Mapping Aggregate Variation

John Nerbonne, University of Groningen

1 Introduction

We proceed from the view that linguistic variation must be examined from an aggregate perspective, i.e. from a perspective which encompasses as much of the variation between language varieties as possible rather than concentrating on single linguistic features. We review the motivation for this position in the first section of the article, and then examine ways in which aggregate variation can be projected to maps, focusing on the questions of how geography influences language variation and how maps can aid our understanding of that question.

1.1 Visualizing Information

Dialect maps are information visualizations. They are often beautiful to behold, but however pleasing to the eye they may be, their key value in dialectology lies in whether and how well they convey information. Good maps convey information well, even to the point of suggesting the answer to research questions and sometimes even stimulating new questions.

A key issue therefore in evaluating the usefulness and even, the veracity of different mapping procedures is the research question one wishes to address. We doubt very much that there is one, single best map to reveal all the issues concerning dialects and their distributions. If this is the case, then we also need to identify the research questions we have in mind as we examine various mapping techniques. The question we focus on here is a classic question within dialectology: granted that all of dialectology proceeds from the assumption that geography influences linguistic variation, what form does this influence take? We examine several sub-questions: First, is variation organized as a continuum in which the degree of change is gradual throughout, or, alternatively, are there sharp discontinuities (borders)? Second, are there dialect areas, i.e., sub-regions of the area in which a language is spoken which are relatively homogeneous themselves, and divided by clear borders? Third, assuming that there exist at least some dialect areas, do these form a partition of the language area, i.e. a set of non-overlapping sub-regions which together form the language area?

1.2 An Aggregate Perspective

Most dialectological research and very nearly all sociolinguistic research focuses on the distribution of single features, e.g. the pronunciation of /r/ in England, the order of verbs within verb clusters in continental West Germanic, or the realization of the third-person singular present in English (normally /s/, but often zero). We have several reasons to prefer aggregate analyses. First, we obviate the difficult question of choosing which features are to be regarded as singling out significant geographic differences, a question which seriously biases further analysis. Second, on the contrary we can approach the question of single features from the perspective of the aggregate analysis, asking how important they are, i.e., we can ask to what degree single features correlate with aggregate analyses. Third, the methodology of proceeding from aggregate analyses restricts the number of hypotheses. While single-feature analyses can choose any of thousands of features to analyze, the aggregating approach is committed to obtaining an aggregate characterization (preferably through a representative sample, but one which will always contain dozens of the single features). Fourth, and most importantly, we position ourselves to examine general tendencies in linguistic variation by working from the aggregating perspective.

The aggregate perspective is closely related to an older dialectological idea, namely that one might identify dialect borders by observing where isoglosses tend to "bundle". Haag (1898) operationalized this idea, proposing to identify significant dialect areas by drawing many isoglosses, darkening the lines where they coincide (Schiltz 1996). Indeed Séguy regarded the counting of differences between sites as equivalent to counting the isoglosses that divide them (1973:14). We mean to go beyond counting isoglosses, however, in our plea for an aggregate perspective. In the first place, we do not have in mind selecting isoglosses that provide the borders we want, but rather using *all* of our sample of language material. Second, we have in mind rather more sophisticated data analysis than mere counting, as will be obvious below.

Third, the "isogloss-counting" procedure needs sharpening in any case. Consider cases where a feature has several values, so that several isoglosses may separate two sites even while the values at the two sites count only as "different". Concretely, we have in mind cases such as the lexicalization of the concept 'girl' in German, discussed by König (1994, 166). Abstracting from many details, the northwest (Low Saxon) area lexicalizes 'girl' as *Deern*, the southeast (Bavarian) as *Diandl*, and the intermediate area(s) as *Mädchen*/*Mädle*. Two isoglosses separate the southeast from the northwest, marking the borders between the three different values of the same lexical variable. In this sort of case, Séguy's own procedure would have regarded the sites as different in a single feature, i.e. *not* as different in the same degree as the number of differing isoglosses. Séguy must have been aware of this, but perhaps he was willing to ignore such details to focus on the key distinction between analytic techniques versus cartographic ones, namely the distinction between analysis and visualization.

The pronunciation difference metric we shall introduce refines the sorts of measures that can be used, but is similar in quantitative spirit to Séguy's (and Goebl's) approach. The mapping procedures we introduce are useful for all maps involving aggregates, whether those aggregates are obtained via simple counting or through the use of more refined procedures.

1.3 The Case Examined

While the mapping issues we address centrally do not depend on the source of the aggregate characterization, it will ease the understanding of what follows if we provide some concrete background.

In what follows we use material from the *Phonetischer Atlas Deutschlands* (PAD), material collected between 1965 and 1991 by Marburg fieldworkers under the supervision of Prof. Joachim Göschel. 201 words from the famous Wenkersätze were recorded in 186 sites throughout Germany (Göschel 1992). The pronunciations in these recordings were subsequently transcribed by a team of professional phoneticians, including Prof. Angelika Braun of Marburg. They used a methodology in which two phoneticians transcribed each pronunciation independently, and later compared results to obtain consensus transcriptions. Researchers from the University of Groningen digitized the handwritten IPA material in X-SAMPA in 2003 (Nerbonne & Siedle 2005). The material concerns pronunciation exclusively, but we maintain that other linguistic levels will show similar patterns vis-à-vis their projection to geography.

We use the same data used in that paper in the present one so that we do not need to repeat all of the details here. The material consists of the pronunciations of 186 words recorded at 201 different data collection sites and subsequently transcribed in IPA (noted above). X-SAMPA is a computer coding of IPA that the data was then converted to for further processing. For each pair of sites in the collection and for each word in the sample, we examine the two pronunciation variants.

2 Arguing for the Aggregate Perspective

Although the focus of this contribution concerns the mapping of aggregate analyses of variation, the focus of the collection is on mapping variation more broadly, so that it may be interesting to note the application of one of the classes of information visualizations discussed by Tufte (1990, 67ff), that of "small multiples". It plays a role in arguing for the importance of the aggregate perspective, since it demonstrates convincingly that the individual features of most non-dialectometric work do not coincide with each other and also that the individual distributions are geographically exception-ridden. The linguistic point was made by Bloomfield (1933, 321ff), who based his presentation on earlier work by Kloeke, so it is hardly novel. But we submit that Tufte's presentational idea makes the point succinctly.

We use features which König (1994) and Niebaum & Macha (2006) have identified as important in German dialectology. There are nine maps in Fig. 1 in total, and we review the features whose geographic distribution they portray here, discussing the three rows from top to bottom and, within each row, discussing each of the three maps from left to right. The leftmost map in the first row shows the degree to which /p,t,k/ series has remained stops (dark colors in the north), contrasted with the light colors in the south reflecting greater degrees of lenition, including affricates /pf, ʦ, kx/ and fricatives /f,s,x/. This is the famous "second sound shift" in German. The middle map in the top row indicates where non-initial /s/ is palatalized to [ʃ] in words such as *Wurst* 'sausage', *fest* 'firm', *gestern* 'yesterday', *ist* 'is' and *selbst* 'self'. The third and rightmost map in the top row shows the distribution of [s] vs. [z] initially in words such as *Sonntag* 'Sunday', *selbst* 'self', *Seife* 'soap', *sie* 'she', *sieben* 'seven', *so* 'so' and *sollen* 'should'.

We turn the discussion to the middle row. The first and leftmost map sketches the distribution of dental stops after nasals. In the majority of sites /t/ and /d/ are pronounced after /n/, but there is a tendency as one moves northward to omit these. The tendency is heard in frequent words such as *unten* 'underneath', which is pronounced e.g. as [untən] and as [unə] or [unn̩] (with final syllabic n), but also in words such as *anderen* ‘others’ and *gefunden* ‘found (part.)’, often pronounced with no trace of an alveolar stop. The same phonological environment is present in the less frequent *Winter* ‘winter’, but the t/d is only rarely suppressed when *Winter* is pronounced. The middle map in the middle row projects the distribution of the variation in the pronunciation of /r/ onto German geography. The apical [r] is projected onto dark shades and contrasted with the lighter-colored, uvular [ʀ]. The contrast is heard throughout the vocabulary. Finally, the last and rightmost map sketches where the final /n/ is omitted in unstressed syllables such as *machen* ‘make’, *treiben* ‘drive’, *trinken* ‘drink’, *wachsen* ‘grow’, and *werden* ‘become’.

The first and leftmost map in the bottom row sketches the lenition of medial t i.e., [t] vs. [s] in *Wasser* ‘water’. This is a special part of the second shift (top left), namely that affecting the medial /t/ vs. /s/ contrast, but note how fragmented the distribution is. The middle map in the bottom row shows where the initial /g/ is lenited to [x] or [ç] (or even [j]) in participles such as *geschlafen* ‘sleep (part.)’ but also in *gut* ‘good’. Finally, the last and rightmost map in the bottom row shows the distribution of the vowel in *Haus* ‘house’. Vowels are pronounced in so many different ways that simple characterizations are perhaps always misleading. We found 322 (!) different vowels (different combinations of base segment and diacritics) in the six words *Haus* ‘house’, *braune* ‘brown’, *verkaufen* ‘sell’, *auch* ‘also’, *Frau* ‘woman’ and *auf* ‘on’. We divided these into vowels with mid to high back onsets, such as [u, ʊ, ʌ, o, ɔ] and [ɤ], and those with low back or front onsets, such as [a, æ, ɛ, ɪ] and [y].

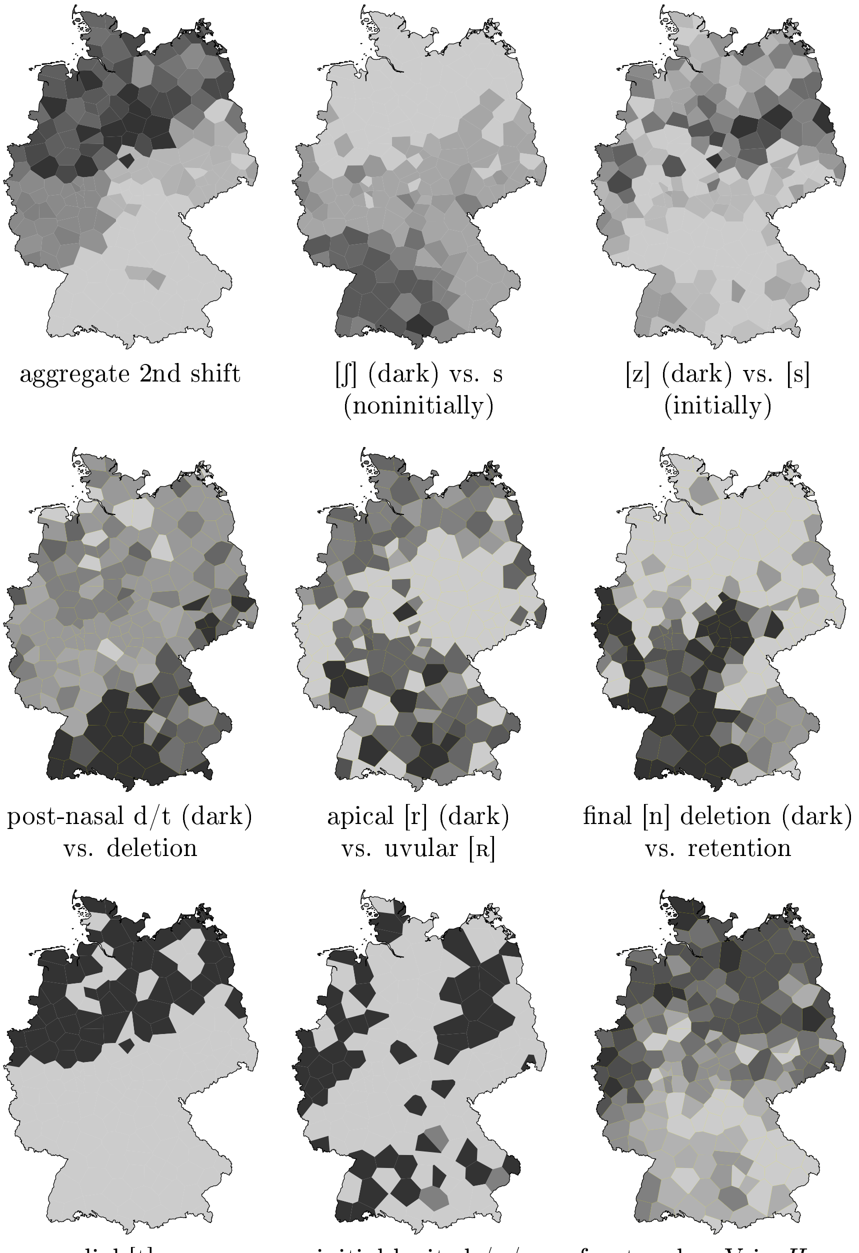
Fig. 1 visualizes the fact that linguistic variation is multifaceted, and that the lines of geographic division do not coincide, both important in the argument that an aggregate analysis is needed if one wishes to make more general statements. But it would be wrong to stop at the observation that the details coincide imperfectly, since Fig. 1 likewise underscores the view that many details of linguistic variation in fact provide similar “signals of provenance.” In spite of the many differences in detail, the multiple small maps of Fig. 1 likewise suggest, at a glance, that the north-south division is signaled by a number of different linguistic options.

2.1 Measuring Pronunciation Distance

We measure the distance or dissimilarity between the two pronunciations using a string comparison procedure known as edit distance, Levenshtein distance or sequence distance. It may be thought of a seeking a transformation of one string into another, keeping track of the number (and “costs”) of operations need to accomplish the transformation. The heart of the procedure is It is illustrated below as applied to two pronunciations of the word *Durst* ‘thirst’ as rendered in IPA:

Aachen	t ʊ ə ʃ	substitute ɔ/ʊ	1
	t ɔ ə ʃ	delete ə	1
	t ɔ ʃ	insert t	1
Vielbrunn	t ɔ ʃ t		
Total			3

The procedure always seeks the least costly set of basic operations capable of performing the transformations, in this case calculating a pronunciation distance of 3 units between these two words. The procedure is admittedly rough,



aggregate 2nd shift

[ʃ] (dark) vs. s (noninitially)

[z] (dark) vs. [s] (initially)

post-nasal d/t (dark) vs. deletion

apical [r] (dark) vs. uvular [ʀ]

final [n] deletion (dark) vs. retention

medial [t] vs. s

initial lenited /g/

front or low V in *Hause*

Figure 1: Individual linguistic differences do not coincide geographically. Tufte's (1990) idea of "small multiples" suggests noting several distinctions together, reinforcing the idea that a more comprehensive view is needed. See text for explanation of features.

but it derives accuracy from the application to large amounts of data. We applied it to each of the 186 pairs of words from each of the 20,100 pairs of sites. For each pair of dialects, we use the mean word pronunciation distance to estimate the aggregate pronunciation distance between the two sites. This procedure has been applied to varieties from about ten different language areas, and is described at length by Heeringa (2004). Heeringa demonstrates *inter alia* that the procedure is consistent when applied to sufficiently large data sets and that it is valid with respect to the judgments of dialect speakers (Gooskens & Heeringa 2004). The procedure used in this paper was updated to take advantage of the improvements described in Heeringa, Kleiweg, Gooskens & Nerbonne (2006), resulting in very little difference in the place-by-place table of aggregate linguistic distances. For technical details we refer to the 2006 paper.

The key for the present article is that the use of Levenshtein distance yields a characterization of the aggregate pronunciation distance between the sites in a given sample. This work thus extends the work of dialectometry (Séguy 1973, Goebl 1984), which likewise provides aggregate characterizations of linguistic distances. Earlier work in dialectometry analyzed the data at a nominal level, where each pair of linguistic items is measured as the same or different, while the application of Levenshtein distance allows numeric characterizations per pair of pronunciations to be obtained. But let us emphasize that the scientific challenge of understanding the aggregate characterizations is the same, no matter how the aggregate characterization is obtained. Whether one examines percentage differences in vocabulary lists or mean Levenshtein distances in pronunciation, the question is then how to proceed with the analysis, and which mapping techniques are most useful and insightful.

Because the procedure is quite abstract, it may be useful to describe it a second time, from a different perspective, from the perspective of alignment. From this perspective, we note that the procedure effects an alignment of the two strings of phonetic segments, identifying corresponding segments as accurately as possible. The procedure aligns the two pronunciations, virtually ensuring that any identical segments are aligned, and disallowing alignments of consonants with vowels. We illustrate the alignment for the two pronunciations of *Durst* used in the example above. Note that corresponding elements are either the results of substitutions or, in the case of insertions and deletions, they involve correspondences between segments and nothing at all:

t	ʊ	ə	ʃ	
t	ɔ		ʃ	t
	1	1		1

The distance is reflected in the alignment as the number of aligned positions which involve non-identical elements, in this case ʊ/ɔ, ə/∅, and ∅/t, where we use '∅' to indicate the empty phonetic symbol.

We emphasize again that the discussion of mapping aggregate distances is logically independent of the means used to obtain those differences. We now turn to mapping.

2.2 Mapping Aggregate Differences

Let's first remark that no one would look forward to a reading a table of 200 × 200 mean varietal distances, even one that is symmetric, allowing the reader to ignore half the entries. That still leaves nearly 20,000 distances. So we

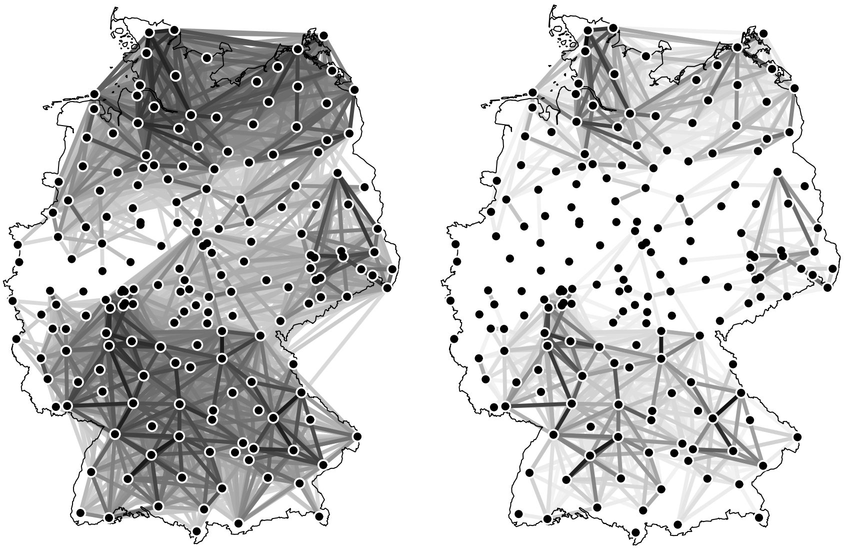


Figure 2: The aggregate pronunciation distances among the 201-site sample in the PAD as portrayed in two network maps drawn from the same data. The map on the left draws the lines darker, revealing more of the geographical structure.

need to visualize the results or to analyze them further, or preferably, both. A "network" map is a simple initial visualization made by drawing lines between data collection sites where the darkness of the line is inversely proportional to the linguistic distance between the sites, so that the less distant (or more similar) the sites are linguistically, the darker the line. In principle there is a line connecting every pair of sites whose darkness is inversely proportional to the linguistic distance between the two sites (and directly proportional to the linguistic similarity). If lines are not visible between some pairs of sites, that indicates that they are linguistically very different. Note that the scale of the darkness of the line influences how well dialectological structure is revealed. At either extreme there is no contrast, and the resulting map is maximally light or dark.

This sort of visualization is only useful if there is sufficient contrast, and one normally adjusts the darkness so that groups of sites make themselves apparent, even while white bands appear along borders. Sometimes the lines in network maps are only drawn if they are shorter than a fixed limit, emphasizing local relations. Fig. 2 shows two line maps made at two scales of darkness. Both of them are "correct", but my own preference is for the map on the left-hand side because it reveals more of the structure in the data.

If we turn to the research question of how geography influences linguistic variation, then the network map is suggestive about the role of areas in structuring linguistic variation. If linguistic variation were perfectly continuous (the way direct geographic distance is), the network map would show dark lines from each point to each of its nearest neighbors, and increasingly light lines to other sites. The overall result would be a map of little contrast. The fact that we see swaths of relatively light color is an indication that there are lines of discontinuity, presumably borders around dialect areas.

We can sharpen our question about continua vs. areas in the following way. It is completely clear that geography influences linguistic variation, but in asking about the role of dialect areas, we are asking whether there are mediating structures (areas) or, alternatively, whether those structures implicit in geography, including especially geographic distance, are sufficient. If dialects are organized as continua, then simple geographic models should suffice to predict linguistic distance. The network map suggests that areas have a role to play.

Let us add that we have used geographic distance "as the crow flies" as a particularly straightforward operationalization of "the chance of social contact", which is presumably the genuine influence. Gooskens (2005), in demonstrating that travel time is a better predictor of linguistic distance than simple distance "as the crow flies", presents convincing indication that social contact is the real variable behind geographic distance.

3 Clustering

In this section we present means of seeking relatively homogeneous groups among the survey sites. The first subsection concerns basic "clustering" techniques, and second a technical improvement which presents its own mapping challenge.

3.1 Basic Clustering Concepts

In order to ask our question about the role of dialect areas more concretely, we need to identify the best candidates for dialects areas in the PAD. For this purpose we apply clustering, more specifically, hierarchical agglomerative clustering, an accepted method for searching for groups in all sorts of data, including dialectological data. Again, we cannot present all the details (and especially not all the variants) of clustering here. We note that clustering is a "greedy" procedure that inputs a distance matrix of the sort we describe above, and, at each step, fuses the two data points that are closest to each other, obtaining fused "clusters" of ever increasing size. If we imagine clustering as creating a tree of similarity, then the step of fusing two points results in a node under which the two fused elements appear as daughters.

We apply clustering to the average pronunciation distances as measured by the Levenshtein procedure with the goal of determining which sites are most similar in pronunciation on average. Note that there is no geographic information which our application of clustering uses as input so that geographic proximity plays no role whatsoever in the results of clustering. We nonetheless expect to see that our application of clustering respects geography, i.e. yield geographically coherent groups. But this expectation arises from the fundamental dialectological postulate that nearby varieties are normally more similar linguistically that distant ones. If clustering indeed produces geographically identifiable groups, this confirms the soundness of our analysis, but other results would be surprising.

There is a great variety of clustering algorithms, depending on which rule is used to calculate the distance from a newly fused element to all the remaining elements in the matrix. We will use an algorithm here that simply takes the mean distance of all elements in the newly fused group to each of the elements

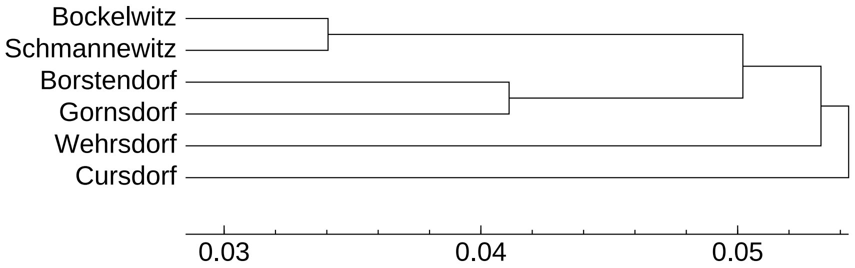


Figure 3: Example dendrogram where cophenetic distance is reflected in the horizontal distance from the leaves to the encompassing node. The cophenetic distance between Bockelwitz and Borstendorf is a bit less than 0.05.

in the remaining matrix (including both groups and individual sites). This is known as "weighted pair group method with average", or WPGMA, and its simplicity recommends it here. Let us not ignore the fact that the choice of cluster technique is controversial. Please consider that we have made a reasonable choice here, but should another ultimately prove to be preferable, the further steps in the analysis would nonetheless be sensible. See Heeringa (2004) and Felsenstein (2004) for basic presentations, and Jain, Murty & Flynn (1999) for details and comparison to other procedures.

It is important for our purposes that the clustering procedures result in a "re-estimation" of the distances between collection sites, the so-called 'cophenetic distance'. The cophenetic distance is the distance between two sites at the point at which they are fused in the clustering process. Cophenetic distances distort the original distance matrix because of the stipulation that the distance between the newly fused nodes and all others be the average of the distances from each component of the fusion to the others. Advanced work seeking to improve clustering techniques tries to ensure as close a fit as possible between the cophenetic distance and the pre-clustered distances. Cophenetic distance may be thought of as the distance according to the best grouping (the one found by the clustering). The groups in a cluster result from specifying a cophenetic distance and then regarding everything joined at a shorter distance as a group. Fig. 3 illustrates a dendrogram in which cophenetic distance is shown and in which groups are defined.

We illustrate cophenetic distance more concretely using the dendrograms in Fig. 3. Wherever two leaves are immediately joined (e.g. Borstendorf and Gornsdorf), the cophenetic distance is just the distance in the input matrix. Otherwise, the cophenetic distance between two leaves is the distance to smallest encompassing node. The cophenetic distance between Bockelwitz and Borstendorf is a bit less than 0.05.

We obtain groups from a dendrogram by examining the groups joined below a threshold. In terms of the diagram above, we draw a vertical line at some point and regard as groups everything joined to the left of that line. If we drew the line at 0.045 in Fig. 3, then we would recognize two groups of two plus two "islands", Wehrsdorf and Cursdorf, four in total.

Recalling that clustering (in the form we have used) makes no use of geographic information, but instead seeks groups defined purely by linguistic similarity, one useful step in understanding the potentially mediating role of areas is to examine the projection of groups found by linguistically based clustering to geography. While there are many ways of doing this, a very simple way is simply to assign a discrete symbol to each recognized group in a dendrogram. Fig. 4 shows the division of the German data into the nine largest groups in the dendrogram. The fact that the groups may be assigned to non-overlapping areas (with the exception of two sites, representing 1% of the data) is an indication of geographic cohesion in the data. The degree to which the dialect groups correspond to geographic areas also suggests the importance of those areas as organizing elements.

Note that the cluster groups do not correspond perfectly to geographic areas, however, especially the single asterisk on the northeast border, and one triangle among +'s and ×'s. Normally exceptional sites such as these two lead to follow-up analyses, involving checks on the data, transcription, digitization and comparison. This illustrates yet another ways in which mapping influences dialectological analysis, as a control for collection and analysis procedures. As we wish to illustrate mapping techniques here, we continue to other topics.

The map using discrete symbols to identify membership in one of the major "clusters" serves the important purpose of facilitating comparison to earlier dialectology, where results were normally presented in terms of dialect areas. But it is an impoverished visualization of the clustering result because fails to represent two significant aspects of the clustering, namely cophenetic distance and whether one cluster is a sub-cluster of another. All the different areas are distinguished discretely so that, if we filled in the dialect areas with different colors, we would also see discrete borders leading to the impression that with "one step, you hear Bavarian". The impression of local speakers and that of dialect experts is that borders are seldom so precise. We shall keep in mind these criticisms as we develop other maps, but, first we re-examine the use of clustering to identify groups.

3.2 Stable clustering

Clustering depends on finding the minimal elements in a matrix, and these may differ very little. This means in turn that small differences in the input data matrix may influence the resulting clusters a great deal. Even if the "instability" of clustering plays little role in the discussion of how to detect groups in dialectological data (Moisl & Jones 2005), we suggest that it should. The established method for counteracting the instability in clustering is to introduce a stochastic element in the process. In our case we apply the "bootstrap", varying the set of words we use to obtain the cluster, and allowing some to be repeated. An alternative is to add random small amounts of noise to clustering, but Nerbonne, Kleiweg, Heeringa & Manni (2008) show that the bootstrap is equivalent to this, leading to a correlation of 0.997 on the same data set used here.

To be concrete, we choose words randomly from our sample, and once we have chosen a word, it is not blocked from being chosen again, i.e. we use selection with replacement. We choose e.g. 200 words from our sample randomly, with replacement, meaning that some words will be chosen more than once, and others not at all. Once we have a selection, we cluster, and then we repeat the process again, with a new selection of words. When we cluster using the bootstrap, we identify groups more robustly, e.g., by focusing only on those groups which are found in at least 50% of the clusterings (runs of the bootstrap

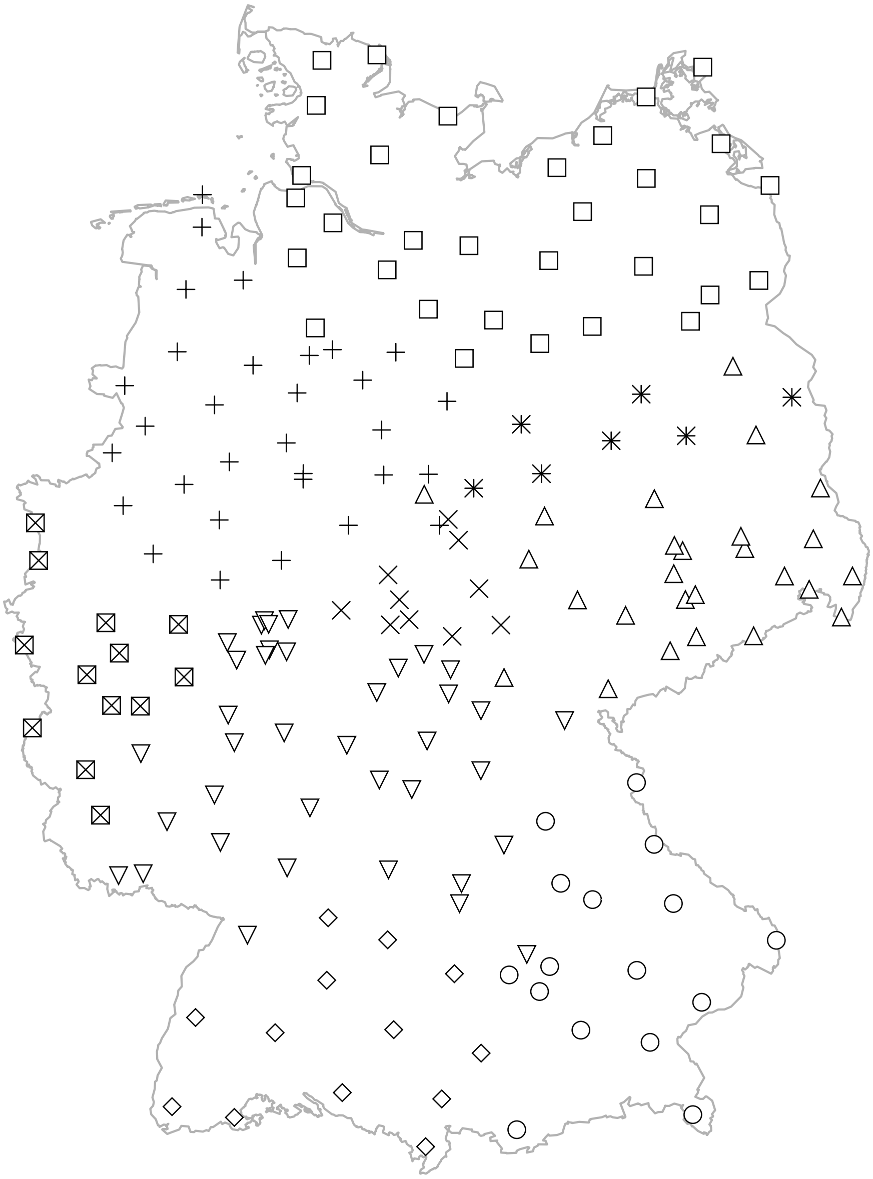


Figure 4: The nine largest groups in the clustered German PAD data. With the exceptions of a single asterisk on the northeast border, and one triangle among +'s and ×'s, this data falls neatly into areas.

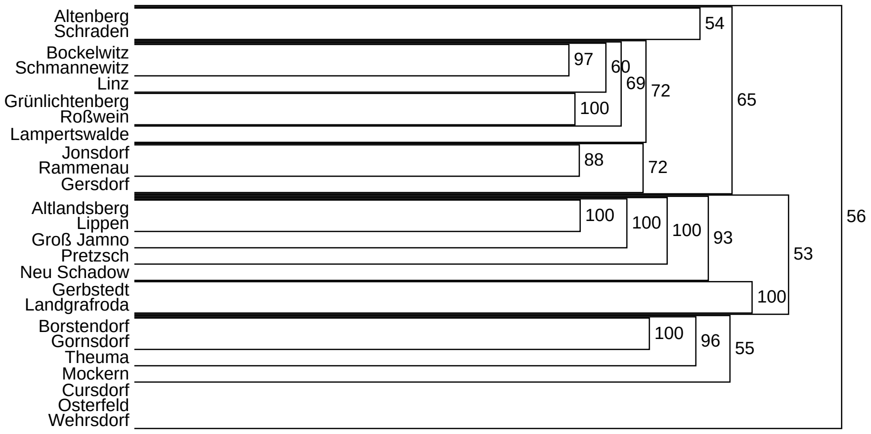


Figure 5: A section of the "consensus dendrogram" resulting from 100 bootstrap runs of the clustering algorithm.

procedure just outlined). The result can be represented in a dendrogram, but now one in which branching is no longer binary.

Regarding the consensus dendrogram in Fig. 5, note the numbers associated with the brackets, indicating how often a group was found. We interpret this as an indication of the reliability of the hypothesis that the items grouped under the bracket form a dialectologically coherent group. The length of the bracket is the mean cophenetic distance found in the runs where the group was found, analogous to the branch length in simple dendrograms.. Note that consensus dendrograms need not be binary-branching even when every contributing dendrogram would have been. An example of this is seen in the first interior group in Fig. 5, which was detected 65 of 100 times (from Altenberg through Gersdorf). It includes three subgroups, found 54, 72 and 72 times, respectively. This almost certainly means that sometimes the first two groups were found together, sometimes the second two, other times the first and third, but no single pair was found a majority of the time.

How might we visualize the results of this process? We might naturally use the same mapping technique used in Fig. 4, in which elements are represented as if they were strictly partitioned, but that mapping technique would need to strain to represent the reliability parameter in the consensus dendrograms. We wish to project the results of clustering onto geography in a way that is more informative than simply introducing discrete symbols for each of the major groups in the consensus dendrograms. Peter Kleiweg developed a "composite cluster map" (Kleiweg, Nerbonne & Bosveld 2004), in which he begins from a standard Voronoi tiling of an area, which divides the map into regions around collection sites (see glossary), i.e. one which optimally divides an area into polygons so that each sampling site within the area is maximally separate from its neighbors within its own polygon. Note that in this tiling there is always a unique border separating any two adjacent sites and perpendicular with respect to the line joining the two sites. Kleiweg's idea was to draw that border, shading the line in inverse proportion to the cophenetic distance between the two points in the consensus dendrogram.

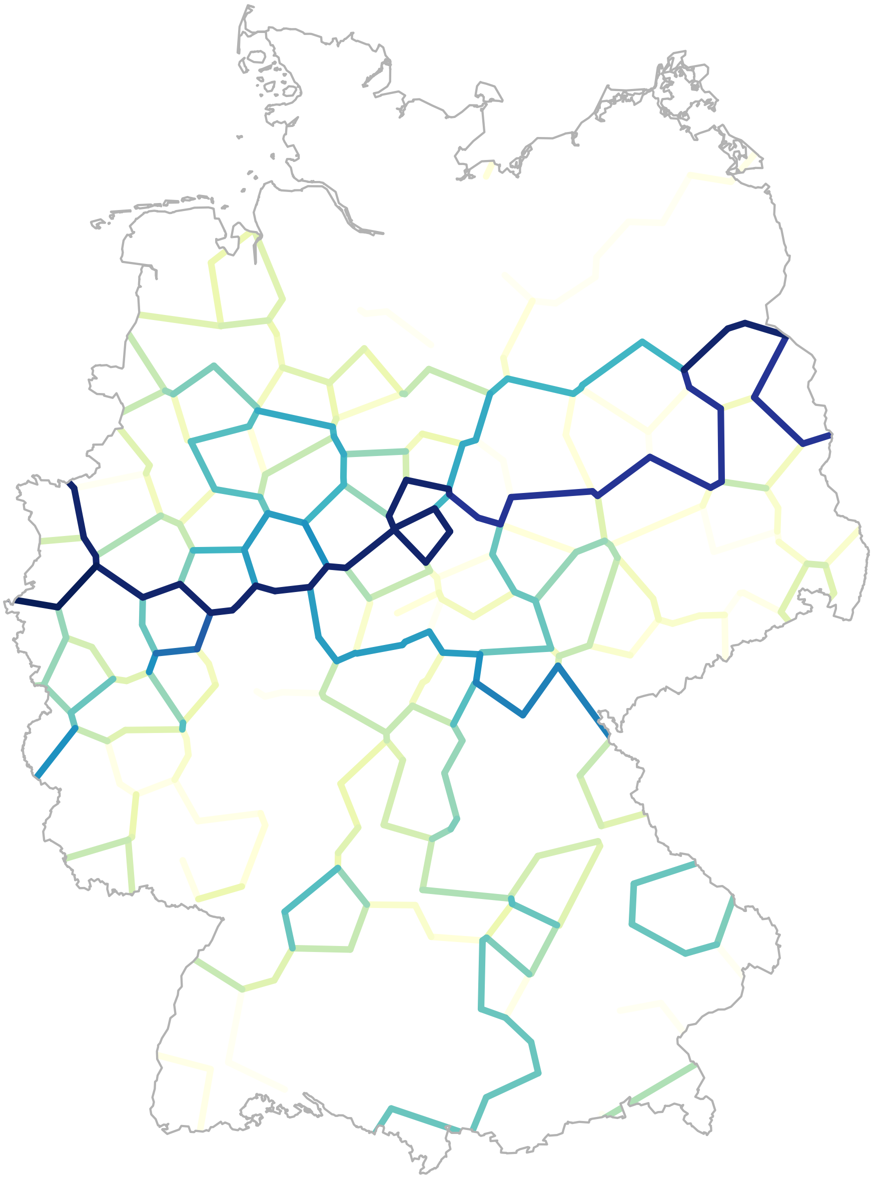


Figure 6: A composite cluster map. The darkness of each border separating two adjacent sites is inversely proportional to the cophenetic distance between them in the consensus dendrogram.

The composite cluster map (Fig.r̃effigv) represents the information in the consensus dendrogram and the clustering procedures more completely and more effectively than the maps which show only the top elements in a discrete partition (Fig. 4). The darkness of the borders reflects the statistical stability of divisions, and the map is not restricted to reflecting only the most important divisions. In both respects the composite cluster map is superior to the discrete partition map. Let us therefore consider whether the composite cluster maps might provide further insight into the dialectal relations among the sites in our data set.

Turning to our research questions, we note that composite cluster maps suggest that there certainly are borders, but, unlike the discrete-partition maps, it also identifies some borders as more important than others. Recalling that we assume that geography influences linguistic variation, we nonetheless ask what form this influence takes. Fig. 6 shows one very clear border, namely the dark border running from east to west and dividing Germany into a northern and a southern area. Other borders are much fainter and do not always enclose coherent areas. More concisely, these maps visualize dialectology as involving borders and therefore areas, but only to a rather limited degree. Finally, the composite cluster map does not reinforce the idea that dialect areas divide a language area exhaustively. Instead composite cluster maps emphasize borders which may fail to result in a clean "partition", or set of non-overlapping subregions which together constitute the language area.

4 Multidimensional Scaling

This section presents an alternative to clustering as an analysis of the aggregate distance matrices produces in dialectometry. The first section presents the basics of multidimensional scaling (MDS), and the second presents means of mapping MDS.

4.1 Reducing Dimensions

The aggregate analyses visualized above are based on large quantities of data and are therefore well positioned to answer questions about the overall speech habits of the (German) language community, including questions about the degree to which these speech habits are reflected in other cultural or biological traits. The aggregate analyses, by examining *all* the signals of provenance which language users emit, are extremely reliable.

But we also would like to determine the most important and reliable of the signals and to verify that it is indeed these signals that matter most to the language community at large. For this reason, we wish to apply techniques capable of isolating the most important and reliable of the signals. Dimension reducing techniques are promising in regard to this purpose.

Another reason motivating linguists to study dialectology is the reflection it provides of general tendencies found in all languages and dialects. Dialects, seen from this perspective, are yet another instantiation of the language capacity which humans universally possess. By examining the set of signals which together indicate a particular provenance, we hope to see whether some of them are strongly associated for structural reasons. For this purpose, as well, it is interesting to distill from the large aggregate of differing speech habit the most important and reliable ones.

Multidimensional scaling (MDS) is a statistical technique aimed at representing very high dimensional data in a smaller number of dimensions. One can imagine it procedurally. Beginning from the symmetric distance matrix we described earlier, we attempt first to assign each element (collection site, in our case) a coordinate along a single dimension. Note that this assignment implicitly reassigns distances to all the pairs of points, namely the difference between the coordinates. We compare this distance implied by the coordinate assignment to the original distances in the input distance matrix. The distances normally correspond poorly, so we then perturb the initial (existing) assignment of coordinates, checking whether the distances implicitly assigned by the new coordinates are an improvement over the original. One coordinate assignment is better than another if the distances it implicitly assigns match those of the input matrix more closely (e.g., as measured by the correlation coefficient). It will normally be impossible to match distances perfectly as the single dimension will not provide enough flexibility. We therefore seek solutions in two and three and more dimensions, now assigning pairs and triples and n-tuples of coordinates, measuring the distance between points using Euclidean distance, and continuing to seek optimal coordinate assignments. Fig. 7 displays the two-dimensional solutions, using the symbols of the discrete partition map (Fig. 4).

Naturally we make use, not only of maps, but also of the plot of MDS coordinates normally supplied as part of the procedure. The plot shown in Fig. 7 suggests that there are two major dialect areas in Germany, north and south, that some further differentiation is possible in the north, that the southern varieties are less sharply distinguished, and that the Palatinate varieties are particularly difficult to pigeon-hole, at least in two dimensions.

Note that the northern and southern groups (see Fig. 4) are mostly well separated in the graph in Fig. 7: there is no overlap whatsoever, and a wide swath of white testifies to the linguistic dissimilarity of the different varieties. The three northern varieties, represented by the boxes, plusses and asterisks, are fairly separate from the southern varieties in this two-dimensional solutions, and they are somewhat separate from each other, unlike the five southern groups, represented by circles, diamonds, crosses, triangles and inverted triangles. Finally, the crossed boxes clearly represent varieties which fit less well in the overall division, the Palatinate and Ripuarian varieties.

MDS was introduced to dialectology by Embleton (1993), citing earlier applications applying MDS to lexico-statistical data from different languages (Black 1976, *inter alia*). We emphasize three aspects of MDS that make it especially attractive for advanced dialectological use. First, MDS is not plagued with the instability we needed to guard against in the use of clustering. While there is an element of estimation in obtaining MDS solutions, the solutions, once obtained, do not change greatly when the input changes in minor ways. Second, we have a means of quantifying the quality of the lower dimensional solutions (something a bit like cophenetic correlation in the case of clustering). If we measure the quality of a solution using the correlation coefficient r, then we may also quantify the degree to which the solution explains the original variation as r^2, the square of correlation coefficient. And third, MDS suggests an approach to exploring the linguistic basis of the aggregate classification. We attempt to interpret the most important reduced dimensions of the MDS solution in order to

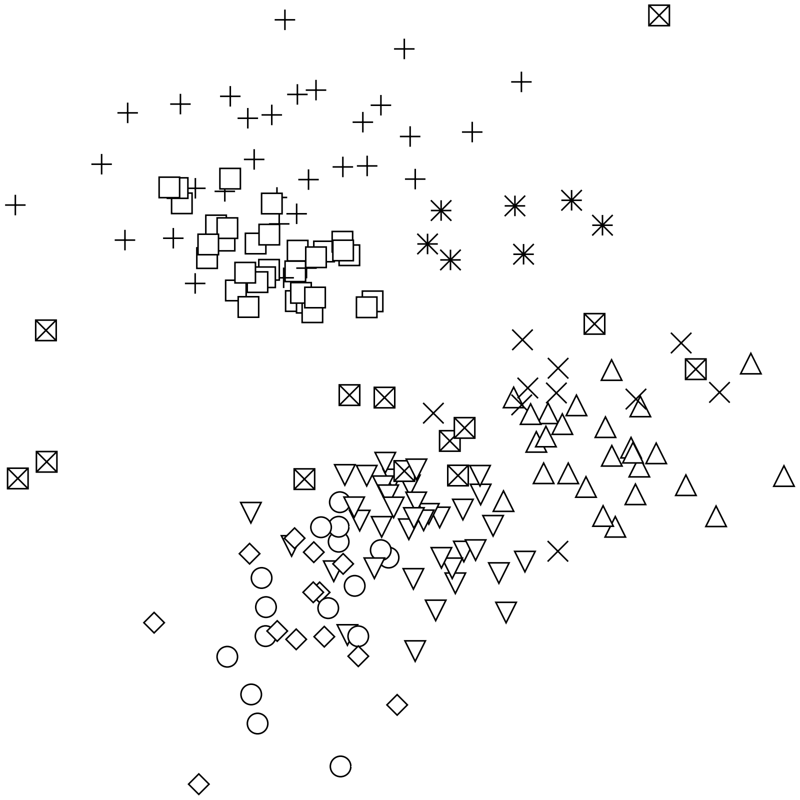


Figure 7: The pronunciation distance data from the PAD in two MDS dimensions, using the same symbols as Fig. 4.

identify their associated linguistic signals. We sketch one approach to doing this below.

Let us note then that the correlations between the distances in the original distance matrix and the distances implicitly assigned via the assignment of coordinates is 0.76 in the one-dimensional solution, 0.88 in the two-dimensional solution, and 0.93 in the three dimensional solution, so that these solutions account for 58%, 77% and 86% of the variation in the data, respectively. With respect to seeking the linguistic basis of aggregate characterizations, we note that this is an active area of research where canonical solutions have not emerged (Prokić 2007), but if we examine the pronunciation distance matrices per word, then we can ask which words correlate best with individual dimensions of the MDS solutions. Examining the most important dimension, it turns out that *Zeiten* 'times', *sein* 'be (inf.)', *bleib* 'stay (imp.)' and *weiße* 'white (infl.)' are at the top of the list, all showing strong correlations ($r > 0.5$), suggesting that the variation in the stressed vowel (standard German [aɪ], but South [i]) is the single strongest indicator of the most importance dimension of variation among the 201 words in our sample.

4.2 Mapping MDS

Although the combination of MDS plots (Fig. 7) in which sites are symbolized in a way that facilitates recognizing their geographic locations (Fig. 4), it is worthwhile trying to project MDS results more directly onto geography. Wilbert Heeringa and Peter Kleiweg attribute the essential idea in the following suggestion to each other. We first normalize the coordinates assigned in the three-dimensional MDS solution by expressing them as a percentage of the maximal value in a given dimension, so e.g. a coordinate of zero on a scale of -2 to $+3$ would be interpreted as 0.4 (or 40%). We then regard these normalized values then as indicators of intensity in a standard three-element color scheme. A site with normalized coordinates of 40%, 10% and 90% would then result in a color with 40%, 10% and 90% of the intensities of green, red and blue. Fig. 8 contains a map constructed in this way based on the same analysis of the PAD data used above. We have added a legend so suggest the sorts of pronunciation differences which are most prominent in the analysis.

Given that the map displays three dimensions which together account for 86% of the variation in the PAD data set, it is fair to say that it gives us a good picture of German dialect variation. If, on the basis of this map, we reflect again on the question of how geography influences linguistic variation, then we see that the notion of dialect continuum emerges here saliently. The existence of this continuum is compatible with the existence of areas, but denies the existence of sharp borders. There have been portrayals of dialect continua based on single features and small numbers of features, but the degree to which these depend on the fortunate choice of the right features has been unclear. The use of an aggregate analysis based on a large sample means that the results are likely to be stable and not to depend on the fortunate choice of material or sampling sites. We believe this therefore to be the first analytically well-founded portrayal of the dialect continuum

Although we thought that we had invented this sort of map in 1999 (see too "Acknowledgments" below), a similar sort of map was published by Cavalli-Sforza, Menozzi & Piazza (1994), illustrating the distribution of genetic vari-

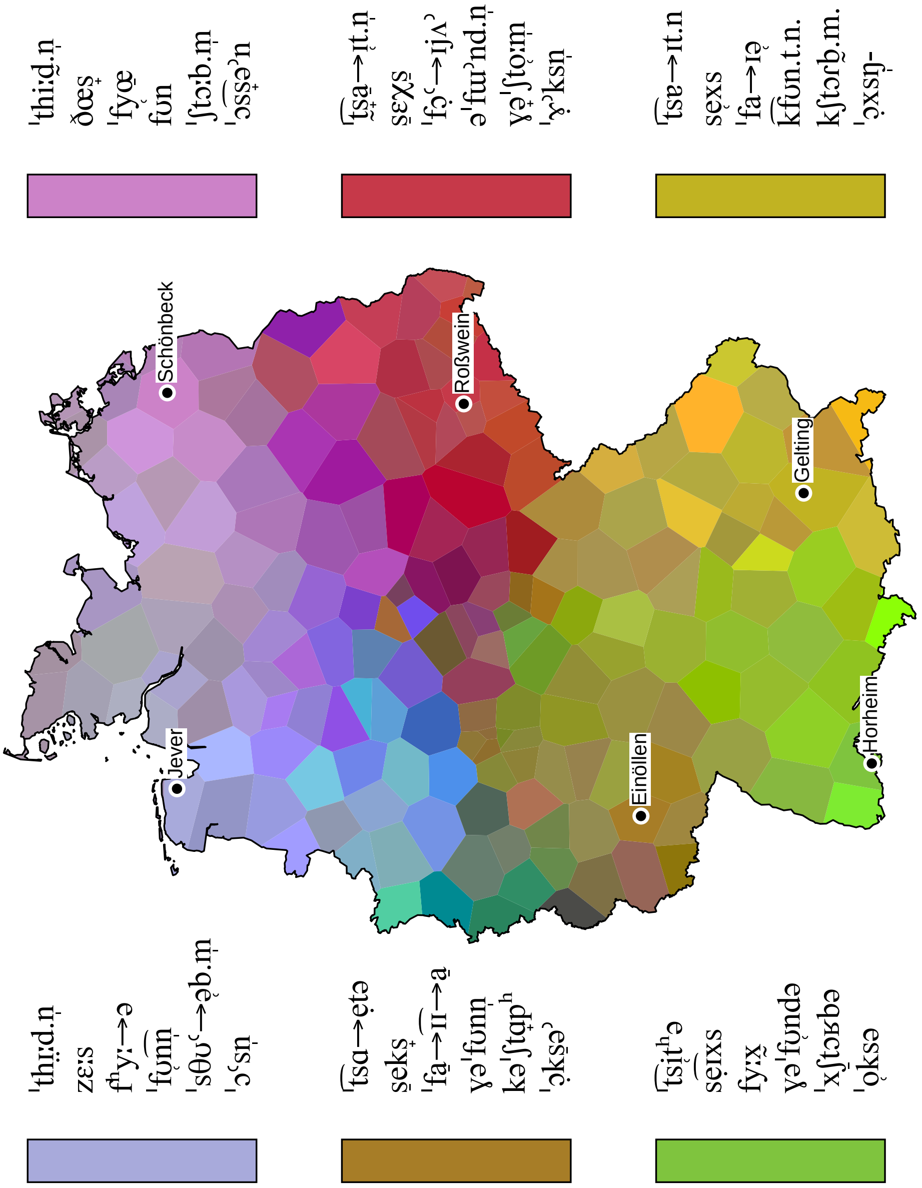


Figure 8: Aggregate pronunciation variation as analyzed via MDS, whence major dimensions are mapped to colors in the red, green, blue color scheme.

ation in the peoples of the world (see e.g. the first map in the color section of Cavalli-Sforza et al. 1994). There are also technical differences in the maps. Cavalli-Sforza et al. used principal component analysis (PCA) to analyze their data, while we used MDS. But PCA and MDS yield nearly identical solutions, even though MDS is the recommended procedure for reducing the dimensionality of distance tables, while PCA works on arbitrary sets of vectors (Tabachnick & Fidell 2001, 582ff). Second, a technical detail: we mapped our dimensions to the computer color scheme red, green, blue ('RGB') (or, for printing, cyan, magenta, yellow, 'CMY'), while the maps in Cavalli-Sforza et al. used yellow, red and green. With enough experimentation, either color scheme should be fine, but the standard schemes are designed to support mixtures that remain distinct to the eye.

Finally, while Cavalli-Sforza et al.'s map shows variation as smooth, the map in Fig. 8 indicates the fineness of the sampling grid, and each polygon is given a unique color that does not necessarily blend smoothly into the color of its neighbor. It is technically straightforward to smooth the colors, by first introducing a number of additional points to the map (the more points, the smoother the color shifts), and secondly determining the color of the additional points by interpolating the colors of the surrounding points (e.g. using an inverse-square weighting). Nerbonne & Siedle (2005, 144) provide a smoothed map also based on the PAD. For the purpose of displaying results to researchers, however, we prefer exposing the grid as in Fig. 8. This illustrates the density of sampling, and can serve the purpose of drawing attention to anomalies in the results-sites which interrupt the smooth flow of the continuum.

Our shift to the use of color means that we need to be cautious about the fact that the human eye is more sensitive to some colors and color distinctions than others. In fact Imhof's *Kartographische Geländedarstellung* recommends as a first rule for the use of color in maps:

> Reine, leuchtende oder sehr kräftige Farben wirken schreiend, unerträglich wenn sie unvermittelt und großflächig nebeneinander stehen [...] Gibt man allen und insbesondere auch den großen Flächen schreiende, satte Farben, so wirken die Bilder bunt, ungeordnet, verwirrend und unschön. (Imhof 1965, 83).

Quoted by Tufte (1990), who offers a translation by (Steward, 1982):

> Pure, bright or very strong colors have loud, unbearable effects when they stand unrelieved over large areas adjacent to each other [...] If one gives all, especially large areas, glaring, rich colors, the pictures have brilliant, disordered, confusing and unpleasant effects.

We have definitely not used "pure" colors in our maps, but some shades are bright and strong. By offering the legend supporting linguistic interpretation, we hope to engage the observer in an exploration of what the distinctions mean dialectally.

Finally, we would like to note that MDS is very sensitive to scale. For example, if we examine only the northernmost section of the distance matrix obtained from the PAD via the edit-distance, and apply MDS to that, the results will emphasize the differences in the restricted area. This increase the flexibility of the MDS-based colored map as a tool for the exploration of dialectometric results. Fig. 9 shows the results of applying MDS to the distance matrix obtained by examining only the sites in the northernmost section of the map. It should be compared to Fig. 8 to appreciate how much more differentiation is present when one restricts one's view to a limited area. The maps Fig. 8 and 9 are based on exactly the same pronunciation difference measurement applied to the data in the PAD. The only difference is that we have restricted MDS to apply only to the sub-table of differences involving the northern dialects. Note how much more finely the map in Fig. 9 is differentiated in contrast to the map in Fig. 8, particularly in the west. The east, by way of contrast, is fairly homogeneous.

5 Conclusions, Discussion and Future Work

We have emphasized the dialectometric perspective, and in particular, its inherent aggregation, in this sketch of opportunities for mapping aggregate linguistic variation. Working at a high level of aggregation, we naturally ask questions about how geography influences variation—whether so-called "dialect areas" mediate in this influence or not. If they do, we expect to find relatively distinct borders reflecting a larger local rate of change. If areas are not important, we expect to see a continuum reflecting a relatively constant rate of change. Seen from this perspective, it is natural to seek (linguistically similar) groups in data and then to check whether these groups are geographically coherent. We applied (hierarchical) clustering for this purpose, confirming that the clusters we found were indeed geographically coherent, and we found it useful to compare the clusters of dialects thus obtained to traditional dialectological scholarship. When we compared the discrete cluster maps to the clustering dendrograms, we were discouraged by paucity of detail in the maps (as opposed to the dendrograms), and we sought means of exposing more detail. At the same time, we turned to the bootstrap to obtain more reliable indications of grouping. The iterations within bootstrap together with the cophenetic distance in the dendrogram suggested the composite cluster map, in which average cophenetic distance between adjacent sites is reflected in the darkness of a border drawn between them. For the researcher, composite cluster maps are much more interesting for the wealth of detailed information they expose.

At this point it is worthwhile reiterating that, even while the aggregate perspective is indebted to the early idea of "bundling isoglosses", it nonetheless surpasses it when combined with modern data analysis possibilities such as bootstrap clustering. For example, bootstrap clustering finds groups of similar linguistic varieties regardless of their geographic relation—while isogloss bundling only makes sense for local distinctions. Aggregation is also less selective about which features to map, and it is valuable even when the purpose is not to find distinctions, but rather to examine how smoothly the distinctions accumulate with respect to distance, the question of the degree to which linguistic variation constitutes a continuum.

As an important linguistic question concerns the nature of the differences we aggregate over, we turned to multi-dimensional scaling (MDS) and opportunities for mapping based on it. An important step in any MDS analysis is the interpretation of the dimensions to which distance tables are reduced, and we examined simple correlations between (tables based on) individual words on the

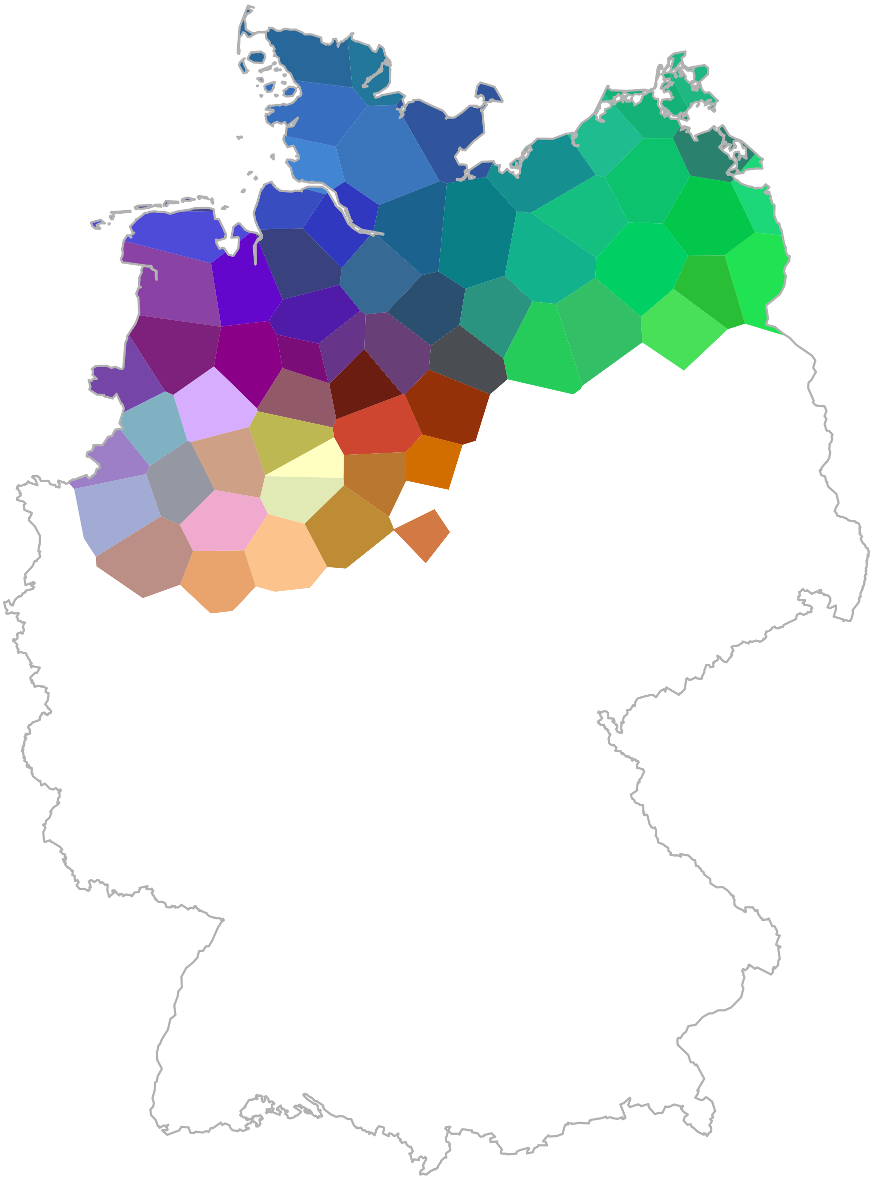


Figure 9: Aggregate pronunciation differences in the PAD, as analyzed via MDS, when attention is restricted to northern Germany.

one hand and aggregate distance tables on the other. The results are shown in map legends. The fact that there exist high-quality, low-dimensional MDS solutions suggests that linguistic variation is less complex than it first appears, and the maps derived from MDS visualize the commonly invoked notion of a dialect continuum in a well-founded and general procedure, we believe for the first time.

Several further interesting opportunities and questions concerning the mapping of aggregate variation analyses suggest themselves. We suggest three. First, although we have shown how to examine the linguistic basis of aggregate analyses in a simple way, searching for words that mimic the important dimensions of variation, linguists are very interest in more abstract levels of linguistic structure, and it would be very helpful to develop techniques to isolate these and mapping techniques to display them. This topic requires more than simply cartographic innovations.

Second, many researchers have been interested in the extra-linguistic correlates of linguistic variation. Gilliéron, the father of French dialectology, was famous for suggesting that the linguistic divisions running roughly East-West across France, corresponded well with established cultural divisions, in particular, the ethnic split between the slightly Romanized Celts in the north, and the thoroughly Romanized non-Celts in the south, the legal division between the common-law north and the Roman law south, and patterns of agriculture and architecture (Chambers & Trudgill 1998, 95–103). Manni, Heeringa & Nerbonne (2006) explore the degree to which family relationships coincide with dialect divisions. There are many more interesting questions. The mapping challenge is to develop analytical techniques and mapping projections that support these explorations.

Third, and technically more challenging, we would like to have interactive maps allowing the exploration of variation at various geographic scales. Since MDS analyses are scale sensitive, as we noticed, in would be intriguing to develop an interactive map for which MDS analyses (and MDS-based) maps were recalculated interactively, based on the user's specification. The ideal visualization would allow one to "zoom" in on interesting areas, and to seek the linguistic bases of the major distinctions.

Acknowledgments

My debt to Peter Kleiweg, who has realized all the maps discussed in this article and who invented one of them and co-invented a second extends well beyond the normal debt owed to a programmer. I also thank Wilbert Heeringa for co-inventing the MDS-based color maps. I originally thought Heeringa and Kleiweg invented it together, but it turns out that each of them credits the other. Our publication was Nerbonne, Heeringa & Kleiweg (1999), the foreword to the re-edition of Sankoff and Kruskal's classic *Time Warps, String Edits and Macromolecules.* Joseph Kruskal was kind enough to send me a Bell Laboratories *Technical Memorandum* dated Jan 15, 1974 which appears to be Black's original paper applying MDS to linguistic data, later published as Black (1976). I am also pleased to thank the members of the computational variationist group in Groningen, Charlotte Gooskens, Sebastian Kürschner, Therese Leinonen and Jelena Prokić for general discussion of the issues here, and the editors of the

present volume for their detailed comments on an initial version of the paper.

References

Black, Paul. 1976. “Multidimensional Scaling applied to Linguistic Relationships.” *Cahiers de l’Institut de linguistique de Louvain* 3:43–92.

Bloomfield, Leonard. 1933. *Language*. New York: Holt, Winehart and Winston.

Cavalli-Sforza, Luigi Luca, Paolo Menozzi & Alberto Piazza. 1994. *The History and Geography of Human Genes*. Princeton: Princeton University Press.

Chambers, J. K. & Peter Trudgill. 1998. *Dialectology*. 2nd ed. Cambridge: Cambridge University Press. [1980].

Embleton, Sheila. 1993. Multidimensional Scaling as a Dialectometrical Technique: Outline of a Research Project. In *Contributions to Quantitative Linguistics*, ed. Reinhard Köhler & Burghard Rieger. Dordrecht: Kluwer pp. 267–276.

Felsenstein, Joseph. 2004. *Inferring Phylogenies*. Sunderland: Sinauer.

Goebl, Hans. 1984. *Dialektometrische Studien: Anhand italoromanischer rätoromanischer und galloromanischer Sprachmaterialien aus AIS und ALF*. Tübingen: Max Niemeyer. 3 Vol.

Gooskens, Charlotte. 2005. “Traveling Time as a Predictor of Linguistic Distance.” *Dialectologia et Geolinguistica* 13:38–62.

Gooskens, Charlotte & Wilbert Heeringa. 2004. “Perceptive Evaluation of Levenshtein Dialect Distance Measurements using Norwegian Dialect Data.” *Language Variation and Change* 16(3):189–207.

Göschel, Joachim. 1992. Deutscher Sprachatlas. Technical report Das Forshungsinstitut für deutsche Sprache Marburg: .

Haag, Karl. 1898. *Die Mundarten des oberen Neckar- und Donaulandes*. Reutlingen: Buchdruckerei Egon Hutzler.

Heeringa, Wilbert. 2004. Measuring Dialect Pronunciation Difference Using Levenshtein Distance PhD thesis University of Groningen.

Heeringa, Wilbert, Peter Kleiweg, Charlotte Gooskens & John Nerbonne. 2006. Evaluation of String Distance Algorithms for Dialectology. In *Linguistic Distances*, ed. John Nerbonne & Erhard Hinrichs. ACL Workshop held at ACL/COLING, Sydney Shroudsburg, PA: ACL pp. 51–62.

Imhof, Eduard. 1965. *Kartographische Geländedarstellung*. Berlin/New York: Mouton De Gruyter.

Jain, Anil K., M. Narasimha Murty & Patrick J. Flynn. 1999. “Data Clustering: A Review.” *ACM Computing Surveys* 31(3):264–323.

Kleiweg, Peter, John Nerbonne & Leonie Bosveld. 2004. Geographic Projection of Cluster Composites. In *Diagrammatic Representation and Inference*, ed. Alan Blackwell, Kim Marriott & Atsushi Shimojima. Number 2980 *in* "Lecture Notes in Artificial Intelligence" Third International Conference, Diagrams 2004. Cambridge, UK, March 2004 Berlin: Springer pp. 392–394.

König, Werner. 1994. *DTV Atlas zur deutschen Sprache*. 1st ed. München: Deutscher Taschenbuchverlag. [1978].

Manni, Franz, Wilbert Heeringa & John Nerbonne. 2006. "To what Extent are Surnames Words? Comparing Geographic Patterns of Surnames and Dialect Variation in the Netherlands." *Literary and Linguistic Computing* 21(4):507–528. Special issue, J. Nerbonne & W. Kretzschmar, Jr. (eds.) Progress in Dialectometry: Toward Explanation.

Moisl, Hermann & Val Jones. 2005. "Cluster Analysis of the Newcastle Electronic Corpus of Tyneside English: A Comparison of Methods." *Literary and Linguistic Computing* 20:125–146.

Nerbonne, John & Christine Siedle. 2005. "Dialektklassifikation auf Grundlage aggregierter Aussprachenunterschiede." *Zeitschrift für Dialektologie und Linguistik* 72(2):129–147.

Nerbonne, John, Peter Kleiweg, Wilbert Heeringa & Franz Manni. 2008. Projecting Dialect Differences to Geography: Bootstrap Clustering vs. Noisy Clustering. In *Data Analysis, Machine Learning, and Applications*, ed. Christine Preisach, Lars Schmidt-Thieme, Hans Burkhardt & Reinhold Decker. Proc. of the 31st Annual Meeting of the German Classification Society Berlin/Heidelberg: Springer. Accepted to appear.

Nerbonne, John, Wilbert Heeringa & Peter Kleiweg. 1999. Edit Distance and Dialect Proximity. In *Time Warps, String Edits and Macromolecules: The Theory and Practice of String Edits*, ed. David Sankoff & Joseph s. Stanford: CSLI pp. v–xv.

Niebaum, Hermann & Jürgen Macha. 2006. *Einführung in die deutsche Dialektologie*. 2te, neubearbeitete auflage ed. Tübingen: Niemeyer. 1st ed. [1999].

Prokić, Jelena. 2007. Identifying Linguistic Structure in a Quantitative Analysis of Dialect Pronunciation. In *Proceedings the ACL 2007 Student Research Workshop, Prague*. Shroudsburg, PA: ACL pp. 61–66.

Schiltz, Guillaume. 1996. German Dialectometry. In *Data Analysis and Information Systems: Statistical and Conceptual Approaches*, ed. Hans-Hermann Bock & Wolgang Polasek. Proc. of 19th Meeting of the Gesellschaft für Klassifikation (Basel, Mar. 8-10, 1995) Berlin/Heidleberg: Springer pp. 526–539.

Séguy, Jean. 1973. "La dialectométrie dans l'Atlas linguistique de Gascogne." *Revue de Linguistique Romane* 37(145):1–24.

Tabachnick, Barbar & Linda Fidell. 2001. *Using Multivariate Statistics*. 4th ed. Boston: Allyn and Bacon.

Tufte, Edward. 1990. *Envisioning Information.* Cheshire, CT: Graphics.

Glossary

Composite cluster map. Map appropriate for visualizing the results of clustering enhanced by the bootstrap or via another technique designed to indicate reliability and resulting in a consensus dendrogram. We begin with a Voronoi tiling of the survey sites of a dialect study. Note that there is an edge in the Voronoi tiling which acts as a border between each pair of adjacent sites. In a composite cluster map the darkness of the border is inversely proportional to the cophenetic distance between the two sites in the consensus dendrogram.

Network map. A simple initial visualization made by drawing lines between data collection sites where the darkness of the line is inversely proportional to the linguistic distance between the sites, so that the less distant (or more similar) the sites are linguistically, the darker the line.

Voronoi tiling, Voronoi tesselation. Given a set of sites on the map, e.g. data collection sites, the Voronoi tiling divides the map into regions, one site per region, where every point in a given region is closer to the site it is built around than to any other site.