Projecting Dialect Distances to Geography: Bootstrap Clustering vs. Noisy Clustering

John Nerbonne[1], Peter Kleiweg[1], Franz Manni[2], and Wilbert Heeringa[1]

[1] Alfa-informatica, University of Groningen
{j.nerbonne,p.c.j.kleiweg,w.j.heeringa}@rug.nl
[2] Musée de l'Homme, Paris manni@mnhn.fr

Abstract. Dialectometry produces aggregate DISTANCE MATRICES in which a distance is specified for each pair of sites. By projecting groups obtained by clustering onto geography one compares results with traditional dialectology, which produced maps partitioned into implicitly non-overlapping DIALECT AREAS. The importance of dialect areas has been challenged by proponents of CONTINUA, but they too need to compare their findings to older literature, expressed in terms of areas.

Simple clustering is unstable, meaning that small differences in the input matrix can lead to large differences in results (Jain et al. 1999). This is illustrated with a 500-site data set from Bulgaria, where input matrices which correlate very highly ($r = 0.97$) still yield very different clusterings. Kleiweg et al. (2004) introduce COMPOSITE CLUSTERING, in which random noise is added to matrices during repeated clustering. The resulting borders are then projected onto the map.

The present contribution compares Kleiweg et al.'s procedure to resampled bootstrapping, and also shows how the same procedure used to project borders from composite clustering may be used to project borders from bootstrapping.

1 Introduction

We focus on dialectal data, examined at a high level of aggregation, i.e. the average linguistic distance between all pairs of sites in large dialect surveys. It is important to seek groups in this data, both to examine the importance of groups as organizing elements in the dialect landscape, but also in order to compare current, computational work to traditional accounts. Clustering is thus important as a means of seeking groups in data, but it suffers from instability: small input differences can lead to large differences in results, i.e., in the groups identified.

We investigate two techniques for overcoming the instability in clustering techniques, bootstrapping, well known from the biological literature, and "noisy" clustering, which we introduce here. In addition we examine a novel means of projecting the results of (either technique involving) such repeated clusterings to the geographic map, arguing that it is better suited to revealing the detailed structure in dialectological distance matrices.

2 Background and Motivation

We assume the view of dialectometry (Goebl, 1984 *inter alia*) that we characterize dialects in a given area in terms of an aggregate distance matrix, i.e. an assignment of a linguistic distance d to each pair of sites s_1, s_2 in the area $D_l(s_1, s_2) = d$. Linguistic distances may be derived from vocabulary differences, differences in structural properties such as syntax (Spruit, 2006), differences in pronunciation, or otherwise. We ignore the derivation of the distances here, except to note two aspects. First, we derive distances via individual linguistic items (in fact, words), so that we are able to examine the effect of sampling on these items. Second, we focus on true distances, satisfying the usual distance axioms, i.e. having a minimum at zero: $\forall s_1 D(s_1, s_1) = 0$; symmetry: $\forall s_1, s_2 D(s_1, s_2) = D(s_2, s_1)$; and the triangle inequality: $\forall s_1 s_2 s_3 D(s_1, s_2) \leq D(s_1, s_3) + D(s_3, s_2)$ (see (Kruskal 1999:22). We return to the issue of whether the distances are ULTRAMETRIC in the sense of the phylogenetic literature below.

We focus here on how to analyze such distance matrices, and in particular how to detect areas of relative similarity. While multi-dimensional scaling has undoubtedly proven its value in dialectometric studies (Embleton (1987), Nerbonne et al. (1999)), we still wish to detect DIALECT AREAS, both in order to examine how well areas function as organizing entities in dialectology, and also in order to compare dialectometric work to traditional dialectology in which dialect areas were seen as the dominant organizing principle.

CLUSTERING is a standard way in which to seek groups in such data, and it is applied frequently and intelligently to the results of dialectometric analyses. The research community is convinced that the linguistic varieties are hierarchically organized; thus, e.g., the urban dialect of Freiburg is a sort of Low Alemannic, which is in turn Alemannic, which is in turn Southern German, etc. This means that the techniques of choice have been different varieties of hierarchical clustering (Schiltz (1996), Mucha and Haimerl (2005)).

Hierarchical clustering is most easily understood procedurally: given a square distance matrix of size $n \times n$, we seek the smallest distance in it. Assume that this is the distance between i and j. We then fuse the two elements i and j, obtaining an $n - 1$ square matrix. One needs to determine the distance from the newly added $i + j$ element to all remaining k, and there are several alternatives for doing this, including nearest neighbor, average distance, weighted average distance, and minimal variance (Ward's method). See Jain et al. (1999) for discussion. We return in the discussion section to the differences between the clustering algorithms, but in order to focus on the effects of bootstrapping and "noisy" clustering, we use only weighted average (WPGMA) in the experiments below.

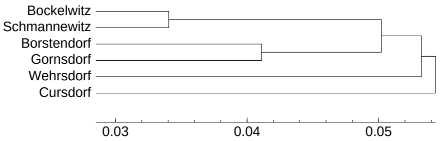

Fig. 1. An Example Dendrogram. Note the cophenetic distance is reflected in the horizontal distance from the leaves to the encompassing node. Thus the cophenetic distance between Borstendorf and Gornsdorf is a bit more than 0.04.

The result of clustering is a DENDROGRAM, a tree in which the history of the clustering may be seen. For any two leaf nodes in the dendrogram we may determine the point at which they fuse, i.e. the smallest internal node which contains them both. In addition, we record the COPHENETIC DISTANCE: this is the distance from one subnode to another at the point in the algorithm at which the subnodes fused.

Note that the algorithms depend on identifying *minimal* elements, which leads to instability: small changes in the input data can lead to very different groups' being identified (Jain et al., 1999). Nor is this problem merely "theoretical". Figure 2 shows two very different cluster results which from genuine, extremely similar data (the distance matrices correlated at $r = 0.97$).

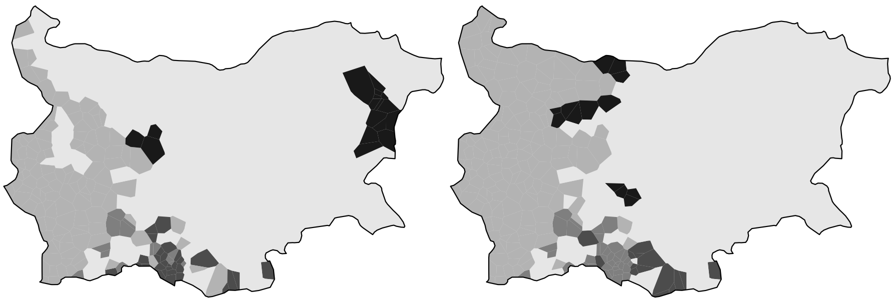

Fig. 2. Two Bulgarian Datasets from Osenova et al. (to appear). Although the distance matrices correlated nearly perfectly ($r = 0.97$), the results of WPGMA clustering differ substantially. Bootstrapping and noisy clustering resolve this instability.

Finally, we note that the distances we shall cluster do *not* satisfy the ultrametric axiom: $\forall s_1 s_2 s_3 D(s_1, s_2) \leq max\{D(s_2, s_3), D(s_1, s_3)\}$ (Page and Holmes (2006:26)). Phylogeneticists interpret data satisfying this axiom temporally, i.e., they interpret data points clustered together as later branches in an evolutionary tree. The dialectal data undoubtedly reflects historical developments to some extent, but we proceed from the premise that the social function of dialect variation is to signal geographic provenance, and that similar linguistic variants signal similar provenance. If the signal is subject to change due to contact or migration, as it undoubtedly is, then similarity could also result from recent events. This muddies the history, but does not change the socio-geographic interpretation.

2.1 Data

In the remainder of the paper we use the data analyzed by Nerbonne and Siedle (2005) consisting of 201 word pronunciations recorded and transcribed at 186 sites throughout all of contemporary Germany. The data was collected and transcribed by researchers at Marburg between 1976 and 1991. It was digitized and analyzed in 2003–2004. The distance between word pronunciations was measured using a modified version of edit distance, and full details (including the data) are available. See Nerbonne and Siedle (2005).

3 Bootstrapping Clustering

The biological literature recommends the use of bootstrapping in order to obtain stable clustering results (Felsenstein, 2004: Chap. 20). Mucha and Haimerl (2005) and Manni et al. (2006) likewise recommend bootstrapping for the interpretation of clustering applied to dialectometric data.

In bootstrapped clustering we resample the data, using replacement. In our case we resample the set of word-pronunciation distances. As noted above, each linguistic observation o is associated with a $site \times site$ matrix M_o. In the observation matrix, each cell represents the linguistic distance between two sites with respect to the observation: $M_o(s, s') = D(o_s, o_{s'})$. In bootstrapping, we assign a weight to each matrix (observation) identical to the number of times it is chosen in resampling:

$$w_o = \begin{cases} n \text{ if observation } o \text{ is drawn n times} \\ 0 \text{ otherwise} \end{cases}$$

If we resample I times, then $I = \sum_o w_o$. The result is a subset of the original set of observations (words), where some of the observations may be weighted as a resulted of the resampling. Each resampled set of words yields a new distance matrix $M_{i \in I}$, namely the average distances of the sites using the weighted set of words obtained via bootstrapping.

We apply clustering to each M_i obtained via bootstrapping, recording for each group of sites encountered in the dendrogram (each set of leaves below some node) both that the group was encountered, and the cophenetic distance of the group (at the point of fusion). This sounds as if it could lead to a combinatorial problem, but fortunately most of the 2^{180} possible groups are never encountered.

In a final step we extract a COMPOSITE DENDROGRAM from this collection, consisting of all of the groups that appear in a majority of the clustering iterations, together with their cophenetic distance. See Fig. 3 for an example.

4 Clustering with Noise

Clustering with noise is also motivated by the wish to prevent the sort of instability illustrated in Fig. 2. To cluster with noise we assume a single distance matrix, from which it turns out to be convenient to calculate variance (among all the distances). We then specify a small noise ceiling c, e.g. $c = \sigma/2$, i.e. one-half standard deviation of distances in the matrix. We then repeat 100 times or more: add random amounts of noise r to the matrix (i.e., different amounts to each cell), allowing r to vary uniformly, $0 \leq r \leq c$.

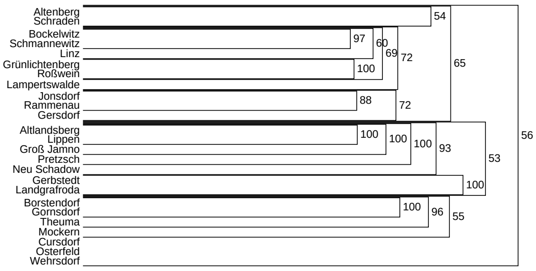

Fig. 3. A Composite Dendrogram where labels indicate how often a groups of sites was clustered and the (horizontal) length of the brackets reflects mean cophenetic distance.

If we let M_i stand in this case for the matrix obtained by adding noise (in the i-th iteration), then the rest of the procedure is identical to bootstrapping. We apply clustering to M_i and record the groups clustered together with their cophenetic distances, just as in Fig. 3.

5 Projecting to Geography

Since dialectology studies the geographic variation of language, it is particularly important to be able to examine the results of analyses as these correspond to geography.

In order to project the results of either bootstrapping or noisy clustering to the geographic map, we use the customary Voronoi tessellation (Goebl (1984)), in which each site is embedded in a polygon which separates it from other sites optimally. In this sort of tiling there is exactly one border running between each pair of adjacent sites, and bisecting the imaginary line linking the two. To project mean cophenetic distance matrices onto the map we simply draw the Voronoi tessellation in such a way that the darkness of each line corresponds to the distance between the two sites it separates. See Fig. 4 for examples of maps obtained by bootstrapping two different clustering algorithms. These largely corroborate scholarship on German dialectology (König 1991:230–231).

Unlike dialect area maps these COMPOSITE CLUSTER MAPS reflect the variable strength of borders, represented by the border's darkness, reflecting the consensus cophenetic distance between the adjacent sites.

Haag (1898) (discussed by Schiltz (1996)) proposed a quantitative technique in which the darkness of a border was reflected by the number of differences counted in a given sample, and similar maps have been in use since. Such maps look similar to the maps we present here, but note that the borders we sketch need not be reflected in *local* differences between the two sites. The clustering can detect borders even where differences are gradual, when borders emerge only when many sites are compared.[3]

6 Results

Bootstrapping clustering and "noisy" clustering identify the same groups in the 186-site German sample examined here. This is shown by the nearly perfect correlation between the mean cophenetic distances assigned by the two techniques ($r = 0.997$). Given the general acceptance of bootstrapping as a means of examining the stability of clusters, this result shows that "noisy" clustering is as effective.

The usefulness of the composite cluster map may best be appreciated by inspecting the maps in Fig. 4. While maps projected from simple clustering (see Fig. 2) merely partition an area into non-overlapping subareas, these composite maps reflect a great deal more of the detailed structure in the data. The map on the left was obtained by bootstrapping using WPGMA.

Although both bootstrapping and adding noise identifies stable groups, neither removes the bias of the particular clustering algorithm. Fig. 4 compares the bootstrapped results of WPGMA clustering with unweighted clustering (UPGMA, see Jain (1999)). In both cases bootstrapping and noisy clustering correlate nearly perfectly, but it is clear that the WPGMA is sensitive to more structure in the data. For example, it distinguishes Bavaria (in southeastern Germany) from the Southwest (Swabia and Alemania). So the question of the optimal clustering method for dialectal data remains. For further discussion see `http://www.let.rug.nl/kleiweg/kaarten/MDS-clusters.html`.

7 Discussion

The "noisy"clustering examined here requires that one specify a parameter, the noise ceiling, and, naturally, one prefers to avoid techniques involving

[3] Fischer (1980) discusses adding a contiguity constraint to clustering, which structures the hypothesis space in a way that favors clusterings of contiguous regions. Since we use the projection to geography to spot linguistic anomalies—dialect islands, but also field worker and transcriber errors—we do not wish to push the clustering in a direction that would hide these anomalies.

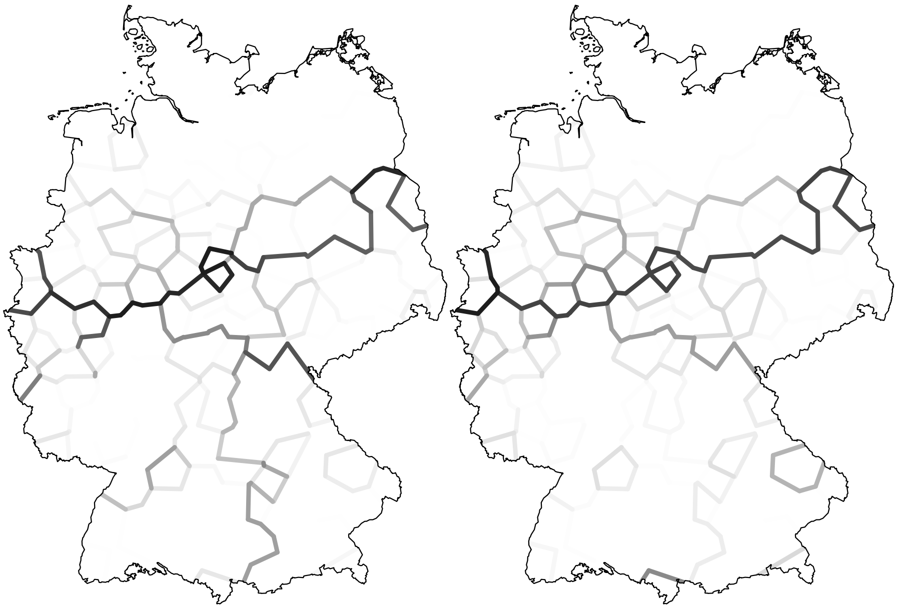

Fig. 4. Two Composite Cluster Maps, on the left one obtained by bootstrapping using weighted group average clustering, and on the right one obtained by unweighted group average. We do not show the maps obtained using “noisy” clustering, as these are indistinguishable from the maps obtained via bootstrapping. The composite distance matrices correlate nearly perfectly ($r = 0.997$) when comparing bootstrapping and “noisy” clustering.

extra parameters. On the other hand it is applicable to single matrices, unlike bootstrapping, which requires that one be able to identify components to be selected in resampling. Both techniques require that one specify a number of iterations, but this is a parameter of convenience. Small numbers of iterations are convenient, and large values result in very stable groupings.

Acknowledgments

We are grateful to Rolf Backofen and Hans Holm for discussion, to Hans-Joachim Mucha and Edgar Haimerl for making their 2007 paper available before GfKl 2007, and to two anonymous GfKl referees.

References

EMBLETON, S. (1987): Multidimensional Scaling as a Dialectometrical Technique. In: R. M. Babitch (Ed.) *Papers from the Eleventh Annual Meeting of the At-*

lantic Provinces Linguistic Association, Centre Universitaire de Shippagan, New Brunswick, 33-49.

FELSENSTEIN, J. (2004): *Inferring Phylogenies.* Sinauer, Sunderland, MA.

FISCHER, M. (1980): Regional Taxonomy: A Comparison of Some Hierarchic and Non-Hierarchic Strategies. *Regional Science and Urban Economics* 10, 503–537.

GOEBL, H. (1984): *Dialektometrische Studien: Anhand italoromanischer, rätoromanischer und galloromanischer Sprachmaterialien aus AIS und ALF* 3 Vol. Max Niemeyer, Tübingen.

HAAG, K. (1898): *Die Mundarten des oberen Neckar- und Donaulandes.* Buchdruckerei Egon Hutzler, Reutlingen.

JAIN, A. K., MURTY, M. N., and FLYNN, P. J. (1999): Data Clustering: A Review. *ACM Computing Surveys* 31(3), 264–323.

KLEIWEG, P., NERBONNE, J. and BOSVELD, L. (2004): Geographic Projection of Cluster Composites. In: A. Blackwell, K. Marriott and A. Shimojima (Eds.) *Diagrammatic Representation and Inference. 3rd Intn'l Conf, Diagrams 2004. Cambridge, UK, Mar. 2004. (Lecture Notes in Artificial Intelligence* 2980). Springer, Berlin, 392-394.

KÖNIG, W. (1991, ^{1}1978): *DTV-Atlas zur detschen Sprache.* DTV, München.

KRUKSAL, J. (1999): An Overview of Sequence Comparison. In: D. Sankoff and J. Kruskal (Eds.) *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison, 2nd ed.* CSLI, Stanford, 1–44.

MANNI, F. HEERINGA, W. and NERBONNE, J. (2006): To what Extent are Surnames Words? Comparing Geographic Patterns of Surnames and Dialect Variation in the Netherlands. In *Literary and Linguistic Computing* 21(4), 507-528.

MUCHA, H.J. and HAIMERL, E. (2005): Automatic Validation of Hierarchical Cluster Analysis with Application in Dialectometry. In: C. Weihs and W. Gaul (Eds.) *Classification—the Ubiquitous Challenge. Proc. of 28th Mtg Gesellschaft für Klassifikation, Dortmund, Mar. 9–11, 2004.* Springer, Berlin, 513–520.

NERBONNE, J., HEERINGA, W. and KLEIWEG, P. (1999): Edit Distance and Dialect Proximity. In: D. Sankoff and J. Kruskal (Eds.) *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison, 2nd ed.* CSLI, Stanford, v-xv.

NERBONNE, J. and SIEDLE, Ch. (2005): Dialektklassifikation auf der Grundlage aggregierter Ausspracheunterschiede. *Zeitschrift für Dialektologie und Linguistik* 72(2), 129-147.

OSENOVA, P., HEERINGA, W. and NERBONNE, J. (to appear): A Quantitative Analysis of Bulgarian Dialect Pronunciation. *Zeitschrift für slavische Philologie.*

PAGE, R.D.M., and HOLMES, E.C. (2006): *Molecular Evolution: A Phylogenetic Approach.* (^{1}1998) Blackwell, Oxford.

SCHILTZ, G. (1996): German Dialectometry. In: H.-H. Bock and W. Polasek (Eds.) *Data Analysis and Information Systems: Statistical and Conceptual Approaches. Proc. of 19th Mtg of Gesellschaft für Klassifikation, Basel, Mar. 8–10, 1995.* Springer, Berlin, 526–539.

SPRUIT, M. (2006): Measuring Syntactic Variation in Dutch Dialects. In J. Nerbonne and W. Kretzschmar, Jr. (Eds.) *Progress in Dialectometry: Toward Explanation.* Special issue of *Literary and Linguistic Computing* 21(4), 493–506.