Finding Dialect Areas by Means of Bootstrap Clustering

Wilbert Heeringa
Fryske Akademy

In dialectometry cluster analysis is a means to find groups given a set of local dialects and their mutual linguistic distances. The weakness of cluster analysis is its instability; small differences in the distance matrix may strongly change the results.

Kleiweg, Nerbonne & Bosveld (2004) introduced composite cluster maps, which are obtained by collecting chances that pairs of neighboring elements are part of different clusters as indicated by the darkness of the border that is drawn between those two locations. Noise is added to the clustering process, which enables the authors to estimate about how fixed a border is. Nerbonne et al. (2008) use clustering with noise and bootstrap clustering to overcome instability. Both the work of Kleiweg, Nerbonne & Bosveld (2004) and Nerbonne et al. (2008) focus on boundaries which may be weaker or stronger.

We introduce a new flavor of bootstrap clustering which generates areas, similar to classical dialect maps. We perform a procedure consisting of four steps. First, we randomly select 1,000 times n items from n items with replacement. For each resampled set of items we calculate the aggregated distances. Second, on the basis of the distances we perform agglomerative hierarchical cluster analysis. We choose nearest neighbor clustering since this method reflects the idea of dialect areas as continua. On the basis of the tree we determine the number of natural groups by means of the elbow method. Third, for each pair of dialects we count the number of times that both dialects are found in the same natural group. Fourth, when two dialects belong to the same group in more than 95% of the cases, we mark them as 'connected.' In this way we will obtain networks which are the groups.

We apply the procedure to distances in the sound components measured with Levenhstein distance between a set of 86 Dutch dialects. We use material which was collected in the period 2008–2011.

Wilbert Heeringa. 2016. Finding dialect areas by means of bootstrap clustering. In Martijn Wieling, Gosse Bouma & Gertjan van Noord (eds.), *From semantics to dialectometry*, 1–11. Groningen: University of Groningen.

1 Introduction

Jain & Dubes (1988: p. 55) define cluster analysis as 'the process of classifying objects into subsets that have meaning in the context of a particular problem.' The goal of clustering is to identify the main groups in complex data. In dialectometry cluster analysis is a means to find groups given a set of local dialects and their mutual linguistic distances. Goebl (1982) introduced cluster analysis in the field of dialectometry (see also Goebl (1984) and Goebl (1993).

The weakness of cluster analysis is its instability; small differences in the distance matrix may strongly change the results (Jain, Murty & Flynn (1999), Nerbonne et al. (2008)). Kleiweg, Nerbonne & Bosveld (2004) introduced *composite cluster maps*, which are obtained by collecting chances that pairs of neighboring elements are part of different clusters as indicated by the darkness of the border that is drawn between those two locations. Noise is added to the clustering process, which enables the authors to estimate how fixed a border is. Given a distance matrix, a random value between 0 and a maximum is added to each distance, and subsequently the dialects are clustered. The maximum may be one or two standard deviations. This is repeated, e.g. 1,000 times, giving 1,000 clusterings, and the number of times that pairs of neighboring elements are part of different clusters in those 1,000 clusterings is counted. The results can be visualized in a map, where the darkness of the border between two locations represents the chance that the locations belong to different clusters.

In addition to *noise clustering* Nerbonne et al. (2008) also introduced *bootstrap clustering* to overcome instability. Given, e.g. a data set with transcriptions of 100 words for each local dialect, 100 words are randomly selected using replacement, Levenshtein distances are calculated between the dialects, and the dialects are clustered on the basis of the Levenshtein distances. When this is repeated, e.g. 1,000 times, the number of times that pairs of neighboring elements are part of different clusters is counted.

Nerbonne et al. (2008) show that noise clustering and bootstrap clustering produce similar results, but bootstrap clustering has the advantage that no noise ceiling needs to be specified.

Both the work of Kleiweg, Nerbonne & Bosveld (2004) and Nerbonne et al. (2008) focus on boundaries which may be weaker or stronger, i.e., they are gradual. This makes it harder to compare the maps with traditional dialect maps where the color distinctions give a visual representation of the borders between different dialect areas, for example, the map of (Te Winkel 1901) and the map of (Daan & Blok 1969).

We introduce a new flavor of bootstrap clustering which generates areas, similar to classical dialect maps. In our approach 1) we consider dialect groups as continua, i.e. each local dialect is not necessarily strongly related to any other local dialect in the same group; the local dialects in a group rather constitute a 'network' and 2) we take into account that not every local dialect can be classified with statistical confidence.

We apply the procedure to distances in the sound components measured with Levenshtein distance between a set of 86 Dutch dialects. Recorded transcriptions of older and younger speakers are used. Thus we are able to show the change of Dutch dialect areas in apparent time.

2 Data

In this paper we use a corpus database of recordings of 86 local Dutch dialects. This database was compiled in the period 2008–2011 by Heeringa & Hinskens (2014). The dialects are evenly spread over the Dutch and Frisian language areas and represent the major dialect regions.

In order to be able to measure dialect change in apparent time, at least two male speakers aged 60 or older, and two or more female speakers aged between 20 and 40 were recorded in each of the 86 locations. The males represent the older phase of a particular variety and the females the newer phase.

An scene of the Charlie Chaplin movie "The Kid" served as the basis of the recordings that were made. The scene can be regarded as a cross-section of plain, simple daily spoken language, and consists of 23 sentences, each containing an average of 7.6 words. We used a selection of 13 sentences for this study, which include a maximum of 125 words in the written standard Dutch version of the text.

Both the older male and the younger female speakers operated in small groups of at least two people. When a small group was being recorded, the individuals were first asked to write down a translation of the text in their own dialect, independently of each other. Then, together they compiled and wrote a consensus text upon which both of them agreed. Finally, they both read the consensus text aloud.

Phonetic transcriptions of the recordings were made. Usually, two recordings of the consensus dialect version of the story were produced by both the older males and the younger females. Since phonetic transcription is time-consuming, only one recording per group was transcribed, where the recording of the speaker who was the most autochthonous, had the clearest voice, and read the text most

fluently was prefered. The transcriptions were made in IPA and digitized in X-SAMPA.

The recordings in our data set were transcribed by one transcriber. To ensure optimal consistency per item, transcriptions are made per sentence instead of per text. The same sentence was played (2 times 86 is) 172 times and transcribed. Subsequently, the next sentence was played 172 times and transcribed, etc.

For more details see Heeringa & Hinskens (2014).

3 Methodology

We perform a procedure consisting of four steps. First, we randomly select n items from n items with replacement 1,000 times. For each resampled set of items we calculate the aggregated distances. Second, on the basis of the distances we perform agglomerative hierarchical cluster analysis. Third, for each pair of dialects we count the number of times that both dialects are found in the same natural group. Fourth, when two dialects belong to the same group in more than 950 of the cases (95%), we mark them as 'connected.' In this way we will obtain networks which are the groups. Below we will discuss each step in more detail.

3.1 Calculating distances and resampling

Distances in the sound components between dialects are measured with the aid of the Levenshtein distance metric (Levenshtein 1966). This algorithm was introduced into dialectology by Kessler (1995). The Levenshtein distance between two strings is calculated as the "cost" of the total set of insertions, deletions and substitutions needed to transform one string into another (Kruskal & Liberman 1999).

The aggregated distance between the two dialects is based on 125 word pairs (fewer if words were missing). We use normalized distance measures, calculating the aggregated distance between two dialects as the sum of a maximum of 125 word pair distances divided by the sum of the alignment lengths that correspond to the word pairs. For more details about the measurements see Heeringa & Hinskens (2014).

Now we calculate 1,000 times aggregated distances between the 86 local dialects on the basis of 125 words randomly chosen from 125 words, using either the transcriptions of the older or younger speakers.

3.2 Nearest neighbor clustering

Once we have obtained 1,000 distance matrices, for each distance matrix we apply agglomerative hierarchical cluster analysis. Each observation starts in its own cluster. The clusters are then sequentially combined into larger clusters, until all elements end up being in the same cluster. We use *single-linkage*, which is also known as *nearest neighbor clustering*. In this kind of clustering at each step, the two clusters separated by the shortest distance are combined. The result is a binary hierarchical tree structure in which the dialect varieties are the leaves, and the branches represent the distances among (clusters or groups of) dialect varieties (Jain & Dubes 1988)

We choose nearest neighbor clustering since this method reflects the idea of dialect areas being continua, where the distance between geographically neighboring dialects is small, and the difference between geographically distant points may be large. This method also agrees with the little arrow method which the map of (Daan & Blok 1969) is based on. Using the little arrow method, locations which have similar dialects according to the speakers are connected by arrows in the map. It then may happen that local dialects A and B are judged as similar, and dialects B and C are judged as similar, but A and C are judged as different by the speakers.

3.3 Count number of times that two dialects share the same group

On the basis of the tree we determine the number of natural groups. Dendrograms are binarily branching trees. Within a dendrogram different levels of detail can be distinguished. Starting at the root, a division into two groups is found. Then, if we delve a little deeper we find that one of the two groups is divided into two further groups. At the bottom of the tree are the leaves, and here we find a classification into the maximum number of groups, in our case 86, with each grouping containing a single variety. We thus have 85 levels, the first suggesting a division into two groups and the 85th suggesting a division into 86. For each division in i groups ($2 \leq i \leq 85$), we compute 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 are distances between the dialects as reflected by the dendrogram. The cophenetic distance between two local dialects is the height of the dendrogram where the two branches that include the two objects merge into a single branch[1] (Sokal & Rohlf 1962).

[1] Definition taken from *Wikipedia*, retrieved August 27, 2016, from https://en.wikipedia.org/wiki/Cophenetic,

In a graph, the variances are plotted against the number of groups as found in each of the 85 divisions. The initial clusters usually explain a great deal of the variance. However, at a certain point the marginal gain will drop, yielding an angle in the graph. This angle provides the number of natural clusters and is known as the "elbow" (Aldenderfer & Blashfield 1984). After the angle, the amount of explained variance in the distances increases much more slowly than before. An example of an elbow plot is shown in Figure 1.

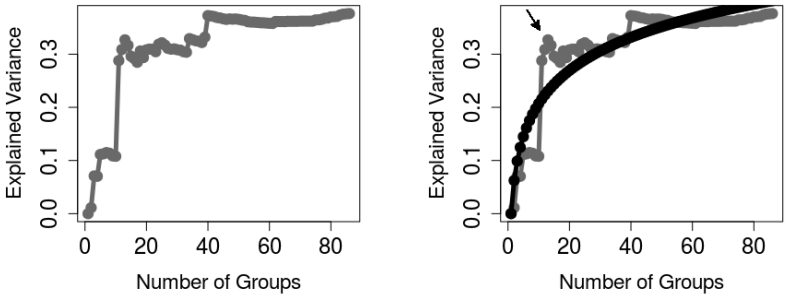

Figure 1: Left: elbow plot in which the variances are plotted against the number of groups. Right: a logarithmic regression curve is added. The elbow is found where the difference between predicted variance and 'real' variance is largest. The elbow is indicated by a little arrow. The number of natural groups is therefore 13.

We perform a linear regression analysis where the logarithmic number of clusters is the predictor and the explained variance the dependent variable. The curve which represents the explained variance predicted by the logarithmic number of clusters is added in the right graph in Figure 1. The elbow is found where the difference between the variance predicted by the logarithmic number of cluster and the 'real' variance is largest. The elbow is indicated by a little arrow and corresponds with 13 groups. The number of natural groups is therefore 13.

Now for each pair of dialects we count the number of times that both dialects are found in the same natural group. The number will vary between 0 (never) and 1,000 (always). The counts are graphically shown in Figure 2.

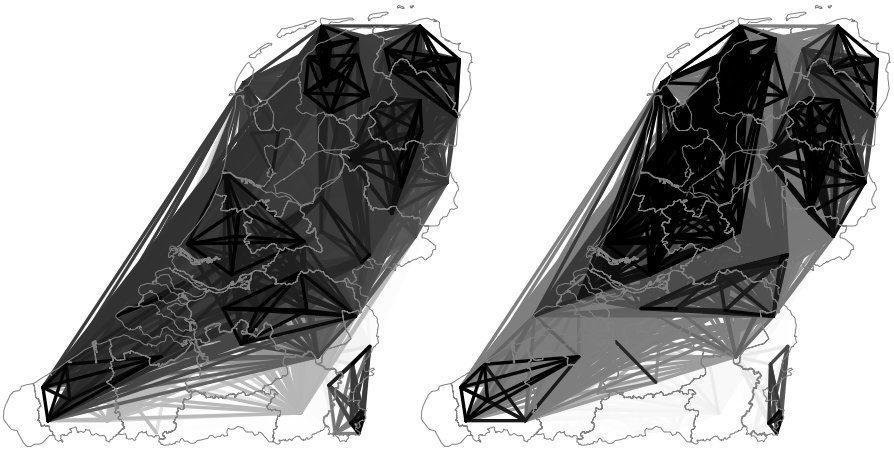

Figure 2: For each pair of dialects the number of times that both dialects are found in the same natural group is counted. The maps show counts obtained on the basis of transcriptions of older males (left) and younger females (right). Higher counts are represented by darker lines.

3.4 Create networks

When two dialects belong to the same group in more than 950 of the 1,000 bootstrap runs (95%), we mark them as being 'connected'. Pairs of dialects which belong to the same group in more than 95% of the cases are connected by a line in the map (see Figure 3).

In this way we will obtain networks. For example, when dialects A and B are connected, and dialects C and D are connected, but dialects A and B are not connected with dialects C and D, we obtain two separated networks, each of which consists of two dialects. We consider each network as a group. Dialects which are a part of the same network, belong to the same group. Dialects which are not connected to any other dialect remain unclassified. Our procedure does not force a local dialect to belong to a group when it is not strongly related to any other local dialect.

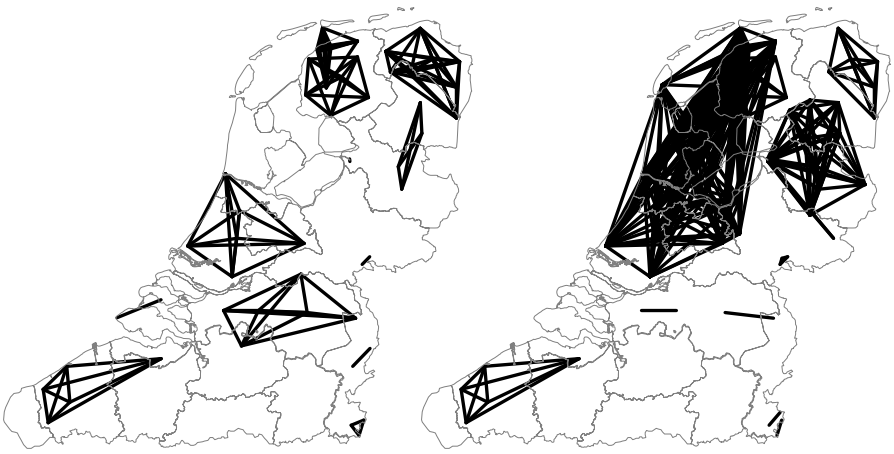

Figure 3: Local dialects that are connected by a line share in 95% of 1,000 bootstrap runs the same groups. Thus we obtain networks which are considered as dialect groups.

4 Results and conclusions

The final classifications are shown in Figure 4. The number of groups has decreased from 12 to 10. The mixed Frisian group (orange on the 'old' map) has been absorbed by the group of Holland dialects (yellow), and the petrol group has expanded and absorbed the dialects of Kampen and IJsselmuiden (lighter green on the 'old' map). The group of two Zeeland varieties (green/brown) has disappeared and the pink group has split into two smaller groups. The northern Limburg group (red/brown) has disappeared, and the group in the Southeast (red dots on the 'old' map) has split into two groups. On the 'new' map the red/brown dots represent transitional Limburg dialects, and the red dots Ripuarian dialects.

The number of unclassified dialects is larger for the older males than for the younger females: 31 versus 24, but the difference is not significant.

We conclude that the new cluster procedure produces clear area maps which are statistically justified and which can be easily compared to older dialect maps. Additionally changes of dialect areas can be clearly visualized by this approach.

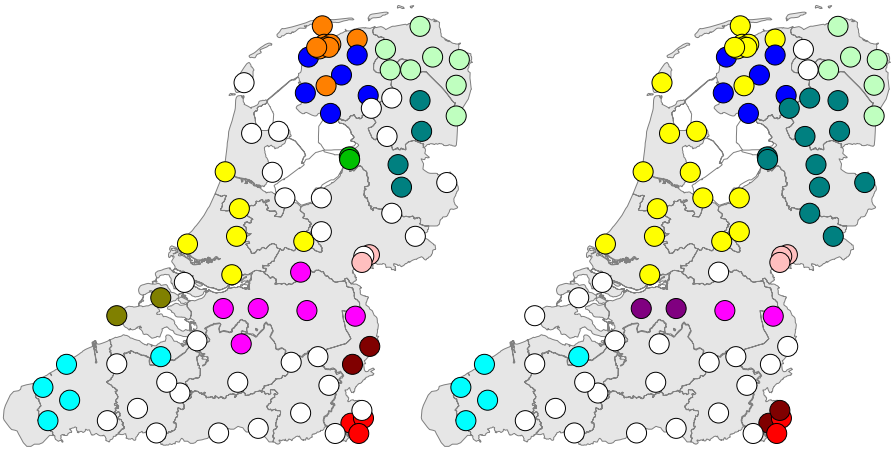

Figure 4: For the older males 12 groups are found (left) and for the younger females 10 groups are found (right).

References

Aldenderfer, M & R. Blashfield. 1984. Cluster analysis. In *Sage University Paper Series on Quantitative Applications in the Social Sciences 07-044*. Newbury Park (CA): SAGE publications.

Daan, J. & D. P. Blok. 1969. *Van Randstad tot Landrand; toelichting bij de kaart: Dialecten en Naamkunde*. Vol. XXXVII (Bijdragen en mededelingen der Dialectencommissie van de Koninklijke Nederlandse Akademie van Wetenschappen te Amsterdam). Amsterdam: Noord-Hollandsche Uitgevers Maatschappij.

Goebl, H. 1982. *Dialektometrie; Prinzipien und Methoden des Einsatzes der numerischen Taxonomie im Bereich der Dialektgeographie*. Vol. 157 (Philosophisch-Historische Klasse Denkschriften). With assistance of W.-D. Rase and H. Pudlatz. Vienna: Verlag der Österreichischen Akademie der Wissenschaften.

Goebl, H. 1984. *Dialektometrische Studien. Anhand italoromanischer, rätoromanischer und galloromanischer Sprachmaterialien aus AIS und ALF*. Vol. 191, 192, 193 (Beihefte zur Zeitschrift für romanische Philologie). With assistance of S. Selberherr, W.-D. Rase and H. Pudlatz. Tübingen: Max Niemeyer Verlag.

Goebl, H. 1993. Probleme und Methoden der Dialektometrie: Geolinguistik in globaler Perspektive. In W. Viereck (ed.), *Proceedings of the International Congress of Dialectologists*, vol. 1, 37–81. Stuttgart: Franz Steiner Verlag.

Heeringa, W. & F. Hinskens. 2014. Convergence between dialect varieties and dialect groups in the Dutch language area. In B. Szmrecsanyi & B. Wälchli (eds.), *Aggregating Dialectology, Typology, and Register Analysis; Linguistic Variation in Text and Speech* (Linguae et Litterae: Publications of the School of Language and Literature), 26–52. Berlin & Boston: De Gruyter.

Jain, A. K. & R. C. Dubes. 1988. *Algorithms for Clustering Data.* Englewood Cliffs, New Yersey: Prentice Hall.

Jain, A.K., M.N. Murty & P.J. Flynn. 1999. Data clustering: a review. *ACM Computing Surveys (CSUR)* 31(3). 264–323.

Kessler, B. 1995. Computational dialectology in Irish Gaelic. In *Proceedings of the 7th Conference of the European Chapter of the Association for Computational Linguistics*, 60–67. Dublin: EACL.

Kleiweg, P., J. Nerbonne & L. Bosveld. 2004. Geographic projection of cluster composites. In A. Blackwell, K. Marriott & A. Shimojima (eds.), *International Conference on Theory and Application of Diagrams*, 392–394. Springer.

Kruskal, J. B. & M. Liberman. 1999. The symmetric time-warping problem: from continuous to discrete. In D. Sankoff & J. Kruskal (eds.), *Time Warps, String edits, and Macromolecules. The Theory and Practice of Sequence Comparison*, 2nd, 125–161. 1st edition appeared in 1983. Stanford: CSLI.

Levenshtein, V. I. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Cybernetics and Control Theory* 10(8). 707–710.

Nerbonne, J., P. Kleiweg, W. Heeringa & F. Manni. 2008. Projecting dialect distances to geography: bootstrap clustering vs. noisy clustering. In C. Preisach, L. Schmidt-Thieme, H. Burkhardt & R. Decker (eds.), *Data Analysis, Machine Learning and Applications*, 647–654. Springer.

Sokal, R. R. & F. J. Rohlf. 1962. The comparison of dendrograms by objective methods. *Taxon* 11. 33–40.

Te Winkel, J. 1901. *Geschiedenis der Nederlandsche taal.* Naar de tweede Hoogduitsche uitgave met toestemming van den schrijver vertaald door Dr. F. C. Wieder. Met eene Kaart. Culemborg: Blom & Olivierse.