A comparison of statistical methods for the aggregation of regional linguistic variation

Jack Grieve

Quantitative Lexicology and Variational Linguistics Research Unit

University of Leuven

May 4th, 2011

Abstract

This paper compares two statistical approaches to the analysis of aggregated regional linguistic variation. In the standard approach to dialectometry, common patterns of regional variation are identified by analyzing the distance between a set of locations based on the values of a set of linguistic variables using statistics such as multivariate scaling.

Alternatively, in a multivariate spatial analysis, common patterns of regional variation are identified by using a factor analysis to analyze correlations between linguistic variables that are first smoothed using a local spatial autocorrelation analysis that identifies underlying spatial patterns in the values of each variable. To compare these approaches, both methods are used to analyze the acoustic vowel data from the Atlas of North American English. It is concluded that the multivariate spatial analysis identifies clearer and more detailed patterns of aggregated regional linguistic variation than the standard approach to dialectometry.

1. Introduction

There are many possible ways to search for regional patterns in the values of a series of linguistic variables measured over a series of locations. In traditional dialectology (e.g. Kurath, 1949), maps charting the values of linguistic variables are analyzed by hand in order to identify individual and common patterns of regional variation. Often this is accomplished by drawing isoglosses that divide the region under analysis into sub-regions where the different values of the linguistic variables predominate. Aggregated patterns of regional linguistic variation are then identified by searching for bundles of isoglosses—subsets of linguistic variables that exhibit similar patterns of regional variation. Alternatively, in dialectometry, the values of linguistic variables are analyzed using statistical methods. The standard approach to dialectometry involves computing the distance (or similarity) between every pair of locations in the dataset based on the complete set of linguistic variables. The resultant linguistic distance (or similarity) matrix is then analyzed using multivariate statistics in order to identify aggregated patterns of regional linguistic variation (Heeringa, 2004; Goebel, 2006; Nerbonne and Heeringa, forthcoming).

The goal of this paper is to compare the standard approach to dialectometry with an alternative method for the statistical analysis of regional linguistic variation, known as a *multivariate spatial analysis* (Grieve, 2009; Grieve et al, forthcoming). This approach to aggregation differs from the standard approach in two ways. First, the individual linguistic variables are initially smoothed using a local spatial autocorrelation analysis so as to identify underlying patterns of spatial clustering in the values of each variable. Second, aggregation is achieved by computing a correlation matrix, as opposed to a distance matrix, which not only allows for common patterns of regional variation to be identified, but which also allows for the linguistic variables associated with each of these common patterns to be identified. By applying both methods to the same dataset, this paper shows that a multivariate spatial

analysis is able to recognize common patterns of regional linguistic variation with greater accuracy and detail than is possible using the standard approach to dialectometry.

2. Data

In order to compare these two methods for the analysis of aggregated regional linguistic variation, both methods were used to analyze the acoustic data from the *Atlas of North American English* (Labov et al, 2006), which consists of the average formant 1 and formant 2 values for 24 vowels measures, based on recordings of linguistic interviews with 439 informants. These recordings include both the language produced by informants while conversing with the interviewers before and during the interview as well as the language directly elicited from the informant through word lists and minimal pair tests.

This study is based on the values of 38 of these 48 vowel formant variables, measured in the language of 402 of these 439 informants, representing 236 cities in the contiguous United States. The other 37 informants were excluded from the analysis for numerous reasons: Canadian informants were excluded to control for the influence of national linguistic variation, Alaskan informants were excluded because they are extreme geographical outliers, and the one informant from Bloomington, Illinois was excluded because he is an extreme outlier on formant 2 for all vowels. Because the statistical methods being applied here require that each location is associated with a single value, each vowel formant variable was also averaged across all informants from the same city. Pooling the data at each location also decreased the overall percentage of missing data. Nonetheless, 10 vowel formant variables (formant 1 and formant 2 for /oy/, /aeh/, /eyr/, /uwr/, and /ah/) were still missing data for over 10% of the 236 locations in the final dataset. These vowel formant variables were therefore excluded from the analysis. In addition, 12 vowel formant variables (formant 1 and formant 2 for /uh/, /u/, /ay/ before voiceless consonants, /iw/, /uw/ word internally, /ohr/) were retained

despite missing data for less than 5% of the locations. These missing values were replaced with the mean value for that vowel formant variable across all locations.

This *regional linguistic data matrix*, which consists of 38 vowel formant variables measured across 236 cities, is the input for both the standard dialectometry analysis and the multivariate spatial analysis. The 15 distinct vowels in the regional linguistic data matrix are listed in Table 1 using the phonetic transcription system from the Atlas. Four of these vowels were analyzed in two different phonemic contexts: /ay/ before voiced (/ayv/) and voiceless (/ayc/) consonants, /uw/ word-internally (/uwc/) and word-finally (/uwf/), /ow/ before /r/ (/owr/) and before all other consonants (/owc/), and /oh/ before /r/ (/ohr/) and in all other contexts (/oh/). In addition, /iy/ and /ey/ were only analyzed word internally (/iyc/ and /eyc/), and /ah/ was only analyzed before /r/ (/ahr/). The mean formant 1 and formant 2 value for each of these 19 vowel measures are also plotted in Figure 1.

In addition, the values of the 38 variables were mapped across the 236 locations. Figure 2 maps formant 1 and formant 2 for 4 of these vowel formant measures: /eyc/, /ae/, /oh/, and /ayv/. While regional patterns are apparent in many of the individual maps, the underlying patterns of regional variation in the dataset appear to be highly complex, as suggested by the variety and the messiness of the patterns exhibited by these 8 vowel formant variables. The goal of this paper is to assess whether a standard dialectometry analysis or a multivariate spatial analysis is better able identify and explain these underlying patterns of regional linguistic variation in this dataset.

Table 1 The English Vowel System

	Short		Long					
			Uplide				Inglide	
			Front Uplide		Back Uplide			
	Front	Back	Front	Back	Front	Back	Unround	Round
High	i (ɪ) <i>bit</i>	u (ʊ) <i>book</i>	iy (i) <i>beat</i> ey (e) <i>bait</i>	ay (ai) <i>bite</i>	iw (u) <i>suit</i>	uw (u) <i>boot</i>	oh (ɔ) <i>caught</i>	
Mid	e(ɛ) <i>bet</i>	uh (ʌ) <i>but</i>				ow (o) <i>boat</i>		
Low	ae (æ) <i>bat</i>	o (a) <i>cot</i>				aw (aʊ) <i>bout</i>		
							ah(a:) <i>balm</i>	

Figure 1 Average Formant 1 and Formant 2 Values for all Vowel Measures

Figure 2 Raw Maps for 8 Vowel Formant Variables

3. Dialectometry Analysis

The statistical analysis of aggregated regional linguistic variation was pioneered by Jean Séguy (1971, 1973), the director of the *Atlas Linguistique de la Gascon*. Séguy's basic approach involved calculating the linguistic distance between neighboring pairs of locations based on the values of a set of linguistic variables, which could then be analyzed and mapped to identify aggregated patterns of regional linguistic variation. For example, if the variants of 60 out of 100 linguistic variables were the same in a pair of neighboring locations, then a linguistic distance of 40% would be assigned to that pair of locations. Séguy's methods have since been adopted and expanded on by a number of other researchers, most notably in Salzburg (e.g. Goebel 1982, 1984, 2006, 2007) and Groningen (e.g. Nerbonne et al, 1996; Nerbonne and Heeringa, 1997, 2001, 2010, forthcoming; Nerbonne and Kleiweg, 2003, 2007; Heeringa, 2004; Heeringa and Nerbonne, 2001). Data analysis in standard modern dialectometry involves calculating the linguistic distance between every pair of locations to produce a *linguistic distance matrix*, which is then analyzed using a variety of multivariate statistics, including cluster analysis and multidimensional scaling, to identify common patterns of regional variation. This section presents a standard dialectometry analysis of the regional linguistic data matrix introduced above.

3.1 Cronbach's Alpha

Before generating a linguistic distance matrix, Cronbach's alpha was used to measure the internal consistency of the linguistic variables in the regional linguistic data matrix (Nerbonne and Heeringa, 1997; Heeringa et al, 2002; Heeringa, 2004; Nerbonne, 2008; Szmrecsanyi, 2008; Spruit et al, 2009). Cronbach's alpha was originally developed to assess if a set of items in a psychometric test measure the same underlying construct based on the scores on the test items for a sample of test takers (Cronbach, 1951). More generally,

Cronbach's alpha is used in the social sciences to measure the extent to which the values of a set of variables for a set of objects follows the same basic pattern of variation, with a Cronbach's alpha larger than .700 considered to indicate a relatively high degree of internal consistency in a dataset (Nunnally, 1978). Measuring Cronbach's alpha for the linguistic data matrix therefore allows for the degree to which a set of linguistic variables follow a similar pattern to be assessed, with a relatively high degree of internal consistency being preferred in dialectometry (Heeringa, 2004).

Cronbach's alpha for the linguistic data matrix was calculated as .678, which is slightly low. However, there are numerous problems with this application of Cronbach's alpha. The main issue is that there is no reason to expect or to require that a set of linguistic variables follows one basic pattern of regional variation, as traditional dialect studies have clearly shown that different variables tend to follow different patterns of regional variation (Chambers and Trudgill, 1998). Looking over the raw maps for the individual vowel formant variables (Figure 2), it is clear that this dataset does not exhibit one single pattern of regional linguistic variation, which is why Cronbach's alpha is relatively low. The generation and analysis of the linguistic distance matrix, which is at the core of a standard dialectometry analysis, also do not require that the linguistic data matrix exhibits a high degree of internal consistency. Multidimensional scaling, for example, is able to identify multiple dimensions of variation in a distance matrix. Another problem with this application of Cronbach's alpha is that while a high value does indicate that a set of linguistic variables follows a similar pattern, it does not indicate that a set of linguistic variables follows a similar *regional* pattern. Furthermore, Cronbach's alpha can underestimate the degree of regional consistency in a linguistic data matrix because the values of the variables are only compared for the same locations, while the values of the variables at other nearby locations are ignored. Cronbach's alpha therefore needs to be interpreted with care. A low Cronbach's alpha only indicates that

there may be more than one regional pattern in a dataset—information that should never dissuade further statistical analysis.

3.2 Linguistic Distance Matrix

In order to identify aggregated patterns of regional linguistic variation, the regional linguistic data matrix was first used to generate a regional linguistic distance matrix. The generation of the linguistic distance matrix is the basic act of aggregation that underlies all standard dialectometry analyses.

In general, given a data matrix consisting of a set of variables measured over a set of objects, an object-by-object distance matrix is generated by computing the distance between every pair objects based on the values of the complete set of variables for those two objects. In dialectometry, the linguistic distance matrix is generated by calculating the distance between every pair of locations in the regional linguistic data matrix based on the complete set of linguistic variables. There are various methods for calculating the distance between objects based on the values of a set of variables, including Euclidean and Manhattan (or city-block) distance. In this case, the linguistic distance matrix was generated for the 236 locations based on Euclidean distance. The linguistic distance matrix was then subjected to a statistical analysis to identify underlying patterns of aggregated regional linguistic variation. The analysis was also repeated using a distance matrix generated using Manhattan distance but this change had very little effect on the results of the analysis.

3.3 Mantel Test

In order to test for an overall pattern of aggregated regional linguistic variation, a Mantel test (based on the Pearson product-moment correlation coefficient) (Mantel, 1967) was used to regress the linguistic distance matrix against a corresponding geographic distant matrix,

which consists of the distances between every pair locations in the dataset based on their longitude and latitude (Heeringa, 2004; Gooskens and Heeringa, 2006; Nerbonne and Heeringa, 2007). A Mantel test is used rather than a simple correlation to correct for the fact that the distances in the matrices are not independent of each other. A significant positive correlation indicates that as the geographic distance between locations increases, so does linguistic distance, so that the language of nearby locations tends to be more similar than the language of distant locations. This general approach has also been used to identify regional patterns in the values of individual variables by regressing a geographic distance matrix against a distance matrix based on a single variable (Szmrecsanyi, 2008).

In this case, the correlation between the linguistic and geographic distance matrices was found to be significant but weak ($r = 0.057$, simulated $p = 0.038$), with geographic distance only accounting for 0.32% of the variation in the linguistic distance matrix. The lack of a clear relationship between the distance matrices can be seen in Figure 3, where linguistic distance is plotted against geographic distance, which notably does not exhibit the type sub-linear pattern that has been identified in previous dialectometry studies (Nerbonne, 2009, 2010b). These results are surprising considering that regional patterns are clearly visible in maps plotting the raw values of many of the individual linguistic variables (Figure 2).

It appears, however, that this application of the Mantel test will generally *underestimate* the degree of spatial dependency in a distance matrix (Legendre and Fortin, 2010). This is because the test only detects the very specific regional pattern represented by a geographic distance matrix, which can be visualized as a rectangular region with four different colors in the corners that gradually fade into each other (Figure 19). There are, however, numerous other types of regional patterns whose corresponding distance matrices would not align with a geographic distance matrix. Consider, for example, a single variable that is patterned from the northeast to the southwest, such as a vowel on formant 1 that is

high in the northeast, low in the southeast, and middling in the northwest and southeast. Even when both longitude and latitude are perfectly correlated with the values of such a variable, the distance matrix associated with this variable would not align with the corresponding geographic distance matrix because locations in the northwest and the southeast, which are geographic distant, are linguistically similar. In fact, it is impossible to derive a geographic distance matrix based on the values of a single variable, as the geographic distance matrix is derived from two independent variables (i.e. longitude and latitude). By extension, it is clear that a linguistic data matrix associated with a set of variables that follow the same pattern would never align with a geographic distance matrix, despite the fact that this is the clearest possible case of aggregated regional linguistic variation. There are possible data matrices whose corresponding distance matrices do align with a geographic distance matrix, for example when half of the variables are distributed from north-to-south and half from east-to-west, but it is incorrect to assume that a linguistic distance matrix that does not align with a geographic distance matrix does not exhibit a commensurable level of spatial dependency.

Figure 3 Linguistic vs. Geographic Distance

3.4 Linguistic Distance Maps

Before conducting further statistical analysis, the linguistic distance matrix was directly mapped in two ways. Individual distance maps were first generated for all 236 locations. These maps plot the linguistic distances between a location and every other location based on one row (or column) of the distance matrix. For example, the distance map for Buffalo, New York is presented in Figure 4. The map shows that the language of Buffalo is most similar to the language of the other Northern Cities—a region whose language is characterized by a chain shift involving a number of the vowel formant variables under analysis here (Labov et al, 2006). Of course, generating 236 distance maps does not represent an aggregation of the 38 maps for the raw linguistic variables. Nonetheless, these distance maps are useful for analyzing individual cities and for identifying outlier locations. For example, the linguistic distance map for Savannah, Georgia, which is reproduced in Figure 5, clearly shows that Savannah is distinct from the rest of the Southeast (seemingly having more in common with Iowa); however, it is unclear if this is due to the fact that the language of Savannah is actually distinct or if the one informant from Savannah is an outlier within his own speech community. Given a larger number of informants per location, however, the interpretation of such maps would be more straightforward.

It is also possible to map the link between each of the cities based on the distance between the cities in the distance matrix using what is called a *beam map* (Goebl, 1982) or a *network map* (Nerbonne and Heeringa, 1997). Often all of the links are mapped (Nerbonne, 2010a), but this approach produced a map that was difficult to analyze in this case. An alternative approach is to only map the links between the nearest neighbors (Szmrecsanyi and Wolk, submitted), for example by triangulating the data. A triangulated network map is presented in Figure 6. This approach makes the regions of transition (light lines) and homogeneity (dark lines) relatively clear, although this approach did not identify a great deal

of spatial structure in this distance matrix, aside from some relatively clear regions of transition between the East and the West, which particularly obvious due to the fact that the links are longer in the West. There also appears to be regions of transition between the North and the South and between the Northeast and the Midwest. This method, however, is also very sensitive to outlying locations, which can be especially problematic when locations are based on a single informant, as is the case here. This method is also sensitive to the density of locations in the dataset.

Figure 4 Linguistic Distance Map for Buffalo, New York

Figure 5 Linguistic Distance Map for Savannah, Georgia

Figure 6 Triangulated Network Map

3.5 Multidimensional Scaling

In dialectometry, multidimensional scaling is often used to identify underlying patterns of aggregated regional variation in a linguistic distance matrix (Heeringa and Nerbonne, 2001; Nerbonne and Kleiweg, 2003; Nerbonne, 2010a; Szmrecsanyi, 2008). Multidimensional scaling is a set of statistical methods used for visualizing patterns of proximity between objects in a distance matrix (Cox and Cox, 2001). These methods generate a preselected number of quantitative dimensions that preserve the distance between the objects in the distance matrix as accurately as possible (Cox and Cox, 2001). As such, multidimensional scaling can also be seen as a method for data reduction, as the analysis generates a data matrix with the same number of objects as the data matrix upon which the distance matrix was based, but with a smaller number of aggregated dimensions that represent the original, larger set of variables. To interpret the patterns identified by the multidimensional scaling, the scores associated with these dimensions can then be plotted in low-dimensional space.

In dialectometry, each dimension extracted by a multidimensional scaling represents an underlying pattern of regional variation in the distance matrix. The scores associated with these dimensions can then be used to plot the locations in low-dimensional space or, more commonly in modern dialectometry, three dimensions are extracted and used to generate a single map representing all three patterns simultaneously using RGB mapping. The extracted dimension scores can also be mapped individually, in order to analyze each individual pattern, although such maps are surprisingly rare in dialectometry.

In this case, the linguistic distance matrix was subjected to a *metric* multidimensional scaling in order to extract three dimensions. A *non-metric* multidimensional scaling was also applied, but this made very little difference in the results of the analysis. Because the search space for non-metric scaling subsumes the search space for metric scaling, the fact that these two analyses produce very similar results indicates that there is little reason to apply non-

metric scaling here. Overall, the three dimensions have a goodness of fit of .61, which is acceptable but low, indicating that these three dimensions approximate the distances between the locations in the distance matrix. The three dimensions are plotted individually in Figures 7-9. Although regional patterns are visible in each map, the analysis has not identified clean patterns of regional variation, with these maps exhibiting a similar level of clarity as the raw maps for the individual variables (Figure 2). All three dimensions identify a similar north-south distinction, although there are notable differences between each of these maps.

Dimension 1 identifies a northern region that includes the northern halves of Pennsylvania, Ohio, Indiana and Illinois, with an area of transition with the South in the southern halves of these states, whereas Dimension 2 includes this area of transition in the North. This Midland region is also identified as a distinct region by Dimension 3, which contrasts the Midland with the regions to the north and to the south, especially on the East Coast. The patterns identified by Dimension 2 and 3, however, are not as strong as the pattern identified by Dimension 1.

The results of the multidimensional scaling can also be visualized by combining all three dimensions to form a single map using RGB mapping (Nerbonne, 2010a; Szmrecsanyi, 2008). In RGB mapping the three dimensions extracted by the multidimensional scaling are mapped simultaneously by associating each location with a color defined by deriving each of the three RGB color parameters (red, blue and green) from the three dimension scores for that location. The RGB map based on the three extracted dimensions is presented in Figure 20.

While the map is not perfectly clear, as there are numerous obvious outliers scattered throughout the map, there is a strong two way division between the North (red/pink) and South (green), basically following the same pattern as Dimension 1 (Figure 7). The Midland, however, which is particularly varied in its composition, appears to be a region of relatively sharp transition. The RGB map also identifies a weak Western region (blue).

Figure 7 Metric MDS Dimension 1

Figure 8 Metric MDS Dimension 2

Figure 9 Metric MDS Dimension 3

3.6 Cluster Analysis

While the multidimensional scaling identifies continuous patterns of aggregated regional linguistic variation, a cluster analysis can be used to produce a discrete classification of the locations, which can then be mapped in order to identify absolute patterns of aggregated regional linguistic variation. In this analysis, the linguistic distance matrix was subjected to a hierarchical cluster analysis (Shackleton, 2005, 2007; Goebel, 2007; Nerbonne & Heeringa, 2009; Prokic & Nerbonne, 2008; Wieling & Nerbonne, 2010). A hierarchical cluster analysis identifies clusters of similar objects in a distance matrix by initially assigning each observation to its own cluster and by then repeatedly combining the two most similar clusters to form larger and larger clusters until all of the objects have been combined to form one large cluster. Various methods exist for measuring the similarity between clusters consisting of multiple observations, but Ward's method (Ward, 1963) was used here because it is a

common approach to clustering, which has been found to perform well in dialectometry (Prokic & Nerbonne, 2008) and which tends to produce clear and compact clusters. The results of the cluster analysis are represented by a tree diagram called a *dendrogram*, which shows the order in which the clusters were formed, and which can be used to identify clusters and sub-clusters of observations in the dataset.

The dendrogram produced by the analysis is presented in Figure 10, which appears to identify two major regions, with one region being clearly divided into two sub-regions. The three regions are plotted in Figure 11. Similar to the results of the multidimensional scaling, the cluster analysis roughly divides the United states in three—the North, the Midland, and the South—although because these regions are not spatially homogenous it is unclear if they can be referred to as *dialect regions*, unless they are seen as approximating underlying patterns obscured by noise in the dataset. It is also important to note that the Midland region, which also stretches into the West, combines with the North to form a northern super-region. The three clusters identified dendrogram could also be divided further, but regional consistency breaks down even more at lower levels of the analysis.

Figure 10 Dialectometry Dendrogram based on Raw Distances using Ward's Method

Figure 11 Cluster Map Based on Raw Distance Matrix

4. Multivariate Spatial Analysis

The standard approach to dialectometry identified aggregated patterns of regional variation in the linguistic data matrix. The standard approach, however, does not allow for regional patterns to be identified in the values of the individual linguistic variables or in subsets of these variables, unlike a traditional analysis where the identification of isoglosses and bundles of isoglosses is central (Kurath, 1949). Séguy initiated the field of dialectometry to overcome the limitations of the isogloss method, but the analysis of individual linguistic variables and the identification of subsets of variables that exhibit similar patterns are still worthwhile. Most dialectometry analyses, however, focus on the linguistic distance matrix from which information about the patterns exhibited by the individual linguistic variables cannot be extracted directly (although for dialectometry research that addresses some of these issues see Shackleton, 2005, 2007; Nerbonne, 2006; Rumpf et al, 2009, 2010; Wieling and

Nerbonne, 2011). These limitations with the standard approach to dialectometry led to the development of an alternative statistical approach to the analysis of regional linguistic variation known as a *multivariate spatial analysis* (Grieve, 2009; Grieve et al, forthcoming). This method differs from the standard approach in two ways. First, the linguistic variables are analyzed using spatial autocorrelation statistics in order to identify individual patterns of regional variation. Second, the smoothed maps produced by the spatial autocorrelation analysis are subjected to a factor analysis in order to identify subsets of variables that follow similar patterns of regional variation. This section describes the results of subjecting the regional linguistic data matrix to a multivariate spatial analysis.

4.1 Global Spatial Autocorrelation Analysis

Spatial autocorrelation was analyzed in order to identify significant patterns of regional variation in the values of each of the 38 vowel formant variables. Spatial autocorrelation statistics are measures of spatial dependency—i.e. the tendency for the values of a variable to depend on their location so that nearby locations tend to either have similar values (i.e. spatial clustering) or different values (i.e. spatial dispersion). Spatial autocorrelation statistics therefore allow for patterns of spatial clustering to be identified in the values of linguistic variables in a regional linguistic data matrix.

By measuring *global* spatial autocorrelation, it possible to test if the values of a linguistic variable exhibit an overall level of spatial clustering that is significantly higher than would be expected by chance. Each of the 38 vowel formant variables was therefore subjected to an analysis of global spatial autocorrelation to identify variables exhibiting significant patterns of spatial clustering, using global Moran's I (Moran, 1948; Odland, 1988), which compares every pair of locations in the distribution of variable in order to determine if nearby locations tend to have similar, different, or random values. In particular,

Moran's *I* returns a significant positive value if nearby locations tend to have similar values at a greater degree than would be expected at random. The calculation of Moran's *I* also requires that a *spatial weighting function* is defined, which is a set of rules that assigns a weight to the comparison of every pair of locations so that comparisons between locations that are close together are given greater weight than comparisons between locations that are far apart (Odland, 1988). The analyses reported here are based on a reciprocal weighting function, which is a common spatial weight function that assigns a weight to a comparison based on the reciprocal of the distance between locations, so that weighting decreases with distance (Odland, 1988). A reciprocal weighting function was selected because it allows for the measure of global spatial autocorrelation to be based primarily on locations that are close together, which is more conservative and also controls to some degree for the uneven distribution of locations in the data matrix.

The analysis of global spatial autocorrelation found that 25 out of the 38 vowel formant variables exhibit significant levels of positive spatial autocorrelation, indicating that these variables are characterized by patterns of spatial clustering, including all of the variables mapped in Figure 2 except /ayv/ on format 1. The value of Moran's *I* was tested for significance by calculating the associated *z*-score under the assumption of randomization (Odland, 1988). A variable was deemed to exhibit significant global autocorrelation if the computed *z*-score was larger than or equal to 3.01—corresponding to a one-tailed .0013 significance level, which was selected based on a Bonferroni correction for 38 variables ($.05/38 = .0013$), which controls for the fact that every time a variable is added to the analysis the likelihood that a significant pattern will be found by chance increases. A one-tailed test was used because the goal of the analysis was to identify spatial clustering but not spatial dispersion. These results clearly show that there is considerable spatial clustering in the regional linguistic data matrix. Although this stage of the analysis is not necessary for the

analysis of aggregated regional variation, verifying that at least some individual variables do exhibit significant regional patterns is justification for further aggregated analysis.

4.2 Local Spatial Autocorrelation

In order to identify the location of clusters in the values of the linguistic variables, each variable was subjected to an analysis of local spatial autocorrelation. Unlike a measure of global spatial autocorrelation, which returns one value indicating the overall level of spatial clustering in the values of a variable, a measure of local spatial autocorrelation returns one value per location indicating the degree to which that particular location is part of a high or low-value cluster. The results of the local spatial autocorrelation analysis can then be mapped in order to identify the locations of these high- and low-value clusters.

In order to measure local spatial autocorrelation, local Getis-Ord G_i^* (Ord and Getis, 1995) was calculated for each location for each variable, using the same spatial weighting function described above. Local Getis-Ord G_i^* returns a z-score indicating the degree to which a location is surrounded by locations with similar values. A significant negative Getis-Ord G_i^* z-score indicates that the location is part of a low-value cluster, whereas a significant positive Getis-Ord G_i^* z-score indicates that the location is part of a high-value cluster. A Getis-Ord G_i^* z-score was interpreted as significant if it was larger than or equal to ± 3.21 , which corresponds to a two-tailed 0.0013 alpha level, based on the Bonferroni correction for 38 variables described above. A two-tailed test was used because the goal of the analysis was to identify both high- and low-value clusters.

The analysis of local spatial autocorrelation identified significant high- and/or low-value clusters for all 25 vowel formant variables found to exhibit significant levels of global autocorrelation. Figure 12 maps the Getis-Ord G_i^* z-scores for the 8 vowel formant variables whose raw values were mapped in Figure 2. A comparison of these two sets of maps shows

that the local spatial autocorrelation analysis identified patterns that are clearly visible in the raw maps of these linguistic variables. Specifically, the local autocorrelation analysis found that /eyc/ distinguishes between the North and the South, /ae/ distinguishes between the Midwest and the rest of the United States, /oh/ distinguishes between the Northeast and the West, and /ayv/ distinguishes between the Upper Midwest and the Midland, although this final pattern is weaker, especially on formant 1. The local autocorrelation maps also appear to be quite accurate. For example, the analysis correctly reproduces the small but important difference between the northern regions visible in the raw maps for /eyc/ and /ae/, with the northern region for /eyc/ reaching the East Coast and transitioning into the South in the Lower Midwest, but with the northern region for /ae/ encompassing the Lower Midwest and not extending past western New York. These eight maps are largely representative of the types of patterns identified by the analysis, although vowel measures do not always pattern similarly on formant 1 and formant 2.

By converting the raw values of the linguistic variables into corresponding Getis-Ord G_i^* z -scores that represent the location of significant high and low value clusters in the raw values of these variables, the local autocorrelation analysis has essentially smoothed the raw variables, removing noise added by other sources of linguistic variation, so that the values of the new smoothed variables only represent the underlying regional signals in the raw values of these variables. This smoothed regional linguistic data matrix is then subjected to further multivariate analysis in order to identify common patterns of regional linguistic variation.

Figure 12 Local Spatial Autocorrelation Getis-Ord G_i^* z-scores for 8 variables

4.3 Factor Analysis

In order to identify common patterns of regional variation in the values of the 38 vowel formant variables, the Getis-Ord G_i^* z -scores for the complete set of variables were subjected to a factor analysis. Given a set of variables measured over a large number of objects, a factor analysis generates a variable-by-variable correlation matrix and then extracts a series of aggregated factors that each represents a common pattern of correlated variation in the dataset, as well as the variables associated with each of these patterns (Hair et al, 2006). Because the local Getis-Ord G_i^* z -scores represent the location of spatial clusters in the values of the vowel formant variables, subjecting the smoothed data matrix to a factor analysis identifies common patterns of spatial clustering, as well as the specific vowel formant variables that are associated with each of these patterns. This application of factor analysis is similar to the use of multidimensional scaling in dialectometry, except that because the factor analysis is based on a correlation matrix, as opposed to a distance matrix, it is possible to determine the variables that define each dimension. While factor analysis and the related technique known as principal component analysis are not as common as multidimensional scaling in dialectometry, these statistics have been used on occasion (Shackleton, 2005, 2007; Nerbonne, 2006; Szmrecsanyi and Wolk, forthcoming), as well as in the *Atlas of North American English* (Labov et al, 2006), although the full power of these techniques is not exploited in any of these studies, nor are any of these analyses based on smoothed variables.

Before conducting a factor analysis, it is necessary to select the number of factors to extract—a choice that can be made by identifying the point where retaining further factors would explain relatively little additional variance. In this case 4 factors were selected because together they accounted for 86% of the variation in the values of the 38 vowel formant variables, with additional factors explaining relatively little additional variation. The final

factor analysis was thus run to extract 4 factors, which were rotated using varimax rotation. Varimax rotation was used to limit the number of factors onto which each of the variables loaded so that the factors would more clearly reflect the spatial patterns instantiated in the individual local autocorrelation maps.

The results of the factor analysis are presented in Table 2, which lists the loadings (larger than .300) of the 38 variables on the 4 factors and the uniqueness values associated with each of the variables. The loadings show the degree to which each vowel formant variable is represented by each factor. The uniqueness values, which are all relatively low, show that the 4 factors account for the regional patterns exhibited by all 38 variables relatively well. The 4 sets of factor scores are also mapped in Figures 13-16, which identify the common patterns of spatial clustering extracted by the factor analysis. Given these maps, the factor loadings can be interpreted further: a vowel with a positive loading on a factor is either higher or backer in the average vowel space for locations in the negative region identified by the factor, whereas a vowel with a negative loading is either lower or fronter in the average vowel space for locations in the positive region.

Factor 1 represents the strongest regional pattern, accounting for 39.0% of the variance in the smoothed data matrix and contrasting the Southeast with the North (Figure 13). Numerous vowel formant variables distinguish between these two regions, as indicated by the large number of variables loading on Factor 1. A majority of these variables are associated with Southern Shift (Labov et al, 2006), including the lowering and backing of /iy/ and /ey/, the rising and fronting of /i/, /e/, /ohr/ and /ahr/, and the fronting of /uw/ and /ow/, all of which characterize the Southern region identified by Factor 1. Factor 2 accounts for 23.4% of the variance and contrasts the Midwest with the rest of the United States (Figure 14). The factor loadings identify numerous vowel formant variables that are associated with the Midwest, many of which are also associated with the Northern Cities Shift (Labov et al,

2006), including the fronting and raising of /ae/, the fronting of /o/, the lowering and backing of /e/ and /i/, and the lowering of /uh/. Factor 3 accounts for 17.8% of the variance and contrasts the Northeast with the West (Figure 12). The vowel that loads most strongly on Factor 3 is /oh/ on both formant 1 and formant 2, with /oh/ tending to be lower and fronter in the West, thereby merging with /o/ to produce the low back merger, which is known to be present in the West and absent in Northeast (Labov et al, 2006). Factor 3 also identifies the fronting of back vowels as characteristic of Western locations (Labov et al, 2006). Finally, Factor 4 accounts for 6.2% of the variance and contrasts the Midland (especially the Lower Midwest) with the Upper Midwest (Figure 13). Only two variables load strongly on Factor 4, with /ayv/ tending to be both slightly higher and backer in the Midland than in the Upper Midwest.

Finally, in order to generate one continuous map of aggregated regional linguistic variation, the four sets of factor scores were mapped using CMYK mapping. CMYK mapping is a variant of RGB mapping, where each four sets of values are associated with one of the four CMYK color parameters (cyan, magenta, yellow, and black), which define one color, which can then be mapped. The results of the CMYK mapping are presented in Figure 21, identifying a clear regional pattern. This aggregate factor map identifies at least four clear clusters, consisting of the Northeast, the Midwest, the Southeast and the West, with areas of transition of varying widths between these clusters. Gradual internal change can also be seen within each of these regions.

Table 2 Factor Loadings and Uniqueness values Add Mean Values

Vowel	Formant	Uniqueness	Factor 1	Factor 2	Factor 3	Factor 4
/i/	1	0.214	0.828			
/i/	2	0.115		0.898		
/e/	1	0.175	0.691		-0.413	-0.37
/e/	2	0.097		0.871		
/ae/	1	0.051		0.927		
/ae/	2	0.206		-0.852		
/o/	1	0.406	0.382	-0.595		
/o/	2	0.140	0.352	-0.839		
/uh/	1	0.156	0.915			
/uh/	2	0.068		0.627	0.714	
/u/	1	0.267	0.827			
/u/	2	0.080		0.496	0.812	
/iyc/	1	0.069	-0.689		-0.522	0.325
/iyc/	2	0.061	0.839			-0.42
/eyc/	1	0.014	-0.866	0.448		
/eyc/	2	0.012	0.98			-0.162
/ayv/	1	0.478				0.65
/ayv/	2	0.214	0.317	-0.428	0.455	0.544
/ay0/	1	0.098	-0.524	0.595	0.498	
/ay0/	2	0.067	0.95			
/iw/	1	0.269			-0.762	0.324
/iw/	2	0.033	-0.758	0.521	0.333	
/uwc/	1	0.276	-0.388	0.736		
/uwc/	2	0.019	-0.786	0.517	0.308	
/uwf/	1	0.180		0.574	-0.642	
/uwf/	2	0.023	-0.618	0.484	0.52	-0.302
/owc/	1	0.147	-0.857			
/owc/	2	0.025	-0.822	0.503		
/aw/	1	0.300	0.644		0.372	-0.348
/aw/	2	0.022	-0.668	0.621	-0.37	
/oh/	1	0.064			0.947	
/oh/	2	0.024	0.315		0.901	
/ahr/	1	0.172	0.753		0.396	
/ahr/	2	0.093	0.761	-0.554		
/ohr/	1	0.148	0.49		0.634	0.395
/ohr/	2	0.183	0.774		0.362	
/owr/	1	0.116	0.676	-0.388	0.483	
/owr/	2	0.092	0.933			

Figure 13 Factor 1

Figure 14 Factor 2

Figure 15 Factor 3

Figure 16 Factor 4

4.4 Cluster Analysis

In addition to aggregating the four sets of factor scores using CMYK mapping to produce an overall map of continuous linguistic variation, it is also possible to identify dialect regions by clustering the locations based on their factor scores using an agglomerative hierarchical cluster analysis, as in dialectometry. In this case, however, the location-by-location distance matrix was based on the four sets of extracted factor scores, instead of the raw variable values. Alternatively, the distance matrix could have been based on the smoothed Getis-Ord G_i^* z -scores for the raw variables, but in this case this change resulted in a very similar classification.

Based on the dendrogram reproduced in Figure 17, the hierarchical cluster analysis identified 5 clear dialect regions, which are mapped in Figure 18: the Northeast, the Southeast, the West, and the Lower and Upper Midwest, with the Lower and Upper Midwest combining to form a Midwestern super-region. Although the cluster analysis also groups the Northeast with the Midwest, and the West with the Southeast, because these two super-regions are formed so late in the cluster analysis these clusters are less reliable. On the other hand, the most distinct internal clusters within each of the 5 main clusters identified above are of some importance, although the clusters do lose spatial consistency as one descends into the lower levels of the dendrogram. In particular, the Northeast is divided into New England and Mid Atlantic sub-regions, the West is divided into Coastal and Inland regions along the Rocky Mountains, and the Lower Midwest is divided into Northern (i.e. the Northern Cities) and Southern sub-regions.

Figure 17 Dendrogram for Multivariate Spatial Analysis (based on 4 factors)

Figure 18 Cluster Analysis Multivariate Spatial Analysis

5. Conclusion

The regional linguistic data matrix, which consists of the values of 38 vowel formant variables measured over 236 American cities based on the acoustic data from the *Atlas of North American English* (Labov et al, 2006), was subjected to two statistical analyses in order to identify aggregated patterns of regional linguistic variation. The data matrix was first analyzed using the standard approach to dialectometry, which identified a north-south distinction. The data matrix was then subjected to a multivariate spatial analysis, which identified a north-south distinction, as well as strong Midwestern, Northeastern and Western clusters.

The multivariate spatial analysis identified more detailed regional patterns than the dialectometry analysis because the linguistic variables were smoothed prior to aggregation. By using a local spatial autocorrelation analysis to transform the raw regional linguistic data matrix into a smoothed regional linguistic data matrix, representing the underlying patterns of spatial clustering in the values of the individual linguistic variable, it was possible to focus the multivariate analysis on the identification of aggregated regional linguistic variation. The standard dialectometry analysis, on the other hand, was forced to identify aggregated regional patterns in the midst of various other forms of linguistic variation, which essentially constitute non-regional noise in the dataset, whose presence confounds the aggregated analysis. In fact, had the dialectometry analysis been based on the smoothed data matrix, then the dimensions extracted by the multidimensional scaling would have been very similar to the factors extracted by the factor analysis as illustrate in Figure 22, which is based on a RGB mapping of three dimensions extracted from the smoothed regional linguistic data matrix using a multidimensional scaling. The increased level of detail identified by the multivariate spatial analysis is therefore due to the autocorrelation smoothing, as opposed to the method of aggregation.

Figure 19 Visualization of a Geographic Distance Matrix

Figure 20 RGB Map of Dimensions 1,2 and 3 from MDS based on Raw Variables

Nonetheless, the use of factor analysis as opposed to multidimensional scaling is an important difference between the multivariate spatial analysis and the standard approach to dialectometry. A factor analysis is based on a variable-by-variable correlation matrix and is therefore able to identify the linguistic variables that are associated with each of the extracted factors. This information is lost in a multidimensional scaling, which is based on a location-by-location distance matrix. It is important to know which variables are associated with which patterns in order to verify the results of the analysis. For example, the four pairs of vowel formant variables mapped in Figure 2 and 12 each load strongly on one of the four factors, and in each case it is clear that the regional pattern identified by the corresponding factor (Figures 13-16) is representative of the regional patterns instantiated in those vowel formant variables. It is also important to know which variables are associated with which patterns in order to explain the patterns identified by the aggregated analysis. For example, the fact that Factor 1 and Factor 2 identify a Southern dialect region and a Midwestern dialect region can be explained by the fact that the vowel formant variables loading on each factor are associated with the Southern Shift and the Northern Cities Shift, respectively.

The advantages of the multivariate spatial analysis can be illustrated by considering the status of the Midland in these two analyses. Based on the dialectometry analysis, it is unclear if the Midland is an area of transition between the North and the South, as suggested by the multidimensional scaling, or a distinct dialect region, as suggested by the hierarchical cluster analysis. The multivariate spatial analysis, however, resolves this dilemma. Factor 4 does identify a distinct Midland dialect region, which contrasts primarily with the Upper Midwest; however, this is the only factor that groups Philadelphia with the cities of the Lower Midwest, which is central to the traditional definition of the Midland dialect region (Kurath, 1949; Labov et al, 2006). Furthermore, Factor 4 only accounts for 6% of the variance in the dataset and only strongly loads the /ayv/ vowel formant variables, which

reflects the fact that the local spatial autocorrelation analysis only identifies a distinct Midland region in the values of these two vowel formant variables. The Midland therefore is clearly not a strong dialect region—far weaker than the Southern, Midwestern, and Northeastern regions with which it overlaps.

The results of the factor analysis also offer an explanation for why the strength of the Midland is overestimated in the dialectometry analysis, while the distinction between the Midwest and the Northeast is largely lost. The strength of the Midland is amplified due to the fact that three different patterns interact in this region: the weak Midland dialect region defined by Factor 4 roughly corresponds to the area of overlap between the South as defined by Factor 1 and the Midwest defined by Factor 2. These three distinct patterns, which are represented by three factors, are conflated by the dialectometry analysis, which essentially identifies one basic north-south pattern in the dataset. In addition, the distinction between the Northeast and the West (Factor 3) is largely lost in the dialectometry analysis because this factor is overwhelmed by the combined strength of these three factors.

Despite apparent success of the multivariate spatial analysis, it is notable that the dialect regions identified by this analysis (Figure 18) differ from the dialect regions identified in the *Atlas of North American English* (Figure 23), which is based in large part on the same acoustic dataset. In fact, the dialect regions identified by the dialectometry analysis (Figure 11) are clearly more in line with the regions identified in the Atlas due to the presence of a Midland dialect region. It appears, however, that the traditional analysis has misidentified the Midland as strong dialect region for the same basic reasons as the dialectometry analysis, in addition to the fact that the Midland has been identified as a strong region in previous research (Kurath, 1949), which may have biased the subjective analysis upon which the Atlas is based. Nevertheless while the results of this multivariate spatial analysis do not align particularly well with the regions identified by the Atlas, these results are supported by a

careful analysis of the individual variables, as argued above. In addition, these results correspond to the results of perceptual dialectology (Preston, 2002), and multivariate spatial analyses of grammatical variation in written American English (Grieve, 2009, forthcoming; Grieve et al, forthcoming), as well as the cultural regions of the United States (Zelinsky, 1973).

Figure 23 Simplified Dialect Regions from the Atlas of North American English

Finally, while the methods applied here were designed for the analysis of regional linguistic variation within one language, these basic methods can also be used to analyze regional variation across language, i.e. in typological research. Specifically, an obvious applications of spatial autocorrelation statistics in typological research is to test for areal diffusion—i.e. to test if languages that are not genetically related or that are distantly genetically related are linguistically similar due to contact. While the Mantel test is occasionally applied for this purpose (Barbujani et al, 1994; Weng and Sokal, 1995), this

method generally underestimates the strength of a geographical pattern in a dataset and therefore should not be applied in typology. Applying a global spatial autocorrelation analysis to the individual variables under consideration is a more precise method for testing for areal diffusion. Once identified, spatial autocorrelation should then be controlled, because areal patterns confound the analysis of genetic patterns. The methods used here could also be used to identify areal groupings of languages independent of their genetic relationships in order to facilitate the comparison of geographic and genetic patterns in languages of the world.

References

- Barbujani, G., G. N. Whitehaed, G. Bertorelle, and I. S. Nasidze, 1994. Testing hypotheses on processes of genetic and linguistic change in the Caucasus. *Human Biology: the International Journal of Population Biology and Genetics* 66: 843–864.
- Chambers, Jack, and Peter Trudgill. 1998. *Dialectology* (2nd Edition). Cambridge, UK: Cambridge University Press.
- Cox, Trevor F. and Michael A. A. Cox. 2001. *Multidimensional Scaling* (2nd Edition). Boca Raton, FL: Chapman & Hall.
- Cronbach, Lee. 1951. Coefficient alpha and the internal structure of tests. *Psychometrika* 16: 297–334.
- Goebel, Hans. 1982. *Dialektometrie; Prinzipien und Methoden des Einsatzes der numerischen Taxonomie im Bereich der Dialektgeographie*. Vienna: Verlag der Österreichischen Akademie der Wissenschaften.
- Goebel, Hans. 1984. *Dialektometrische Studien: Anhand italoromanischer, rätoromanischer und galloromanischer Sprachmaterialien aus AIS und ALF*. Tübingen: Niemeyer.
- Goebel, Hans. 2006. Recent Advances in Salzburg Dialectometry. *Literary and Linguistic Computing* 21: 411-435.
- Goebel, Hans. 2007. On the geolinguistic change in Northern France between 1300 and 1900: a dialectometrical inquiry. In John Nerbonne, T. Mark Ellison, and Grzegorz Kondrak (eds.), *Computing and Historical Phonology*. Proceedings of the Ninth Meeting of the ACL Special Interest Group in Computational Morphology and Phonology (Prague). 75–83.
- Gooskens, Charlotte, and Wilbert Heeringa. 2006. The relative contribution of pronunciation, lexical and prosodic differences to the perceived distances between Norwegian dialects. *Literary and Linguistic Computing* 21: 477-492.

- Grieve, Jack. 2009. *A Corpus-Based Regional Dialect Survey of Grammatical Variation in Written Standard American English*. Ph.D. Dissertation. Manuscript. Northern Arizona University
- Grieve, Jack. Forthcoming. A statistical analysis of regional variation in adverb position in a corpus of written Standard American English. *Corpus Linguistics and Linguistic Theory*.
- Grieve, Jack, Dirk Speelman, and Dirk Geeraert. Forthcoming. A statistical method for the identification and aggregation of regional linguistic variation. *Language Variation and Change*.
- Hair, Joseph, Bill Black, Barry Babin, Rolph E. Anderson, and Ronald L. Tatham. 2006. *Multivariate Data Analysis* (6th edition). Englewood Cliffs, NJ: Prentice-Hall.
- Heeringa, Wilbert. 2004. Measuring Dialect Pronunciation Differences using Levenshtein Distance. Ph.D. Dissertation. Manuscript. University of Groningen.
- Heeringa, Wilbert, and John Nerbonne. 2001. Dialect areas and dialect continua. *Language Variation and Change* 13: 375–400.
- Heeringa, Wilbert, John Nerbonne, and Peter Kleiweg. 2002. Validating dialect comparison methods. In Wolfgang Gaul and Gunter Ritter (eds.), *Classification, Automation, and New Media. Proceedings of the 24th Annual Conference of the Gesellschaft für Klassifikation*. Berlin: Springer. 445-452.
- Kurath, Hans. 1949. *Word Geography of the Eastern United States*. University of Michigan Press.
- Labov, William, Sharon Ash, and Charles Boberg. 2006. *Atlas of North American English: Phonetics, Phonology, and Sound Change*. New York: Mouton de Gruyter.

- Legendre, Pierre and Marie-Josée Fortin. 2010. Comparison of the Mantel test and alternative approaches for detecting complex multivariate relationships in the spatial analysis of genetic data. *Molecular Ecology Resources* 10: 831-844.
- Mantel, Nathan. 1967. The detection of disease clustering and a generalized regression approach. *Cancer Research* 27: 209–220.
- Moran, Patrick A. P. 1948. The interpretation of statistical maps. *Journal of the Royal Statistical Society, Series B* 37: 243- 251.
- Nerbonne, John. 2006. Identifying linguistic structure in aggregate comparison. *Literary and Linguistic Computing* 21: 463-476.
- Nerbonne, John. 2009. Data-driven dialectology. *Language and Linguistics Compass* 3: 175-198.
- Nerbonne, John. 2010a. Measuring the diffusion of linguistic change. In Alfred Lameli, Ronald Kehrein, and Stephan Rabanus (eds.), *Language and Space*. Berlin: Mouton De Gruyter. 476-495.
- Nerbonne, John. 2010b. Measuring the diffusion of linguistic change. *Philosophical Transactions of the Royal Society B: Biological Sciences* 365: 3821-3828.
- Nerbonne, John, and Wilbet Heeringa. 1997. Measuring dialect distance phonetically. In John Coleman (ed.), *Workshop on Computational Phonology, Special Interest Group of the Association for Computational Linguistics*. 11-18.
- Nerbonne, John, and Wilbert Heeringa. 2001. Computational comparison and classification of dialects. *Dialectologia et Geolinguistica* 9: 69–83.
- Nerbonne, John, and Wilbert Heeringa. 2007. Geographic distributions of linguistic variation reflect dynamics of differentiation. In Sam Featherston and Wolfgang Sternefeld (eds.), *Roots: Linguistics in Search of its Evidential Base*. Berlin: Mouton De Gruyter. 267-297.

- Nerbonne, John, and Wilbert Heeringa. 2009. Measuring dialect differences. In Jürgen Erich Schmidt, and Peter Auer (eds.), *Language and Space: Theories and Methods*. Berlin: Mouton De Gruyter. 550-567.
- Nerbonne, John, and Wilbert Heeringa. 2010. Measuring dialect differences. In Peter Auer and Jürgen Erich Schmidt (eds.), *Language and Space: An international Handbook of Linguistic Variation. Volume 1: Theories and Methods*. Berlin: Mouton De Gruyter. 550-567.
- Nerbonne, John and Wilbert Heeringa. Forthcoming. Dialectometry. In Frans Hinskens and Johan Taeldeman (eds.), *Language and Space. An International Handbook of Linguistic Variation, Volume III: Dutch*. Berlin: Mouton de Gruyter.
- Nerbonne, John, Wilbert Heeringa, Eric van den Hout, Peter van de Kooi, Simone Otten, and Willem van de Vis. 1996. Phonetic distance between Dutch dialects. In G. Durieux, W. Daelemans, and S. Gillis (eds.), *Proceedings of the Sixth CLIN Meeting*. Antwerp: Centre for Dutch Language and Speech. 185-202.
- Nerbonne, John, and Peter Kleiweg. 2003. Lexical Distance in LAMSAS. *Computers and the Humanities* 37: 339-357.
- Nerbonne, John, and Peter Kleiweg. 2007. Toward a Dialectological Yardstick. *Journal of Quantitative Linguistics* 14: 148-166.
- Nunnally, J. C. 1978. *Psychometric Theory*. McGraw-Hill, New York.
- Odland, John D. 1988. *Spatial Autocorrelation*. Sage Publications.
- Ord, J.K., and Arthur Getis. 1995. Local Spatial Autocorrelation Statistics: Distributional Issues and an Application. *Geographical Analysis* 27: 286-306.
- Preston, Dennis. 2003. Language with attitude. In Jack Chambers, Peter Trudgill, and Natalie Schilling-Estes (eds.), *The Handbook of Language Variation and Change*. Malden, MA: Blackwell. 40-66.

- Prokic, Jenna, and John Nerbonne. 2008. Recognizing groups among dialects. *International Journal of Humanities and Arts Computing* 1: 153-172.
- Rumpf, Jonas, Simon Pickl, Stephan Elspass, Wener Koenig, and Volker Schmidt. 2009. Structural analysis of dialect maps using methods from spatial statistics. *Zeitschrift für Dialektologie und Linguistik* 76: 280-308.
- Rumpf, Jonas, Simon Pickl, Stephan Elspass, Wener Koenig, and Volker Schmidt. 2010. Quantification and statistical analysis of structural similarities in dialectological area-class maps. *Dialectologia et Geolinguistica* 18: 73–100.
- Séguy, Jean. 1971. La relation entre la distance spatiale et la distance lexicale. *Revue de linguistique romane* 35: 335-357.
- Séguy, Jean. (1973). La dialectometrie dans l'Atlas linguistique de la Gascogne. *Revue de linguistique romane* 37: 1-24.
- Shackleton, Robert G. 2005. English-American speech relationships: a quantitative approach. *Journal of English Linguistics* 33: 99-160.
- Shackleton, Robert G. 2007. Phonetic variation in the traditional English dialects: a computational analysis. *Journal of English Linguistics* 35: 30-102.
- Spruit, Marco René, Wilbert Heeringa, and John Nerbonne. 2009. Associations among Linguistic Levels. *Lingua* 119: 1624-1642.
- Szmrecsanyi, Benedikt. 2008. Corpus-based dialectometry: aggregate morphosyntactic variability in British English dialects. *International Journal of Humanities and Arts Computing* 2: 279-296
- Szmrecsanyi, Benedikt and Christoph Wolk. Submitted. Holistic corpus-based dialectology. Available at <http://www.benszm.net/publications.html>

- Weng, Z. and R. R. Sokal. 1995. Origins of Indo-Europeans and the spread of agriculture in Europe: comparison of lexicostatistical and genetic evidence. *Human Biology: the International Journal of Population Biology and Genetics* 67: 577–594.
- Wieling, Martijn, and John Nerbonne. 2010. Hierarchical bipartite spectral graph partitioning to cluster dialect varieties and determine their most important linguistic features. *TextGraphs-5 Workshop on Graph-Based Methods for Natural Language Processing* 16: 33-41.
- Wieling, Martijn and John Nerbonne. 2011. Bipartite spectral graph partitioning for clustering dialect varieties and detecting their linguistic features. *Computer Speech and Language* 25: 700-715.
- Zelinsky, Wilbur. (1973). *Cultural Geography of the United States*. Englewood Cliffs, NJ: Prentice-Hall.