Handbook of Dialectology

Statistics for Aggregate Variationist Analyses

John Nerbonne and Martijn Wieling

1. Aggregation and perspectives from the aggregate

Dialect geographers have viewed the distribution of most individual linguistic features (pronunciations, allomorphy, or lexical choice) as complex, and moreover, as non-overlapping. Features often overlap very poorly, resulting in isoglosses that do not "bundle". Nerbonne (2009) maps nine features often discussed in German dialectology, showing how poorly they overlap in detail, but noting that they mostly distinguish the north from the south. Bloomfield's (1933) discussion of dialects came to a similar conclusion, leading him to quote Grimm's (1819: IV) famous dictum that "every word has its own history." The view has been disputed, as Schirmunksi's (1962:78ff) account of Wenker's reception among the neo-grammarians documents, but it is now accepted. The conclusion that individual features have very noisy geographic distributions has been seen to threaten the dialectological enterprise. Gaston Paris concluded on this basis that "there was no geography of *dialects*, only of individual linguistic features" (quoted by Goebl, Ch. 7).

Séguy (1971, 1973) introduced the term 'dialectometry' as he took the liberating step of AGGREGATING over large sets of features (over 400) to then examine how well geographic neighbors agreed; poor aggregate agreements indicated dialect boundaries. The simple step of examining aggregate distributions proved to be enlightening. Goebl (1982) introduced local perspectives on the larger dialect space, examining the distribution of linguistic differences (or equivalently, similarities) from each site in his (Italian) data collection, pointing out *inter alia* that sites with skewed distributions might be transition zones. We also refer to differences as distances, noting that the measures used in dialectometry generally satisfy the conditions imposed on mathematical distances (symmetry, zero between identical elements, and the "triangle inequality": $d(a,b) \leq d(a,c)+d(c,b)$, for all c). Seguy (1971) examined the distribution of aggregate differences (in Gascogne) as a function of geography, displaying a sub-linear curve, which Nerbonne (2010) argues contradicts Trudgill's (1974) gravity theory of dialect divergence.[1] These early works motivate the aggregate perspective.

There are various ways of probing linguistic data to obtain a measure of difference between sites (or samples, in case we are interested in non-geographic differences as well[2]), e.g. measuring the percentage of concepts that is realized in different ways lexically, or by using the percentage of syntactic features that are realized differently (see Ch. 20, this volume, for more sophisticated methods). Equivalently we may always measure similarities and convert these to differences. In the interest of brevity we shall abstract away from the concrete details of how the distances we analyze below are obtained. The presentation below proceeds from the assumption that we have a table of aggregate differences between pairs of sites in our data. We will use a small example from Heeringa (2004:146):

[1] See Chap. 7 (Goebl) and Chap. 20 (Heeringa and Prokić) above for more on the theory and computational underpinnings of dialectometry, respectively.

[2] We emphasize geographical differences in this chapter (but see Sec. 4.2), but aggregating techniques may also be applied to social varieties or social dialectology (Boberg 2005).

Table 1. An example table of aggregate differences between Dutch sites. The cells in the diagonal are empty, with the implicit value zero since there are no differences between a site and itself. The cells below and to the left of the diagonal are blank, because these would be the same as those above and to the right, since the differences between sites *a* and *b* are the same as those between *b* and *a*. This also means that the row for the last site in the table can be empty. We omit this as the example is developed further.

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

2. Clustering

A traditional question in dialect geography is whether there are DIALECT AREAS, i.e. regions within which variation is low, but which are relatively distinct with respect to varieties outside the region. There are many CLUSTERING techniques that search for groups in data (Kaufmann and Rosseeuw 2005; Manning et al. 2008:Ch.16-17); we focus here on the ones popular in dialectology. All the techniques we examine search for groups in the linguistic data without reference to location.[3] If, on the basis of aggregate differences, a group is identified and confirmed, and if it indeed constitutes a geographic region, then this is already a modest success.

Some clustering algorithms, such as k-means clustering (Manning et al. 2008:515ff), aim to partition the data, i.e., determine groups such that each variety (or sampling point) belongs to exactly one. These "flat" algorithms thus assume that there is no hierarchical structure in the grouping, where a given variety might be assigned not only to Vologda, but also to North Russian, which is part of Russian, which in turn finally belongs to East Slavic (see Zhobov and Alexander, Ch. 31, this volume). It is uncontroversial, however, that dialect groups may be hierarchically structured, which has leads to a focus on HIERARCHICAL CLUSTERING ALGORITHMS, which we will concentrate on. A cross-cutting distinction separates algorithms which work bottom-up, or agglomeratively, from those which work top-down, or divisively. We focus on agglomerative algorithms, which have enjoyed the most dialectological attention, but we will return to divisive algorithms at the end of this section.

2.1 Hierarchical agglomerative clustering

Johnson (1967) realized that different hierarchical agglomerative algorithms can be characterized in similar ways. They all begin with a sample × sample table of aggregate differences such as that in Table 1 (below). They all then seek the two (groups of) sites for which the differences are minimal and fuse these two to create a cluster. The table is then revised so that the two minimally different groups are eliminated while the result of fusing them is included. This leaves the table with the difference values of several cells unspecified (see Table 2).

Table 2. The (partial) table after fusing Haarlem and Delft but before determining the differences between the recently fused element and the other elements. The missing values are signalled by question marks, and following value is the mean of the distances from two components of the fusion to the other site. From Heeringa (2004:147).

	Grouw	Haarlem & Delft	Hattem	Lochem
Grouw		? (43)	46	47
Haarlem & Delft			? (37)	? (39)
Hattem				21

[3] See Chap. 25 (Grieve) for geo-statistical techniques that include a bias for geographically coherent areas.

Once we have obtained a filled-in table, we are in a position to iterate, again choosing the closest (groups of) sites, fusing them, and assign the necessary distance to the new elements (and eliminating the components of the fusion). Johnson (1967) showed that several sorts of hierarchical agglomerative clustering approaches could be characterized by the function used to determine the new distances in the step immediately following the fusion. One may use the arithmetic mean between the two components (and shown in parentheses in Table 2), which is referred to as the 'Unweighted Pair-Group Method using Arithmetic averages' (UPGMA). Heeringa (2004) also discusses a version which uses a mean weighted by group sizes (WPGMA), and a pair of methods ((un-)weighted by group size) which determine a centroid in the abstract space of differences. In addition, one may simply assign the new element either the smallest difference value available (among all the pairs with one element in the new group and one in the old one), which is referred to as "nearest neighbor clustering" or "single-link", or assign the new element the largest distance, which is referred to as "furthest neighbor" or "complete-link". Finally, Ward's method proceeds from the insight that assigning a single distance to the elements newly fused introduces a kind of "error" in treating the fused elements as the same. It then is designed to minimize the error. Prokić and Nerbonne (2008) review other popular techniques,.

Figure 1 shows the output of clustering, a dendrogram, i.e., a tree with the varieties as leaves, which are joined to form more substantial branches reflecting the fusion process. Dendrograms are popular not only for illustrating groups, but also for showing COPHENETIC DISTANCES. The cophenetic distance between any two sites is their common distance to the first encompassing node. In Fig.1 the cophenetic distance between Hattem and Lochem is 21.

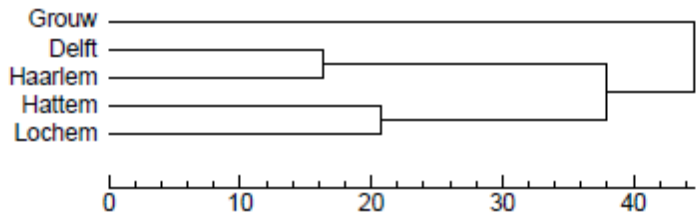

Figure 1. A dendrogram displaying the results of UPGMA clustering on the distance matrix in Table 1. From Heeringa (2004:147). The horizontal distance from the leaves on the left to a branching point shows the cophenetic distance.

2.2 Stochastic clustering

All the popular hierarchical agglomerative algorithms suffer from instability, meaning that small differences in the input table can sometimes lead to large differences in the output dendrogram, occasionally very large (Prokić and Nerbonne 2008).

In order to overcome the inherent instability in clustering, two approaches have been used, bootstrap clustering and noisy clustering (Nerbonne et al. 2008), both of which add a STOCHASTIC element.[4] In bootstrap clustering, clustering is repeated a fixed number of times, usually 100 or 1000 times, while the choice of elements in the aggregate sample is allowed to vary. A common choice is to fix the number of elements in the aggregate at the total number available, and then to choose the elements randomly with replacement. If one begins with 200 words, then sample size is fixed at 200. When

[4] The rest of the presentation assumes the discussion of statistics in the introduction to this section of the handbook.

selecting elements with replacement, one may select the same element more than once, which will then forces other elements to be omitted. Another option is to repeat clustering adding different small amounts of noise (e.g., 0.3 standard deviations) to the distance matrix at each iteration. In either case the result is a dendrogram where each internal node is associated with a percentage indicating how often the sites below the node (its leaves) emerged during the stochastic process. One may be fairly confident of clusters that emerge 90% or more of the time.

2.3 Other clustering techniques

Since dialectologists often present their results in maps of dialect areas, and since clustering produces groups that normally project onto areas, clustering thus facilitates comparison to earlier work.

Before closing this section, we note that there are perhaps hundreds of alternative clustering algorithms that have not yet been used on dialect data (see the NIPS conferences, http://nips,cc), meaning that there is clearly room for further experimentation. Divisive algorithms begin with the entire set of samples, seeking a split that leaves the subgroups as internally similar as possible. Manning et al. (2008:Ch.17.6) claim that divisive algorithms are computationally more effective and that they may also be more accurate, since they base their decisions on the entire data set, not just on local evidence (pairs of varieties or groups of varieties). Given their potential advantages and the scant attention they have received in dialectology, more work on this topic would be desirable.

Two principles are important in further experiments. First the potentially hierarchical structure of dialect relations should not be ruled out (as in k-means clustering, see Manning et al. 2008:515ff); and second, the algorithm be required to assign a reliability to the clusters it proposes.

3. Dimension reduction

3.1 Multidimensional Scaling

Multi-dimensional scaling (MDS, Kruskal and Wish 1978) was introduced to dialectology by Embleton (1993). It inputs an input distance matrix such as Table 1 and assigns coordinates to each element in a small number of dimensions. Given the MDS coordinates we may derive assigned distances using the Euclidean formula, and the "dimension reduction" is successful to the degree that the derived distances agree with the input distances. This success is indicated by the STRESS in the solution, where less stress is better, or by the correlation of the input distances with the MDS-derived distances, where greater correlation coefficients are naturally preferred. MDS analyses must be accompanied by one of these two numbers if they are to be published. One should also attempt to interpret the dimensions, and in the case of dialectological applications we prefer to see both geographic interpretations (how are the dimensions projected onto a map) and linguistic ones (what linguistic features do these sites share). See Wieling et al. (2007:Fig. 6), , Prokić (2010:§3.5.1) and Nerbonne (2010a: Map 2405) for examples.

We also need to ask how many dimensions to use in a solution. Each additional dimension will reduce stress (but less than all the previous dimensions) and improve the correlation with the input matrix, so a common way to determine the optimal number of dimensions to be used is to plot, e.g., stress as a function of the number of dimensions in what is called a scree plot.

Since each additional dimension reduces stress less than previous ones, there will always be a point where the curve in the scree plot begins to flatten out, so that additional dimensions in the flat portion of the curve no longer account for much variance (Johnson 2008:209), indicating that further dimensions may be ignored. A further question concerns whether one may use metric versions of

MDS or whether one should stick to a non-metric version, given the categorical nature of linguistic data. But since it is fine to treat large aggregates as numerical data, i.e. metrically, we have no compunction about using the (simpler) metric version.

Applications of MDS to dialect data typically represent the data well in two or maximally three dimensions, and this has led to a popular innovation in dialect mapping where each dimension is assigned a color – usually red, green or blue – and each site is assigned a color intensity corresponding to its coordinate in the MDS solution (Nerbonne et al.1999), arguably the first representation of dialect continua based on exact techniques.[5]

MDS does not partition dialect sites into non-overlapping dialect areas, and it also does not suffer from the instability we noted in non-stochastic clustering routines. Moreover it analyzes the same sort of distance matrix which is input to clustering. This means that it can be used to examine clustering results in more detail.

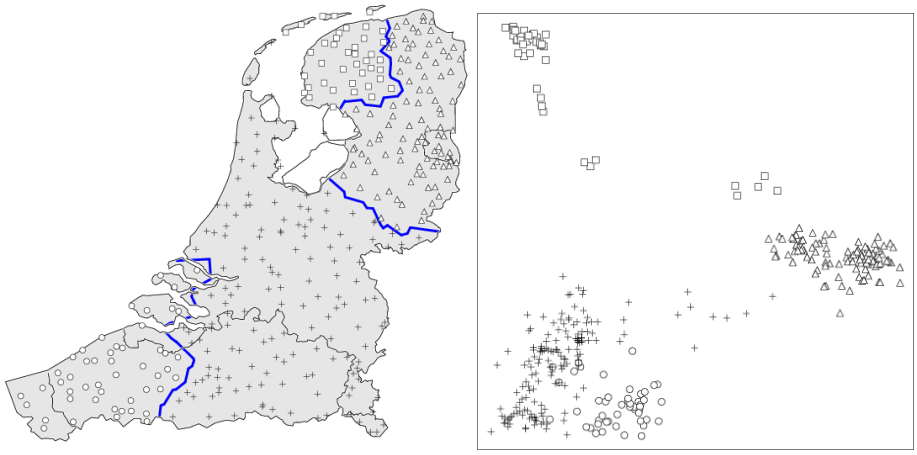

Figure 2 Left a map of the Netherlands showing the largest four clusters obtained using Ward's method. We shall discuss the Frisian area in the Northwest, the Lower Saxon area in the Northeast, the West Flemish (including Zealand) area in the Southwest, and the large Franconian area in the middle. On the right a projection of the same data into two dimensions using MDS (r = 0.76). The MDS graph shows that there is an "archipelago" of Frisian sites (boxes) that are quite similar to Lower Saxon, and that Lower Saxon is otherwise fairly distinct, while the West Flemish and Franconian sites are less well distinguished. The MDS perspective adds to the clustering. From Gabmap, www.gabmap.nl

3.2 Principal Component Analysis and Factor Analysis

PRINCIPAL COMPONENT ANALYSIS (PCA) and FACTOR ANALYSIS (FA) are two related dimension-reducing techniques, which input not distance matrices, but rather matrices of sites and numeric features. They differ in that PCA attempts to account for all the variance in the matrix, while FA is more modest, taking aim only at the variance that is shared among the input variables. This means that noise and variables that share no variance with others are ignored in FA.

Following Tabachnik and Fidell (2001:Ch.13), we shall (mostly) discuss PCA and FA together, although we will naturally give some examples of each. In particular, we use the term FACTOR to refer both to PCA components and FA factors. This eliminates some awkward formulations.

The need to provide numeric features naturally entails representing linguistic data numerically. This may be straightforward, as when Labov (2001: 286ff, 345ff) applies PCA to formant frequencies, but Shackleton (2010, Ch.3 & App.B) translates the vowels in *The Pronunciation of English in the*

[5] We do not show colored maps in this book in order to keep the costs of printing low.

Atlantic States (Kurath and McDavid 1961) and *The Dialect Structure of Southern England* (Kurath and Lowman 1970) into numerical values on a set of features in order to apply PCA. This involves coding each token of a vowel numerically, so that the vowel in *wool* was <5,3,2,1>, representing maximal height, maximal backness, maximal roundness, and non-rhoticity. See Shackleton (2010: 188, Table B.1). Using FA, on the other hand, Clopper and Paolillo (2006) use formant frequencies and vowel duration for fourteen American vowels, while Nerbonne (2006) translates transcribed vowels into (numeric) feature vectors somewhat as Shackleton does. Grieve et al. (2011) identify lexical alternatives and then use their relative frequencies in an analysis using FA, but see Grieve (this volume) as well for several other crucial steps in that analysis. Pickl (2013) also uses relative lexical frequencies as input to FA.

The result of applying the analysis is a set of factors smaller than the original set of variables. The goal is to interpret one's data not on the basis of, say, twenty vowels, or two hundred or more phoneme tokens, but rather in terms of a much smaller number of factors in the solution. We note two further interesting outputs of PCA and FA. On the one hand we obtain for each factor and variable, the LOADING, i.e. the degree to which the two correlate, which facilitates in interpreting the variable. If a factor correlates highly with [i,ɪ,e,ɛ] and [æ] (which therefore have high loadings for this factor), but not with other vowels, the factor may be interpreted straightforwardly, giving dimension reduction the potential to contribute to the deeper linguistic analysis of aggregate analyses. Tabachnik and Fidell (2001) suggest the loadings should have values greater than 0.32 if one is to interpret them. On the other hand, PCA and FA also provide SCORES for each factor and site in the input matrix, contributing to the geographical interpretation.

Turning to the question of how many factors to retain in analyses, several considerations may play a role. First, just as in MDS, we want to interpret the factors we decide to retain in the solution. The geographic or social interpretation is important, but given that we analyze individual variables jointly in PCA and FA (via factor loadings), it is most intriguing to seek linguistic interpretation, especially when it suggests that a more abstract level of linguistic structure might be influential. Second, each factor in PCA and FA is associated with an EIGENVALUE derived from the input matrix which represents variance, in particular where an eigenvalue of 1.0 is roughly the variance associated with a single case (site or speaker). This is the source of a common rule of thumb to disregard eigenvalues less than 1.0.[6] Scree plots are used for PCA and FA, just as in MDS, and in a similar way. Since factors are ordered in importance, one uses the scree plot to determine which initial sequence of factors is to be used by examining the decreasing sequence of eigenvalues in a given solution (Tabachnik and Fidell 2001:621). A third desideratum in choosing which factors to retain is the wish to explain a good deal of the variance, where the rule of thumb is to strive for 70% explained variance. But this can clash with the wish to identify and interpret the factors and the total variance explained.

Clopper and Paolillo (2006) show that two factors account for 73% of the variance in their vowel pronunciation data, while Nerbonne (2006), using hundreds of vowel tokens as variables, was able to explain only 35% of variance using three factors. Leinonen (2010:109) can account for 60.1% of the variance using ten factors, and Pickl (2013) retains twenty factors in order to reach 59.3% explained variance.

Before closing the discussion of PCA and FA, we note that the factors may be ROTATED in order to improve interpretability (Tabachnik and Fidell 2001: Ch.13). It is most common to use a rotation in which the first factor explains a maximum of variance, and each successive factor accounts for a

[6] Although Costello and Osborne (2005) find this one of the least reliable criteria.

maximum of remaining variance. Geometrically, each successive factor is orthogonal to the others in this so-called VARIMAX rotation. Alternatively, one may choose to use OBLIQUE rotations, especially when this might facilitate interpretation, in which case there will be overlap in the variance claimed by different factors. All of the dialectology studies examined have opted for the varimax rotation.

FA may also be used in a hypothesis-testing fashion, but since this "confirmatory" FA has not been used much in dialectology, it will not be presented here. Due to the opportunity to hypothesize about factors, FA is also preferred when there is some theory relating the factors under study (Brown 2009). See however, Pickl (2013), who uses confirmatory FA to test whether Bach's (1950) conjecture that the areal extent of the use of a form correlates with its frequency. PCA is always used in an exploratory fashion.

Leinonen (2008) shows the advantage of deciding on PCA and FA on a case-by-case basis: she first uses PCA on the band-filtered vowel spectra of 19 vowels of over one thousand Swedish speakers to obtain a representation of vowel quality, and found that PC1 and PC2 were readily interpretable as the first and second formants.[7] She ignores the first filter and it turns out that she can compensate for sex differences by ignoring the women's second filter while performing separate PCAs for men and women. In a second step, Leinonen uses FA to uncover more abstract dimensions in Swedish geographic and social variation, for example, a factor that could be interpreted as lower vowel height for the two long vowels [œ:] and [æ:] in positions before /r/. Leinonen's work suggests that one should not regard PCA and FA as simple alternatives but rather as techniques for specific purposes.

3.3 Related techniques

Cichocki (2006) experiments with correspondence analysis (CA), first developed as a counterpart to PCA for categorical data by Benzécri (1992). Just as in PCA the input is a matrix of sites × data, but the data may be categorical in the case of CA. Uiboaed et al. (2013) use CA to analyze the geographic distribution of lexically specific constructions (so-called "collostructions") in Estonian.

Leino and Hyvönen (2008) experiment with a range of more advanced "components" and conclude notably "if in doubt, start with factor analysis" (p.186). Prokić and Van de Cruys (2010) experiment with a three dimensional matrix (site × site × 20 phonetic-correspondences), which they reduce to a set of most important correspondences using the tensor-reduction techniques PARAFAC. The interesting technique requires a substantial amount of data.

Wieling and Nerbonne (2011) report on bipartite spectral graph clustering (BSGC), which, despite its name, is less indebted to clustering than to the dimension-reducing techniques reported on in this section. It combines dimension reduction with techniques from graph theory to provide a similar sort of result to PCA and FA, namely a sketch of the affinities of cases (sites) with one another and also of affinities between cases and linguistic features. Space limitations prevent a longer discussion here, but we should note that Wieling and Nerbonne (2011) proposed a numeric evaluation of the features identified by BSGC which involves measuring how representative they are within an area and also how distinctive they are with respect to other areas. Prokić et al. (2012) generalize this work to include numeric features such as edit distance.

4. Regression models

[7] But obviating the need for formant tracking, which has a higher rate of failure.

REGRESSION analyses seek to explain or predict a single dependent variable on the basis of one or more independent variables. In textbook cases weight is predicted on the basis of height, or university success on the basis of high school success, aptitude and discipline. In such cases both the independent (predictor) variables and the dependent (or response) variable are numeric, but we are free to construe potential categorical predictors numerically, e.g. as taking the values zero and one. LOGISTIC REGRESSION, in which categorical variables are predicted, is the subject of the previous chapter in this handbook.

As the bulk of dialectological work aims at characterizations of the relations among varieties and strives to characterize those both at an aggregate linguistic level and also with respect to the linguistic details involved, i.e. the components of the aggregate, it is not surprising that regression, with its focus on a single dependent variable, has played a lesser role (than clustering or dimension-reducing techniques). But regression analyses have played a role in characterizing the relation between geographic distance and aggregate linguistic differences, and new regression techniques have been used in conjunction with aggregate analyses which will be presented in Sec. 4.1 and 4.2.

Trudgill (1974) urged more attention for the theoretical question of how geography and demography influence linguistic variation, and his paper has sparked a stream of studies in the intervening years. Nerbonne and Heeringa (2007) studied the effect of distance and population size on aggregate pronunciation differences in Lower Saxon dialects using a regression analysis. Not to anyone's surprise, they found a robust relation, where more distant settlements had more different varieties, but they also noted that the response variable (aggregate pronunciation differences) failed to have a linear relationship with geographic distance. Instead it was sub-linear, so that they were able to show a linear relation between the logarithm of geographic distance and linguistic distance. Nerbonne (2010) showed that the same sub-linear relation holds for pronunciation in six other language areas, namely American English on the Eastern seaboard, the entire Dutch area, Germany, Gabon Bantu, Norway and Bulgaria. Spruit et al. (2009) show that the same sub-linear relation holds for vocabulary, but perhaps not for syntax, where a linear model was marginally better. Szmrecanyi (2012) finds completely different results using morpho-syntactic frequencies in corpus data. Because Seguy's (1971) paper also graphed the influence of geography sub-linearly, Nerbonne (2010) proposed to call this SEGUY'S CURVE.

4.1 Mixed-effects models

In standard regression models, we assume that the independent variables are non-random, i.e. fixed. The sex or gender of participants in a dialect survey is a clear example of a FIXED EFFECT which is normally controlled for in survey design. Only two sexes are polled and any future work would use exactly these two, so that they are repeatable. In contrast, the words in a data set are assumed to be a random sample from a much larger population of potential words. If we repeated the survey, we might use new words. The choice of words is therefore a RANDOM EFFECT. Taking into account the structural variability associated with these random effects allows for generalizable results with *p*-values which are not over-confident (Baayen et al. 2008). Regression models using both fixed and random effects are known as MIXED-EFFECTS MODELS (Pinheiro and Bates 2000).[8]

Keune et al. (2005) presented a very early use of mixed-effects models for language variation, in which the authors contrast the use of Dutch adjectives and adverbs using the suffix *-lijk* and its reduction in speech, and they treat the choice of words as a random variable. In one of the studies

[8] Clark (1973) had criticized that language differences were often treated as fixed effects even though the words used in studies and experiments were actually a random sample of a larger population of potential words, and therefore should be treated as a random effect (as is done in the mixed-effects regression framework).

reported, the dependent variable is the lexical frequency with which these words appear in newspapers in the Netherlands and Flanders with different registers ("quality", national or regional). As the authors note, the mixed-effects analysis effectively builds a model for each individual word, so that one can note not only that there is a general tendency for *-lijk* to appear more frequently in the Netherlands and in more formal newspapers, but also that some words go against the grain, and moreover that that there is an interaction between register and country. In the study on phonological reduction using the *Corpus of Spoken Dutch* (http://lands.let.ru.nl/cgn/), it is shown that reduction is more common in the Netherlands than in Flanders, more common among men than women, more common when the word was predictable, and less common at the end of utterances. But some words go against the grain with respect to reduction as well. Keune and colleagues further note that the mixed-effects analysis has the welcome consequence that sociolinguists no longer need to identify alternatives to serve in variable rule analysis of the Varbrul sort (see Paolillo's chapter, this volume). It becomes instead possible to examine a large number of linguistic phenomena simultaneously.

Tagliamonte and Baayen (2012) examine the *was/were* alternation in York, where *was* is often used in plural existential sentences. They analyze 300 tokens from only 40 individuals, and turn to mixed-effects modeling treating speakers as a random effect, eliminating problems connected with imbalanced numbers of tokens per individual. Their re-analysis shows that one influence on the choice between *was* and *were*, the polarity of the utterance (whether it is used in construction with negation) needed to be re-thought due to its interactions with other predictors.

The following section (4.2) discusses further studies in which mixed-effects analysis has been used in combination with generalized additive modeling. Baayen (2008:Ch.7) is a step-by-step presentation of how to conduct mixed effects analyses in the R lme4 package with several examples. Jäger (2008) presents mixed-effects logistic regression in a general comparison of regression designs vs. categorical designs. Johnson (2008) presents Rbrul, a package by the author, and a detailed comparison of how mixed-effects analyses compare to the popular logistic regression package, Varbrul (Sankoff and Labov 1979). Winter (2013) provides a tutorial in mixed-effects regression modeling for linguists.

Although mixed-effects models have been clearly gaining in popularity, Paolillo (2013) has criticized their use, arguing for including speakers as a fixed effect (rather than a random effect), since "the sample of individuals [may be] nonrandom" and modeling it as a random factor would be in error. His fixed-effects approach, however, is "limit[ing] the scope of inference" (Bolker et al., 2009), i.e. undercutting the ability of the analysis to generalize from sample to population. We therefore recommend mixed-effects regression as the appropriate method to analyze linguistic data involving participants responding to multiple items.

4.2 Generalized additive modeling

In contrast to methods from geo-statistics, which focus mainly on identifying the aggregate geographical pattern in the data (see Grieve, this volume), another approach, generalized additive modeling (GAM, Hastie & Tibshirani 1990, Wood 2006), is able to *simultaneously* detect the aggregate geographical pattern, while also identifying the importance of other relevant social and lexical predictors.

A GAM is an extension of a generalized linear regression model. As noted above, the response variable is assumed to have a linear relationship with one or more predictor variables in linear regression. The response variable must be numerical (such as pronunciation distance from a certain reference variety), whereas the predictor variables can be either numerical (such as the speaker's age)

or categorical (such as the speaker's gender). The generalized linear regression model is a generalization of the linear regression model in such a way that the response variable (transformed via a link function) has a linear relationship to the predictor variables. For example, logistic regression is a form of a generalized linear regression model which uses the log-odds (logit) link function. Logistic regression (see Paolillo, this volume) is the appropriate form to analyze binary data and is frequently used in sociolinguistics (e.g., Tagliamonte and Baayen 2012).

A generalized additive model extends the generalized linear regression model by allowing the (possibly transformed) response variable to have a non-linear relationship with one or more predictors. The non-linear relationships are modeled via smooths which can have different basis functions and basis dimensions. The basis function indicates how the smooths are built up. For example, a cubic spline basis function is constructed by connecting sections of cubic functions. The basis dimension indicates the upper limit for how complex the smooth can be (i.e. how many degrees of freedom are invested). The higher this number, the more wiggly the smooth can become. To prevent overfitting (i.e. too wiggly curves), GAMs are estimated using penalized likelihood estimation and cross-validation. The best, computationally feasible, smoothing basis for a single predictor or multiple isotropic predictors (i.e. predictors having the same scale) is the thin plate regression spline (Wood 2003). This basis is constructed by a weighted sum of geometrically simpler curves (or surfaces in the multidimensional case). When multiple predictors need to be combined (i.e. interact) which are on a different scale, a tensor product can be used combining different smoothing bases. Importantly, a mixed-effects regression approach (see Chapter 7 of Baayen 2008; Baayen et al. 2008) which is necessary for taking into account the structural variability associated with the random-effect factors (such as speakers and words) is also possible within the GAM framework. In that case, random-effect factors are modeled as smooths as well (see Chapter 6 of Wood 2006).

Rather than constructing a regression model in which geography is simplified as geographical distance (e.g., Nerbonne and Heeringa 2007), a GAM can model complex geographical patterns directly via a two-dimensional smooth of longitude and latitude. The first study to use a generalized additive modeling approach for the aggregate analysis of dialect variation was conducted by Wieling, Nerbonne, and Baayen (2011). They only used a generalized additive model to represent geography and used the fitted values of this model as a new predictor in a linear mixed-effects regression model. Their dataset consisted of pronunciation distances (compared to standard Dutch) for 562 words in 424 locations in the Netherlands. Besides finding support for a complex geographical pattern of pronunciation distances from standard Dutch with greater pronunciation distances from standard Dutch in the peripheral areas (see Figure 3), they simultaneously identified the importance of various social and lexical features. For example, larger communities were found to use pronunciations closer to standard Dutch, and more frequent words were more different from standard Dutch (indicating resistance to standardization).

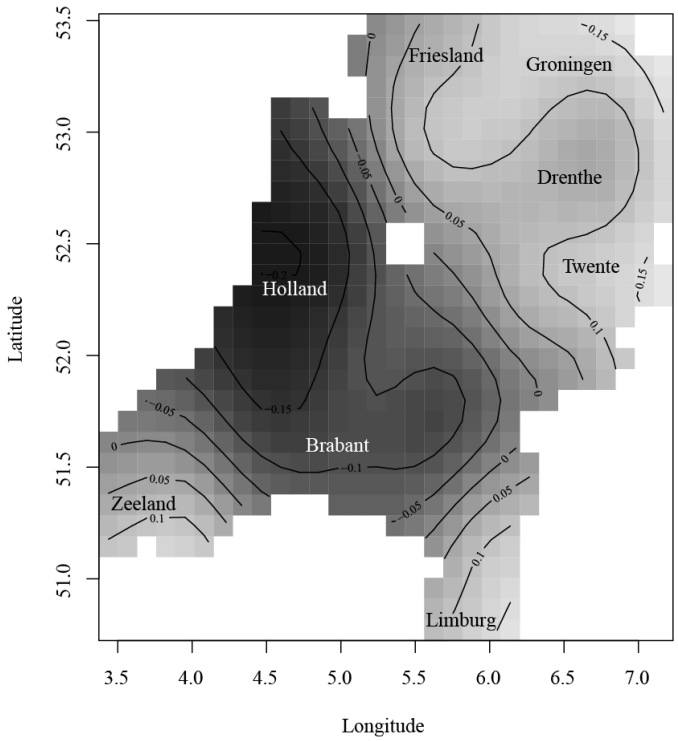

Figure 3. Contour plot obtained with a generalized additive model. The contour plot shows a regression surface of pronunciation distance (from standard Dutch) as a function of longitude and latitude obtained with a generalized additive model using a thin plate regression spline. The (black) contour lines represent aggregate distance isoglosses, darker shades of gray indicate smaller distances closer to the standard Dutch language, lighter shades of gray represent greater distances. Note that the empty square indicates the location of the IJsselmeer, a large lake in the Netherlands. Reprinted (including caption) from Wieling (2012) with permission.

Wieling et al. (submitted; see also Wieling 2012, Ch. 7) created a single generalized additive mixed-effects regression model to determine pronunciation distances from standard Catalan. The Catalan dataset (Valls, Wieling, and Nerbonne, 2013) consisted of 357 words in 40 dialectal varieties located in Catalonia, Andorra and Aragon. In each location 8 speakers (born between 1930 and 1996) were interviewed. Similar to the results of Wieling, Nerbonne, and Baayen (2011), geography was found to be a highly important, non-linear predictor. Furthermore, a clear border effect was observed between Aragon and the other two regions: younger speakers in Catalonia and Andorra (where Catalan is an official language), but not in Aragon (where Catalan is not an official language) had pronunciations closer to the standard Catalan language than the older speakers. In addition, Wieling et al. (submitted) found support for other social and lexical variables, such as word category and the year of birth of the speakers.

The generalized additive modeling approach can also be applied to study other types of aggregate variation. Wieling et al. (2014) investigated Tuscan lexical variation in a dataset consisting of 170 concepts for 2060 speakers (in 213 localities). Their response variable was binary, with a 1 indicating that the speaker used a non-standard Italian form and a 0 indicating the use of the standard Italian form. Consequently, they used a generalized additive mixed-effects *logistic* regression model for this data. They also used a more sophisticated approach to modeling geography: they allowed the geographical pattern to vary depending on concept frequency and speaker age. In effect, they used a four-dimensional tensor product smooth (longitude × latitude × concept frequency × speaker age). Their results highlighted the potential of the generalized additive modeling approach and showed distinctive differences in the geographical patterns associated with speaker age and concept frequency. For example, whereas younger speakers, especially in the area around Florence, were more likely to

use standard Italian lexical forms than older speakers, this did not appear to be the case for the low-frequency concepts. In that situation the younger speakers used more general definitions which did not coincide with the specific (old-fashioned) standard Italian form (e.g., they used 'swine' rather than 'boar' to denote a male pig).

As attempting to use a relatively new and complex statistical method might seem daunting, paper packages for the studies of Wieling et al. (2011), Wieling et al. (submitted) and Wieling et al. (2014) have recently been made available at the Mind Research Repository (http://openscience.uni-leipzig.de). These paper packages include all data and R commands (using the package 'mgcv'; Wood 2006) needed to fit the generalized additive models and replicate the results of these studies.

In sum, the advantage of the generalized additive modeling approach is that it allows one to directly incorporate the complex influence of geography on the aggregate patterns, while simultaneously considering the importance of other social and lexical variables. Furthermore, the availability of paper packages enables other researchers to easily apply these methods to their data as well.

5. Conclusions and prospects

Dialectal material is often analyzed today using a variety of multivariate techniques. This chapter has been slanted to methods for the analysis of large aggregates and to techniques that seek to identify groups together with their common speech habits. In the case of geographical data, these might be areas and their lexical peculiarities, but we have sketched how more advanced techniques are poised to combine the analysis of geographical and social variables, contributing technically to the Chambers-Trudgill program of understanding variationist linguistics – dialectology and sociolinguistics – in a unified way.

We would like to close this chapter by identifying a challenge. As attractive as the GAMs and mixed-effects models are, they remain regression models, focused on the prediction of a single criterion variable. This works quite well for research that can be formulated in such a focused way, e.g. on the relative proximity of local varieties (including different cross-cutting social distinctions) to a single standard language. By contrast, the techniques in the first part of the chapter, including clustering, MDS, but also factor analysis and related techniques, did not require the identification of a single criterion variable, but instead could be used fairly directly on a dialectologist's table of sites and linguistic variables (or on the site × site summary of that tables differences). These techniques provide a more global picture of the landscape of language varieties, but they are (now) poorly equipped to include a range of other variables such as class, education, gender and the like. The challenge thus is to find a way to combine the virtues of the two approaches.

References

Baayen, R. Harald. 2008. *Analyzing Linguistic Data. A Practical Introduction to Statistics Using R*. Cambridge: Cambridge University Press.

Baayen, R. Harald, Doug J. Davidson, and Doug M. Bates. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59(4): 390–412.

Bach, Adolf. 1950. *Deutsche Mundartforschung: ihre Wege, Ergebnisse und Aufgaben*. Heidelberg: Winter.

Benzécri, Jean-Paul. 1992. *Correspondence analysis handbook*. New York: Marcel Dekker.

Boberg, Charles. 2005. The North American regional vocabulary survey: New variables and methods in the study of North American English. *American Speech*, 80(1), 22-60.

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

Bolker, Benjamin M., Mollie E. Brooks, Connie J. Clark, Shane W. Geange, John R. Poulsen, M. Henry H. Stevens, and Jada-Simone S. White. 2009. Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in ecology & evolution* 24(3): 127-135.

Brown, James. 2009. Principal components analysis and exploratory factor analysis–definitions, differences, and choices. *Japan Association for Language Teaching, Testing and Evaluation Special Iinterest Group Newsletter* 13.1: 26-30.

Cichocki, Wladyslaw. 2006. Geographic variation in Acadian French /r/: What can correspondence analysis contribute toward explanation? *Literary and Linguistic Computing* 21.4: 529-541.

Clark, Herbert H. 1973. The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of verbal learning and verbal behavior* 12.4: 335-359.

Clopper, Cynthia G., and John C. Paolillo. 2006. North American English vowels: A factor-analytic perspective. *Literary and Linguistic Computing* 21.4: 445-462.

Costello, Anna and Jason Osborne. 2005. Best practices in exploratory factor analysis: four recommendations for getting the most from your analysis. *Practical Assessment, Research and Evaluation* 10: 1-9.

Embleton, Sheila. 1993. Multidimensional scaling as a dialectometrical technique: Outline of a research project. In: Reinhard Köhler and Burghardt Rieger (eds.) *Contributions to quantitative linguistics*. 267-276. Dordrecht: Kluwer.

Goebl, Hans. 1982. *Dialektometrie; Prinzipien und Methoden des Einsatzes der numerischen Taxonomie im Bereich der Dialektgeographie*. (*Philosophisch-Historische Klasse Denkschriften* 157) Vienna: Verlag der Österreichischen Akademie der Wissenschaften.

Grieve, Jack, Dirk Speelman, and Dirk Geeraerts. 2011. A statistical method for the identification and aggregation of regional linguistic variation. *Language Variation and Change* 23(2): 193-221.

Grimm, Jacob. 1819. *Deutsche Grammatik. 1*. Theil, Göttingen: Dieterich.

Hastie, Trevor J., and Robert J. Tibshirani. 1990. *Generalized Additive Models*. London: Chapman & Hall/CRC.

Heeringa, Wilbert J. 2004. *Measuring dialect pronunciation differences using Levenshtein distance*. Ph.D. thesis, University of Groningen.

Jaeger, T. Florian. 2008. Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of memory and language* 59(4): 434-446.

Johnson, Keith. 2008. *Quantitative methods in linguistics*. Malden, MA:Blackwell.

Johnson, Stephen C. 1967. Hierarchical clustering schemes. *Psychometrika* 32.3: 241-254.

Kaufman, Leonard, and Peter J. Rousseeuw. 2005, [1]1990. *Finding groups in data: an introduction to cluster analysis*. Vol. 344. Hoboken: Wiley.

Johnson, Daniel Ezra. 2009. Getting off the GoldVarb standard: Introducing Rbrul for mixed-effects variable rule analysis. *Language and Linguistics Compass* 3: 359–383.

Joseph B. Kruskal, and Myron Wish. 1978. *Multidimensional scaling*. Vol. 11. London: Sage.

Keune, Karen, Mirjam Ernestus, Roeland van Hout, and R. Harald Baayen (2005). Variation in Dutch: From written MOGELIJK to spoken MOK. *Corpus Linguistics and Linguistic Theory* 1(2): 183-223.

Kurath, Hans, and Guy S. Lowman. 1970. *The Dialectal Structure of Southern England*. Tuscaloosa: University of Alabama Press.

Kurath, Hans, and Raven McDavid. 1961. *The Pronunciation of English in the Atlantic States: Based upon the Collections of the Linguistic Atlas of the Eastern United States*. Ann Arbor: University of Michigan Press.

Labov, William. 2001. *Principles of linguistic change Vol. 2: Social factors*. (Language in Society 29). Oxford: Oxford University Press.

Leino, Antti, and Saara Hyvönen. 2008. Comparison of component models in analysing the distribution of dialectal features. *International Journal of Humanities and Arts Computing* 2(1-2): 173-187.

Leinonen, Therese. 2008. Factor analysis of vowel pronunciation in Swedish dialects. *International Journal of Humanities and Arts Computing* 2(1-2): 189-204.

Leinonen, Therese. 2010. *An acoustic analysis of vowel pronunciation in Swedish dialects*. Ph.D. thesis, University of Groningen.

Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to information retrieval*. Cambridge: Cambridge University Press.

Nerbonne, John. 2006. Identifying linguistic structure in aggregate comparison. *Literary and Linguistic Computing* 21(4): 463-475.

Nerbonne, John. 2009. Data-Driven Dialectology. *Language and Linguistics Compass* 3(1): 175-198. DOI: 10.1111/j.1749-818x2008.00114.x

Nerbonne, John 2010. Measuring the Diffusion of Linguistic Change. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365: 3821-3828. DOI: 10.1098/rstb.2010.0048.

Nerbonne, John. 2010a. Mapping aggregate variation. In: Alfred Lameli, Ronald Kehrein and Stephan Rabanus (eds.) *Language and Space. International Handbook of Linguistic Variation* 2: 476-495. Berlin: Mouton De Gruyter.

Nerbonne, John, and Wilbert Heeringa. 2007. Geographic distributions of linguistic variation reflect dynamics of differentiation. In *Roots: Linguistics in Search of its Evidential Base*, edited by Sam Featherston and Wolfgang Sternefeld, 267-297. Berlin: Mouton De Gruyter.

Nerbonne, John, Wilbert Heeringa, and Peter Kleiweg. 1999. Edit distance and dialect proximity. In David Sankoff and Joseph Kruskal (eds.) *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison, 2nd ed*., Stanford: CSLI Press.

Nerbonne, John, Peter Kleiweg, Wilbert Heeringa, and Franz Manni. 2008. Projecting dialect distances to geography: Bootstrap clustering vs. noisy clustering. In: Christine Preisach, Lars Schmidt-Thieme, Hans Burkhardt and Reinhold Decker (eds.) *Data Analysis, Machine Learning and Applications*., 2008. 647-654. Berlin: Springer.

Paolillo, John C. 2013. Individual effects in variation analysis: Model, software, and research design. *Language Variation and Change* 25(1): 89-118.

Pickl, Simon. 2013. *Probabilistische Geolinguistik: Geostatistische Analysen lexikalischer Variation in Bayerisch-Schwaben*. Stuttgart: Franz Steiner.

Pinheiro, José C. and Douglas M. Bates (2000). *Mixed-effects models in S and S-PLUS.* New York: Springer.

Prokić, Jelena. 2010. *Families and resemblances*. Ph.D. thesis, University of Groningen.

Prokić, Jelena, Çağri Çöltekın, and John Nerbonne. 2012. Detecting shibboleths. *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*. 72-80. Shroudsburg, PA: Association for Computational Linguistics.

Prokić, Jelena and John Nerbonne. 2008. Recognising groups among dialects. *International Journal of Humanities and Arts Computing* 2.1-2: 153-172.

Prokić, Jelena, and Tim Van de Cruys. 2010. Exploring dialect phonetic variation using PARAFAC. *Proceedings of the 11th Meeting of the ACL Special Interest Group on Computational Morphology and Phonology*. Shroudsburg, PA: Association for Computational Linguistics.

Sankoff, David, and William Labov. 1979. On the uses of variable rules. *Language in Society* 8(2-3): 189-222.

Schirmunski, Viktor. 1962. *Deutsche Mundartkunde. Vergleichende Laut-und Formenlehre der deutschen Mundarten.* Berlin: Akademie Verlag.

Séguy, Jean. 1971. La relation entre la distance spatiale et la distance lexicale. *Revue de Linguistique Romane* 35(138):335-357.

Séguy, Jean. 1973. La dialectométrie dans l'Atlas linguistique de Gascogne. *Revue de Linguistique Romane* 37(145):1-4.

Shackleton Jr, Robert G. 2010. *Quantitative assessment of English-American speech relationships.* Ph.D. thesis, University of Groningen.

Spruit, Marco René, Wilbert Heeringa and John Nerbonne. 2009. Associations among linguistic levels. *Lingua* 119(11): 1624-1642.

Szmrecsanyi, Benedikt 2012. Geography is overrated. In: Sandra Hansen, Christian Schwarz, Philipp Stoeckle and Tobias Streck (eds.) *Dialectological and Folk Dialectological Concepts of Space.* 215-231. *(Current Methods and Perspectives in Sociolinguistic Research on Dialect Change 17)* Berlin: De Gruyter.

Tabachnick, Barbara and Linda S. Fidell. 2001. *Using multivariate statistics*. Boston: Allyn and Bacon.

Tagliamonte, Sali, and R. Harald Baayen. 2012. Models, forests and trees of York English: Was/were variation as a case study for statistical practice. *Language Variation and Change*, 24(2): 135–178.

Trudgill, Peter. 1974. Linguistic Change and Diffusion: Description and Explanation in Sociolinguistic Dialect Geography. *Language in Society* 2: 215–246.

Uiboaed, Kristel, Cornelius Hasselblatt, Liina Lindström, Kadri Muischnek and John Nerbonne. 2013. Variation of verbal constructions in Estonian dialects. *LLC: Journal of Digital Scholarship in the Humanities* 28.1: 42-62.

Valls, Esteve, Martijn Wieling, and John Nerbonne. 2013. Linguistic advergence and divergence in north-western Catalan: A dialectometric investigation of dialect leveling and border effects. *LLC: The Journal of Digital Humanities Scholarship*, 28(1): 119-146.

Wieling, Martijn. 2012. *A Quantitative Approach to Social and Geographical Dialect Variation*. PhD dissertation. Groningen: University of Groningen.

Wieling, Martijn, Wilbert Heeringa and John Nerbonne . 2007. An aggregate analysis of pronunciation in the Goeman-Taeldeman-van Reenen-Project data. *Taal en Tongval* 59: 84-116.

Wieling, Martijn, Simonetta Montemagni, John Nerbonne, and R. Harald Baayen. 2014. Lexical differences between Tuscan dialects and standard Italian: Accounting for geographic and socio-demographic variation using generalized additive mixed modeling. *Language 90(3)*: 669-692.

Wieling, Martijn and John Nerbonne. 2011. Bipartite spectral graph partitioning for clustering dialect varieties and detecting their linguistic features. *Computer Speech & Language* 25.3: 700-715.

Wieling, Martijn, John Nerbonne, and R. Harald Baayen. 2011. Quantitative Social Dialectology: Explaining Linguistic Variation Geographically and Socially. *PLOS ONE*, 6(9): e23613.

Wieling, Martijn, Esteve Valls, R. Harald Baayen, and John Nerbonne. submitted. Border effects among Catalan dialects. In *Mixed Effects Regression Models in Linguistics*, edited by Dirk Speelman, Kris Heylen and Dirk Geeraerts (eds.). New York: Springer.

Winter, Bodo. 2013. Linear models and linear mixed effects models in R with linguistic applications. *arXiv preprint arXiv:1308.5499*.

Wood, Simon N. 2003. Thin plate regression splines. *Journal of the Royal Statistical Society: Series B*, 65: 95–114.

Wood, Simon N. 2006. *Generalized Additive Models: An Introduction with R*. London: Chapman & Hall/CRC.