

Multidimensional Scaling (MDS)

Lubomir Zlatkov
ReMa Linguistics

Outline

1. Overview
2. Procedures
 - Classical MDS
 - Kruskal's non-metric MDS
 - Sammon's Non-linear Mapping
3. Dialectometry Example



Multidimensional Scaling

- Geometric representation of the structure of distance data



Multidimensional Scaling

- Geometric representation of the structure of distance data
- Optimal coordinate system based on distances between data points



Multidimensional Scaling

- Geometric representation of the structure of distance data
- Optimal coordinate system based on distances between data points
- Multidimensional space in each case scaled down to a coordinate in a 2D/3D space



Multidimensional Scaling (2)

- Original distance between elements from the data matrix = Euclidian distance between their coordinates in MDS representation



Multidimensional Scaling (2)

- Original distance between elements from the data matrix = Euclidian distance between their coordinates in MDS representation
- Shows meaningful underlying dimensions used to explain differences in data



Procedures

- Metric (classical) MDS – (Togerson 1952)
- Non-metric MDS
 - Kruskal's non-metric MDS (Kruskal 1964, Kruskal and Wish 1978)
 - Sammon's non-linear mapping (Sammon 1969)



Algorithm

1. Initial state (random or classical MDS)
2. Calculation of the Euclidean distances between the elements
3. Comparison between the Euclidean distances and the original distances using STRESS function
4. Adjustments



STRESS function

Metric MDS

$$STRESS = \sum_{i < j} (d_{ij} - D_{ij})^2$$

STRESS function

Metric MDS

Euclidean distance

$$STRESS = \sum_{i < j} (d_{ij} - D_{ij})^2$$



Original distance

STRESS function

Kruskal

$$STRESS = \sqrt{\frac{\sum_{i < j} (f(d_{ij}) - D_{ij})^2}{\sum_{i < j} D_{ij}^2}}$$



STRESS function

Kruskal

Maps original distances to values best preserving rank order

$$STRESS = \sqrt{\frac{\sum_{i < j} (f(d_{ij}) - D_{ij})^2}{\sum_{i < j} D_{ij}^2}}$$

Euclidean distance



STRESS function Sammon

$$STRESS = \frac{\sum_{i < j} \frac{(d_{ij} - D_{ij})^2}{d_{ij}}}{\sum_{i < j} d_{ij}}$$



STRESS function

Sammon

Euclidean distance

$$STRESS = \frac{\sum_{i < j} \frac{(d_{ij} - D_{ij})^2}{d_{ij}}}{\sum_{i < j} d_{ij}}$$

Original distance

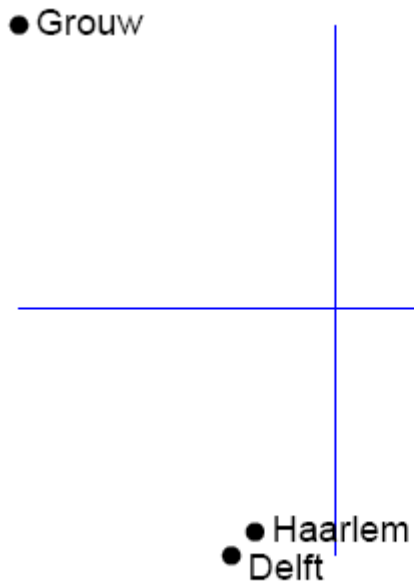


Comparison (Heeringa 2004)

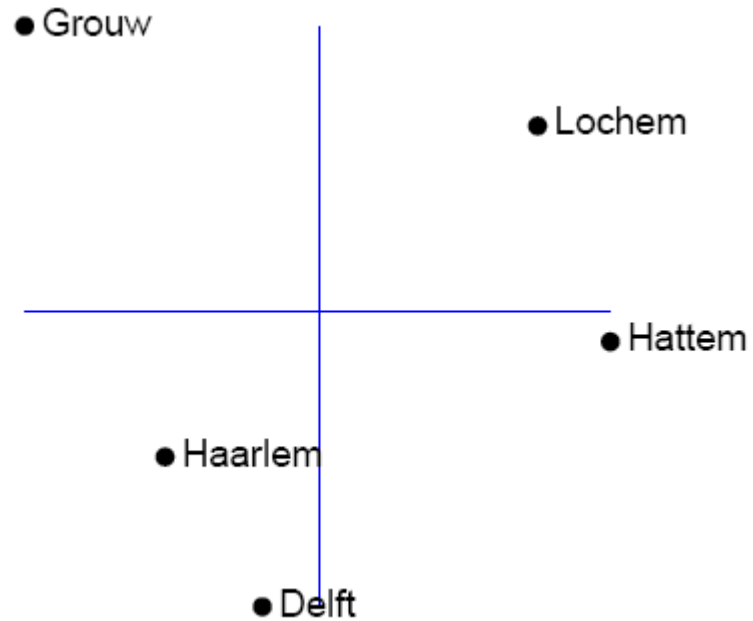
	Grouw	Haarlem	Delft	Hattem	Lochem
Grouw		42	44	46	47
Haarlem			16	36	38
Delft				38	40
Hattem					21
Lochem					



Comparison (Heeringa 2004)



Kruskal



Sammon



Dialectometry Example

- In a 3D MDS represent each dimension as a color (red, green, blue)



Dialectometry Example

- In a 3D MDS represent each dimension as a color (red, green, blue)
- Determine the color for each site based on its coordinates

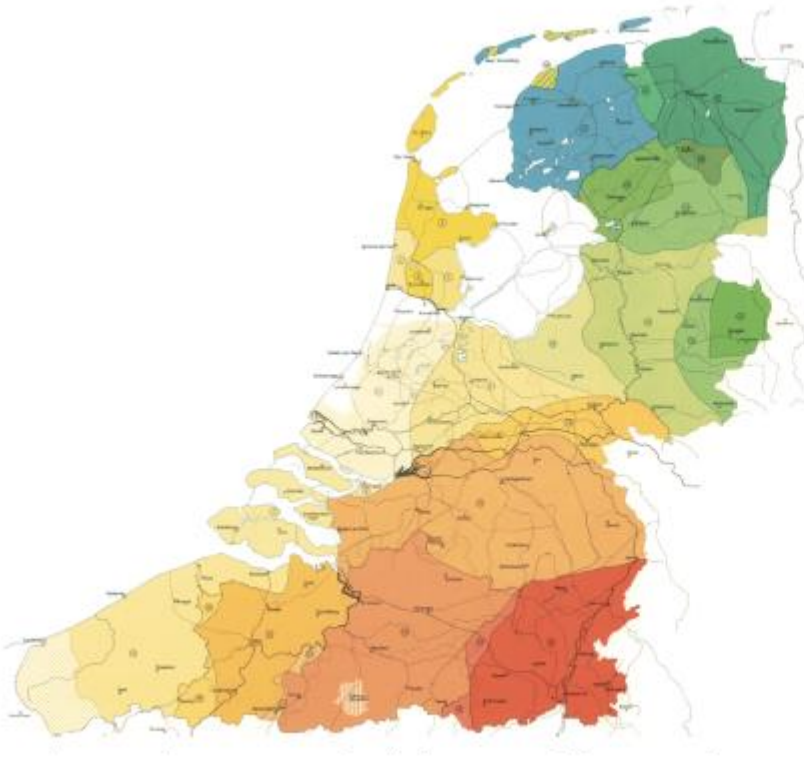


Dialectometry Example

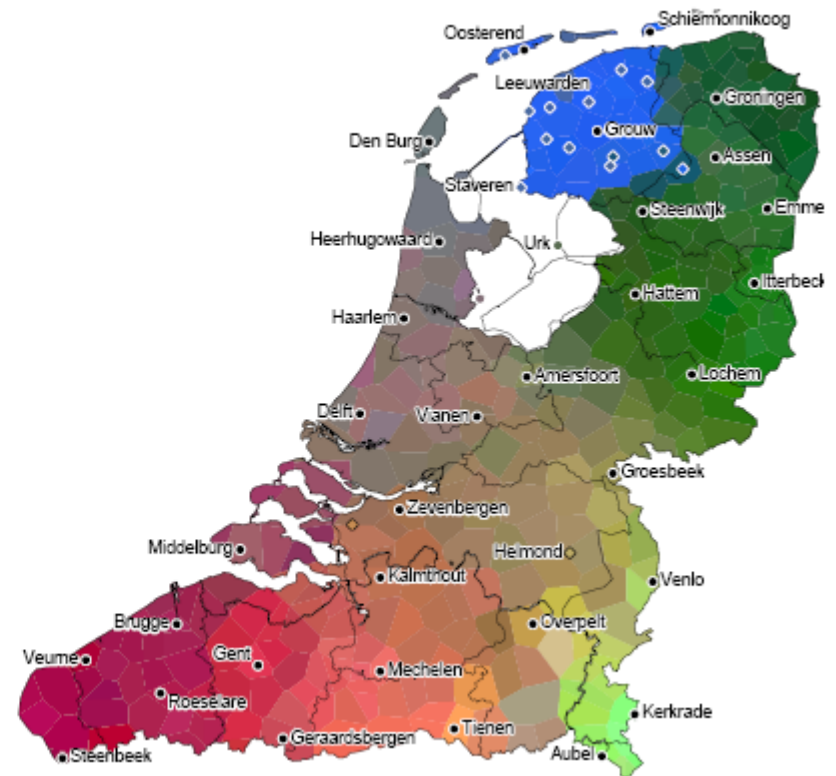
- In a 3D MDS represent each dimension as a color (red, green, blue)
- Determine the color for each site based on its coordinates
- Color the whole map
 - Delannay triangulation
 - Interpolation



Dialectometry Example (2)



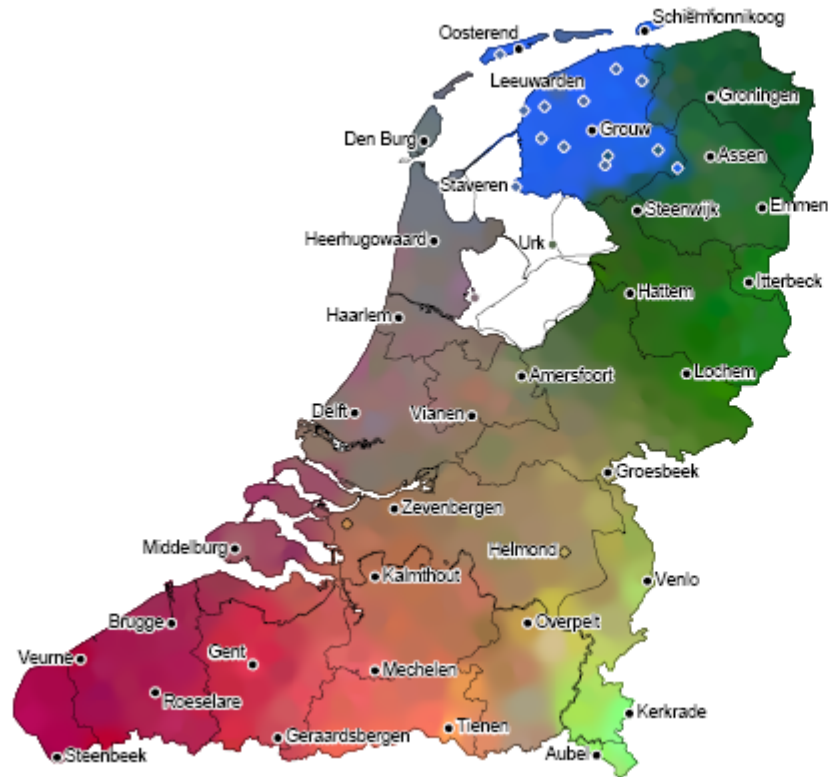
- Daan and Blok 1969



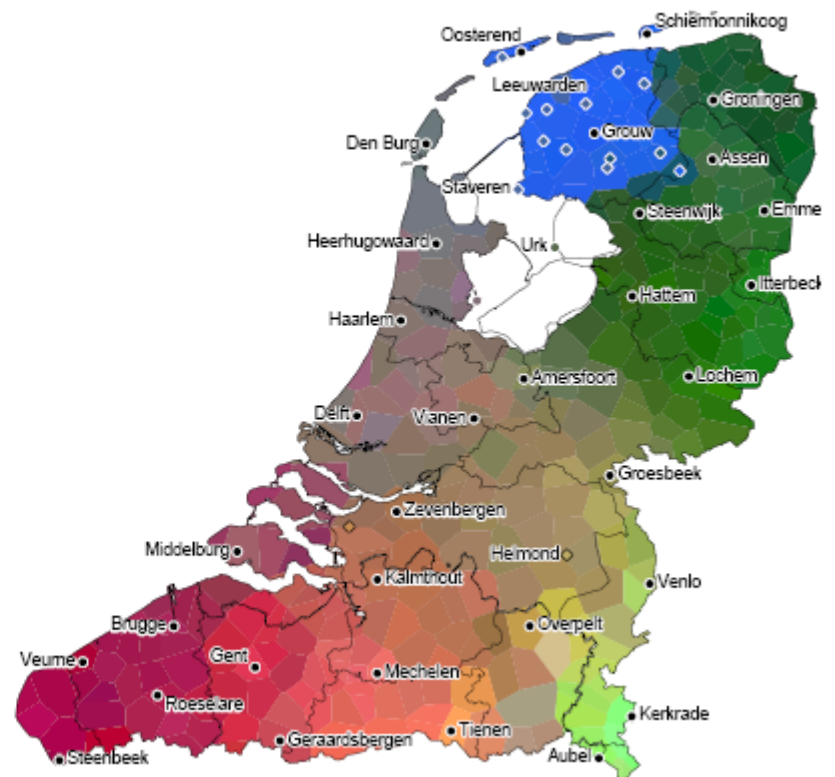
- Heeringa 2004



Dialectometry Example (3)



Heeringa (2004)



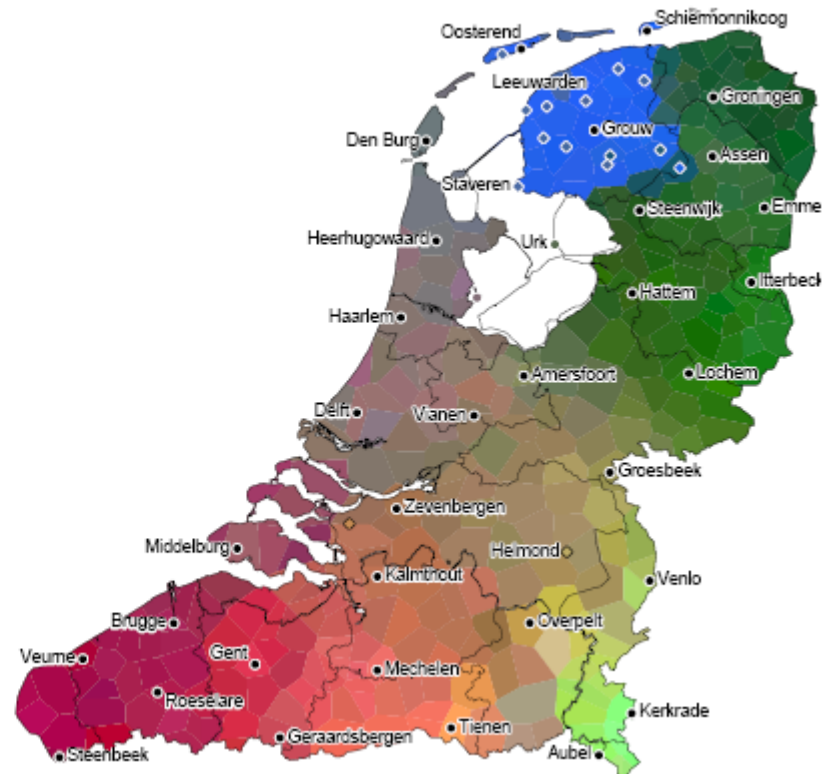
Heeringa (2004)



Dialectometry Example (4)



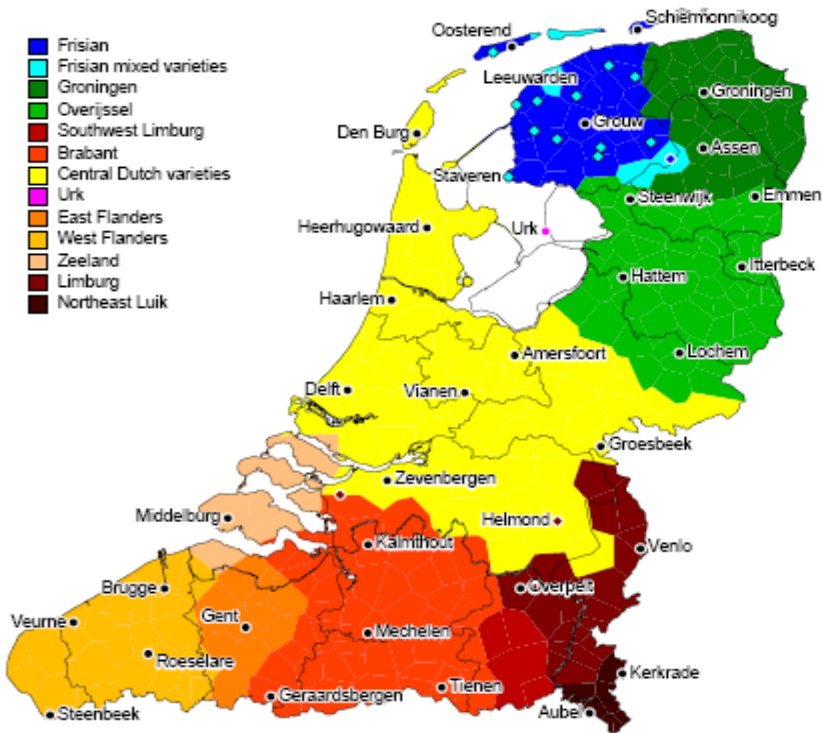
Spruit (2006)



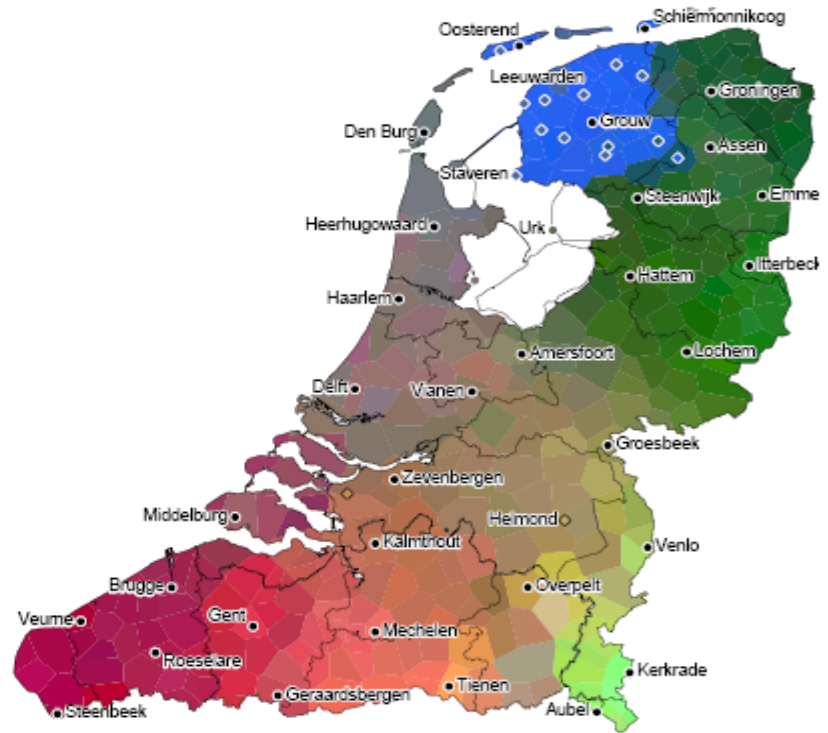
Heeringa (2004)



Dialectometry Example (5)



Heeringa (2004)



Heeringa (2004)



References

- Daan, J. and D.P. Blok (1969). *Van Randstad tot Landrand; toelichting bij de kaart: Dialecten en Naamkunde, volume XXXVII of Bijdragen en mededelingen der Dialectencommissie van de Koninklijke Nederlandse Akademie van Wetenschappen te Amsterdam*. Noord-Hollandsche Uitgevers Maatschappij, Amsterdam.
- Heeringa, W.. (2004). *Measuring Dialect Pronunciation Differences using Levenshtein Distance*. PhD thesis, Rijksuniversiteit Groningen.
- Johnson, K. (2008). *Quantitative Methods in Linguistics*. Oxford, UK: Blackwell.
- Kruskal, J. B. (1964). Multidimensional Scaling by Optimizing Goodness-of-Fit to a Nonmetric Hypothesis. *Psychometrika*, 29:1–28.
- Kruskal, J. B. and Wish, M. (1978). *Multidimensional Scaling*. Number 07-011 in Sage University Paper Series on Quantitative Applications in the Social Sciences. Sage Publications, Newbury Park.
- Legendre P, Legendre L (1998) *Numerical ecology*, 2nd English edn. Elsevier, Amsterdam
- Sammon, J. W. (1969). A nonlinear mapping for data structure analysis. *IEEE Transactions on Computers*, C 18:401-409.
- Spruit, M.R. (2006). Measuring syntactic variation in Dutch dialects. In Nerbonne, J., Kretzschmar, W. (eds), *Literary and Linguistic Computing, special issue on Progress in Dialectometry: Toward Explanation*, 21(4): 493–506
- Togerson, W. S. (1952). Multidimensional scaling. i. Theory and method. *Psychometrika*, 17:401-419.



Thank You!



university of
groningen