A Statistical Comparison of Regional Phonetic and Lexical Variation in American English

Jack Grieve

j.grieve1@aston.ac.uk

Centre for Forensic Linguistics

School of Languages and Social Sciences

Aston University

Aston Triangle

Birmingham UK, B4 7ET

Submitted November 22nd, 2011

Resubmitted with major revisions June 29th, 2012

Accepted July 27th, 2012

Resubmitted with minor revisions September 16th, 2012

Abstract

This paper presents a statistical comparison of common patterns of regional phonetic and lexical variation in American English based on the results of two previous dialect studies. Because these two studies are based on datasets that represent different cities, this paper also introduces a general method for comparing dialect maps that are based on different sets of locations. This method of comparison consists of two steps. First, the dialects maps are defined across a shared set of reference locations through ordinary kriging. Second, these normalized maps are correlated to each other in order to estimate the similarity between the original dialect maps. The results of this comparison show that regional phonetic and lexical variation follow similar patterns in Modern American English.

Keywords

American English, Dialectometry, Lexical Variation, Ordinary Kriging, Phonetic Variation, Regional Dialects, Variogram Analysis

Acknowledgements

I would like to thank Dirk Geeraerts, Kris Heylen, John Nerbonne, Tom Ruette, Robert Shackleton, Dirk Speelman, Joeri Theelen, Emily Waibel, and two anonymous reviewers for their comments on this paper. I would also like to thank Charles Boberg, Wilbert Heeringa, and Martijn Wieling for their comments on a preliminary version of this paper, which was presented at Methods in Dialectology 14 at the University of Western Ontario, August 2nd, 2011.

1. Introduction

Regional linguistic variation in American English has not been adequately compared across linguistic levels. Kurath designed the questionnaire for *The Linguistic Atlas of the United States and Canada* to elicit data from many levels of linguistic analysis, but the *Atlas* was never completed and aside from New England (Kurath et al, 1939) Kurath and his colleagues presented their lexical (Kurath, 1949), morphological (Atwood, 1953), and phonological (Kurath & McDavid, 1961) analyses for the eastern United States separately. All three of these studies did find similar regional patterns but the data was never explicitly compared across these three linguistic levels. Similarly, later surveys that mapped regional linguistic variation in American English focused specifically on lexical (Cassidy, 1985; Carver, 1987), phonetic/phonological (Labov et al, 2006), or lexico-grammatical variation (Grieve, 2009, 2011, 2012; Grieve et al, 2011). Although all of these surveys compared their results to the results of previous American dialect surveys, these comparisons were always informal because the datasets upon which the previous surveys were based were not available for analysis. The lack of a general method for comparing dialect maps based on different sets of locations has also complicated the comparison of regional variation across linguistic levels.

This study introduces a general statistical method for the comparison of dialect maps based on different sets of locations and uses this method to compare common patterns of regional phonetic and lexical variation in American English, as identified in two previous studies. In particular, this study compares the results of a quantitative analysis of vowel formant data (Grieve, forthcoming; Grieve et al, submitted) based on the data from *The Atlas of North American English* (Labov et al, 2006) to the results of a quantitative analysis of lexical variation in a corpus of written American English (Grieve et al, 2011). In both studies, a series of dialect maps representing common patterns of regional linguistic variation were generated through a multivariate spatial analysis. In this study, these two series of aggregated

dialect maps are compared to each other in order to assess the similarity of regional phonetic and lexical variation in American English. Before presenting the results of this comparison, however, the results of these two previous studies are described and the method for the comparison of dialect maps is introduced.

2. Data

This paper compares the common patterns of regional phonetic and lexical variation in American English identified in two previous studies. In particular, the phonetic analysis is presented in Grieve (forthcoming) and Grieve et al (submitted) and the lexical analysis is presented in Grieve et al (2011). In both studies, the values of a series of quantitative linguistic variables measured over a series of locations from across the contiguous United States were subjected to a multivariate spatial analysis to identify common patterns of regional linguistic variation. The four common patterns of regional phonetic variation are mapped in Figures 1-4 and the three common patterns of regional lexical variation are mapped in Figures 5-7. This section briefly describes how these two series of maps were obtained.

The phonetic analysis was based on the acoustic vowel formant data from the *Atlas* of North American English (Labov et al, 2006), which was collected during the 1990s through linguistic interviews conducted over the telephone with 439 informants from across the United States and Canada. In particular, the maps under analysis here are based on the average formant 1 and formant 2 values for 19 contextualized vowel phonemes measured across 236 cities in the contiguous United States. These 38 vowel formant variables are listed in Table 1 using the transcription and classification systems from the *Atlas* along with their IPA equivalents and their average formant 1 and formant 2 values across the entire dataset.

The lexical analysis was based on a 26 million word corpus representing the letter to the editor register as written between 2000-2010 in 206 cities from across the United States (see also Grieve, 2009, 2011, 2012). In particular, the maps under analysis here are based on the values of 40 lexical alternation variables measured across the 206 cities in the corpus (Grieve et al, 2011), where each alternation was measured as the proportion of one high frequency function word or adverb relative to a synonymous high frequency function word or adverb. These 40 lexical alternations and their variants are listed in Table 2 along with the average proportion of the first variant across the entire corpus. Although most of these variables are based on words that are generally interchangeable, some variables are based on words that are only interchangeable in certain contexts and some variables involve the same words counted in different contexts, as noted in Table 2.

In order to identify common patterns of regional linguistic variation in American English, both the phonetic and the lexical datasets were subjected to a multivariate spatial analysis, which consists of two basic steps. First, the individual linguistic variables were subjected to a local spatial autocorrelation analysis using Getis-Ord *Gi* (Ord & Getis, 1995; Grieve, 2011, 2012) to identify patterns of spatial clustering. Second, the Getis-Ord *Gi z*-scores for the complete set of linguistic variables were subjected to a factor analysis to identify a series of factors that each identify a unique and common pattern of spatial clustering. By mapping the factor scores (Figures 1-7), it is possible to visualize the common patterns of spatial clustering identified by the factor analysis, and by inspecting the factor loadings (Tables 1-2), it is possible to identify the variables that are represented by each of these maps.

The multivariate spatial analysis of the phonetic data identified four common patterns of regional variation, which account for 86% of the variance in the Getis-Ord *Gi z*-scores for the 38 phonetic variables. These four sets of factor scores are mapped in Figures 1-4 and the

variable loadings for each of these factors are presented in Table 1. Factor 1 represents the strongest pattern of regional variation in the dataset, accounting for 39% of the variance and contrasting the Southeast with the rest of the United States (Figure 1). A majority of the variables loading on Factor 1 are associated with the Southern Shift (Labov et al. 2006). Factor 2 accounts for 23% of the variance and contrasts the Midwest with the rest of the United States (Figure 2). A majority of the variables loading on Factor 2 are associated with the Northern Cities Shift (Labov et al. 2006). Factor 3 accounts for 18% of the variance and contrasts the East with the West (Figure 3). A majority of variables loading on Factor 3 appear to be associated with a Western shift. Finally, Factor 4 accounts for 6% of the variance and contrasts the Midland and the West with the rest of the United States (Figure 4), although only two variables load strongly on this factor.

The multivariate spatial analysis of the lexical data identified three common patterns of regional variation, which account for 54% of the variance in the Getis-Ord *Gi z*-scores for the 40 lexical variables. This is considerably less variance than was explained in the phonetic analysis; however, accounting for 54% of the variance in the values of 40 linguistic variables with 3 factors is still substantial. These three sets of factor scores are mapped in Figures 5-7 and the variable loadings for each of these factors are presented in Table 2. Factor 1 accounts for 24% of the variance and contrasts the East Coast and the West Coast with the rest of the United States (Figure 4). A majority of the variables loading on Factor 1 are associated with an opposition between more formal forms, which are more common on the Coasts, and more informal forms, which are more common in the rest of the United Sates. Factor 2 accounts for 18% of the variance and contrasts the Northeast and especially the Midwest with the rest of the United States (Figure 5). A majority of the variables loading on Factor 2 are associated with an opposition between standard American forms, which are more common in the Midwest, and non-standard forms, which are more common in the rest of the United Sates.

Finally, Factor 3 accounts for 13% of the variance and contrasts the Southeast with the rest of the United States. The linguistic interpretation of Factor 3 is not as clear as the other factors identified in these analyses, but it was retained because it accounted for a relatively large amount of variance and because when mapped it exhibits a clear regional pattern.

The goal of this study is to compare these four patterns of phonetic variation to these three patterns of lexical variation in order to gauge the similarity of regional phonetic and lexical variation in Modern American English. The rest of this paper describes the statistical procedure through which these factor maps were compared and presents the results of this comparison.

Table 1 Vowel Formant Variables and Factor Loadings

Vowel	IPA	Formant	Mean (Hz)	Factor 1	Factor 2	Factor 3	Factor 4
ae	/æ/	1	744		0.927		
		2	1869		-0.852		
ahr	/a/ + /r/	1	721	0.753		0.396	
		2	1227	0.761	-0.554		
aw	/aʊ/	1	804	0.644		0.372	-0.348
		2	1600	-0.668	0.621	-0.37	
ay0	/ai/ + voiceless	1	777	-0.524	0.595	0.498	
	consonant	2	1481	0.95			
ayv	/ay/ + voiced	1	813				0.65
	consonant	2	1462	0.317	-0.428	0.455	0.544
e	/ɛ/	1	653	0.691		-0.413	-0.37
		2	1826		0.871		
eyc	/e/ word	1	584	-0.866	0.448		
	internal	2	2017	0.98			-0.162
i	/ɪ/	1	517	0.828			
		2	1933		0.898		
iw	/u/	1	425			-0.762	0.324
		2	1843	-0.758	0.521	0.333	
iyc	/i/ word	1	422	-0.689		-0.522	0.325
	internal	2	2322	0.839			-0.42
o	/a/	1	822	0.382	-0.595		
		2	1334	0.352	-0.839		
oh	/ɔ/ + non-	1	755			0.947	
	rhotic cons.	2	1177	0.315		0.901	
ohr	/ɔ/ + /r/	1	548	0.49		0.634	0.395
		2	925	0.774		0.362	
owc	/o/ + non-	1	625	-0.857			
	rhotic cons.	2	1267	-0.822	0.503		
owr	/o/ + /r/	1	533	0.676	-0.388	0.483	
		2	906	0.933			
u	/ʊ/	1	552	0.827			
		2	1425		0.496	0.812	
uh	/ʌ/	1	702	0.915			
		2	1447		0.627	0.714	
uwc	/u/ word	1	456	-0.388	0.736		
	internal	2	1373	-0.786	0.517	0.308	
uwf	/u/ word final	1	452		0.574	-0.642	
		2	1787	-0.618	0.484	0.52	-0.302

Table 2 Lexical Alternation Variables and Factor Loadings[1]

Variant 1	Variant 2	Mean	Factor 1	Factor 2	Factor 3
about	around	0.8772			0.661
about	on	0.4108	-0.481		
actually	in fact	0.61	0.654	-0.393	-0.386
amid	amidst	0.037		0.308	
amongst	among	0.0275		-0.588	0.584
anyone	anybody	0.9362		0.767	-0.565
as well as	in addition to	0.837	0.742		
be going to	will	0.0449	0.638		
because of	due to	0.6728	0.768		
below	under	0.0882			0.774
clearly	obviously	0.5577	-0.784	0.387	
em	them	0.0062		-0.872	
especially	particularly	0.7942	0.873		
everyone	everybody	0.9158		0.851	
Firstly, secondly, thirdly, fourthly, fifthly, lastly	First, second, third, fourth, fifth, last	0.1436			
forward, backward, upward, downward	forwards, backwards, upwards, downwards	0.9609			
have to	must	0.4708	0.809		
however	nonetheless, nevertheless	0.957			
if	whether	0.793		-0.596	
may	might	0.7761	-0.677	0.646	
maybe	perhaps	0.5171	0.842		
no one	nobody	0.8197		0.483	-0.642
of genitive	‘s genitive	0.6984	0.409		-0.458
ought	should	0.0149	0.312		0.825
shall	will	0.0147	0.423	-0.362	
so as to	in order to	0.064			
someone	somebody	0.9415	-0.424	0.694	-0.484
that	which	0.9708			
therefore	thus	0.5317	0.391	0.603	-0.515
though	although	0.5163	0.831	-0.463	
to	toward, towards	0.9835			
toward	towards	0.9058		0.639	0.397
until	till, 'til	0.971	-0.62	-0.253	
usually	normally	0.8104		0.402	
whatsoever	at all	0.0955	-0.496	-0.461	
which	that	0.6802			-0.662
whilst	while	0.0005		-0.765	
who	that	0.9294	-0.592	0.642	
who	that	0.8654	-0.624		
whom	who	0.5837	-0.715		

Figure 1 Phonetic Factor 1

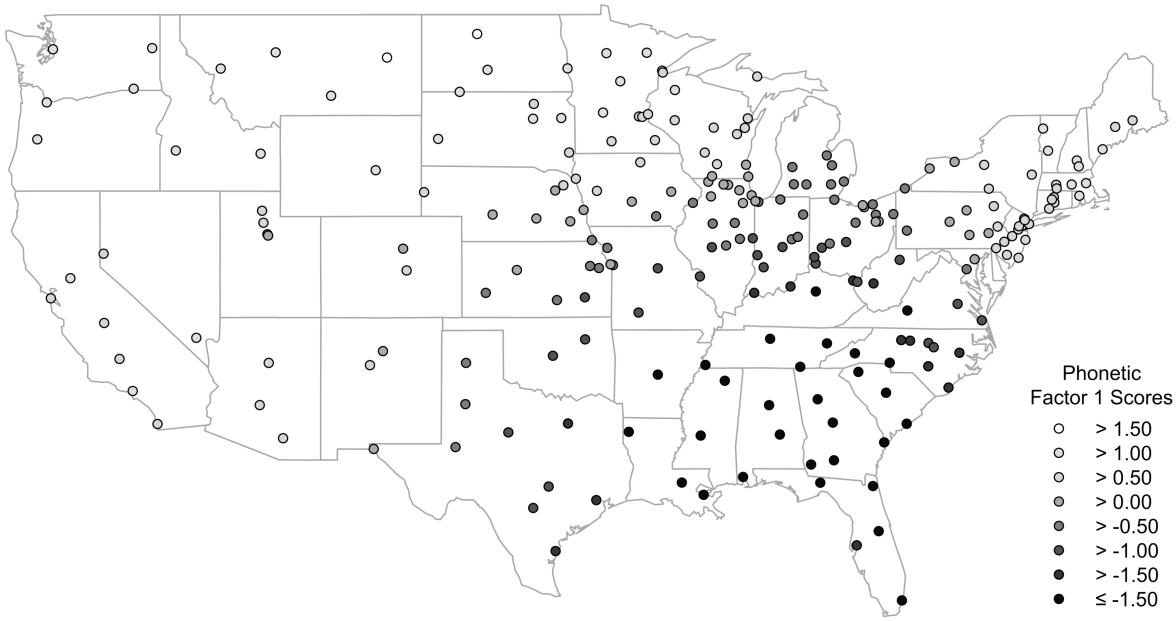


Figure 2 Phonetic Factor 2

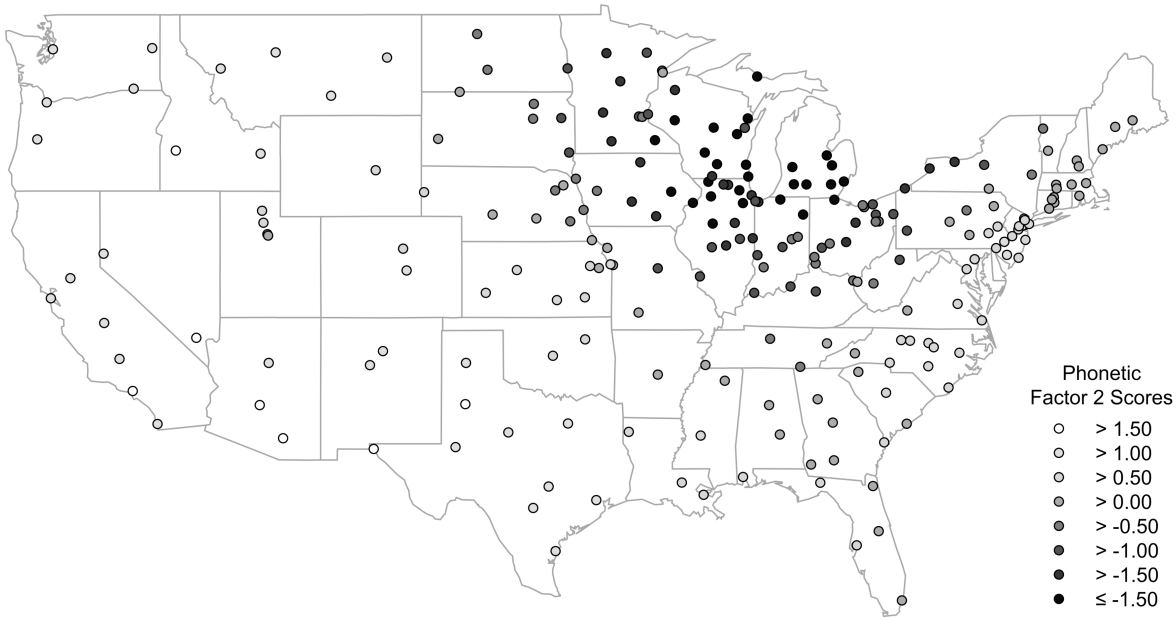


Figure 3 Phonetic Factor 3

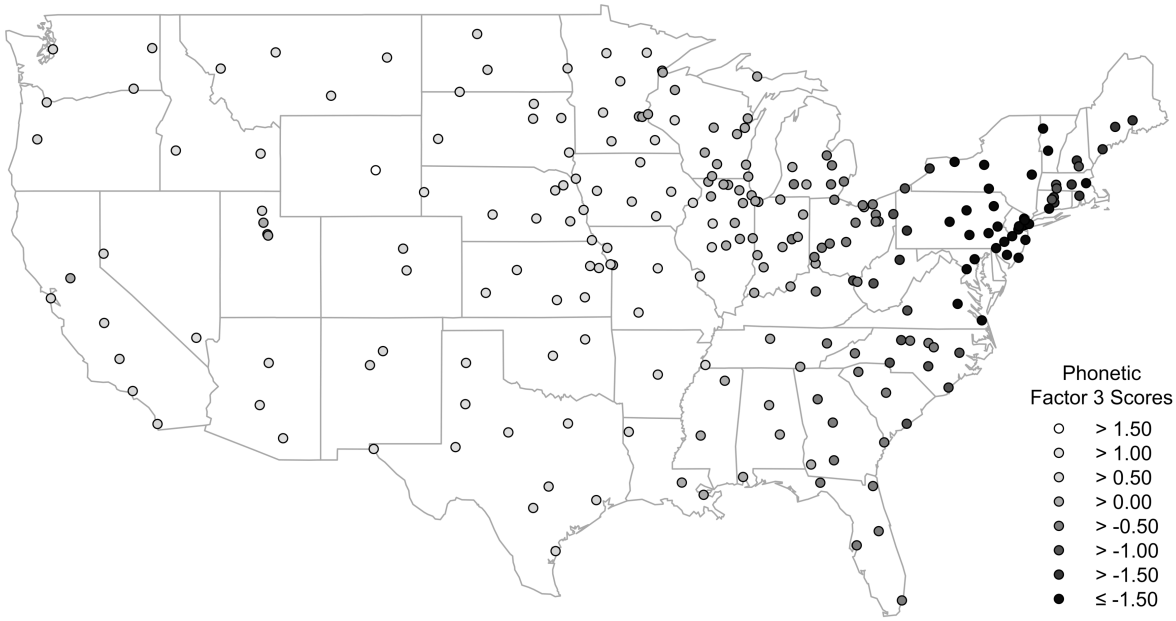


Figure 4 Phonetic Factor 4

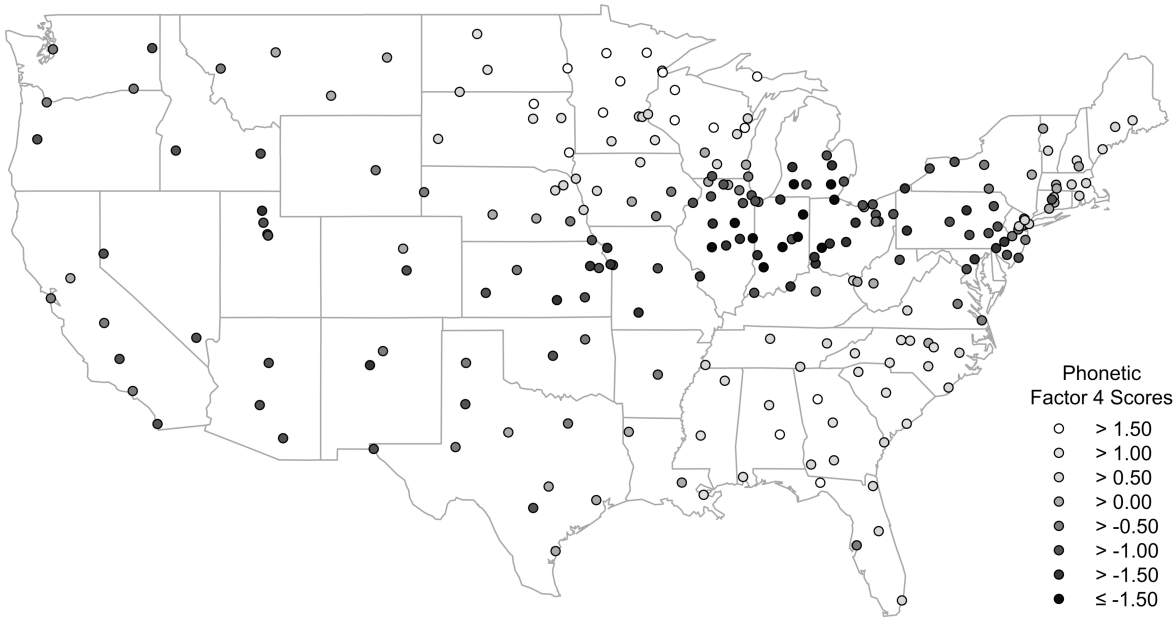


Figure 5 Lexical Factor 1

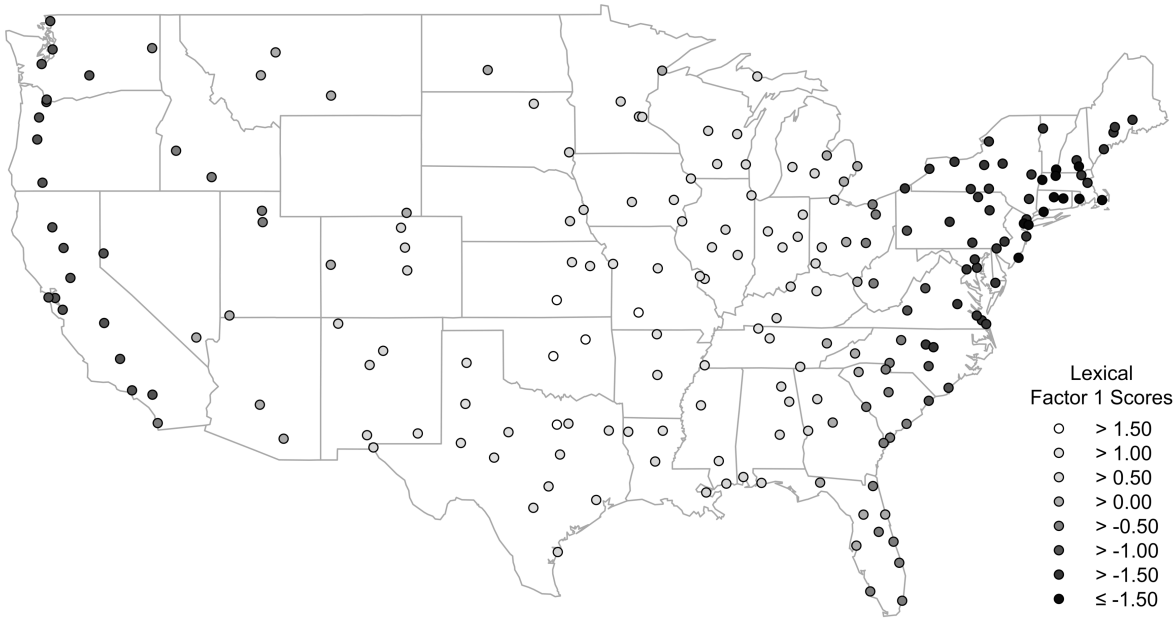


Figure 6 Lexical Factor 2

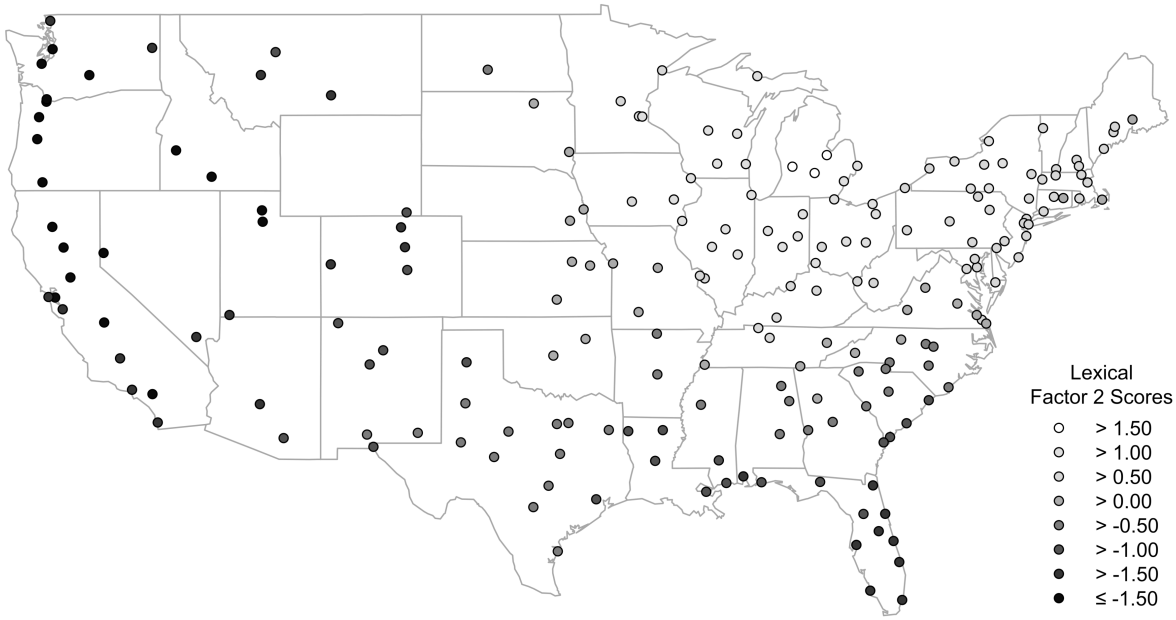
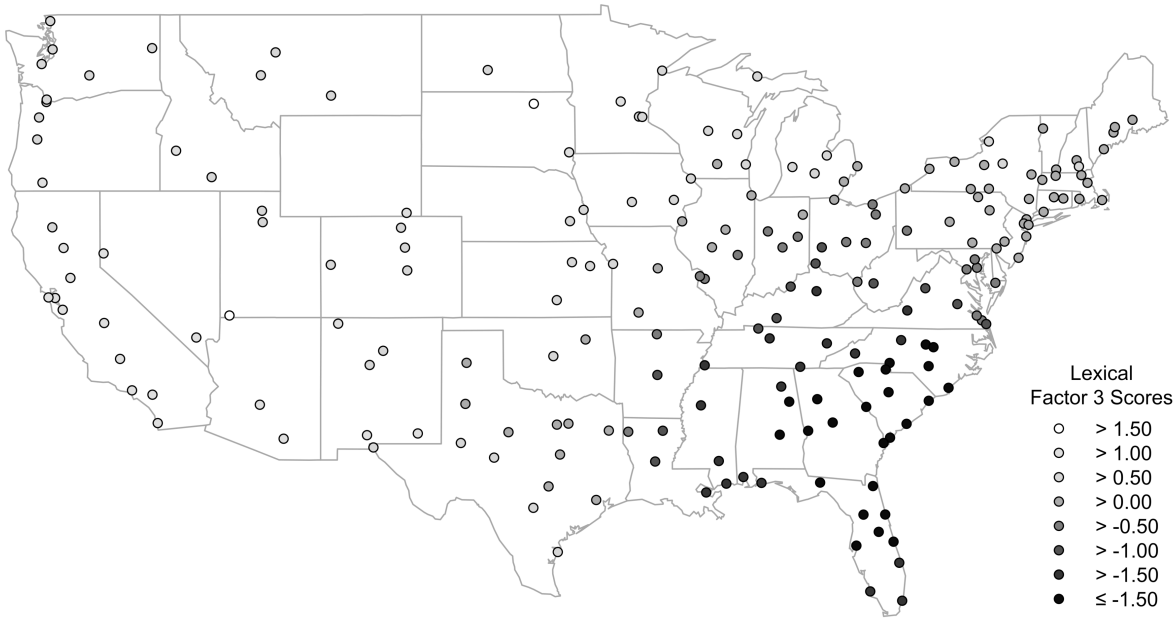


Figure 7 Lexical Factor 3



3. Map Interpolation

In order to compare the four common patterns of phonetic variation to the three common patterns of lexical variation, each factor was first interpolated over a consistent grid of reference locations using ordinary kriging, which is a geostatistical technique that estimates the values of a variable at unobserved locations based on the values of the variable at observed locations.[2] It was necessary to interpolate the factor scores before comparison because the phonetic and lexical datasets are based on two different sets of locations. However, before describing the interpolation of the seven sets of factors scores, this section introduces variogram analysis, which must be conducted before ordinary kriging is used to estimate the value of a variable at an unobserved location.

3.1 Variogram Analysis

In order to use ordinary kriging to interpolate the values of a variable it is necessary to define a theoretical variogram for that variable. This is achieved by computing an empirical variogram based on the observed values of the variable and by then fitting a theoretical variogram to the empirical variogram.

An empirical variogram is a function that describes the amount of spatial variability in the observed values of a regional variable (Isaaks and Srivastava, 1989; Wackernagel, 2010; Bachmaeir and Backes, 2008). In particular, given a variable z observed over locations $x_1,...,$ x_n, the value of the empirical variogram γ^* for a particular distance h is calculated as

$$(1) \qquad \gamma^*(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} (z(x_i + h) - z(x_i))^2$$

where h is a distance interval measured here in Euclidean distance, $N(h)$ is the total number of pairs of locations that are separated by a distance that falls within that distance interval, and $z(x_i+h)$ and $z(x_i)$ are the observed values of the variable z at the i^{th} pair of these locations. The value of an empirical variogram for a particular distance interval is therefore equal to the variance of the differences between the values of all unique pairs of locations separated by a distance that falls within that distance interval. In other words, an empirical variogram shows how the variance in the values of a variable change as the distance between locations increases.

In this study, empirical variograms were computed for the seven sets of factor scores based on 35 equally spaced distance intervals consisting of approximately 80 miles each.[3] The empirical variograms were computed for these distance intervals because 80 miles is the approximate average distance between a location and its three nearest neighbors in the two datasets. For each of these 35 distance intervals, Equation 1 was used to calculate the variance for all pairs of locations separated by a distance that falls within this distance

interval. The empirical variogram for each factor is plotted across these 35 distance intervals in Figures 8-14, as white circles in the foreground of the main graph. Note that in these variograms are based on Euclidean distance, which is marked on the lower *x*-axis; however, the approximate geographic distance in miles is also marked on the upper *x*-axis for reference.

The basic pattern exhibited by a variogram can be described by three values. The nugget is the value of the variogram at distance zero. In almost all of the variograms analyzed here, the nugget appears to be very close to zero, indicating that nearby locations tend to have very similar values. All of these variograms then increase quickly with distance until they plateau at a certain variance, which is known as the sill. For example, the variogram for Phonetic Factor 1 presented in Figure 17 has a sill of approximately 1.3. Finally, the distance at which the sill is reached is known as the range. Phonetic Factor 1, for example, has a range of approximately 10. To facilitate the estimation of these values, which can generally be identified based on the variogram at relatively small distances, a magnified version of each empirical variogram at small distances is also plotted in the inset of Figures 8-14.

Given an empirical variogram for a variable, which is defined for certain distance intervals, it is possible to define a theoretical variogram γ for that variable, which is defined for all possible distances, by fitting a function to the empirical variogram. In this case, the theoretical variograms for each factor were defined for all possible distances h by fitting a Gaussian function

$$(2) \qquad \gamma(h) = (c - c_0)(1 - \exp(-\frac{h^2}{a^2})) + c_0$$

where a represents the range, c represents the sill, and c_0 represents the nugget. The fitted Gaussian function for each factor is plotted in the inset in Figure 8-14. The Gaussian function was used because it is the common variogram models that provides the best approximation of the empirical variograms generated here. This is because all of these variograms approach the

origin parabolically, which is indicative of highly continuous regional patterns (Isaaks and Srivastava, 1989). Other functions were also tested, including the exponential function, but the fit with the empirical variogram was not as good, although varying the function had relatively little effect on the results of the ordinary kriging.

By fitting a Gaussian function to the empirical variogram it is possible to estimate the values of the three variogram parameters discussed above. For example, based on the fitted Gaussian function, the variogram for phonetic Factor 1 presented in Figure 17, was found to have a sill of 1.24, a nugget of 0, and a range of 7.60, which are similar to the manual estimates presented above. The fitted parameter values for all of the variograms are presented in Table 3. Although automatically fitting the Gaussian function to the factor scores resulted in the nugget being estimated at 0 in most cases, to avoid unstable krigings, which are possible when using a Gaussian function, a consistent but very small nugget of .01 was imposed for all factors, except for Phonetic Factor 4, for which a larger nugget was fitted automatically. In addition, the Gaussian function was only fitted to the empirical variogram for the first 13 distance intervals for Lexical Factor 2 (approximately the first 1000 miles) because the empirical variogram for this factor exhibits two separate plateaus.

In addition to plotting the empirical variograms and the theoretical variograms for each of the seven factors, variogram clouds are also plotted in the background of the main plots presented in Figures 8-14. A variogram cloud displays the dissimilarity between all pairs of locations as a function of the distance between those locations (Bivand et al, 2008; Gaetan and Guyon, 2010; Wackernagel, 2010). In particular, given a variable z observed at two locations x_i and x_j, the value of the variogram cloud γ_{cloud} for the distance between those two locations x_i-x_j is

$$\gamma_{cloud}(x_i - x_j) = \frac{(z(x_i) - z(x_j))^2}{2} \quad (3)$$

The variogram cloud therefore consists of the set of $n(n\text{-}1)/2$ points representing every unique pair of locations in the dataset.

Although a variogram cloud is not required for ordinary kriging, it is presented here for two reasons. First, a variogram cloud is essentially the basis for an empirical variogram, which is calculated for a particular distance interval by computing the average value of all pairs of locations in the variogram cloud that are contained within that distance interval, as can be seen by comparing Equation (1) and Equation (3). It is therefore useful to plot the variogram cloud so as to have a more complete picture of the underlying pattern of regional variation in the values of a variable. Second, the variogram cloud is closely related to a type of graph that has been presented in previous dialectometry studies (Séguy, 1971; Nerbonne 2009, 2010), where location-by-location linguistic distance matrices (based on sets of regional linguistic variables) are plotted against the corresponding geographic distance matrices to visualize patterns of regional linguistic variation. Aside from the fact that these plots are generally based on the values of multiple variables, these plots are very similar to a variogram cloud. The basic difference between these two approaches to visualization involves how dissimilarity is measured: in a variogram analysis dissimilarity is measured as variance, whereas in dialectometry dissimilarity is measured as Euclidean distance or based on some other distance metric. Nerbonne (2010) also fits a logarithmic curve to the plots of linguistic distance, which is very similar to the modeling of a theoretical variogram.

Finally, although the main goal of this variogram analysis was to allow for ordinary kriging to proceed, as this is the first time a variogram analysis has been conducted in dialectology, it is important to note that an empirical variogram reveals additional information about the spatial patterns exhibited by a regional linguistic variable. As noted above, an increasing slope at relatively small distances indicates that a variable is spatially patterned. As would be expected, given the appearance of the factor maps, all of the variables

analyzed here exhibit an initially increasing slope. However, the steepness of this slope also reveals how gradually the values of a variable change across space, with a steep slope being associated with more sudden changes and a moderate slope being associated with more gradual changes. For example, Phonetic Factor 3 has a relatively gradual initial slope and in turn has a relatively large number of middling factor scores separating the East from the West, whereas Phonetic Factor 4 has a steeper initial slope and in turn has sharper changes between regions. All of the variograms also start very near to the origin, reflecting the fact that adjacent locations tend to have similar factor scores, which indicates that all of the factors exhibit a continuous pattern of regional variation (Oliver, 2010).

In addition to analyzing the initial section of the variogram, there is also considerable information that can be deduced from the rest of the variogram. For example, a variogram that continues to trend upward after the initial sill has been reached indicates that the values of the variable are changing in a consistent direction across the entire map, as illustrated by the variogram for Phonetic Factor 3 (Figure 10), which when mapped (Figure 3) exhibits a simple progression in value from east to west. A wave-like pattern in a variogram, on the other hand, is indicative of a periodic pattern (Oliver, 2010). For example, the variogram for Phonetic Factor 4 exhibits a relatively wavy pattern and the map for this factor identifies numerous high and low value clusters that alternate across the map. Phonetic Factor 1 and Phonetic Factor 2 exhibit a slower wave, reflecting the fact that both of these maps are characterized by a central cluster resulting in distant locations on either coast having very similar factor scores, which is why the variance for these factors becomes smaller at larger distances. Similarly, the variogram for Lexical Factor 1 is shaped like an arch because the entire central region is contrasted with both coasts.

An empirical variogram can therefore be used to extract a great deal of information about a regional linguistic variable. The variogram, however, does not allow for the similarity

between the patterns of spatial clustering exhibited by two variables to be assessed. For example, Phonetic Factor 1 and Phonetic Factor 2 have relatively similar variograms, yet looking at the corresponding maps, it is clear that these factors exhibit quite different spatial patterns. Quantifying the similarity between the factors is the goal of the final stage of this analysis. First, however, the theoretical variograms defined here must be used to interpolate the factor maps over a grid of reference locations in order to allow for these two sets of factor scores to be compared directly.

Figure 8 Variogram for Phonetic Factor 1

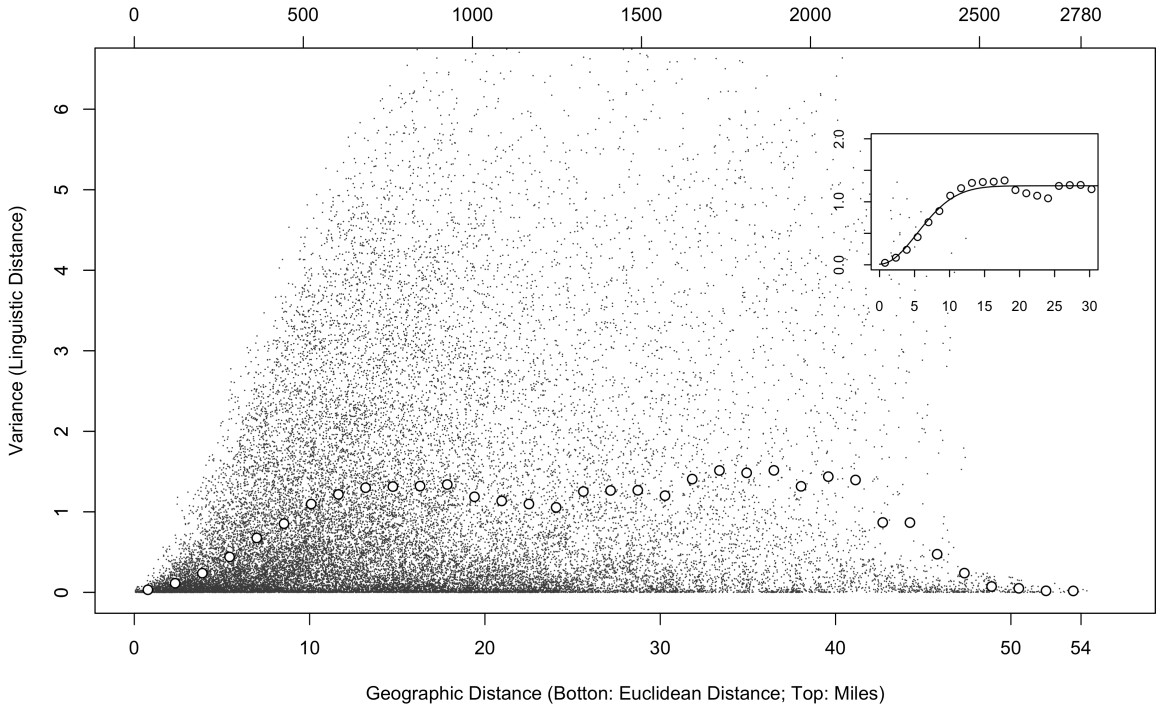


Figure 9 Variogram for Phonetic Factor 2

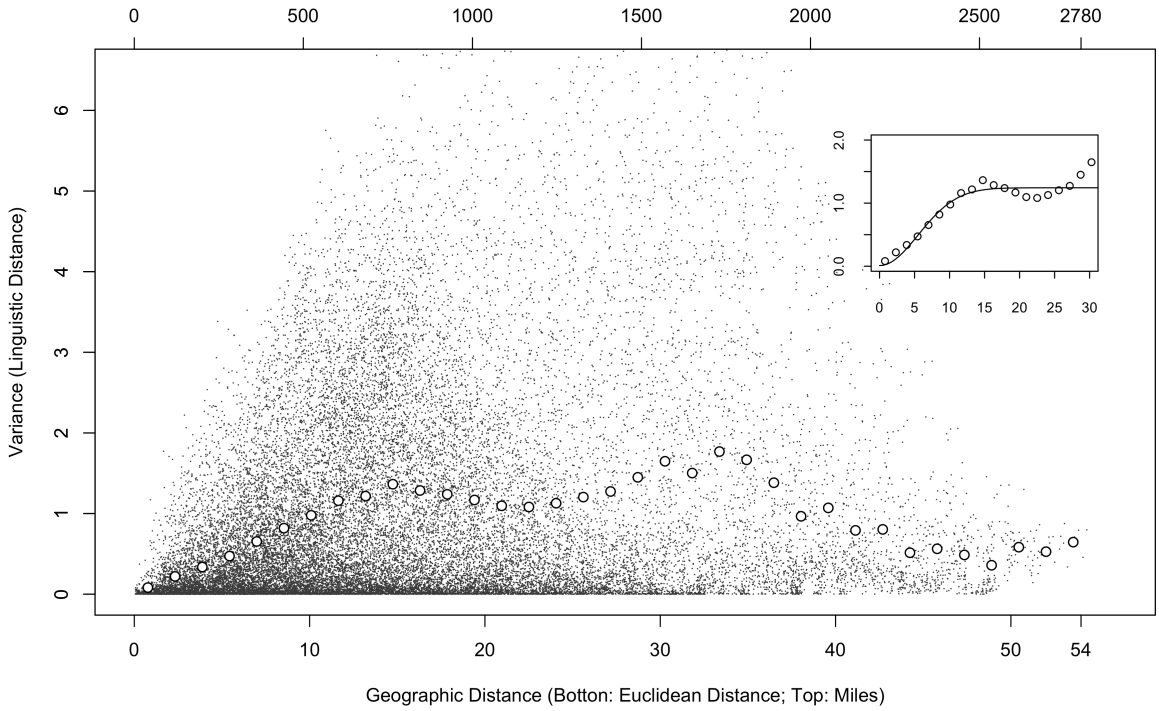


Figure 10 Variogram for Phonetic Factor 3

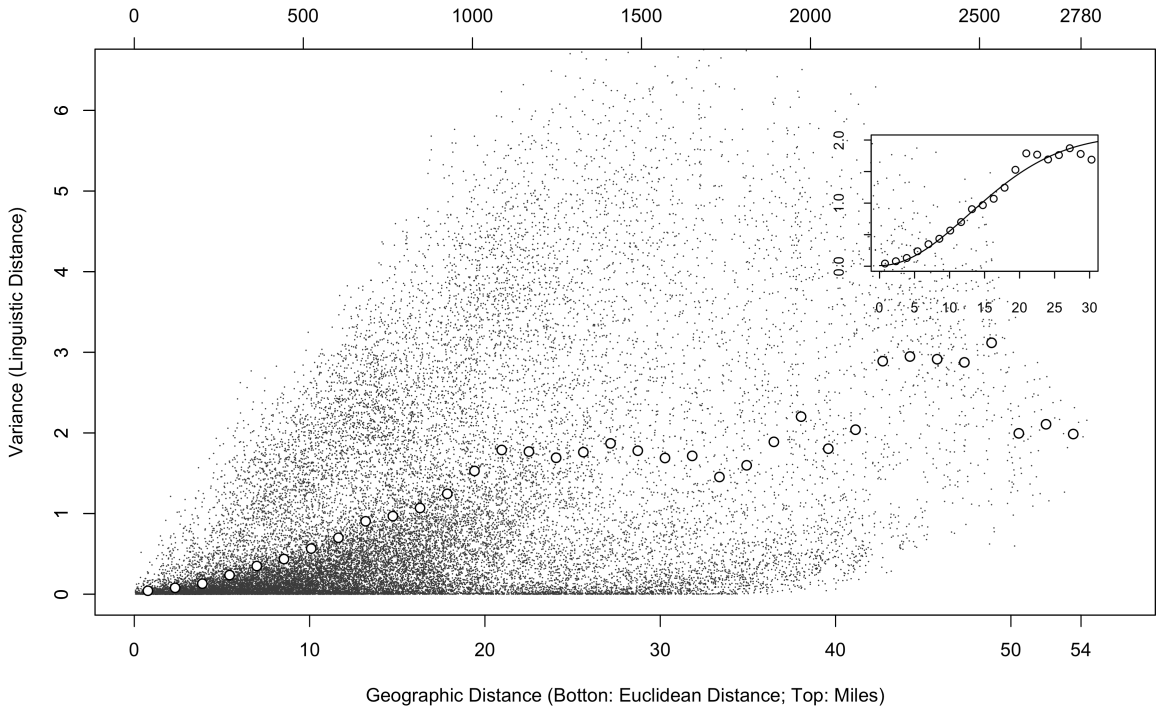


Figure 11 Variogram for Phonetic Factor 4

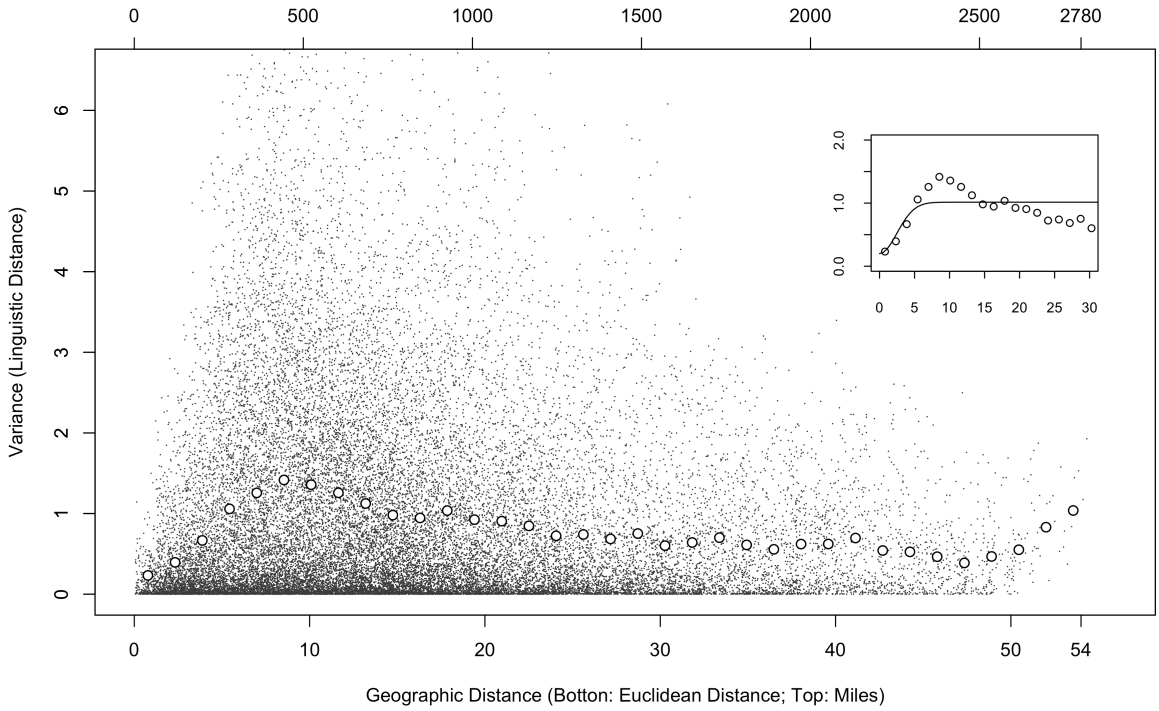


Figure 12 Variogram for Lexical Factor 1

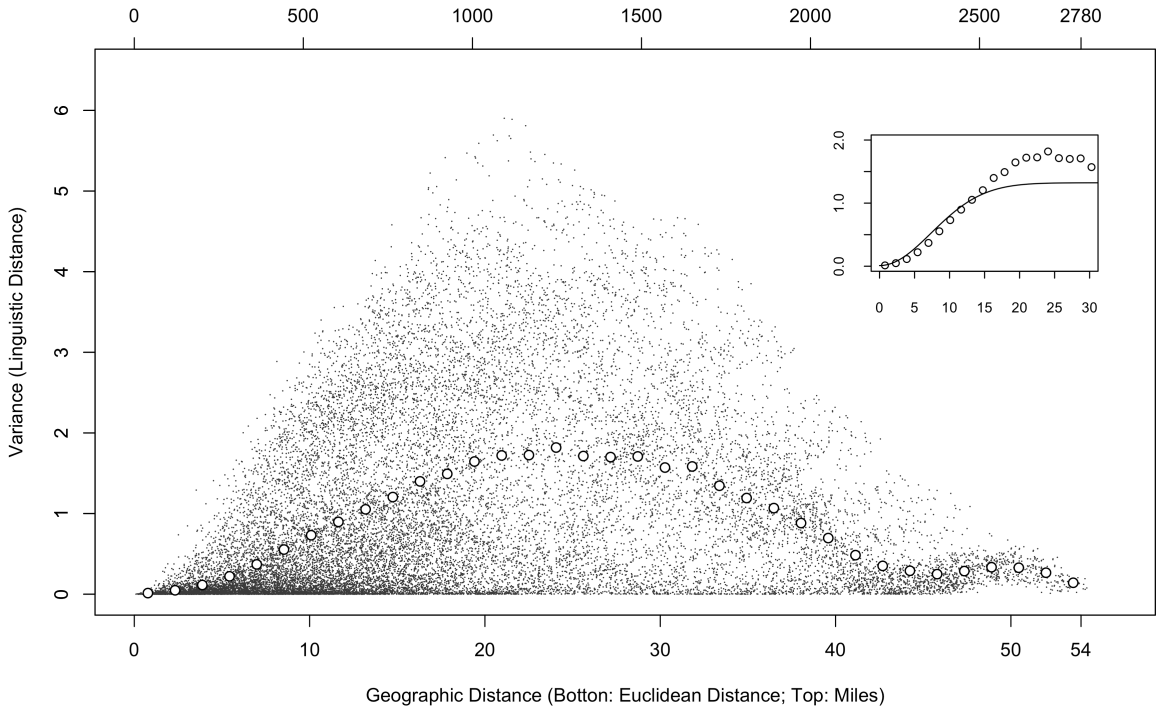


Figure 13 Variogram for Lexical Factor 2

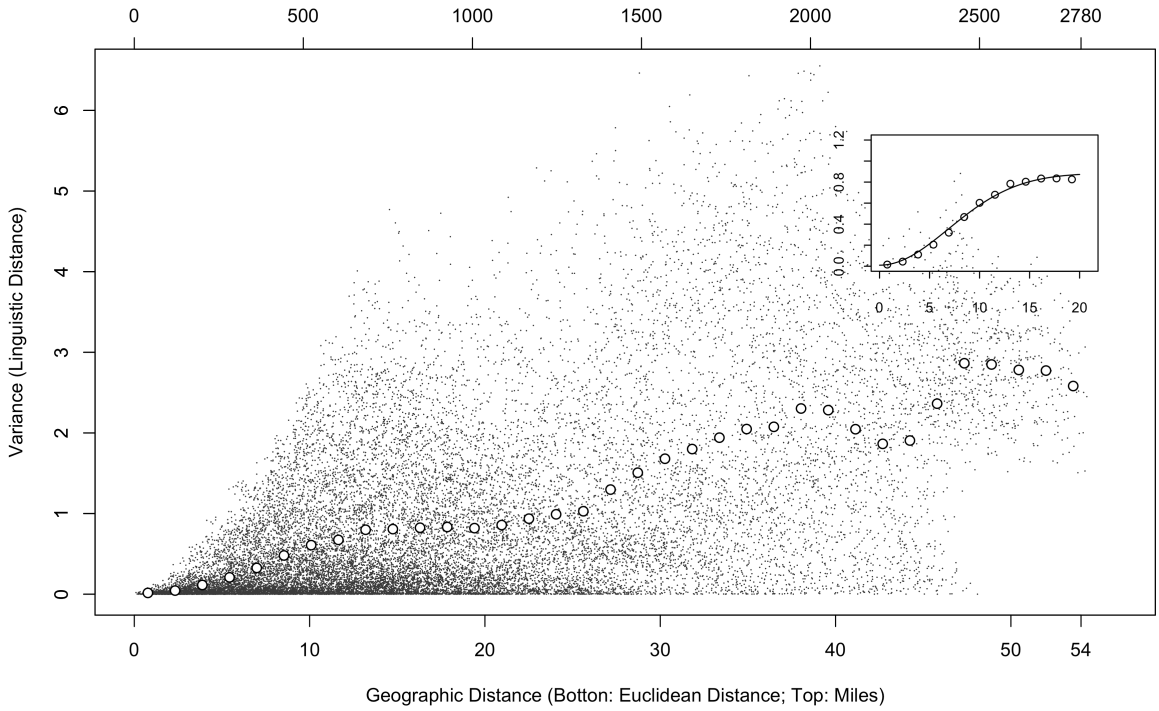


Figure 14 Variogram for Lexical Factor 3

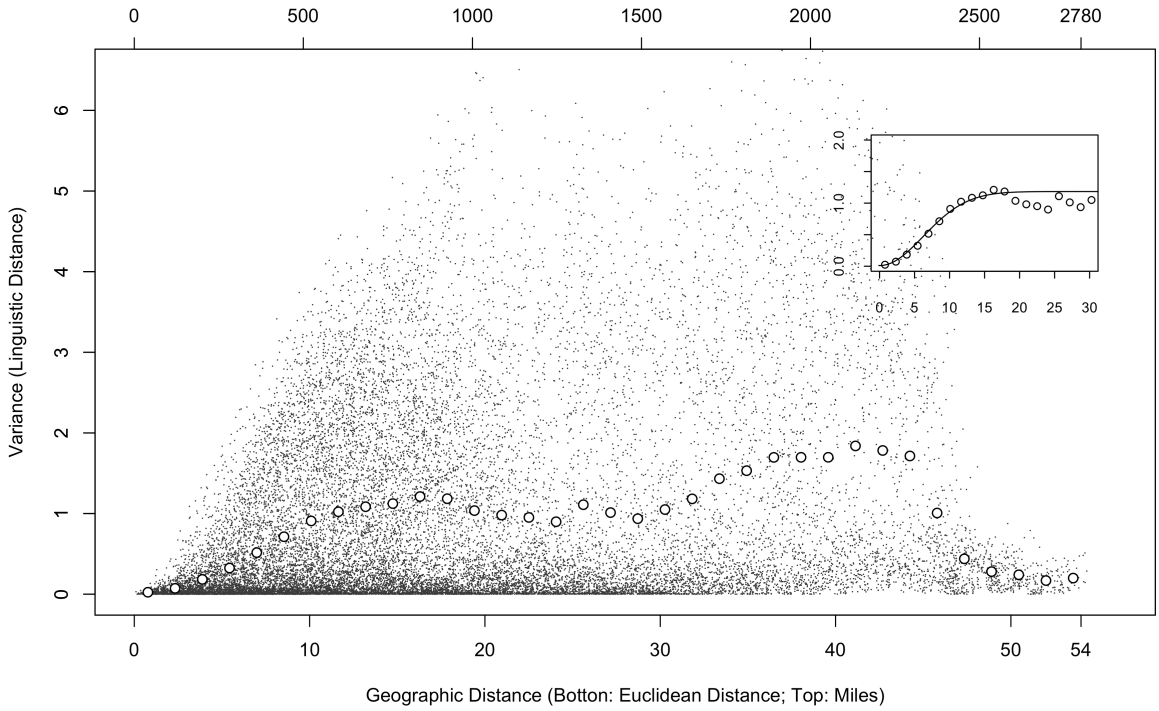


Table 3 Estimated Theoretical Variogram Parameters for fitted Gaussian Function

Variable	Nugget (c_0)	Sill (c)	Range (a)
Phonetic Factor 1	0.01	1.24	7.6
Phonetic Factor 2	0.01	1.23	7.79
Phonetic Factor 3	0.01	2.07	18.09
Phonetic Factor 4	0.2	0.81	3.49
Lexical Factor 1	0.01	1.31	10.43
Lexical Factor 2	0.01	0.88	9.73
Lexical Factor 3	0.01	1.17	8.81

3.2 Ordinary Kriging

Because the two sets of factor maps are not based on the same set of locations, each factor was interpolated over a consistent grid of reference locations before being compared. Interpolation was achieved using ordinary kriging (Isaaks and Srivastava, 1989; Bivand et al, 2008; Wackernagel, 2010), which is the most common approach to interpolation in geostatistics (Wackernagel, 2010).

Ordinary kriging is a method for estimating the values of a variable at unknown locations based on the values of the variable at observed locations. Specifically, the estimated value of the random variable Z at unobserved location x_0 is computed as an unbiased linear combination of the weighted values of the random variable Z across n observed locations

$$(4) \qquad Z^*(x_0) = \sum_{i=1}^{n} w_i Z(x_i)$$

Basically, ordinary kriging estimates the value of a variable at an unobserved location by taking a weighted average of the values of the variable at observed locations, where these weights are based on both the distance separating the locations and the theoretical variogram for that variable. If the weights were only based on the distance between locations, then the weights associated with two variables measured over the same set of locations would be identical, even if the two variables exhibited different theoretical variograms, which is not ideal. For example, if the variogram for the first variable increases more quickly at small distances than the variogram for the second variable, then the weights for the first variable should be stronger at small distances. By basing the interpolation of a variable on its theoretical variogram, which provides a model of how the values of a variable change across space, ordinary kriging allows for the value of a variable to be estimated at an unknown location with greater accuracy than would be possible if only the distance between locations was taken into consideration.

The first stage of ordinary kriging is therefore to compute a theoretical variogram for the variable being kriged. In this case, theoretical variograms were obtained by fitting a Gaussian function to the empirical variograms for each factor, as described in Section 3.1. A theoretical variogram is required for kriging rather than the empirical variogram upon which the theoretical variogram is based because in order to compute the variable weights it is necessary to calculate the value of the variogram for the distances between the unobserved location and each of the observed locations, which may not be instantiated in the empirical variogram, which is only defined for the distances between the observed locations.

In particular, given a theoretical variogram, the kriging weights are computed by solving the ordinary kriging equation system (Wackernagel, 2010)

$$(5)\qquad \begin{pmatrix} \gamma(x_1 - x_1) & \gamma(x_1 - x_2) & \dots & \gamma(x_1 - x_n) & 1 \\ \gamma(x_2 - x_1) & \gamma(x_2 - x_2) & \dots & \gamma(x_2 - x_n) & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \gamma(x_n - x_1) & \gamma(x_n - x_2) & \dots & \gamma(x_n - x_n) & 1 \\ 1 & 1 & \dots & 1 & 0 \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \\ \mu \end{pmatrix} = \begin{pmatrix} \gamma(x_1 - x_0) \\ \gamma(x_2 - x_0) \\ \vdots \\ \gamma(x_n - x_0) \\ 1 \end{pmatrix}$$

which, through matrix multiplication, can alternatively be expressed as

$$(6)\qquad \begin{cases} \sum_{j=1}^{n} w_j \gamma(x_i - x_j) + \mu = \gamma(x_i - x_0) \qquad for\ i = 1, \dots, n \\ \sum_{j=1}^{n} w_j = 1 \end{cases}$$

where $\gamma(x_i - x_j)$ is the value of the theoretical variogram for the distance separating the observed locations x_i and x_j, $\gamma(x_i - x_0)$ is the value of the theoretical variogram for the distance separating the observed location x_i and unobserved location x_0, and μ is the Lagrange parameter.

The first n equations in the ordinary kriging system equate the value of the theoretical variogram for the distance between the unobserved location and one observed location to a linear combination of the weighted values of the theoretical variogram for the distances between that observed location and every other observed location. The final equation in the ordinary kriging system requires that the kriging weights sum to 1 in order to minimize the mean estimation error, thereby fulfilling what is known as the *unbiasedness condition*. In addition, because this equation system contains one more equation than unknowns, the Lagrange parameter is introduced into the system in order to convert what would be a constrained minimization problem into an unconstrained minimization problem. This allows the ordinary kriging system to be solved so that the kriging weights can be obtained. Interpolation can then proceed using Equation (4) by summing the weighted observed values of the variable at each location. This process can be repeated across a series of unobserved locations to interpolate the value of a variable across an entire region.

In this case, the 7 factors were kriged over a grid of 207 centrally-symmetric reference locations evenly spread across the contiguous United States. The analysis focused on a 207 location grid because, of all possible centrally-symmetric grids of the contiguous United States, 207 locations is most similar to the number of locations in the original maps, although grids of 97, 290, 408 and 506 locations were also generated so as to allow for the effect of varying this parameter on the results of the comparisons to be evaluated. Because the grid is regular, in certain regions such as the Northeast the sampling density is less dense than in the original maps, while in other regions such as the North Central States the sampling density is denser than in the original maps. The comparisons will therefore be insensitive to fine-grained patterns that may exist in the more densely sampled regions. Normalizing the sampling density, however, compensates for the inconsistent density of the original maps, allowing for the comparison to be evenly weighted across the region under analysis.

The seven factors interpolated over the 207 reference locations are plotted in Figures 15 to 21, where each horizontal line of locations in a grid has the same longitude and each vertical line of locations has the same latitude. Comparing these maps to the original maps, it is clear that ordinary kriging has produced maps that are representative of the basic patterns exhibited by the original maps. These kriged factors, which are defined for the same set of locations, can now be directly compared in the next stage of the analysis in order to asses the similarity between common patterns of regional phonetic and lexical variation in American English.

Figure 15 Kriged Phonetic Factor 1 (207 Locations)

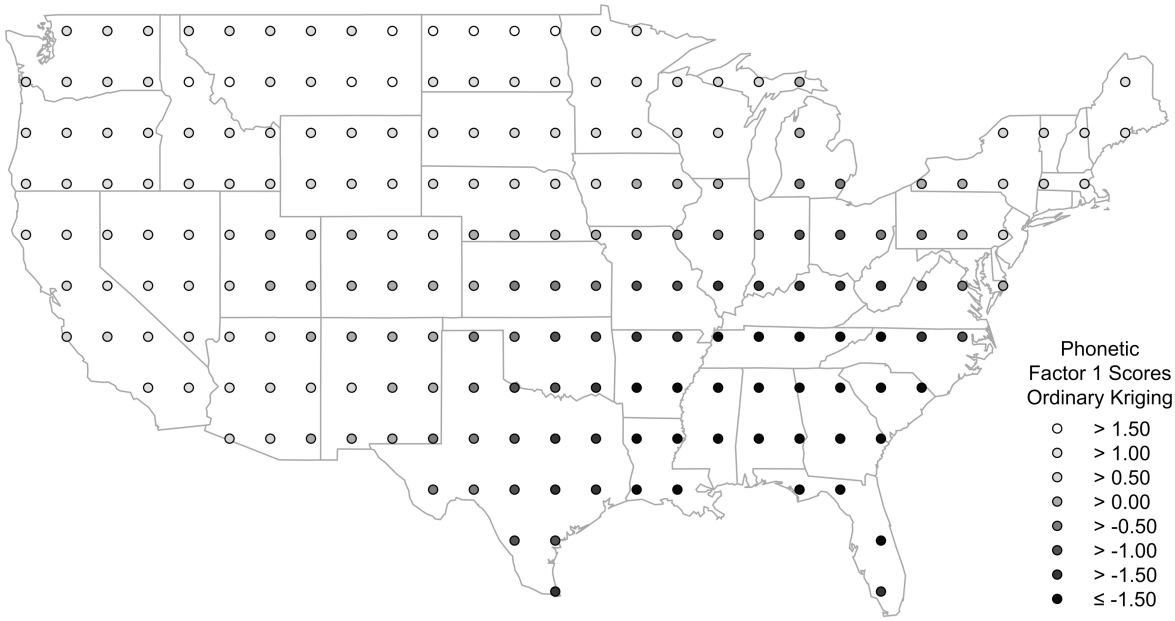


Figure 16 Kriged Phonetic Factor 2 (207 Locations)

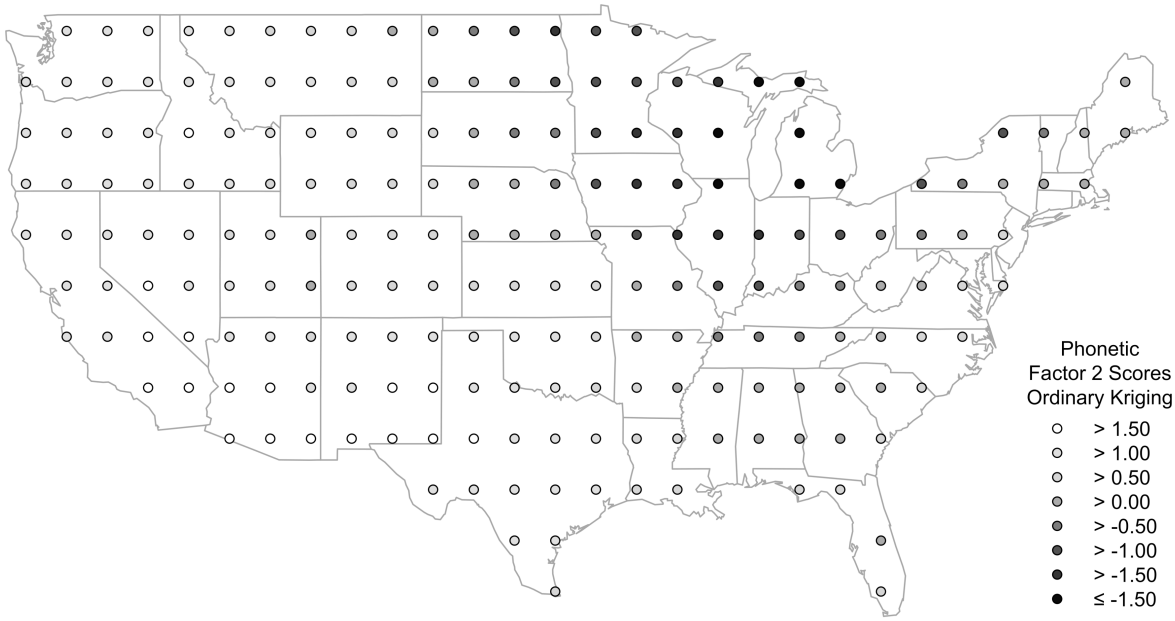


Figure 17 Kriged Phonetic Factor 3 (207 Locations)

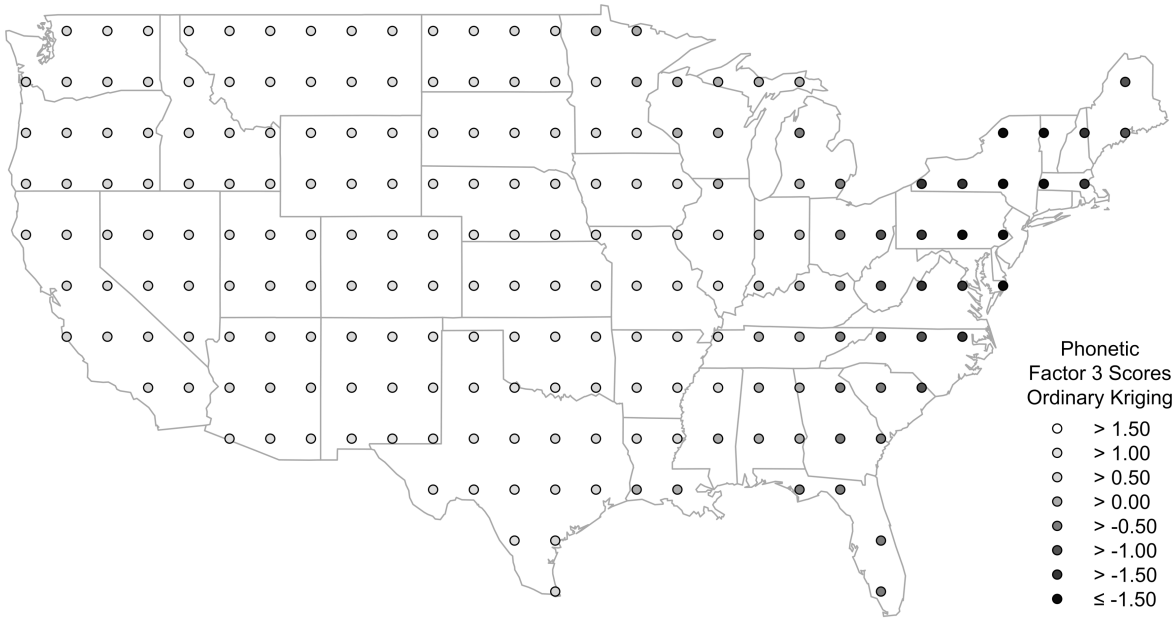


Figure 18 Kriged Phonetic Factor 4 (207 Locations)

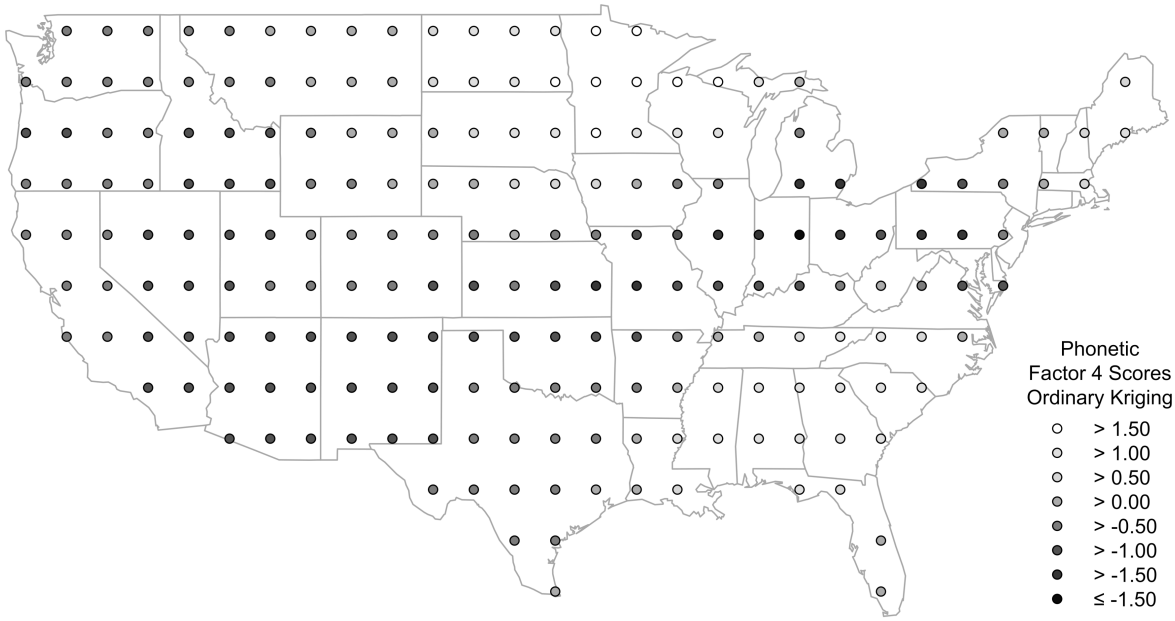


Figure 19 Kriged Lexical Factor 1 (207 Locations)

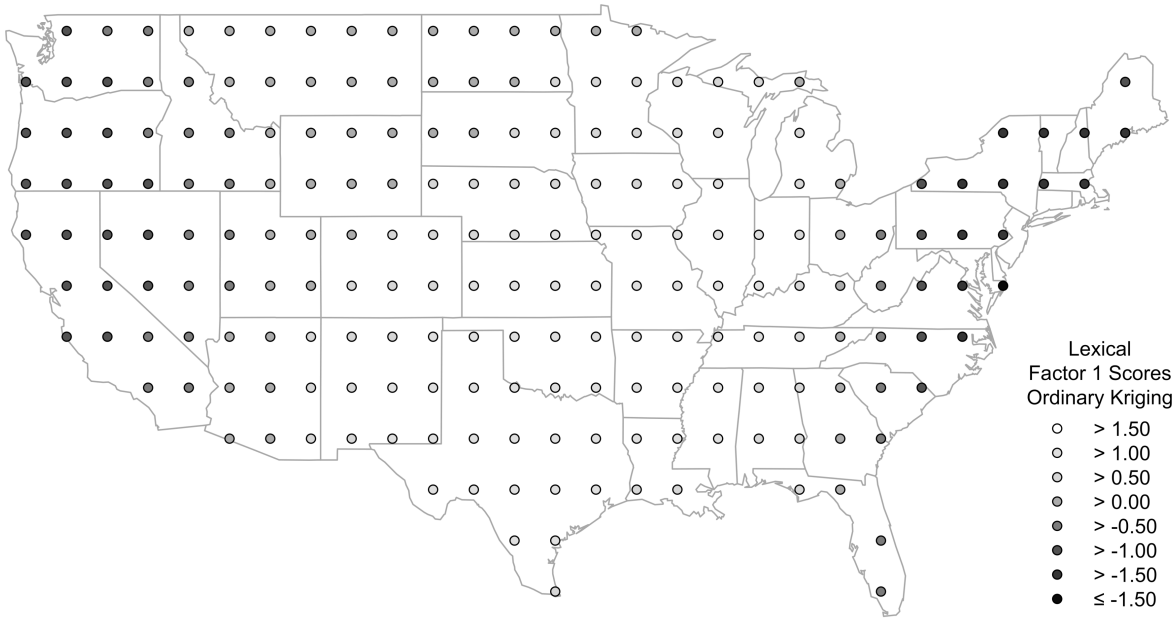


Figure 20 Kriged Lexical Factor 2 (207 Locations)

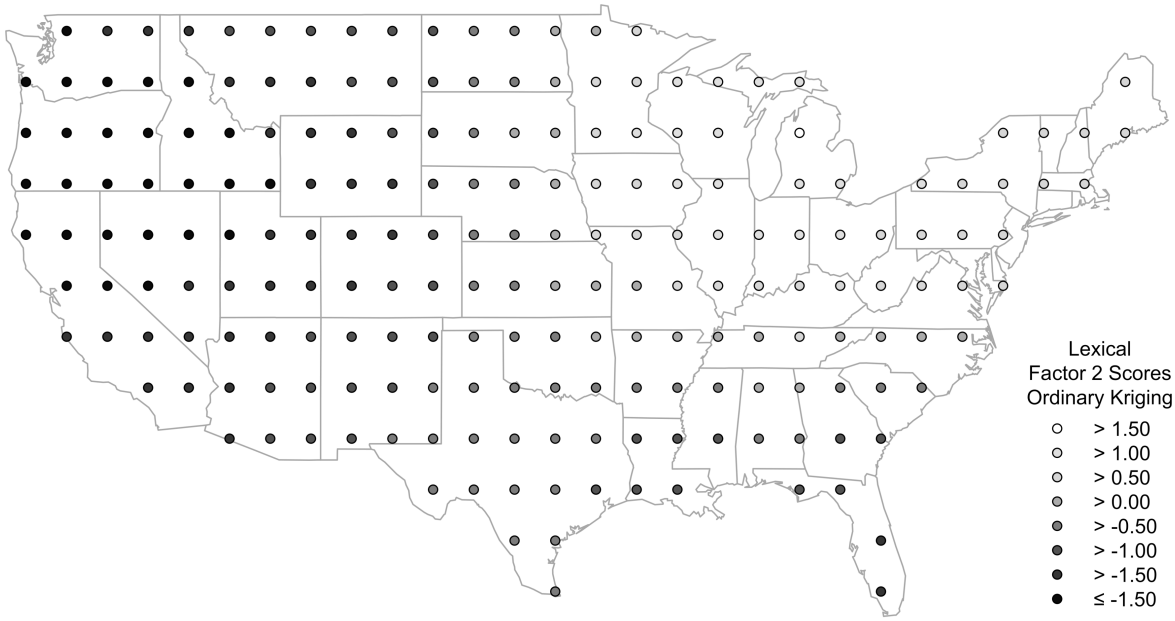
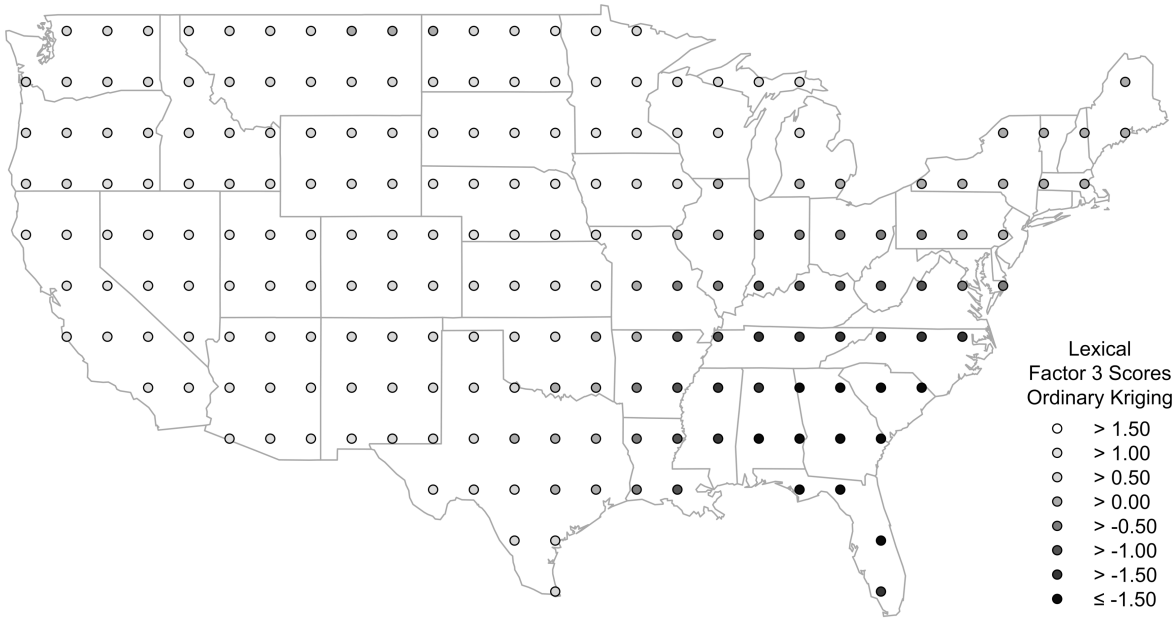


Figure 21 Kriged Lexical Factor 3 (207 Locations)



4. Map Comparison

As discussed above, the two sets of dialect maps being compared in this study represent common patterns of regional phonetic (Figures 1-4) and lexical (Figures 5-7) variation in Modern American English, as identified in two previous studies. Although it is relatively clear that these maps do align to a certain extent, the goal of this study is to quantitatively compare these patterns in order to assess the similarity of regional variation across these two linguistic levels. However, because the two sets of dialect maps were based on different sets of locations, before being compared these maps were first interpolated over a consistent set of reference locations using ordinary kriging (Figures 15-21), as described in Section 3. In this section, the comparison of these kriged phonetic and lexical factor maps is presented. The analysis is based primarily on the factor maps interpolated over the 207 location grid, as this

grid is closest to the number of locations in the original datasets, although the effect of varying grid density is briefly discussed as well.

The interpolated phonetic and lexical factor scores were first plotted against each other across the 207 locations in order to visualize the relationship between these two sets of factor scores. The scatter plot matrix is presented in Figure 22, which displays a scatter plot for every pair of factors. Of particular interest are the scatter plots in the lower left (or the upper right) quarter of the matrix, which plot the 3 lexical factors against the 4 phonetic factors. Overall, these scatter plots identify clear linear relationships between two pairs of kriged phonetic and lexical factor scores, indicating that these pairs of factors identify similar regional patterns. In particular, there is a clear linear relationship between Phonetic Factor 1 and Lexical Factor 3, which contrast the southeastern United States with the rest of the United States, and between Phonetic Factor 2 and Lexical Factor 2, which contrast the Northeast and especially the Midwest with the rest of the United States, indicating that these common patterns of phonetic and lexical regional variation in American English are closely related.[4] In addition, there is a relatively clear linear relationship between Phonetic Factor 3 and Lexical Factor 1, which both contrast the East Coast of the United States with the Central United States, indicating that these factors also exhibit similar regional patterns. The linear relationship, however, between this pair of factors is not as strong as the relationship between the other two pairs of factors because Lexical Factor 1 also clusters the West Coast with the Eastern United States. Finally, Phonetic Factor 4, which exhibits a Midland pattern, does not appear to align with any of the lexical factors. It is also notable that while the other factor pairs are not linearly related, none of these scatter plots exhibit random patterns. This is because all of these comparisons involve factors that represent clear patterns of spatial clustering.

Next, the strengths of the linear correlations between the interpolated phonetic and lexical factors scores were measured by calculating the Pearson correlation coefficients for each pair of factors. The Pearson (r) correlation coefficients for the comparisons between the 4 phonetic factors and the 3 lexical factors are presented in Table 3, along with the amount of variance explained by each linear correlation (r^2). Of the 12 comparisons, the two pairs of factor maps discussed above were identified as exhibiting very strong linear correlations. First, a very strong linear correlation (r = .83) was identified between Phonetic Factor 1 and Lexical Factor 3, accounting for 68% of the variance in these two sets of factor scores. Both of these factors identify a southeastern region. Second, a very strong linear correlation (r = -.81) was identified between Phonetic Factor 2 and Lexical Factor 2, accounting for 66% of the variance in these two sets of factor scores. Both of these factors identify a northeastern region. A strong moderate linear correlation (r = .60) was also identified between Phonetic Factor 3 and Lexical Factor 1, accounting for 36% of the variance in these two sets of factor scores. Both of these factors contrast the East Coast with the Central United States; however, this correlation is weaker because the two factors differ in terms of the status of the West Coast, which is clustered with the East Coast on Lexical Factor 1. In addition, Phonetic Factor 3 was moderately correlated with both Lexical Factor 2 (r = -.49) and Lexical Factor 3 (r = .5), which reflects the fact that Lexical Factor 2 and Lexical Factor 3 also largely contrast the East with the West. On the other hand, Phonetic Factor 4 was not found to correlate even moderately with any of the lexical factors. The correlation analysis was also repeated for the other reference grids. For all analyses, the strength of the linear correlations remained relatively stable, even when as few as 98 or as many as 506 reference locations were used for kriging.

In addition to measuring the strength of the correlation between each pair of phonetic and lexical factors, the four phonetic factors and the three lexical factors were compared to

each other simultaneously to obtain an overall measure of similarity. This was achieved by computing separate phonetic and lexical distance matrices based on the two sets of kriged factor scores, which consist of the Euclidean distance between every (207 x 207 =) 42,849 pairs of locations calculated based on the values of the four phonetic or three lexical kriged factors. The resulting phonetic and lexical distance matrices were then correlated using the Mantel test (Mantel, 1967), following the method for comparison presented in Spruit et al (2009)[5]. A Mantel test was used to correlate the distance matrices, rather than a simple Pearson correlation coefficient, to correct for the fact that the linguistic distances in the matrices are not independent of each other.

The Mantel test identified a strong correlation (r = .73) between the two distance matrices, accounting for 54% of the variance in these two distance matrices, indicating that overall the common patterns of phonetic and lexical patterns in American English identified in two previous studies are strongly correlated. This relationship is also visualized in Figure 23, where the two distance matrices are plotted against each other (i.e. the phonetic and lexical distances for every unique pair of locations). This scatter plot shows a clear linear relationship between the two distance matrices. In addition, when the Mantel test was calculated for a phonetic distance matrix based only on the first three factors, the strength of the correlation rose slightly (r = .77), accounting for 60% of the variation in these two distance matrices.

Figure 22 Phonetic and Lexical Scatter Plots (207 Locations)

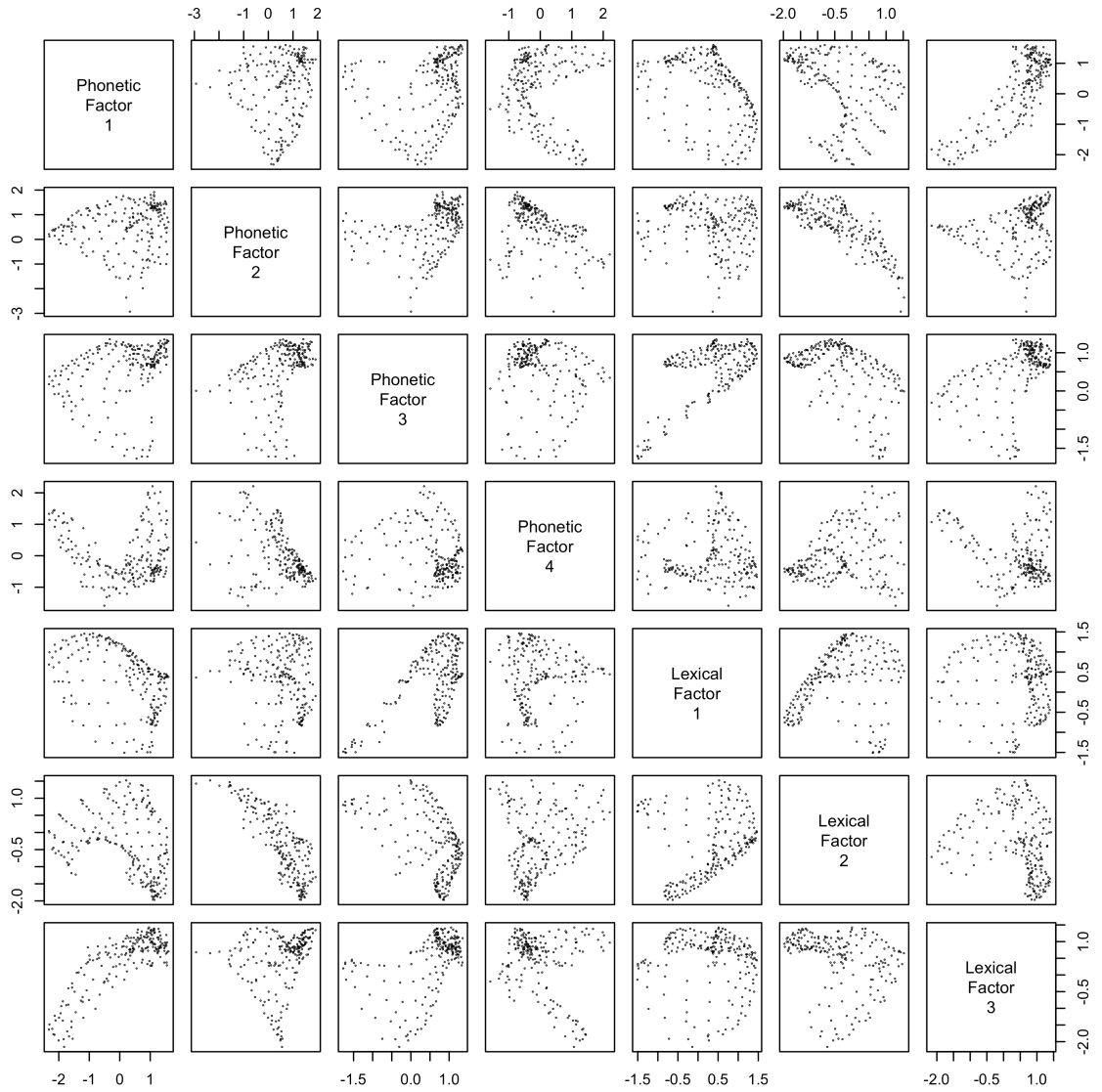
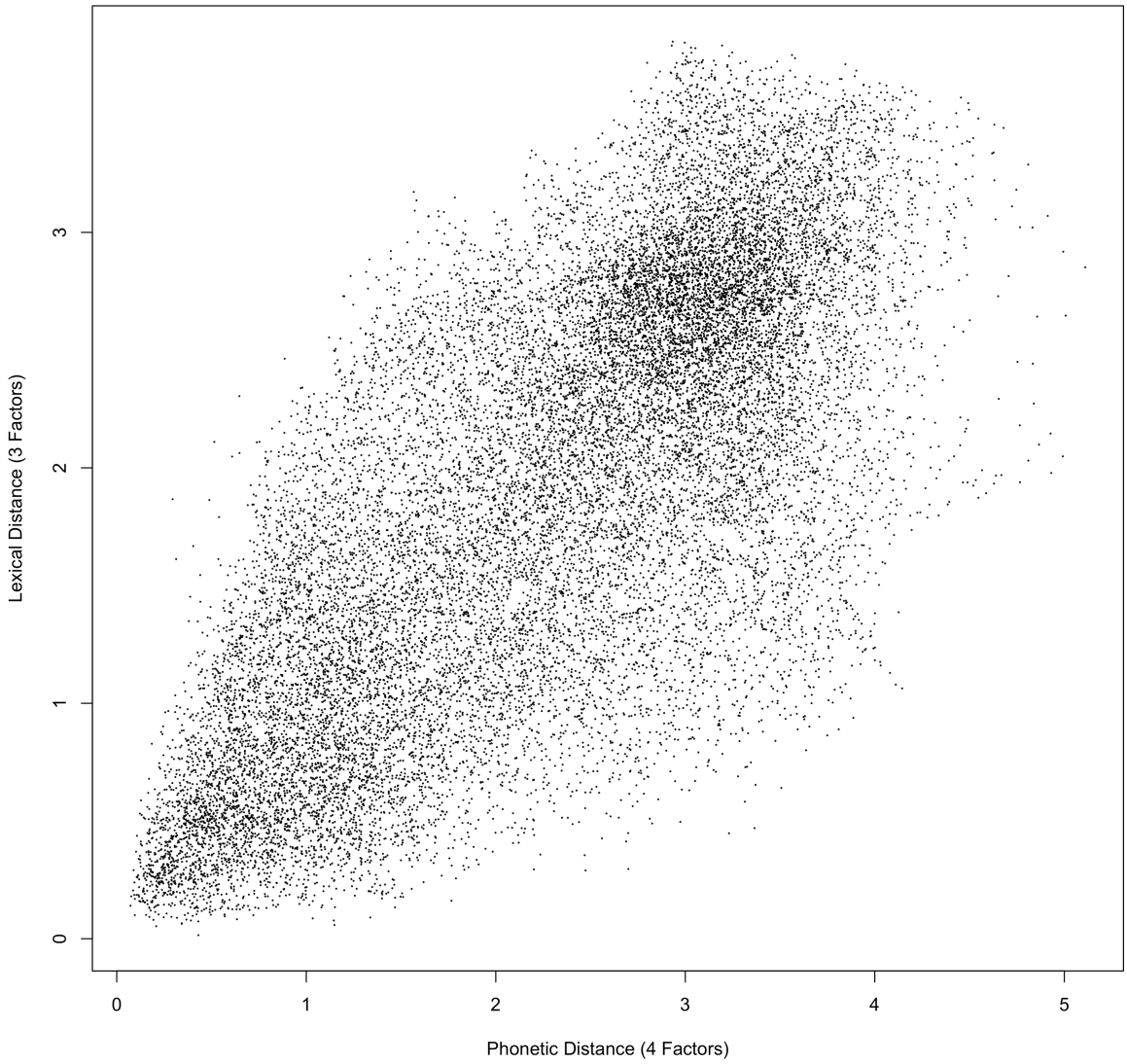


Table 3 Pearson Correlations: Phonetic and Lexical Factor Scores (207 Locations)

		Lexical		
Phonetic		Factor 1	Factor 2	Factor 3
Factor 1	*r*	-0.41	-0.32	0.83
	r^2	16.90%	10.20%	68.30%
Factor 2	*r*	-0.14	-0.81	0.21
	r^2	1.90%	65.50%	4.30%
Factor 3	*r*	0.6	-0.49	0.5
	r^2	36.40%	23.50%	24.60%
Factor 4	*r*	0.01	0.25	-0.27
	r^2	0.00%	6.10%	7.40%

Figure 23 Lexical Distance (3 Factors) vs Phonetic Distance (4 Factors)



5. Discussion

The goal of this study was to compare the common patterns of regional phonetic and lexical variation in American English identified in two previous studies (see Grieve et al, 2011, forthcoming). In these previous studies, common patterns of regional variation were identified through multivariate spatial analyses of a phonetic dataset, based on 38 acoustic vowel formant variables measured over 236 American cities from *The Atlas of North American English* (Labov et al, 2006), and a lexical dataset, based on 40 word alternation variables measured over 206 American cities in a corpus of letters to the editor (Grieve et al, 2011). The statistical analyses identified four common patterns of phonetic variation (Figure 1-4) and three common patterns of lexical variation (Figures 5-7). In this study, these two sets of factor maps were compared in order to gauge the similarity between these common patterns of phonetic and lexical variation. However, because these two sets of factor maps are based on two different sets of locations, the factor scores were first interpolated over a consistent grid of reference locations using ordinary kriging, which requires that each factor be subjected to a variogram analysis (Figures 8-14). The kriged factor maps (Figures 15-21) were then plotted against each other (Figure 22) and correlated (Table 3).

This analysis identified two strong correlations across the two linguistic levels: both phonetic and lexical variation are characterized by similar southeastern clusters (Phonetic Factor 1 and Lexical Factor 3) and similar northeastern clusters (Phonetic Factor 2 and Lexical Factor 3) in Modern American English. In addition, the two sets of kriged factors were used to generate two distance matrices that were then correlated to each other using a Mantel test in order to assess the overall similarity between the two sets of maps. This analysis identified a strong overall correlation between lexical and phonetic variation (Figure 23).

These results offer evidence that there is a strong correlation between regional phonetic and lexical variation in American English. The patterns are not identical, but it is clear that the main patterns of phonetic variation in particular are also attested in the lexical data and that overall the two sets of factor maps show a strong degree of correlation. The major division between the Northeast and the Southeast, and then by extension the West, was clearly present in both datasets. Furthermore, in both datasets a third factor map (i.e. Phonetic Factor 3, Lexical Factor 1) also split the East Coast from the Central United States. The correlation between Phonetic Factor 3 and Lexical Factor 3, however, is of only moderate strength because the status of the West Coast differs across these two maps, with only the lexical factor grouping the two coasts together (see discussion below). Nonetheless, it is clear that the two sets of factor maps exhibit similar patterns of spatial clustering.

Despite the overall degree of similarity between the phonetic and lexical factor scores, there is still considerable variation that goes unexplained, especially considering that the factor maps themselves only account for a proportion of the spatial variation in the original datasets. Furthermore, it is important to acknowledge that the original datasets are not generally representative of phonetic or lexical variation in American English. The data under analysis here, however, is some of the only dialect data available on regional phonetic and lexical variation in Modern American English, and although the correlations are not perfect, the overall similarity between the two datasets is strong. Further research is certainly required, but the results of this study show that regional phonetic and lexical variation are similar in Modern American English.

Given that similarities exist between the two sets of factors scores, it is important to consider explanations for these alignments. Although the variables loading on the individual factors generally have linguistic explanations, there are no clear linguistic explanations for the specific pairs of phonetic and lexical factors identified in this study. In particular, while

the phonetic factors primarily identify chain shifts and the lexical factors primarily identify functional-grammatical patterns, aside from the Northeast appearing to be the most linguistically conservative region of the United States in terms of both phonetics and lexis, there are no obvious connections between the phonetic and lexical variables loading on the paired factors. For example, the Northern Cities Shift (Phonetic Factor 2) and the frequent use of standard American vocabulary items (Lexical Factor 2), which characterize the language of the Midwest, do not seem to be linguistically related. It is possible that with more data, especially with more lexical data, clear connections across linguistic levels could be established, but given the current data, no straightforward explanations for the linguistic variables that pattern similarly across linguistic levels is apparent.

There is no reason, however, to require that linguistic explanations be provided for these alignments. It is entirely possible that extra-linguistic determinants of regional variation are solely responsible for the similarity in regional patterns across linguistic levels and that ultimately the association between the specific linguistic and regional patterns is arbitrary. In traditional dialect studies, for example, variables whose isoglosses bundle together often show no linguistic similarity at all. Extra-linguistic explanations for patterns of regional linguistic variation are therefore usually considered sufficient in American dialectology and the results obtained here are readily explainable in these terms. In particular, given the consistent division between the Northeast, the Southeast, the Midwest and the West in both datasets, it appears that modern patterns of regional linguistic variation in American English follow basic American cultural patterns across linguistic levels (see Grieve, 2009).

Aside from the identification of a Midland region by Phonetic Factor 4, the only major difference between the two datasets involves the status of the West Coast, which clusters with the East coast on Lexical Factor 1, but clusters apart from the East Coast on Phonetic Factor 3. There are several possible explanations for this difference. It may be that

regional phonetic and lexical variation simply differ in this way or it may be that register or demographic differences between the two datasets are responsible for this misalignment. It is also possible, however, that this difference represents a change in progress—an ongoing linguistic convergence between the East and West Coasts. This is a potential explanation because the phonetic dataset represents American English of the 1990s, whereas the lexical dataset represents American English of the 2000s. This hypothesis is also consistent with the finding that that dialect regions align with contemporary cultural patterns because the East and West Coasts have been converging politically and culturally over recent decades. More research, however, is clearly needed to test this hypothesis.

It is also important to note that the strong correlations between the phonetic and lexical factors are not only evidence of the similarity of regional variation across these two linguistic levels, but of the similarity of regional linguistic variation in speech and writing. Similarly, because these two datasets were compiled using different approaches to data collection, with the phonetic dataset being based on recordings of linguistic interviews and the lexical dataset being based on a corpus of written language, these results also suggest that both of these approaches to data collection are valid and that corpora of written data, which are far easier to obtain than spoken interview data, can be the basis of dialect studies, not just of written registers, but of language in general.

This paper has also introduced a new method for statistically comparing dialect maps based on different sets of locations. In this case, the method was used to compare aggregated factors extracted by a multivariate spatial analysis, but this method can also be used for comparing other types of dialect maps. For example, the method could be used to compare the similarity between two maps of the same linguistic variable as measured by two different dialect surveys or at two different times. Describing this method also required that two basic concepts in geostatistics be introduced to dialectology: variogram analysis and ordinary

kriging. The introduction of variogram analysis is particularly important because it is one the basic descriptive techniques in geostatistics, which allows for the patterns of spatial dependency exhibited by a regional variable to be better understood. The variogram analysis can also be extended in numerous ways, for example by modeling directional variograms, which can be incorporated into the kriging procedure to obtain more accurate interpolations. Such topics deserve additional research within dialectology. Ordinary kriging can also be used for other purposes, such as for predicting the value of a variable at a particular location of interest or for interpolating the value of a variable across a region of interest at a very fine level of detail. Because the general method for map comparison introduced here involves the analysis of variograms and the application of ordinary kriging, a secondary contribution of this paper is therefore the introduction of these two fundamental geostatistical techniques to the field of dialectology.

End Notes

1 Ordinal alternation was counted in sentence initial position before commas. *Though/although* was not counted following *as* or before commas or periods. *If/whether* was counted following forms of the verbs *wonder, care, question, determine, see, consider, ask, know, debate, tell,* and *decide*. *About/around* was counted before numbers. *About/on* was counted following forms of the nouns *research, comment, article, impact, letter, report, information, story, debate, opinion, column, view, editorial*, and *book*. *To/toward(s)* was counted following forms of the nouns *contribution, gratitude, threat, respect, responsibility, commitment, devotion, donation,* and *courtesy*. *Whom/who* was counted following prepositions. Nonrestrictive *which/that* was counted following commas preceded by nouns. Restrictive *which/that* was counted following nouns. *Who/that* was counted both following personal nouns and compound pronouns. *Be going to/will* was counted before verbs.

2 All of the analyses described in this section were conducted in R, specifically using functions from the *maptools*, *sp* and *geoR* packages (see Bivand et al, 2008)

3 Although none of the factors are normally distributed, and generating a variogram based on a highly skewed variable can be problematic, none of the factors exhibit a skewness coefficient greater than +1 or smaller than -1, and so factor scores were subjected directly to a variogram analysis (Oliver, 2010).

4 The fact that the relationship between Phonetic Factor 1 and Lexical Factor 3 is positive, whereas the relationship between Phonetic Factor 2 and Lexical Factor 2 is negative is irrelevant as it depends on which variants of the lexical alternation variables the proportions were calculated for, which was essentially an arbitrary decision.

5 Spruit et al (2009) compared regional variation in Dutch across three linguistic levels based on distance matrices generated for pronunciation, lexical and grammatical data; however,

unlike this study, comparisons were made by analyzing the locations that were present in both datasets rather than through interpolation.

References

Atwood, E. B. (1953). *A Survey of Verb Forms in the Eastern United States*. Ann Arbor: University of Michigan Press.

Bachmaier, M. and Backes M. (2008). Variogram or semivariogram? Understandin.g the variances in a variogram. *Precision Agriculture,* 9: 173-175.

Bivand R. S., Pebesma E. J. and Gomez-Rubio V. (2008). *Applied Spatial Data Analysis with R*. New York: Springer.

Carver, C. M. (1987). *American Regional Dialects*. Ann Arbor, Michigan: University of Michigan Press.

Cassidy, F. G. (1985.) *Dictionary of American Regional English*. Cambridge, Massachusetts: Harvard University Press.

Gaetan, C. and Guyon, X. (2010). *Spatial Statistics and Modeling*. Berlin: Springer-Verlag.

Grieve, J. (2009). *A Corpus-Based Regional Dialect Survey of Grammatical Variation in Written Standard American English*. Ph.D. Dissertation, Northern Arizona University

Grieve, J. (2011). A regional analysis of contraction rate in written Standard American English. *International Journal of Corpus Linguistics,* 16: 514-546.

Grieve, J. (2012). A statistical analysis of regional variation in adverb position in a corpus of written Standard American English. *Corpus Linguistics and Linguistic Theory,* 8: 39-72.

Grieve, J. (Forthcoming). A comparison of statistical methods for the aggregation of regional linguistic variation. In Szmrecsanyi, B. and Walchli, B. (eds), *Aggregating Dialectology and Typology: Linguistic Variation in Text and Speech, Within and Across Languages*. Berlin: Walter de Gruyter.

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

Grieve J, Speelman D, Geeraerts D. (Submitted). A multivariate spatial analysis of vowel formants in American English. Submitted to the *Journal of Dialect Geography*.

Isaaks E. H. and Srivastava R. M. (1989). *An Introduction to Applied Geostatistics*. Oxford University Press.

Kurath, H. (1949). *Word Geography of the Eastern United States*. Ann Arbor, Michigan: University of Michigan Press.

Kurath, H., Hansen, L., Bloch, B. and Bloch, J. (1939). *Handbook of the Linguistic Geography of New England*. Providence, Rhode Island: Brown University Press.

Kurath, H. and McDavid R. I. (1961). *The Pronunciation of English in the Atlantic States*. Ann Arbor, Michigan: University of Michigan Press.

Labov W., Ash S. and Boberg C. (2006). *Atlas of North American English: Phonetics, Phonology, and Sound Change*. New York: Mouton de Gruyter.

Moran, P. A. P. (1948). The interpretation of statistical maps. *Journal of the Royal Statistical Society, Series B,* 37: 243- 251.

Nerbonne, J. (2009). Data-driven dialectology. *Language and Linguistics Compass,* 3:175–198.

Nerbonne, J. (2010). Measuring the diffusion of linguistic change. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365: 3821-3828.

Odland, J. D. (1988). *Spatial Autocorrelation*. Los Angeles: Sage Publications.

Oliver, M. A. (2010). *Geostatistical Applications for Precision Agriculture*. New York: Springer.

Ord, J. K. and Getis, A. (1995). Local spatial autocorrelation statistics: distributional issues and an application. *Geographical Analysis,* 27: 286-306.

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

Spruit, M. S., Heeringa , W. and Nerbonne, J. (2009). Associations among linguistic levels. *Lingua,* 119: 1624-642.

Wackernagel, H. (2010). *Multivariate Geostatistics* (Third Edition). Berlin: Springer-Verlag.

Zelinsky, W. (1973). *Cultural Geography of the United States*. Englewood Cliffs, NJ: Prentice-Hall.