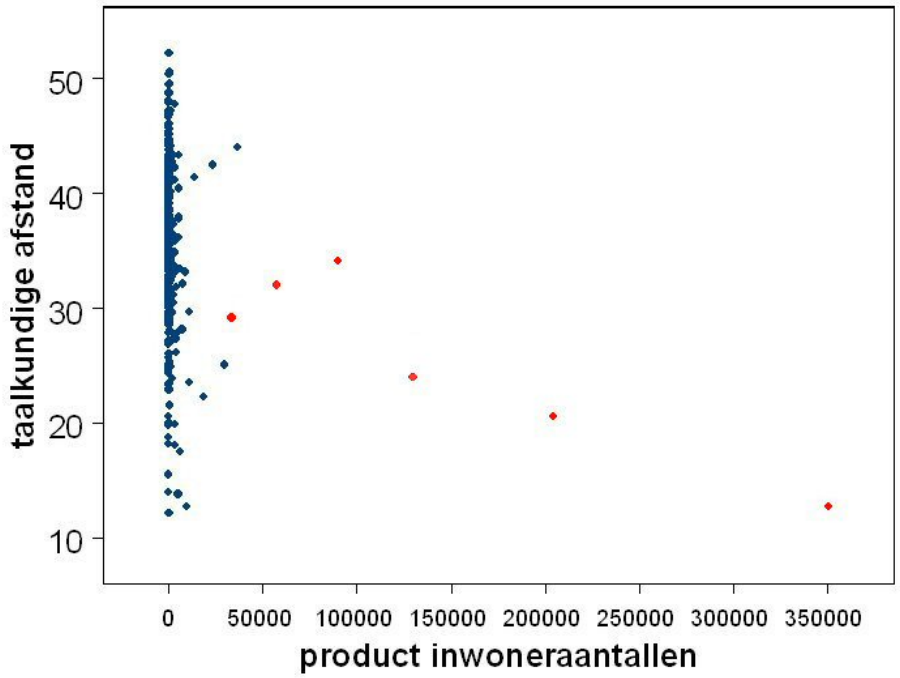


We meten de producten van de inwoneraantallen. Bijvoorbeeld Amsterdam-Deelen: $742780 \times 60 = 44566800$ inwoner-paren.



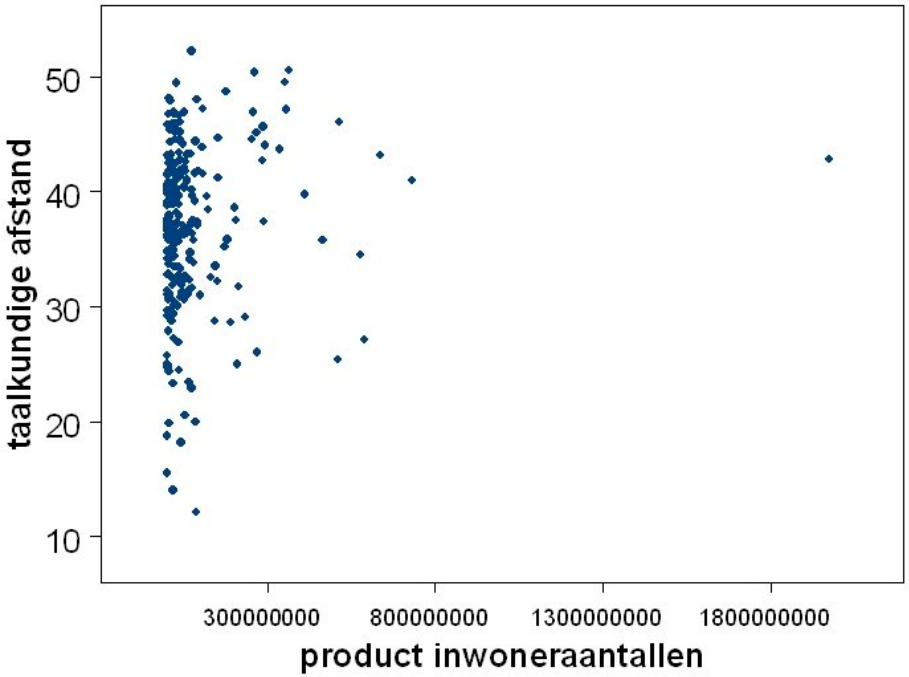
Links: gemiddelde Levenshtein-afstanden, rechts: producten van inwoneraantallen. Correlatie r=-0.24.

Taalkundige afstanden vs. producten inwoneraantallen



Correlatie r=-0.24. De rode stippen representeren afstanden tussen de vier grote steden. Producten inwoneraantallen zijn gegeven in miljoenen.

Taalkundige afstanden vs. producten inwoneraantallen



De afstanden tussen en t.o.v. de grote steden zijn weggelaten. De stip helemaal rechts is het product van de inwoneraantallen van Kerkrade en Etten-Leur. Correlatie: r=0.12. Dus: vooral attractie tussen en t.o.v. de grote steden, en differentiatie tussen en t.o.v. de middelgrote steden.

Zwaartekrachtmodel

- Zwaartekrachtmodel: producten van de inwoneraantallen worden gedeeld door de gekwadrateerde geografische afstanden.
- Alternatief: producten van de inwoneraantallen delen door de logaritmische geografische afstanden.

transformatie geografie	correlatie geografie	correlatie inw. prod.	zwaarte-kracht-model
kwadratisch	0.49	-0.24	-0.22
logaritmisch	0.62	-0.24	-0.24

- Zwaartekrachtmodel is hier niet het juiste model.

Meervoudige regressieanalyse

- Regressieanalyse: statistische techniek die het verband tussen variabelen zo nauwkeurig mogelijk in een formule uitdrukt.
- Idee: taalkundige afstanden kunnen voorspeld worden op basis van geografische afstanden en de producten van de inwoneraantallen.

transformatie geografie	correlatie geografie	correlatie inw. prod.	meerv. regr. model
kwadratisch	0.49	-0.24	0.53
geen	0.58	-0.24	0.60
wortel	0.61	-0.24	0.63
logaritmisch	0.62	-0.24	0.65

- In geen van de vier gevallen geeft het meervoudige regressiemodel een significante verbetering.

Meervoudige regressieanalyse

- Geografie verklaart 33% (lineair) of 39% (logaritmisch) van de variatie in de 27 Nederlandse variëteiten.
- Producten van de inwoneraantallen verklaren 6% van de variantie.
- Producten van inwoneraantallen hebben geen toegevoegde waarde t.o.v. geografie als verklaring voor dialectvariatie.
- Alternatieven voor meting sociaal contact: verkeersstromen, frequentie OV-verbindingen.
- Andere factoren: historische en politieke verschillen.
- Kempken 2005 gebruikte Levenshtein voor meting variatie in de spelling van middeleeuwse documenten. Verklarende factoren?
- Onderzoek ook belangrijk voor tekstanalytisch onderzoek van bijv. middeleeuwse documenten.

Ten slotte

We danken:

- Centraal Bureau voor de Statistiek (inwoneraantallen Nederland)
- Belgische gemeenten (inwoneraantallen België)
- Peter Kleiweg (visualisatie-programmatuur)

Meer over dialectometrie in Groningen en Amsterdam kan gevonden worden via:

- `http://www.dialectometry.net/`

R*u*G/L^{04} *software for dialectometrics and cartography* is ontwikkeld door Peter Kleiweg. Dit pakket kan geladen worden via:

- `http://www.let.rug.nl/~kleiweg/L04`