Finding dialect areas
by means of bootstrap clustering

Wilbert Heeringa

University of Groningen, Faculty of Arts, Applied Linguistics

Maps and Grammar
Meertens Institute Amsterdam
September 18, 2014
The goal of clustering is to identify the main groups in complex data.

Goebl (1982) introduced cluster analysis in the field of dialectometry.

Goal: finding groups given a set of local dialects and their mutual linguistic distances.
Clustering with noise

- Cluster analysis is instable: small differences in the distance matrix may strongly change the results (Nerbonne et al. 2008).
- Kleiweg et al. (2004) introduced clustering with noise which is visualized in composite cluster maps.
- Add noise to the distance matrix (maximal one or two standard deviations), and cluster the dialects.
- Repeat this e.g. 1000 times and count the number of times that pairs of neighbouring elements are part of different clusters.
- The darkness of the border that is drawn between those two locations represents the chance that the locations belong to different clusters.
Bootstrap clustering

- Nerbonne et al. 2008 also used **bootstrap** clustering to overcome instability.

- Given e.g. 100 words, randomly select 100 words using replacement, calculate distances between dialects, and cluster the dialects.

- Repeat this e.g. 1000 times, and count the number of times that pairs of neighbouring elements are part of different clusters.

- Noise clustering and bootstrap clustering produce similar results, but for bootstrap clustering no noise ceiling needs to be specified.
Application

- We apply clustering with noise to linguistic distances between 86 local Dutch dialects.
- Material collected in the period 2008-2011.
- Recorded transcriptions of male speakers (60 years or older) and young female speakers (between 20 and 40 years old), 125 words per speaker.
- Thus we study dialect change in apparent time.
- Levels: lexical level, morphological, sound components.
- Do dialect groups fuse or come apart, or do their mutual relationships otherwise change?
Parameters

- Distances are increased by a random value between zero and 1 sd, where sd is the standard deviation of all the original values.
- Number of runs: 50
- Given 86 local dialects, from the dendrogram 85 partitions can be derived representing a division in 2 groups, 3 groups, ..., 86 groups.
Parameters

- We do no choose a particular number of groups, but consider all 85 partitions in each run.
- Number of partitions: $85 \times 50 = 4250$ partitions
- In the map the darkness of a border between two neighboring dialects represents the number of partitions that the dialects are found in different groups.
- Varies between 0 and 4250.
Sound components
Gradual boundaries or areas?

- The work of Kleiweg et al. (2004) and Nerbonne et al. (2008) focus on boundaries which may be weaker or stronger.
- Traditional dialect maps show areas, for example the maps created by Te Winkel (1901) and Jo Daan (published in 1969).
Map of Te Winkel
Finding areas

- We introduce a new kind of bootstrap clustering which generates **areas**, comparable to classical dialect maps.
- We consider dialect groups as **continua**, i.e. a local dialect belongs to a group when it is strongly related to at least one local dialect in the same group, but not necessarily to all other local dialects in the group.
- The local dialects in a group rather constitute a **network**.
- We take into account that not every local dialect can be classified with statistical confidence.
Procedure

1. Select 1000 times randomly \( n \) items from \( n \) items with replacement. For each resampled set of items calculate the aggregated distances between the local dialects.

2. Perform **nearest neighbour clustering**.

3. Determine the number of natural groups by means of the **elbow method**.

4. For each pair of dialects count the number of times that both dialects are found in the same natural group. The number will vary between 0 (never) and 1000 (always).

5. When two dialects belong to the same group in more than 950 of the cases (95%), mark them as ‘connected.’ In this way we obtain **networks**.
1. Resampling, calculate distances

Calculate 1000 times aggregated distances between the 86 local dialects, on the basis of 125 words randomly chosen from 125 words.
2. Nearest neighbour clustering

- For each of the 1000 distance matrices perform nearest neighbour clustering (single-link).
- Choose the smallest distance:

\[ d_{k[ij]} = \text{minimum}(d_{ki}, d_{kj}) \]

- Reflects the idea of dialect areas as continua.
3. Elbow method

- Within a dendrogram different levels of detail can be distinguished.
- Starting at the root, a division into two groups is found. If we delve a little deeper we find that one of the two groups is divided into two further groups. At the bottom we find a division into 86 groups.
3. Elbow method

• For each division in $i$ groups $2 \leq i \leq 85$, we computed the variance in the original distances, as explained by the cophenetic distances of the part of the tree that gives a division in $i$ groups.

• Cophenetic distances: distances between local dialects as suggested by the dendrogram.

• Elbow plot: the variances are plotted against the number of groups.
3. Elbow method
3. Elbow method

- The initial clusters usually explain a great deal of the variance. The elbow is the point where the marginal gain will drop.
- We perform a linear regression analysis where the logarithmic number of clusters is the predictor and the explained variance the dependent variable. The number of groups corresponding with the largest residue is the number of natural groups (13).
3. Elbow method
3. Elbow method

![Elbow method graph]

- X-axis: Number of Groups
- Y-axis: Explained Variance

The graph shows the explained variance against the number of groups, indicating the elbow point where the explained variance starts to level off, suggesting the optimal number of groups.
For each pair of dialects count the number of times that both dialects are found in the same natural group. The number will vary between 0 (never) and 1000 (always).
5. Create networks

Pairs of dialects which belong to the same group in more than 95% of the cases are connected by a line.
Dialects which are a part of the same network, belong to the same group.
Application

- Do dialect groups fuse or come apart, or do their mutual relationships otherwise change?
Sound components
Thanks!

Dendrograms and maps in this paper are created with RuG/L04 which has been developed by Peter Kleiweg.