A multivariate spatial analysis of vowel formants in American English

Jack Grieve

Centre for Forensic Linguistics

Aston University

Dirk Speelman and Dirk Geeraerts

Quantitative Lexicology and Variational Linguistics Research Unit

University of Leuven

December 10, 2012

Jack Grieve

Lecturer in Forensic Linguistics

Centre for Forensic Linguistics

School of Languages and Social Sciences

Aston University

Aston Triangle, B4 7ET

Birmingham, UK

44 (0)121 2043784

j.grieve1@aston.ac.uk

Short Title: An analysis of vowel formants in American English

Abstract

This paper presents the results of a multivariate spatial analysis of 38 vowel formant variables measured in 236 cities from across the contiguous United States, based on the acoustic data from the *Atlas of North American English* (Labov, Ash & Boberg, 2006). The results of the analysis both confirm and challenge the results of the *Atlas.* Most notably, while the analysis identifies similar patterns as the *Atlas* in the West and the Southeast, the analysis finds that the Midwest and the Northeast are distinct dialect regions that are considerably stronger than the traditional Midland dialect region identified in the *Atlas*. The analysis also finds evidence that a vowel shift is actively shaping the language of the Western United States.

Acknowledgments

We would like to thank William Labov for his comments on this analysis and for providing the data upon which this analysis is based, along with Sharon Ash and Charles Boberg. We would also like to thank the members of the QLVL Research Unit at the University of Leuven, Emily Waibel, and four anonymous reviewers for their comments on this paper.

1. Introduction

Most research on regional linguistic variation in American English has been based on the subjective analysis of linguistic survey data (e.g. Kurath, 1949; Carver, 1987; Labov et al, 2006). The traditional and standard approach to data analysis in American dialectology involves manually analyzing maps that plot the values of numerous linguistic variables across a region to identify individual and common patterns of regional variation. Usually isoglosses are drawn to divide each map into the regions where the different values of the linguistic variable predominate. Common patterns of regional linguistic variation are then identified by searching for linguistic variables with isoglosses that follow similar paths. Finally, dialect regions are identified based on how these bundles of isoglosses divide the region.

Although the traditional approach to the analysis of regional linguistic variation follows a logical series of steps, each stage of the analysis ultimately relies on the judgment of the dialectologist. Descriptive statistics and replicable procedures are sometimes used to help guide the analysis, but key decisions such as the selection of the linguistic variables that define a particular region (e.g. Carver, 1987) or the design of the algorithm that generates isoglosses (e.g. Labov et al, 2006) are still based on the judgment of the dialectologist, making it difficult to replicate analyses and to choose between competing theories of dialect regions. A statistical approach to the analysis of regional linguistic variation can help to resolve issues such as these. Despite the advantages of such an approach, statistical analysis is uncommon in American dialectology, perhaps because the statistical methods commonly used to analyze language variation and change do not allow for regional variation to be analyzed following the same series of steps as the traditional approach.

For example, the types of statistical methods commonly used in variationist sociolinguistics to analyze the relationships between linguistic variables and social variables are unsuitable to analyze the relationships between linguistic variables and regional variables such as longitude and latitude because these spatial relationships are often non-linear. Linear patterns such as a change from the northeast to southwest can be identified using standard variationist methods, but other types of patterns, such as a single central cluster, cannot. While variationist methods are of limited use in regional dialectology, a quantitative approach to the analysis of regional linguistic variation known as *dialectometry* is common in Europe (see Séguy 1971, 1973; Goebl, 1982, 2006; Heeringa, 2004; Nerbonne, 2006). In dialectometry, patterns of aggregated regional linguistic variation are identified using replicable and statistically justified methods; however, because dialectometry does not follow the same steps as a traditional analysis, it does not allow for regional patterns to be identified in a way that is satisfactory to many dialectologists. In particular, dialectometry studies do not generally analyze individual linguistic variables or identify subsets of linguistic variables that exhibit similar patterns of regional variation.

Although the standard statistical methods used in variationist sociolinguistics and dialectometry cannot replace the traditional approach to the analysis of regional linguistic variation, a statistical approach known as a *multivariate spatial analysis* has recently been introduced that follows the same series of steps as a traditional analysis (Grieve, 2009, 2013a; Grieve et al, 2011). This new approach to the analysis of regional linguistic variation is based on spatial autocorrelation statistics (Grieve 2009, 2011, 2012), which allow for significant patterns of spatial clustering to be identified in the values of individual linguistic variables in a manner that is similar to plotting isoglosses. The results of the spatial autocorrelation analysis are then subjected to a factor analysis to identify common patterns of regional variation in a manner that is similar to identifying bundles of isoglosses. A multivariate spatial analysis therefore identifies both individual and common patterns of regional linguistic variation in a way that is similar to the traditional approach to the analysis of regional linguistic variation.

This paper describes a multivariate spatial analysis of the acoustic vowel formant data from the *Atlas of North American English* (Labov et al, 2006). As one would expect, given the results of the *Atlas*, this analysis finds that most vowel formant variables are regionally patterned in American English. However, although the analysis identifies similar regional patterns as those presented in the *Atlas*, the method also identifies additional patterns that had previously gone unnoticed, which in some cases challenge traditional taxonomies and theories of American dialect regions.

2. Data

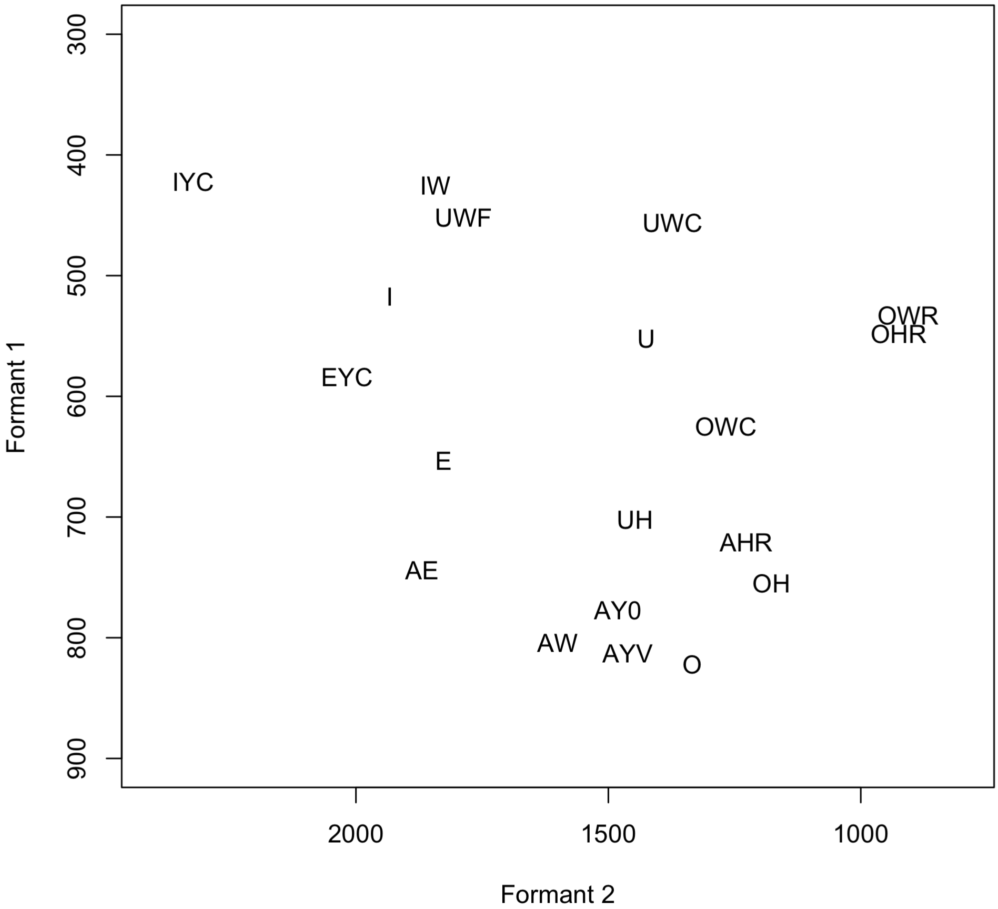
The multivariate spatial analysis reported in this paper was based on the acoustic vowel data gathered for the *Atlas of North American English* (Labov et al, 2006). Data collection for the *Atlas* took place during the 1990s through linguistic interviews conducted over the telephone with informants in cities from across English-speaking North America. On average 2 to 4 informants were selected per city, with each informant being interviewed for approximately 30 to 45 minutes. In total 762 informants were interviewed for the *Atlas.* The recordings for 439 of these informants were then subjected to an acoustic analysis in order to measure the values of 48 vowel formant variables (see Chapter 10 of the *Atlas* for the raw maps), which consist of the average formant 1 and formant 2 values for 24 distinct vowel measures, including a number of vowels measured in different phonological contexts.

The analysis reported here was restricted to 38 of the 48 acoustic vowel formant variables from the *Atlas* (i.e. average formant 1 and formant 2 values for 19 vowel measures). Ten vowel formant variables (formant 1 and formant 2 for /oy/, /aeh/, /eyr/, /uwr/, /ah/) were excluded from this analysis because they were missing data for over 10% of the locations in the dataset. Twelve of the remaining vowel formant variables (formant 1 and formant 2 for /uh/, /u/, /ay0/, /iw/, /uwc/, /ohr/) were also missing data, but because this missing data accounted for less than 5% of the locations in the dataset, these variables were retained and the missing data were replaced by the mean value for that variable across all locations.[1] The 19 vowel measures analyzed in this study are listed in Table 1, including the phonetic symbol from the *Atlas,* the IPA equivalent, and an example of the vowel in context. In addition, Table 1 is organized based on the system of vowel categorization used in the *Atlas.* This system distinguishes between three levels of height and between front and back and round and unrounded vowels, as well as between short and long vowels, with the long vowels being further divided into ingliding vowels and both front and back upgliding vowels. In addition, the 19 vowel measures are plotted in Figure 1 based on their average formant 1 and formant 2 values across the entire dataset.

Table 1 Vowel Measures

Vowel Measure	IPA Vowel	Context Restrictions	Example	Length	Position	Height	Glide Type
/i/	/ɪ/		*bit*	Short	Front	High	
/e/	/ɛ/		*bet*	Short	Front	Mid	
/ae/	/æ/		*bat*	Short	Front	Low	
/u/	/ʊ/		*book*	Short	Back	High	
/uh/	/ʌ/		*but*	Short	Back	Mid	
/o/	/ɒ/		*cot*	Short	Back	Low	
/iyc/	/i/	Word internally	*beat*	Long	Front	High	Front Upglide
/eyc/	/eɪ/	Word internally	*bait*	Long	Front	Mid	Front Upglide
/ayv/	/aɪ/	Before voiced consonants	*bide*	Long	Back	Low	Front Upglide
/ay0/	/aɪ/	Before voiceless consonants	*bite*	Long	Back	Low	Front Upglide
/iw/	/u/		*suit*	Long	Front	High	Back Upglide
/uwc/	/u/	Word internally	*boot*	Long	Back	High	Back Upglide
/uwf/	/u/	Word finally	*boo*	Long	Back	High	Back Upglide
/owr/	/oʊ/	Before /r/	*boar*	Long	Back	Mid	Back Upglide
/owc/	/oʊ/	Before other consonants	*boat*	Long	Back	Mid	Back Upglide
/aw/	/aʊ/		*bout*	Long	Back	Low	Back Upglide
/ohr/	/ɔ/	Before /r/	*north*	Long	Back	Mid	Rounded Inglide
/oh/	/ɔ/	Before other consonants	*caught*	Long	Back	Mid	Rounded Inglide
/ahr/	/ɑ/	Before /r/	*start*	Long	Back	Low	Unrounded Inglide

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



The analysis reported here was also restricted to 402 of the 439 informants whose recorded interviews were subjected to an acoustic analysis. In particular, Canadian informants were excluded from the analysis to control for the influence of national linguistic variation and Alaskan informants were excluded from the analysis because as extreme geographical outliers their inclusion would confound the spatial analysis. In addition, the one speaker from Bloomington, Illinois was also excluded from the analysis because he is an extreme outlier on formant 2 for all vowels. As well as removing these informants, the dataset analyzed here was also pooled across all informants from the same city, reducing the number of cases in the dataset from 402 informants to 236 cities. Pooling the data by location is required for a multivariate spatial analysis but has several additional advantages. Most important, maps based on pooled data are easier to interpret, especially when there is considerable variation in the number of informants per location, as is the case here. Pooling data also reduces the proportion of missing data and controls to some extent for variation in the gender and age of the informants.

The final dataset analyzed in this study therefore consists of 38 vowel formant variables measured across 236 cities from across the contiguous United States. Before subjecting this regional linguistic data matrix to a multivariate statistical analysis, the raw values of the 38 vowel formant variables were mapped across the 236 locations. Examples for four of the variables are presented in Figures 2-5. Figure 2 shows that /eyc/ tends to be raised in the North and lowered in the South and the West. Figure 3 shows that /ae/ tends to be raised in the Midwest and lowered in the Northeast and the West. Figure 4 shows that /oh/ tends to be backed on the East Coast and fronted in the West. Finally, Figure 5 shows that /ayv/ tends to be backed in the Midland and fronted across most of the rest of the United States. In each of these cases a regional pattern is discernible in the raw maps; however, because these patterns are not absolute, their significance is unclear.

Figure 2 Raw Value Map for non-word-final /ey/ (e.g. *bait*) on Formant 1

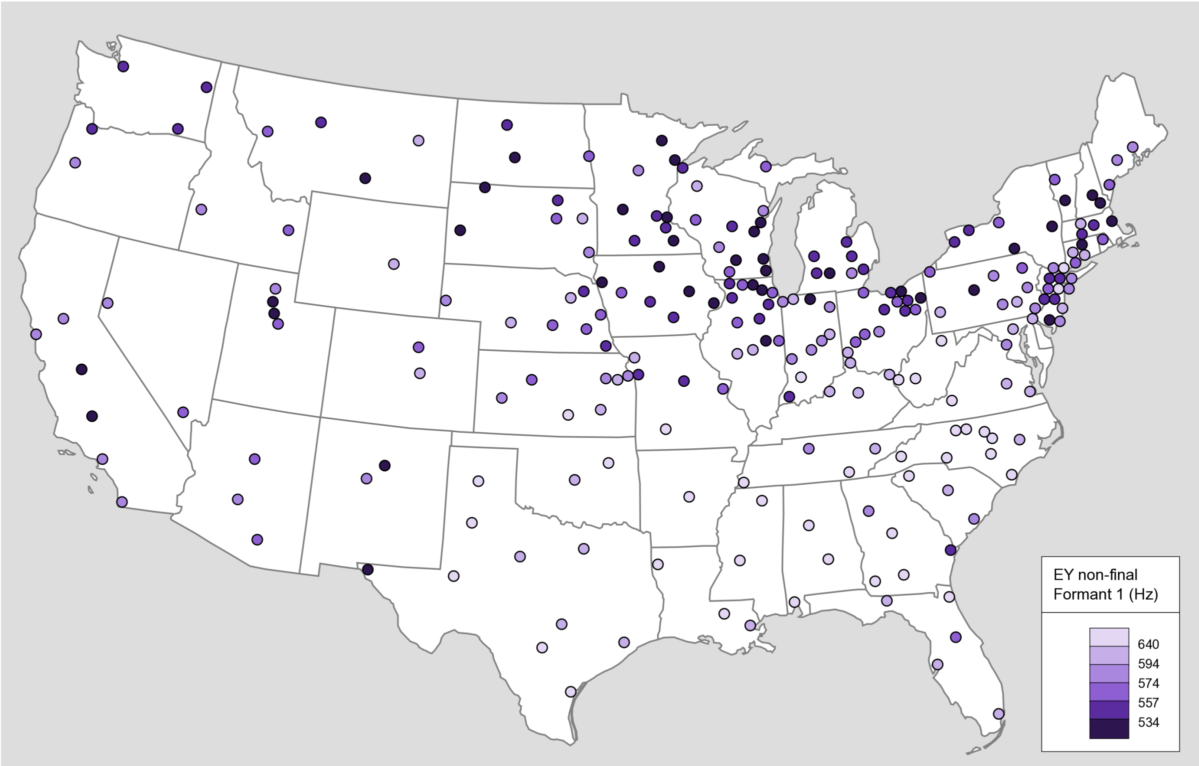


Figure 3 Raw Value Map for /ae/ (e.g. *bat*) on Formant 1

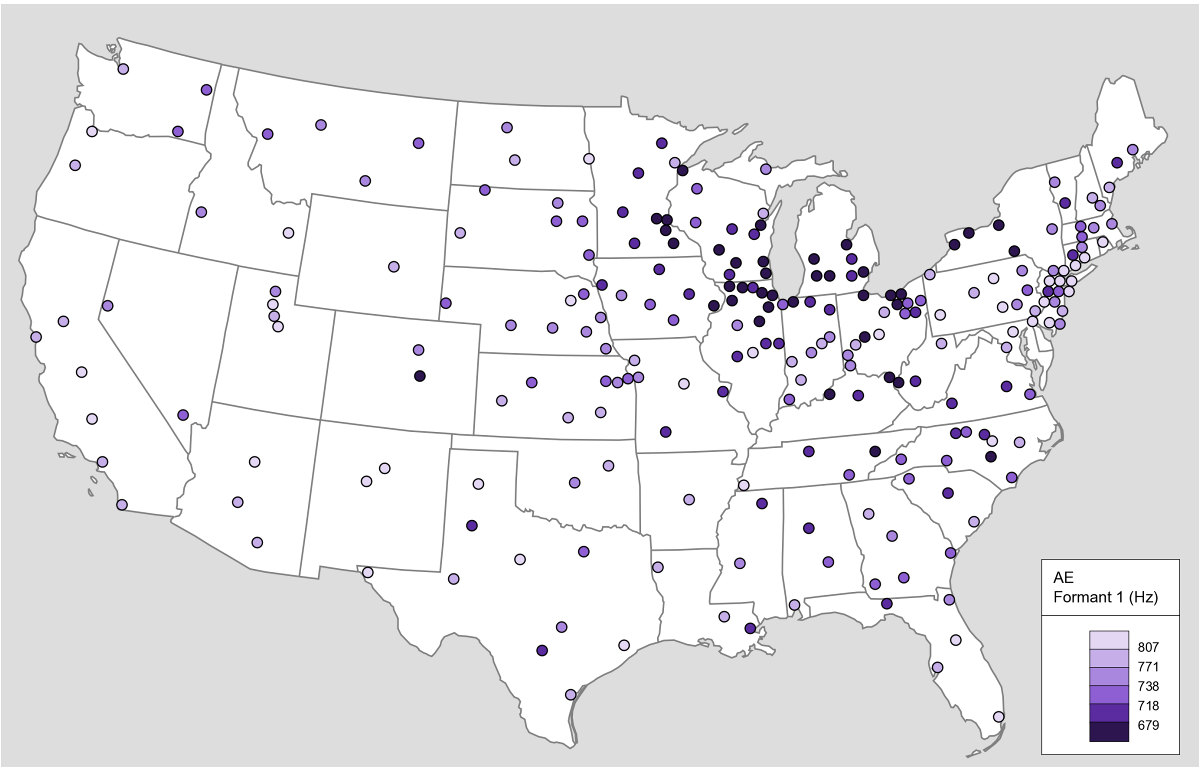


Figure 4 Raw Value Map for /oh/ (e.g. *caught*) on Formant 2

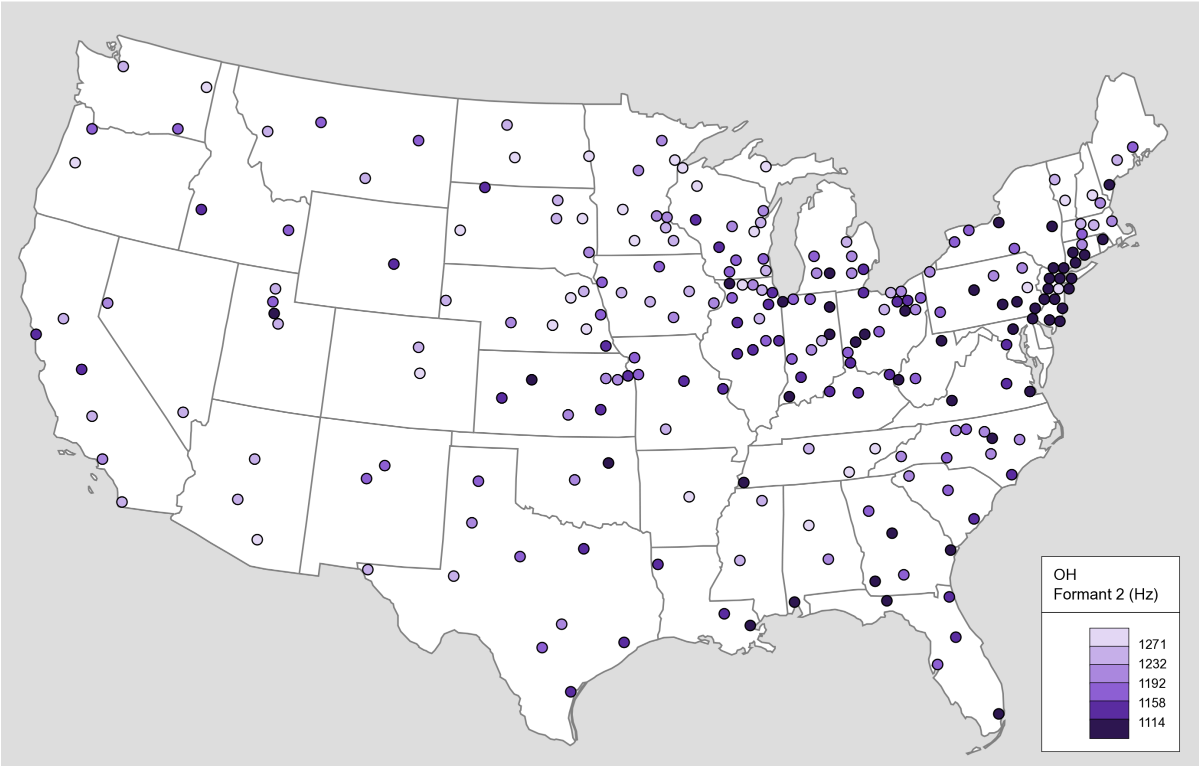
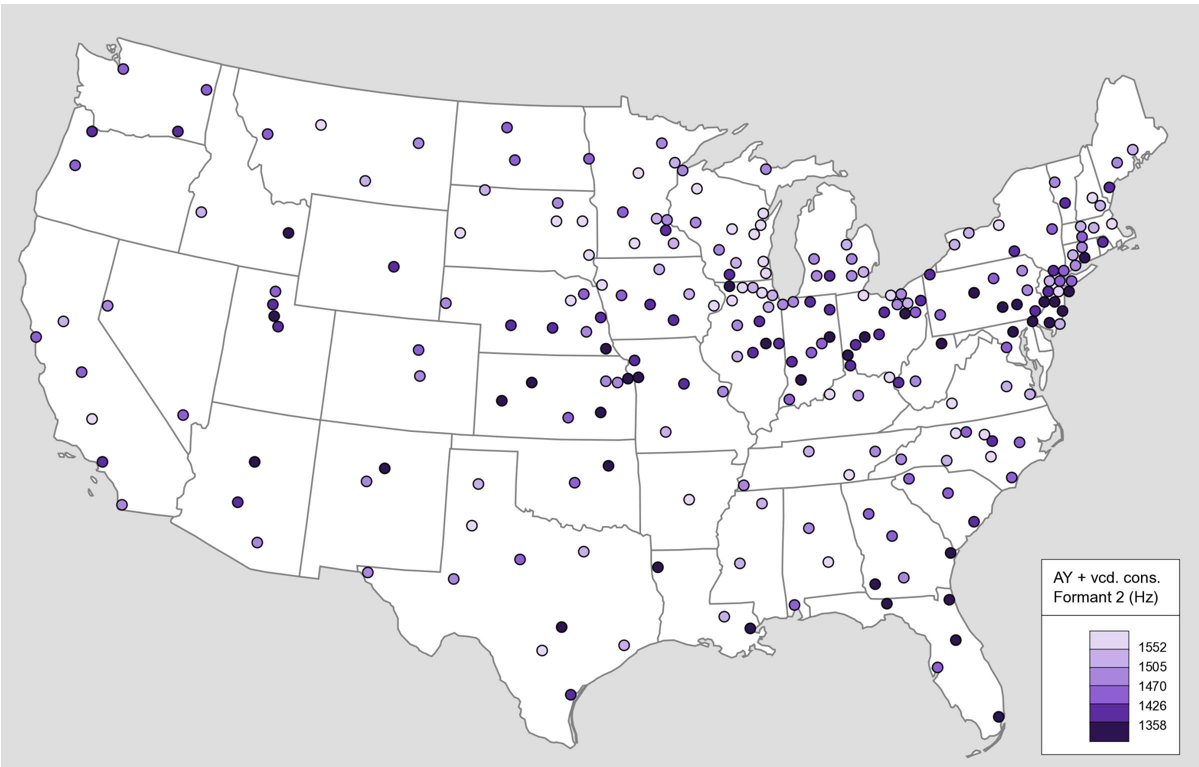


Figure 5 Raw Value Map for /ayv/ before voiced consonants (e.g. *bide*) on Formant 2



3. Spatial Autocorrelation Analysis

In order to identify statistically significant patterns of regional variation in the values of the 38 individual vowel formant variables, each variable was subjected to a spatial autocorrelation analysis (Grieve, 2011, 2012; Grieve et al, 2011).[2] First, each variable was subjected to an analysis of global spatial autocorrelation using global Moran's *I* (Moran, 1948; Odland, 1988) to identify variables exhibiting significant levels of spatial clustering. Second, each variable was subjected to an analysis of local spatial autocorrelation using local Getis-Ord *Gi* (Ord and Getis, 1995) to identify the locations of any high- and low-value clusters.

To calculate both spatial autocorrelation measures, it is necessary to define a *spatial weighting function,* 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; Grieve, 2011, 2012; Grieve et al, 2011). In this study, a reciprocal spatial weighting function was used, which assigns a weight to each pair of locations based on the reciprocal of the distance between the locations so that the weight decreases with distance. A reciprocal weighting function was selected because it allows for the measures of spatial autocorrelation to be based primarily on the closest locations. This is important because some of the dialect regions identified by the *Atlas* are relatively narrow, such as the Midland dialect region or the extension of the Northern Cities region around the Great Lakes. A weighting function that focuses primarily on nearby locations is most suitable for replicating these types of results.

In order to interpret the significance of Moran's *I*, a standardized *z*-score was obtained under the assumption of randomization (Odland, 1988). This *z*-score was then interpreted based on a one-tailed test of significance, because the goal of the analysis was to identify positive spatial autocorrelation. 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). A Bonferroni correction 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 significant positive value for Moran's *I* indicates that the values of the variable exhibit spatial clustering, where nearby locations tend to have similar values at a greater degree than would be expected by chance. The results of the global spatial autocorrelation analysis are presented in Table 2, which for each vowel formant variable lists its mean value, Moran's *I*, the associated *z*-score, and the associated 1-tailed *p*-value. Based on the global spatial autocorrelation analysis, 25 out of the 38 variables were found to exhibit significant levels of spatial clustering at the adjusted .0013 significance level.[3]

In addition to conducting a global spatial autocorrelation analysis, a local spatial autocorrelation analysis was conducted using local Getis-Ord *Gi* to identify the locations of spatial clusters in the values of each vowel formant variable. For each location, Getis-Ord *Gi* returns a *z*-score indicating the degree to which that location is surrounded by locations with similar values. A significant negative Getis-Ord *Gi z*-score indicates that the location is part of a low-value cluster, whereas a significant positive Getis-Ord *Gi z*-score indicates that the location is part of a high-value cluster. A Getis-Ord *Gi 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 described above, although in this case a two-tailed test of significance was used instead of a one-tailed test of significance because the goal of the analysis was to identify both high- and low-value clusters.

In order to visualize the patterns of spatial clustering identified by the local spatial autocorrelation analysis, the Getis-Ord *Gi z*-scores for each variable were plotted over the cities in the dataset. This is essentially a statistical method for plotting isoglosses.These local spatial autocorrelation maps identify clear high and low value clusters in the majority of the vowel formant variables. Local autocorrelation maps are provided for four variables in Figures 6-9, corresponding to the raw maps presented in Figures 2 through 5. These maps support the interpretation of the raw maps presented above: /eyc/ on formant 1 contrasts the South and West with the North (Figure 6), /ae/ on formant 1 contrasts the Midwest with the Northeast and the West (Figure 7), /oh/ on formant 2 contrasts the Northeast with the West (Figure 8), and /ay/ on formant 2 contrasts the Midland with the rest of the United States (Figure 9).

Finally, it is important to emphasize that conducting a local spatial autocorrelation analysis results in the loss of fine details present in the raw maps. For example, the value of a location that differs from the values of surrounding locations will be smoothed over by the local spatial autocorrelation analysis. That outlier location could be noise introduced through data collection or it could represent a true regional pattern, but in either case this variation will be lost. This is not a flaw in the local spatial autocorrelation analysis. The goal of the local autocorrelation analysis is to identify patterns of spatial clustering based on the regional linguistic dataset. The local spatial autocorrelation cannot identify a particular regional pattern if the dataset does not include a sufficient number of locations in that region to allow for that pattern to be identified. A regional linguistic dataset has a certain level of resolution and the local spatial autocorrelation maps must be interpreted accordingly.

Table 2 Global Autocorrelation Results (Reciprocal Weighting Function, N = 236)

Vowel	Formant	Mean (Hz)	Moran's I	*z*-score	*p* (1-tailed)
/oh/	1	755	0.446	30.56	< 0.0001
/oh/	2	1177	0.303	20.82	< 0.0001
/aw/	2	1600	0.251	17.3	< 0.0001
/uwc/	2	1373	0.227	15.62	< 0.0001
/uwf/	2	1787	0.202	13.95	< 0.0001
/ae/	1	744	0.199	13.79	< 0.0001
/owc/	2	1267	0.195	13.47	< 0.0001
/ahr/	2	1227	0.169	11.76	< 0.0001
/o/	2	1334	0.169	11.74	< 0.0001
/iw/	2	1843	0.143	9.98	< 0.0001
/u/	2	1425	0.134	9.35	< 0.0001
/eyc/	1	584	0.134	9.34	< 0.0001
/uh/	2	1447	0.116	8.14	< 0.0001
/eyc/	2	2017	0.105	7.4	< 0.0001
/ae/	2	1869	0.098	6.9	< 0.0001
/uwf/	1	452	0.075	5.35	< 0.0001
/e/	2	1826	0.067	4.85	< 0.0001
/i/	2	1933	0.067	4.8	< 0.0001
/ay0/	1	777	0.066	4.75	< 0.0001
/ay0/	2	1481	0.054	3.96	< 0.0001
/ahr/	1	721	0.054	3.94	< 0.0001
/ayv/	2	1462	0.052	3.81	< 0.0001
/e/	1	653	0.047	3.48	0.0003
/iyc/	1	422	0.047	3.46	0.0003
/ohr/	2	925	0.041	3.08	0.001
/o/	1	822	0.039	2.95	0.0016
/owr/	1	533	0.037	2.8	0.0026
/owr/	2	906	0.032	2.47	0.0068
/uh/	1	702	0.032	2.45	0.0071
/owc/	1	625	0.03	2.33	0.0099
/aw/	1	804	0.026	2.08	0.0188
/iyc/	2	2322	0.021	1.71	0.0436
/uwc/	1	456	0.02	1.63	0.0516
/i/	1	517	0.012	1.09	0.1379
/iw/	1	425	0.011	1.04	0.1492
/ohr/	1	548	0.01	0.95	0.1711
/u/	1	552	0.002	0.42	0.3372
/ayv/	1	813	0	0.27	0.3936

Figure 6 Local Autocorrelation Map for non-word-final /ey/ on Formant 1

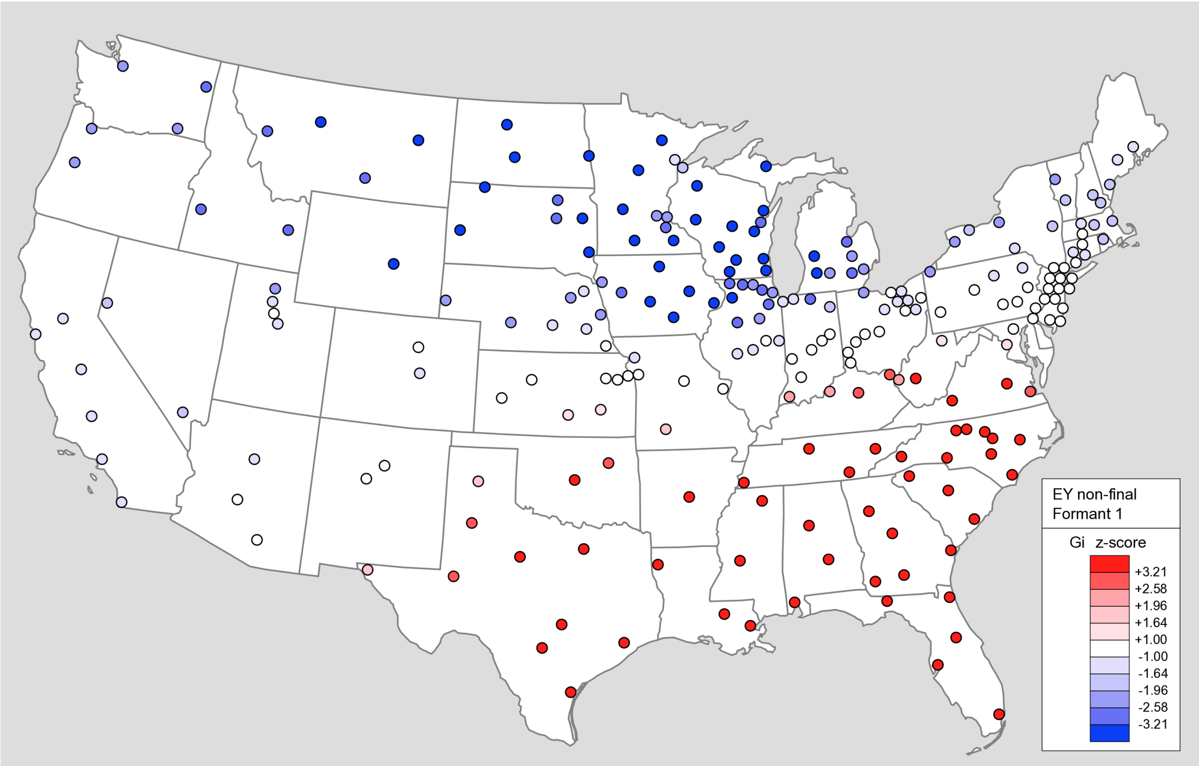


Figure 7 Local Autocorrelation Map for /ae/ on Formant 1

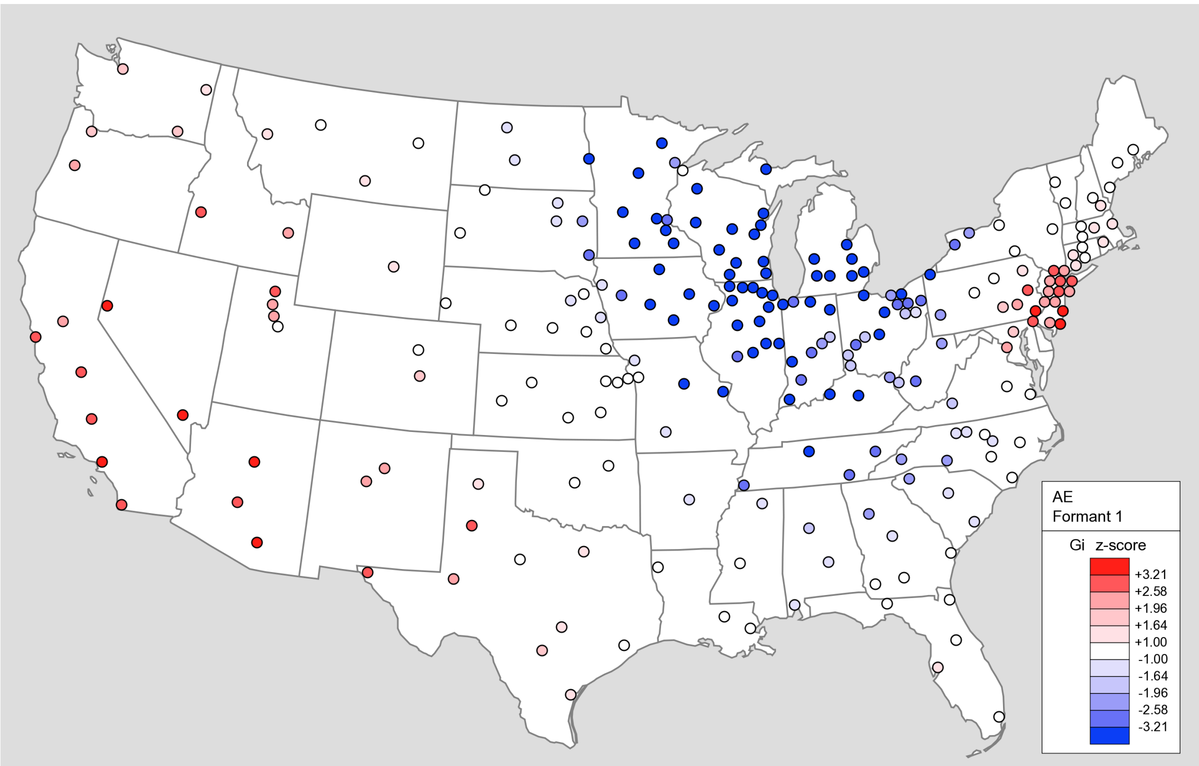


Figure 8 Local Autocorrelation Map for /oh/ on Formant 2

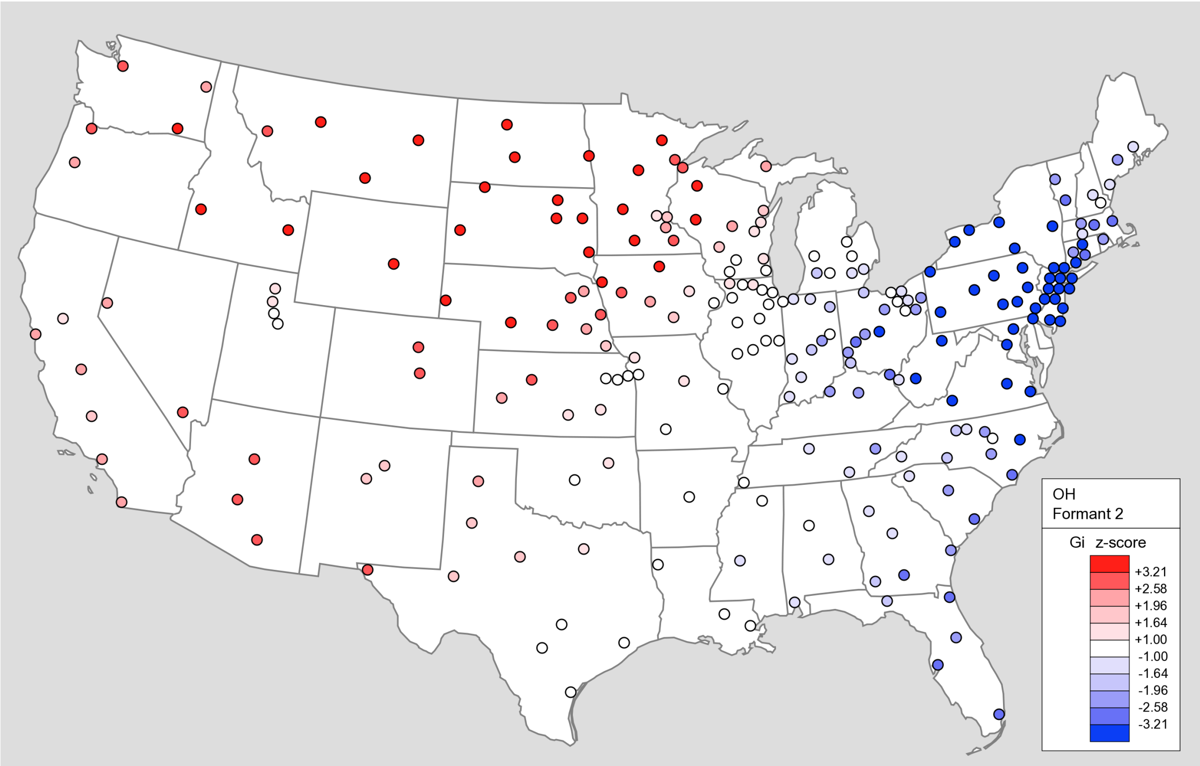
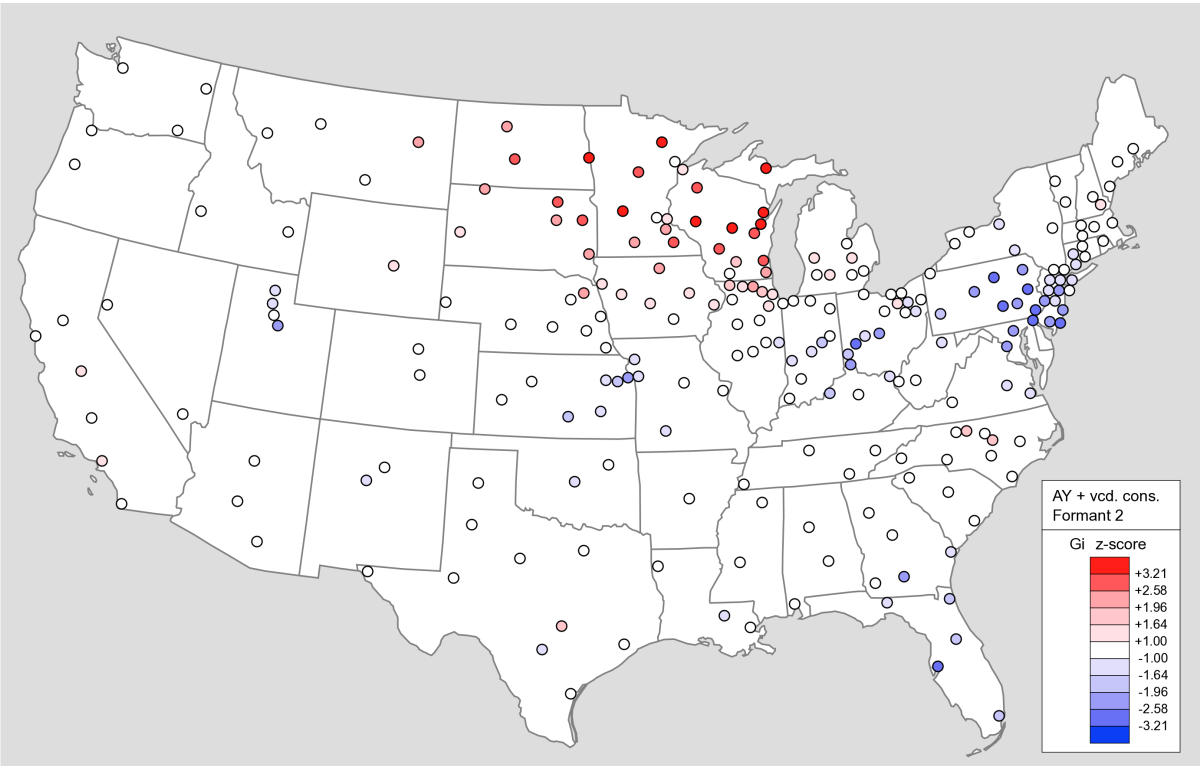


Figure 9 Local Autocorrelation Map for /ay/ before voiced consonants on Formant 2



4. Factor Analysis

In order to identify common patterns of regional variation in the values of the 38 vowel formant variables, the Getis-Ord *Gi* *z*-scores for the complete set of vowel formant variables were subjected to a factor analysis (Grieve et al, 2011). Given a set of variables measured over a set of cases, a factor analysis extracts a series of factors that each represents a common pattern of variation in that dataset, as well as the variables associated with each of these patterns (Hair et al, 2006). Because the local Getis-Ord *Gi* *z*-scores represent the location of spatial clusters in the values of the vowel formant variables, subjecting this dataset 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.[4] Subjecting the results of the local spatial autocorrelation analysis to a factor analysis is therefore essentially a statistical method for identifying bundles of isoglosses.

Before conducting a factor analysis, it is necessary to select the number of factors to be extracted, which can be determined by identifying the point where retaining further factors would explain relatively little additional variance. In this case 4 factors were selected because together these factors accounted for 86% of the variance in the values of the 38 vowel formant variables, with additional factors explaining relatively little additional variance, indicating that the regional patterns exhibited by the 38 vowel formant variables can largely be accounted for by these 4 basic patterns. The final factor analysis was thus run to extract 4 factors. In addition, the 4 factors were rotated using varimax rotation to limit the number of factors onto which each of the variables load, causing the factors to more clearly reflect the spatial patterns visible in the local autocorrelation maps for the individual linguistic variables.

Each factor was analyzed in three ways. First, the factor scores for each factor were mapped across the 236 cities in the dataset in order to visualize the common patterns of spatial clustering represented by each factor. These factor maps, which are presented in

Figures 10-13, contrast regions with positive factor scores (in magenta) to regions with negative factor scores (in green). Second, the factor loadings, which are presented in Table 3, were inspected. A high positive or negative factor loading indicates that the vowel formant variable exhibits the basic pattern of spatial clustering represented by that factor. In addition, for a formant 1 variable, a positive factor loading indicates that the vowel measure is lowered in regions with positive factor scores and raised in regions with negative factor scores, whereas a negative factor loading indicates that the vowel measure is raised in regions with positive factor scores and lowered in regions with negative factor scores. Similarly, for a formant 2 variable, a positive factor loading indicates that the vowel measure is fronted in regions with positive factor scores and backed in regions with negative factor scores, whereas a negative factor loading indicates that the vowel measure is backed in regions with positive factor scores and fronted in regions with negative factor scores. Table 3 also lists the uniqueness values for each vowel formant variable, which in all cases are relatively low (far below .800) indicating that the 4 factors account relatively well for the regional patterns exhibited by all 38 variables. Third, the positions of the 19 vowel measures were plotted based on the average formant 1 and formant 2 values for the vowel measures in cities with the highest (larger than 1.00) and lowest (smaller than -1.00) factor scores for each factor. In essence, these plots, which are presented in Figure 14, show the vowel spaces for the average informants in each of the opposing regions identified by the 4 factors.[5]

Factor 1 accounts for 39.0% of the variance in the dataset and contrasts the Southeast with the North and the West, with the approximate area of transition between these two regions running along the northern borders of the Virginias, through northern Ohio, Indiana and Illinois, and central Iowa (Figure 10). Numerous vowel formant variables distinguish between these two regions, as indicated by the large number of variables that load strongly on Factor 1. Most notably, nine long vowels load strongly on Factor 1 for both formant 1 and formant 2, with /eyc/, /iyc/ and /ay0/ tending to be lowered and backed in the Southeast, /ahr/, /owr/ and /ohr/ tending to be raised and backed in the Southeast, /uwc/ and /owc/ tending to be lowered and fronted in the Southeast, and /aw/ tending to be raised and fronted in the Southeast. Formant 2 values for two additional long vowels also load strongly on Factor 1, with /iw/ and /uwf/ tending to be fronted in the Southeast. Finally, format 1 values for four short vowels also load strongly on Factor 1, with /i/, /e/, /uh/ and /u/ all tending to be raised in the Southeast. In addition, other variables also load weakly on Factor 1 and shift slightly in the opposing vowel spaces for this factor, but these variables all load and shift to greater degrees on other factors.

Factor 2 accounts for 23.4% of the variance in the dataset and contrasts the Midwest with the rest of the United States, especially the West and the Mid-Atlantic (Figure 11). The Midwest region identified here encompasses the core Midwestern states, but also stretches into New York, Pennsylvania, West Virginia, Kentucky, Tennessee, Missouri, Iowa, Nebraska, and the Dakotas. The factor loadings identify numerous vowel formant variables that distinguish between the Midwest and the rest of the United States. Most notable, all six short vowels load on Factor 2. In particular, two short vowels load strongly on Factor 2 for both formant 1 and formant 2, with /ae/ tending to be raised and fronted in the Midwest and with /o/ tending to be lowered and fronted in the Midwest. Four additional short vowels also load strongly for formant 2, with /i/, /e/, /uh/ and /u/ all tending to be backed in the Midwest. Seven long vowel formant variables also load strongly on Factor 2, with /iw/, /owc/ and /aw/ tending to be backed in the Midwest, /uwc/ and /uwf/ tending to be raised and backed in the Midwest, /ahr/ tending to be fronted in the Midwest, and /ay0/ tending to be raised in the Midwest. In addition, other variables also load weakly on Factor 2 and shift slightly in the opposing vowel spaces for this factor, but these variables all load and shift to greater degrees on other factors.

Factor 3 accounts for 17.8% of the variance in the dataset and contrasts the Northeast, which extends into eastern Ohio and Michigan as well as the Virginias and the Carolinas, with the West, especially the Great Plains and the Mountain States (Figure 12). Ten long vowel measures load strongly on Factor 3. Most notably, /oh/ for both formant 1 and formant 2 loads strongly on Factor 3, with /oh/ tending to be lowered and fronted in the West. Similarly, /ohr/, /owr/ and /ay0/ also load strongly on Factor 3, with /ohr/ tending to be lowered and fronted in the West, and with /owr/ and /ay0/ tending to be lowered in the West. In addition, /iw/, /uwf/, /uwc/ and /iyc/ load strongly on Factor 3, with /iw/ and /uwf/ tending to be fronted and raised slightly in the West, with /uwc/ tending to be fronted in the West, and with /iyc/ tending to be very slightly raised in the West. Finally, three short vowels also load strongly on Factor 3, with both /u/ and /uh/ tending to be fronted in the West, and with /e/ tending to be raised in the West. In addition, other variables also load weakly on Factor 3 and shift slightly in the opposing vowel spaces for this factor, but these variables all load and shift to greater degrees on other factors.

Finally, Factor 4 accounts for 6.2% of the variance in the dataset and contrasts the Midland with the rest of the country, especially the Southeast and the Upper Midwest (Figure 13). The Midland as identified by Factor 4 stretches from Philadelphia, Baltimore and southern New Jersey through all of Pennsylvania and western New York State, and into northern West Virginia and Kentucky, southern Michigan, and all of Ohio, Indiana, Illinois, Missouri, and Kansas, and to a lesser extent across the West. Only two vowel formant variables load strongly on Factor 4, with /ayv/ tending to be both raised and backed in the Midland. In addition, other variables also load weakly on Factor 4 and shift slightly in the opposing vowel spaces for this factor, but these variables all load and shift to greater degrees on other factors. Overall, the position of all of the vowels have changed far less in the average vowel spaces for the Midland and the non-Midland regions identified by Factor 4, than in the other opposing vowel spaces identified by the factor analysis. This is not surprising given the relatively small amount of variance explained by this factor.

In addition to the individual factor maps, the four sets of factor scores were mapped simultaneously using CMYK mapping, which is presented in Figure 15. In particular, a hue was defined for each location representing the scores of all four factors at that location by associating each factor with one of the four CMYK color parameters (cyan, magenta, yellow, black) (see Grieve, 2013a). These hues were then mapped across the 236 cities to produce a single overall picture of continuous regional linguistic variation in the dataset. Figure 15 shows a clear regional pattern, plainly derived from the four factor maps reproduced in Figures 10-13. This aggregated factor map identifies at least four clear clusters, consisting of the Northeast, the Midwest, the Southeast and the West. Texas also stands out as a region of transition between the West and the South. In addition, within the Northeast there is a clear distinction between New England and the Mid-Atlantic States, while within the Midwest there is a clear distinction between the Lower and Upper Midwest.

Table 3 Factor Loadings (larger than .300) and Uniqueness 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	-0.370
/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			-0.420
/eyc/	1	0.014	-0.866	0.448		
/eyc/	2	0.012	0.980			
/ayv/	1	0.478				0.650
/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	0.950			
/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	0.520	-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	-0.370	
/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	0.490		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 10 Factor 1

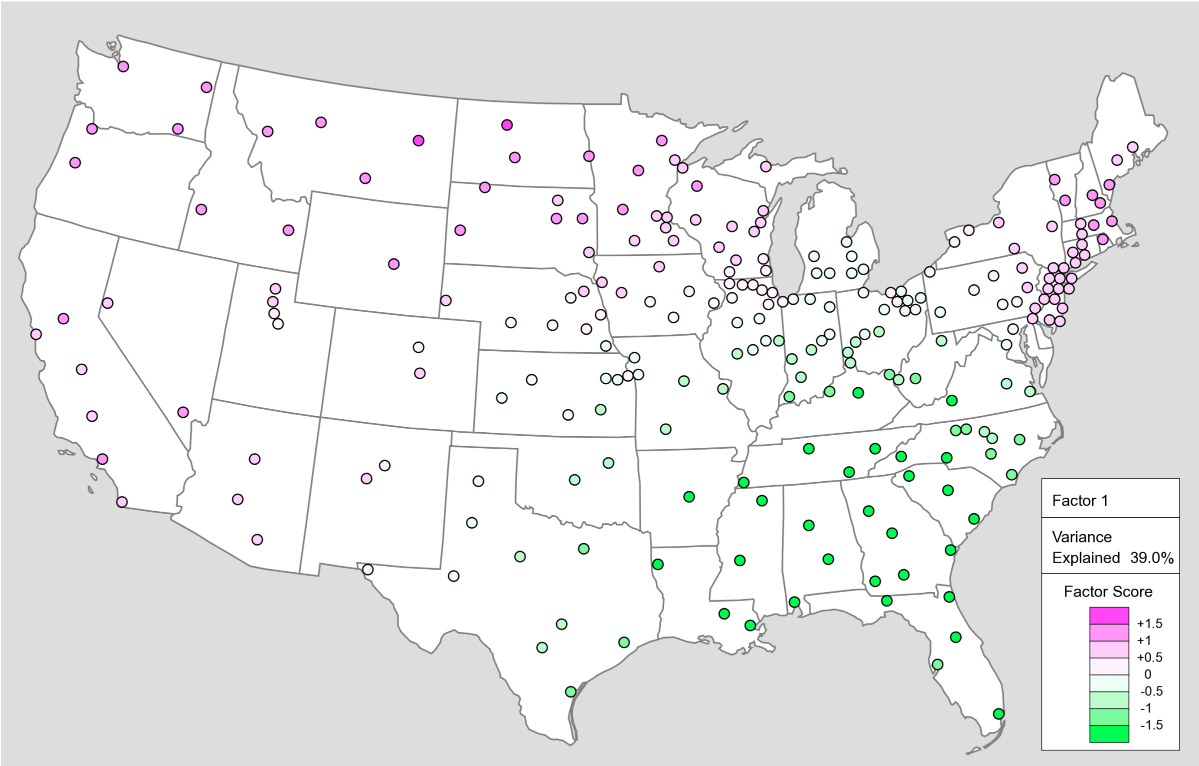


Figure 11 Factor 2

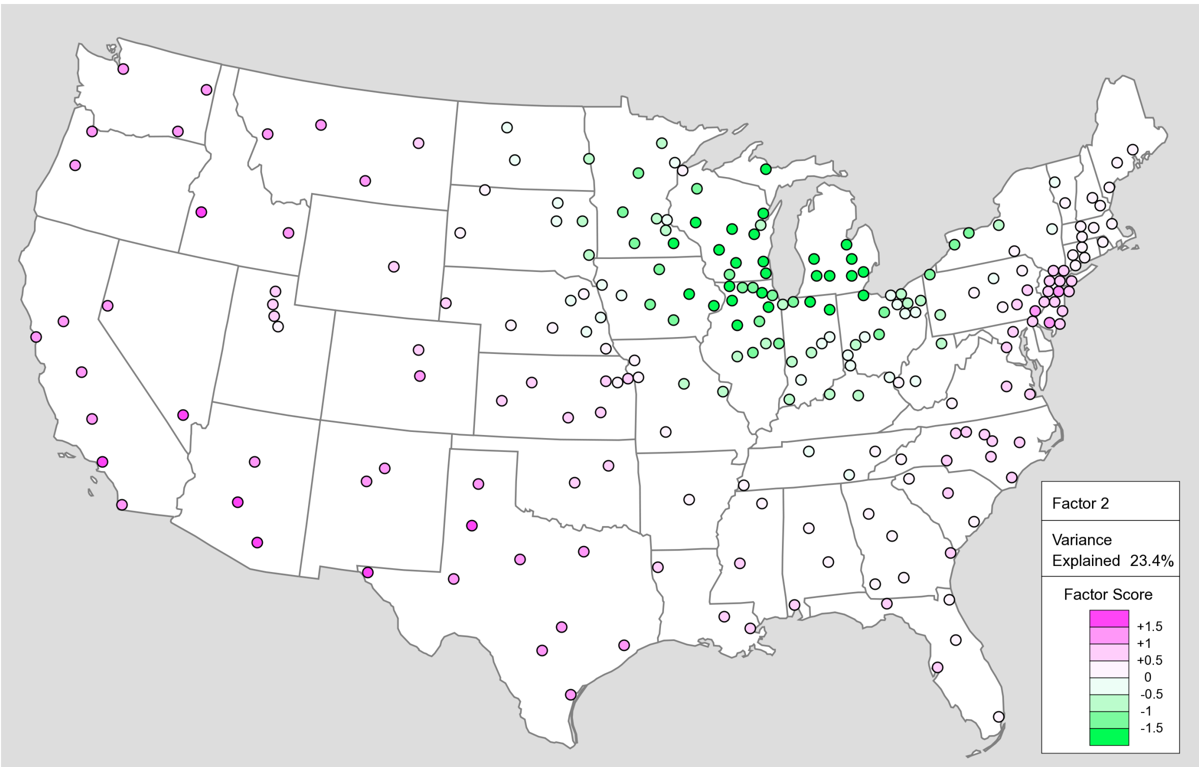


Figure 12 Factor 3

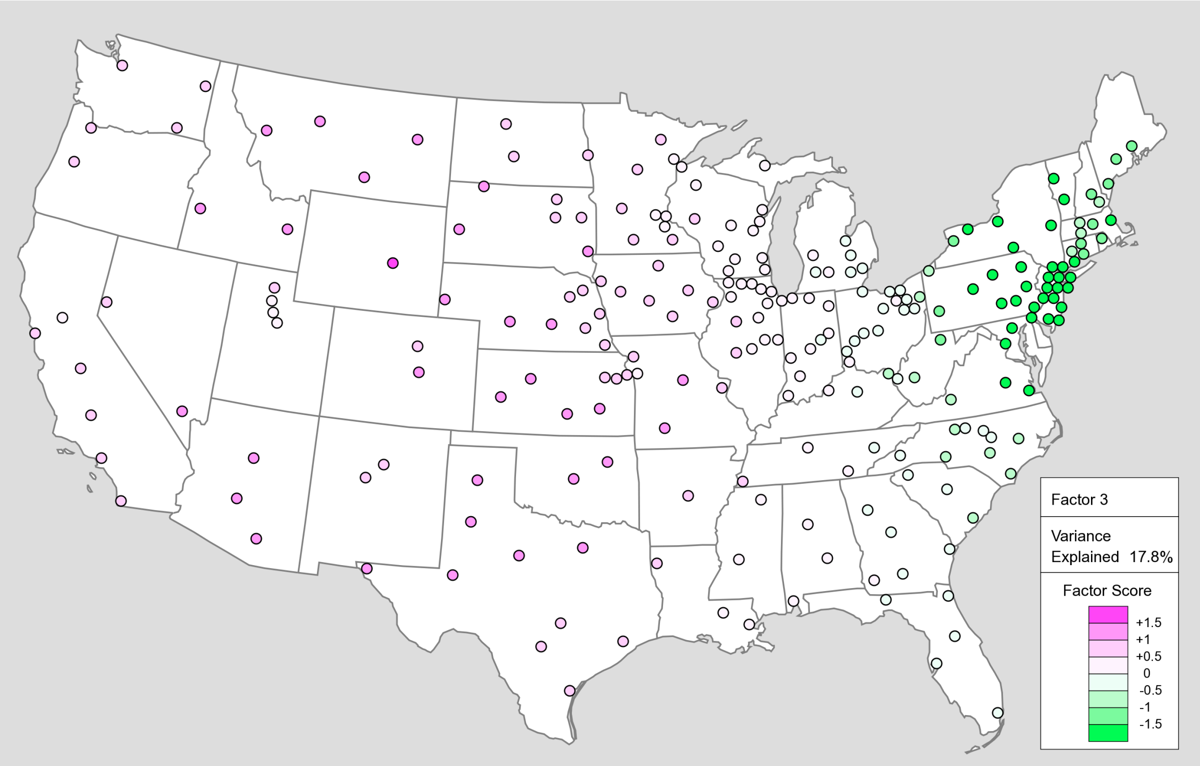


Figure 13 Factor 4

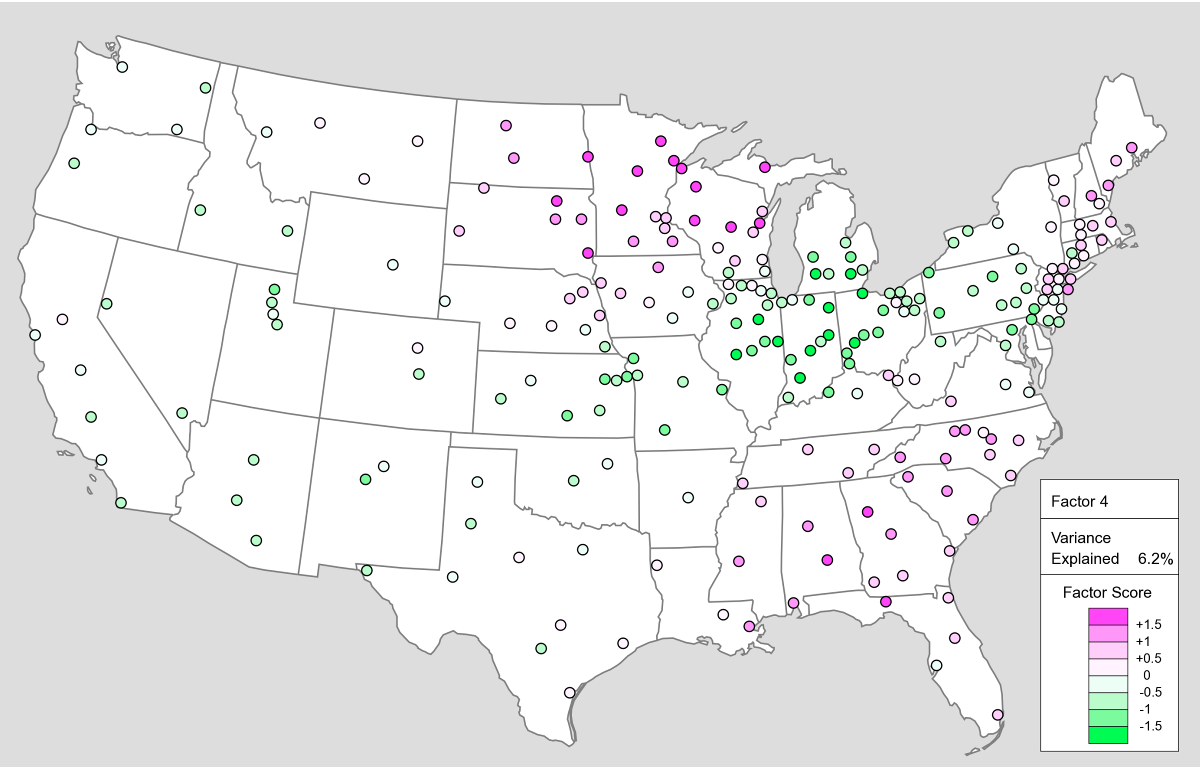


Figure 14 Average Vowel Positions for Locations with High and Low Factor Scores

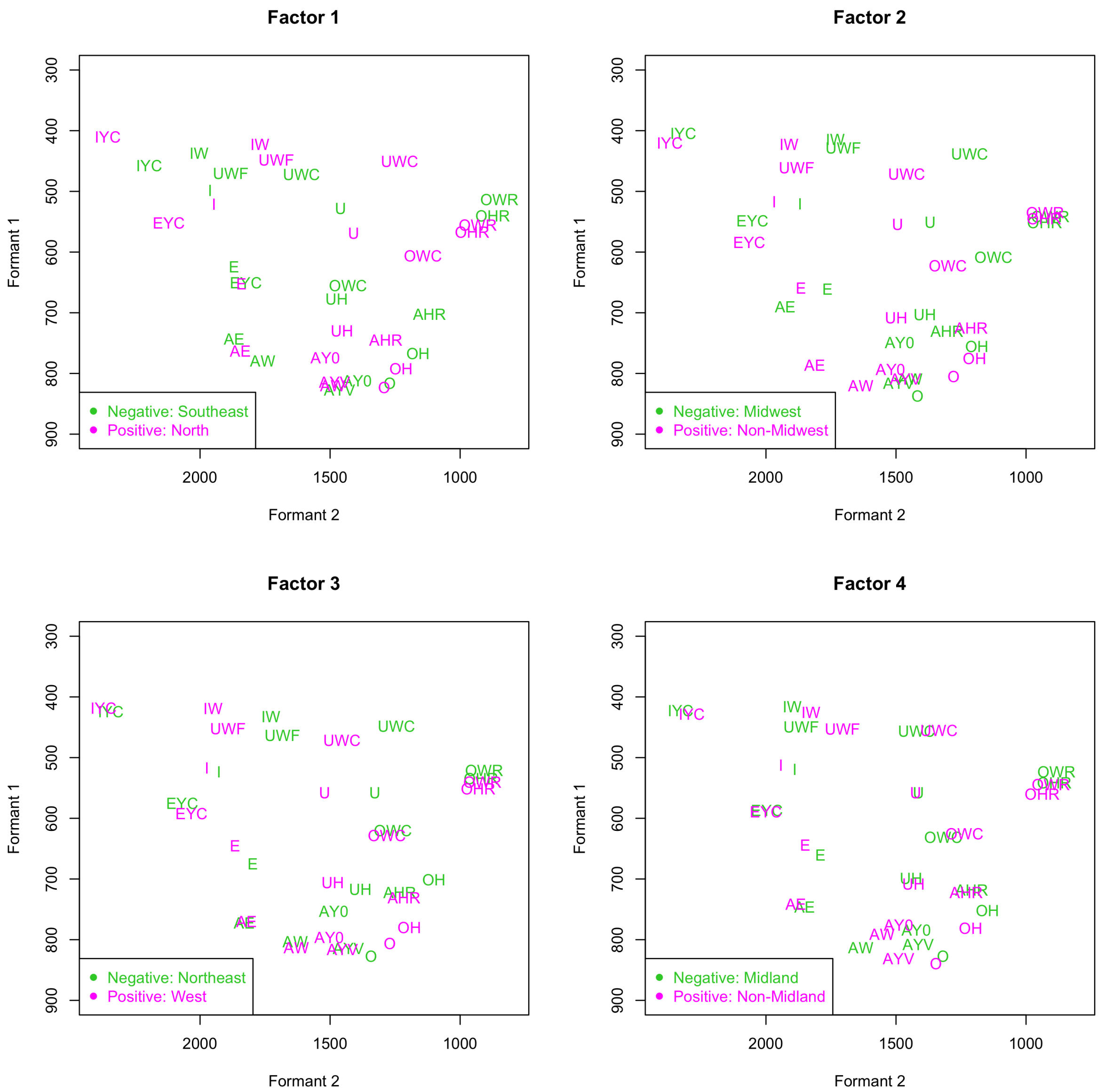
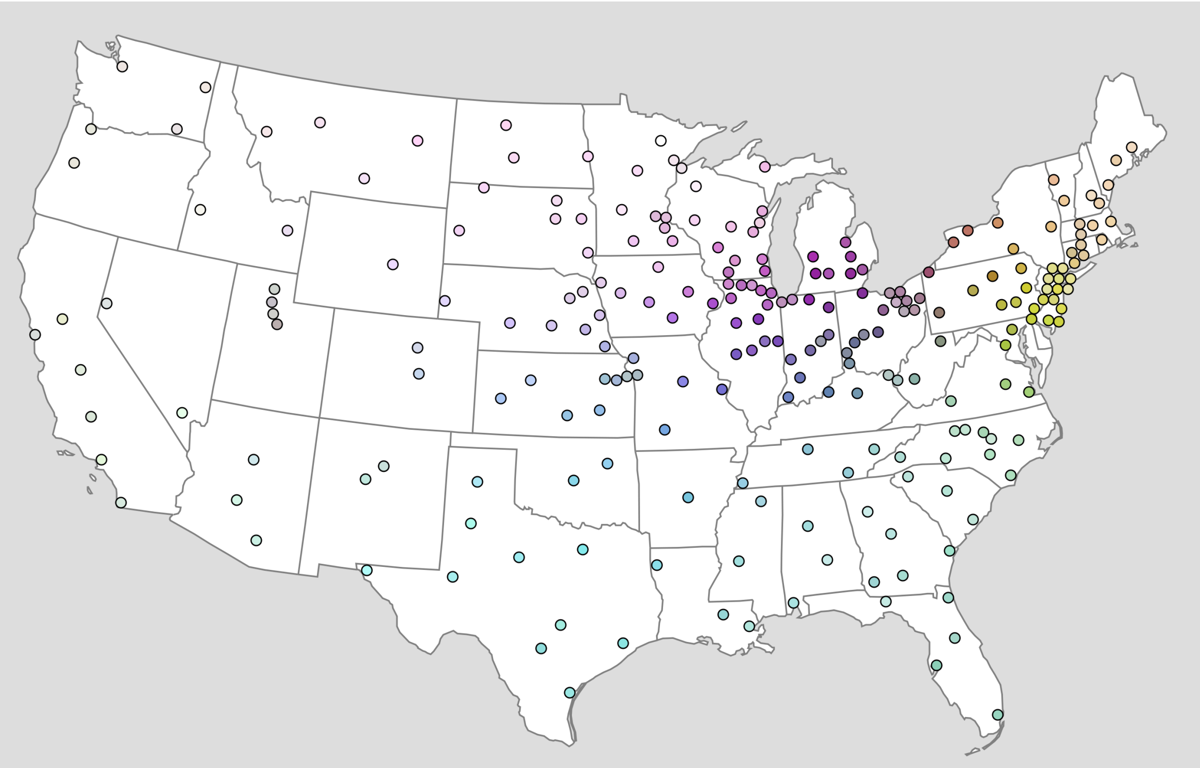


Figure 15 CMYK Map for Factors 1, 2, 3 and 4



5. Cluster Analysis

In addition to aggregating the four sets of factor scores using CMYK mapping to produce an overall map of continuous regional 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 (Hair et al, 2006), as is common in dialectometry (e.g. see Goebl, 2007; Nerbonne and Heeringa, 2009; Prokic and Nerbonne, 2008; Wieling and Nerbonne, 2010; Grieve et al, 2011). In particular, Ward's method for hierarchical clustering (Ward, 1963) was used because it tends to identify the clearest dialect regions and because it is one of the most common methods for hierarchical clustering in dialectometry.[6] 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. These clusters can then be mapped as dialect regions.

Subjecting the results of the factor analysis to a cluster analysis is therefore essentially a statistical method for identifying dialect regions based on how bundles of isoglosses divide a region.

Based on the dendrogram reproduced in Figure 16, the hierarchical cluster analysis identified 5 clear dialect regions: the Northeast, the Southeast, the West, the Lower Midwest, and the Upper Midwest, with the Upper and Lower Midwest further combining to form a Midwestern super-region. These 5 dialect regions are mapped in Figure 17 and the average vowel spaces for these five dialect regions are plotted in Figure 18. Although the cluster analysis also groups the Northeast with the Southeast, and the Midwest with the West, because these two super-regions are formed so late in the cluster analysis they are not particularly meaningful. On the other hand, the most distinct internal clusters within the major clusters identified above are important, although as one descends further down the dendrogram the clusters begin to lose spatial consistency. First, Texas and the South Central States are separated from the rest of the West. Second the Lower Midwest is divided into northern and southern sub-regions. Third, the Northeast is divided into New England and the Mid-Atlantic States. Fourth, the Upper Midwest is divided into northern and southern sub-regions.

Figure 16 HCA Dendrogram based on Factors 1, 2, 3 and 4 (Ward's Method)

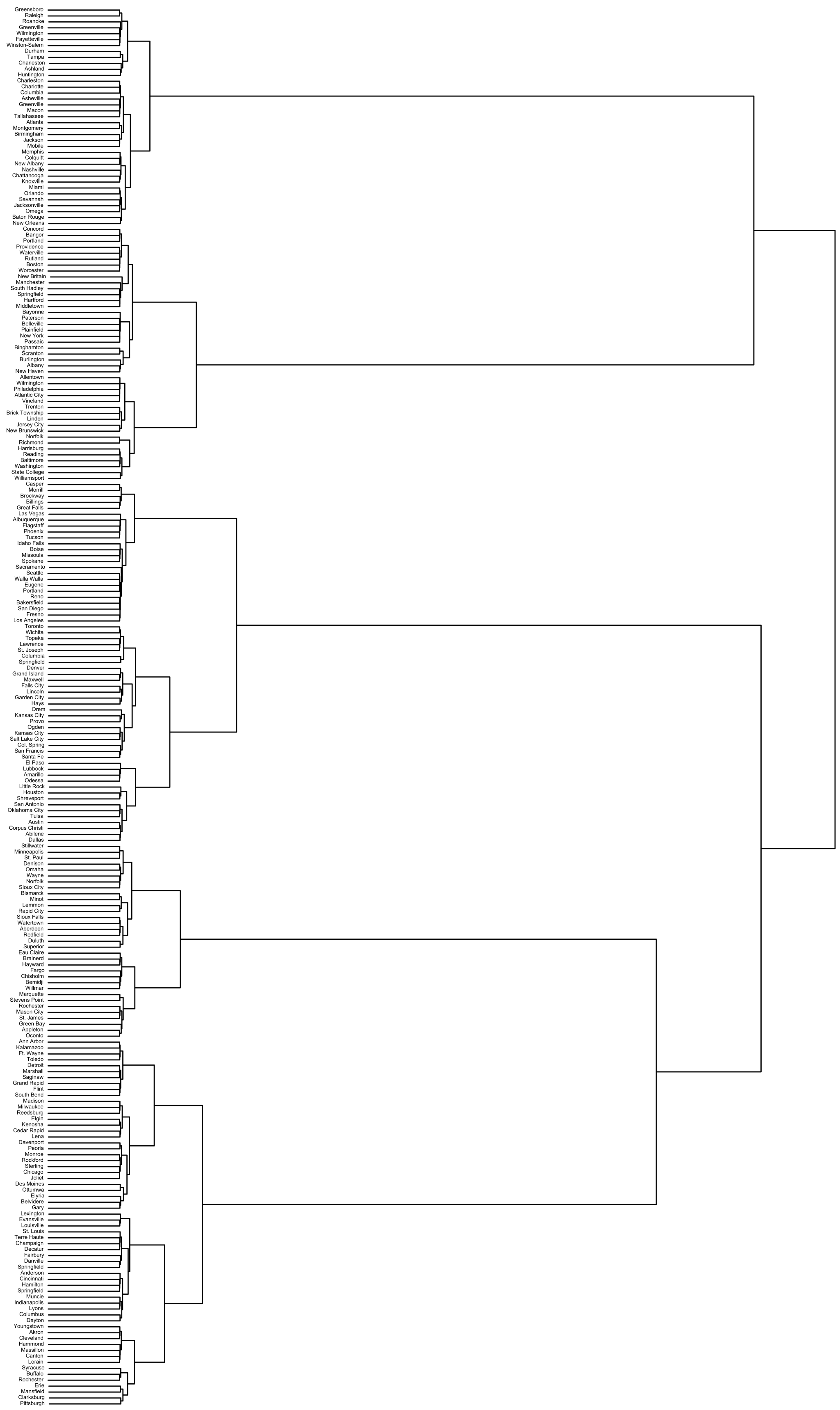


Figure 17 HCA 5 Cluster Map based on Factors 1, 2, 3 and 4

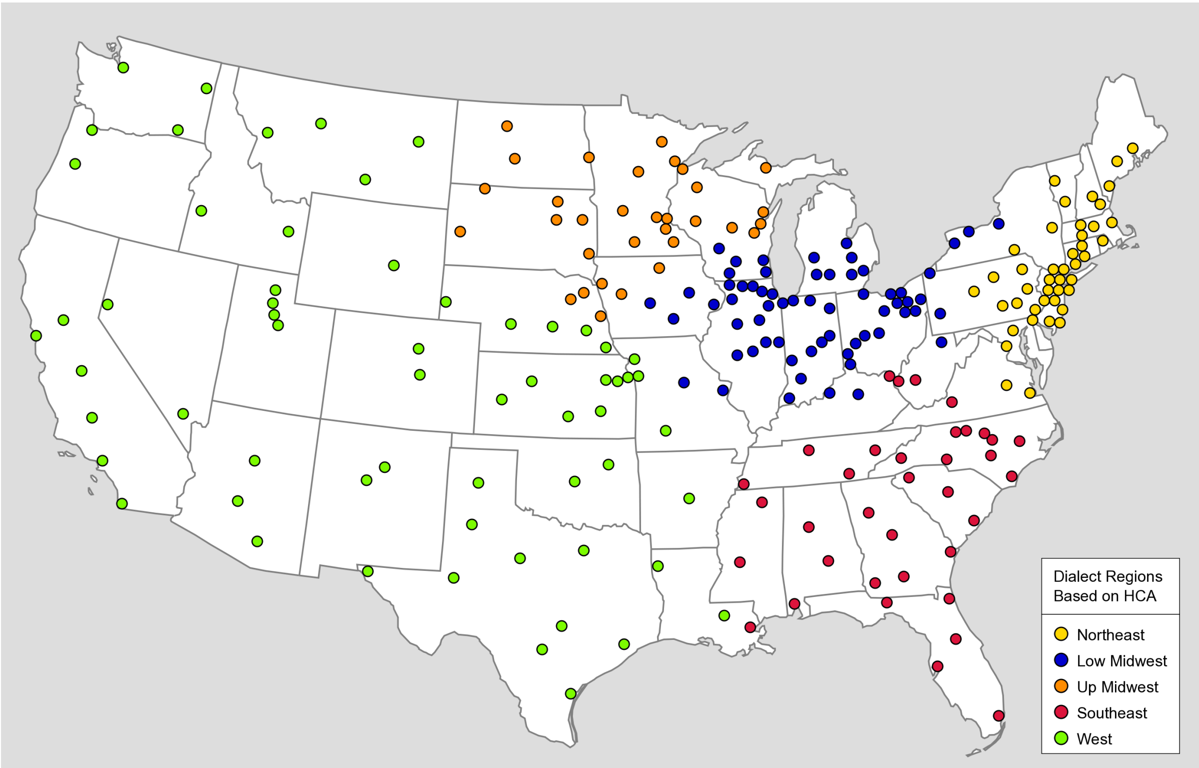
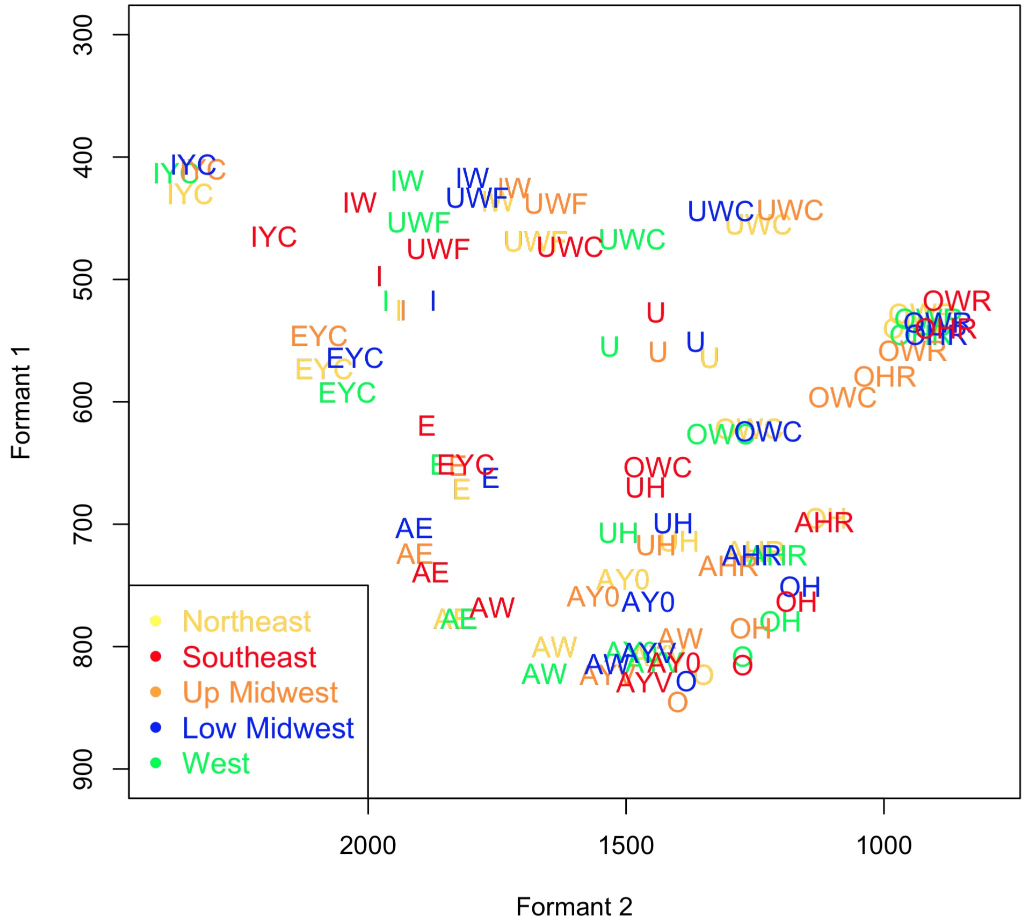


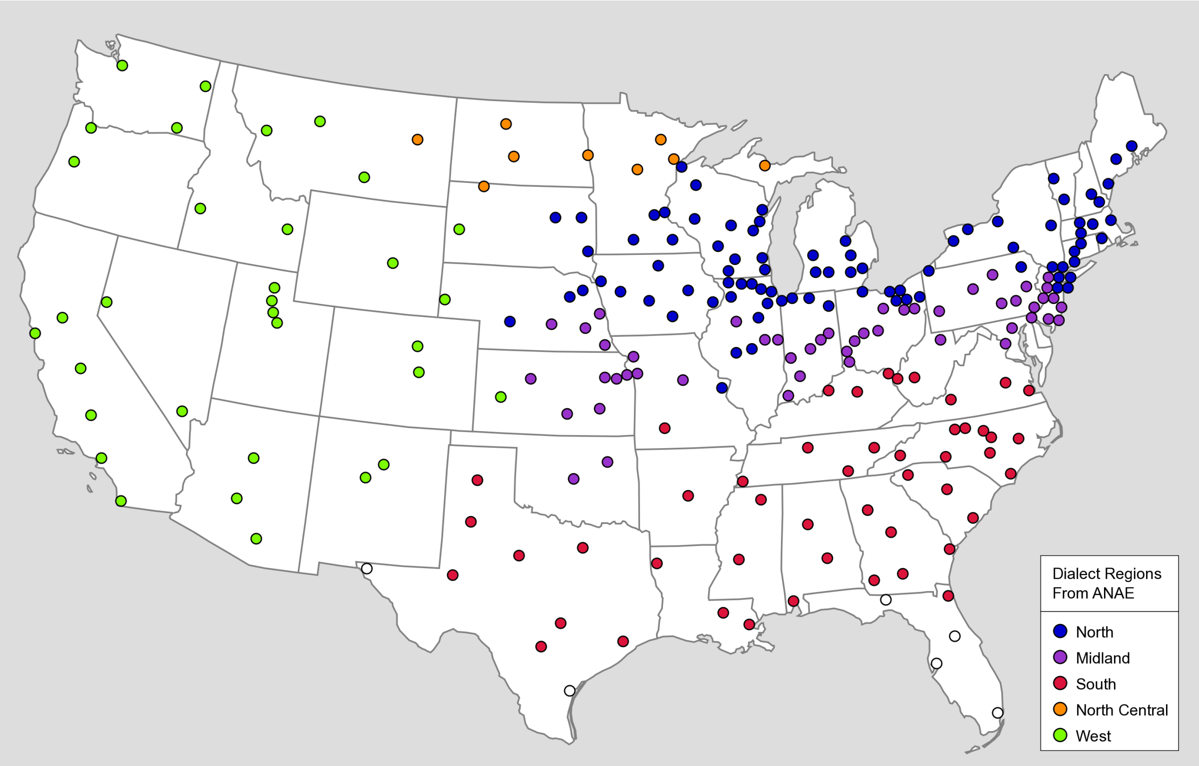
Figure 18 Average Vowel Spaces for the Five Clusters Identified by the Cluster Analysis



6. Discussion

The multivariate spatial analysis of the 38 vowel formant variables identified clear and consistent patterns of regional variation in American English. This basic result was to be expected—the *Atlas* has already shown that vowel formant variables are regionally patterned in American English—but the analysis presented here has identified a somewhat different picture of American dialect regions than was presented in the *Atlas*. For comparison, the approximate dialect regions identified in the *Atlas* are presented in Figure 19 based on Map 11.5 and Figure 11.9 from the *Atlas*.

Figure 19 Dialect Regions in the Atlas



The multivariate spatial analysis identified four major patterns of regional variation. Factor 1 contrasts the Southeast with the North (see Figure 10). The Southeastern region is similar to the Southern region identified in the *Atlas*, although the region identified here extends further north. The Northern region is also similar to the Northern region identified in the *Atlas*, although the region identified here is larger, stretching from the East Coast to the Southwest. Factor 1 also identifies an inland zone of transition that is similar to the Midland as identified in the *Atlas*. Factor 2 contrasts the Midwest with the East Coast and the West (see Figure 11). Unlike the Northern region identified by Factor 1, the Midwest stretches further south and does not extend to either coast. The *Atlas* does not identify a distinct Midwestern region, but the Inland North region identified in the *Atlas* is at the core of the Midwestern region identified here. Factor 3 contrasts the Northeast with the West, especially the Great Plains and the Mountain States (see Figure 12). The Western region is very similar to the Western region identified in the *Atlas*; however, no Northeastern region is recognized by the *Atlas*, which divides that part of the United States between the North and the Midland. Finally, Factor 4 contrasts the Midland with the rest of the United States (see Figure 13). The Midland region is similar to the Midland region identified in *Atlas,* although the region identified here is somewhat larger.

Based on these four common patterns of spatial clustering, two maps of American dialect regions were generated. First, a continuous picture of American dialect regions was generated through a CMYK mapping of the four sets of factor scores (Figure 15). Second, a categorical picture of American dialect regions was generated through a cluster analysis based on the four sets of factor scores (Figure 17), which is easier to interpret and is especially useful for comparing the results of this analysis to the dialect regions identified in the *Atlas* (see Figure 19). Overall, there are numerous similarities between these maps. Most notably, the West and the South are identified as dialect regions in both analyses and have very similar dimensions. The main difference between the two analyses lies in the Northeastern quarter of the country. The *Atlas* first divides the region north-to-south, identifying Northern and Midland dialect regions, whereas the multivariate spatial analysis first divides the region east-to-west.

The analysis presented here does not identify the Midland as a strong dialect region because three of the four factors, which together account for 80% of the regional variance in the dataset, separate the Mid-Atlantic from the Lower Midwest. Only Factor 4, which accounts for 6% of the regional variance in the dataset, combines the Mid-Atlantic with the Lower Midwest, which is the defining characteristic of the traditional Midland dialect region (e.g. see Kurath, 1949). Because this analysis is not based on the complete dataset analyzed in the *Atlas*[7]*,* the strength of the Midland may be underestimated here; however, most English vowels have been included in this analysis and it therefore appears that the distinction between the Northeast and the Midwest is stronger than the distinction between the North and the Midland.[8] Nevertheless, a weak Midland signal is present in this dataset. Given the consistent identification of a strong Midland dialect region in previous American dialect surveys, this result suggests that the traditional Midland dialect region is in the process of being replaced by Northeastern and Midwestern dialect regions.

While the patterns of regional linguistic variation identified by the multivariate spatial analysis differ in some ways from the patterns identified in the *Atlas*, it appears that these results can largely be explained by chain shifts (Gordon, 2002) as proposed in the *Atlas* *(*Labov, 2004; Labov et al, 2006). In particular, the first two common patterns of regional variation identified by the factor analysis, which account for a majority of the regional variance in the dataset, identify both the regions and the vowel formant variables associated with the Southern Shift and the Northern Cities Shift respectively.

The Southern Shift explains the majority of vowel formant variables that load on Factor 1 (see Table 3). As described in the *Atlas,* the Southern Shift begins with the fronting of /ay/, followed by the lowering and backing of /ey/ and /iy/, the raising and fronting of /i/, /e/ and /ae/, and the raising of /ahr/. In addition, the fronting of /uw/ and /ow/ and the raising of /ohr/ are sometimes associated with this shift. Aside from /ae/, all of these vowel formant measures load on Factor 1, although sometimes only on one of the predicted formants. The Southeastern region identified by Factor 1 is also characterized by a vowel space where all of these vowels have shifted as predicted by the Southern Shift, with the exception of /ay/ (see Figure 14). While three of the four /ay/ vowel formant variables do load on Factor 1, /ay/ is not fronted in the Southeastern vowel space as predicted by the Southern Shift, but rather it is lowered and backed. This is particularly surprising given that /ay/ fronting is the first step of the Southern Shift. In addition to these predicted vowel measures, /aw/ also loads on Factor 1 and is fronted in the Southeast. Although this vowel is not associated with the Southern Shift, its movement appears to be in line with the shift nonetheless. In addition, /u/ and /uh/ both load on Factor 1, with both vowels being higher in the Southeast, perhaps filling the space left behind by the fronting of /uw/.

Similarly, the Northern Cities Shift explains the majority of the vowel formant variables that load on Factor 2 (see Table 3). As described in the *Atlas*, the Northern Cities Shift begins with the fronting and raising of /ae/, followed by the fronting of /o/ and the lowering of /oh/, and the backing of /e/, /uh/ and /i/. Aside from /oh/, all of the vowel formant variables involved in the Northern Cities Shift load on Factor 2, although sometimes only on one of the predicted formants. The Midwestern region identified by Factor 2, which is centered around the Northern Cities, is also characterized by a vowel space where all of these vowels have shifted as predicted by the Northern Cities Shift, with the exception of /oh/, which is slightly higher in the Midwest (see Figure 14). Several additional vowel measures also load on Factor 2. In particular, /uwc/, /uwf/, /iw/, /u/, and /owc/ are all backed in the Midwestern vowel system. Although none of these variables are associated with the Northern Cities Shift, these movements appear to be related to the backing of /e/, /uh/, and /i/. In addition, /aw/ is backed, /ay0/ is raised, and /ahr/ is fronted in the Midwestern vowel system; however, the relationship, if any, between these changes and Northern Cities Shift is unclear.

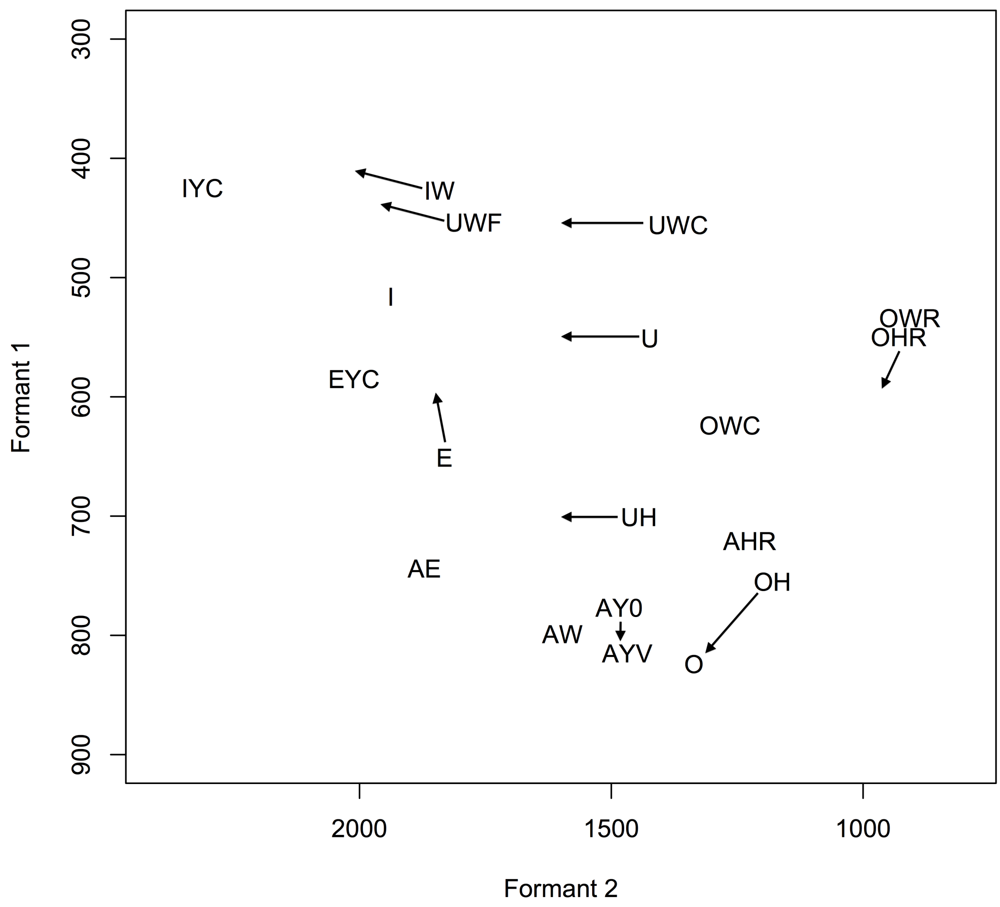
While the Southern and Northern Cities Shifts described in the *Atlas* explain the majority of vowel formant variables that load on the first two factors, no chain shift discussed in the *Atlas* can explain the variables that load on Factor 3 (see Table 3), which contrasts the West, especially the Great Plains and the Mountain States, with the Northeast. Most notably, the West is associated primarily with the lowering and fronting of /oh/ (resulting in the low-back merger with /o/) as well as the lowering and fronting of /ohr/, the fronting and slight raising of /iw/ and /uwf/, the lowering of /ay0/, the raising of /e/, and the fronting of /uh/, /u/ and /uwc/. All of these movements can be seen by comparing the average vowel space for the

Western region identified by Factor 3 to the average vowel space for the Northeastern region identified by Factor 3 (see Figure 14). Some of these features are identified by the *Atlas* as being characteristic of western speech, specifically the fronting of /uw/ (as well as the lack of fronting of /ow/ and /aw/, which accompany /uw/ fronting in the Southeast) and the fronting and lowering of /oh/ (resulting in the low back merger with /o/). However, the other vowel formant variables loading on Factor 3 are not listed as western features in the *Atlas* nor is a vowel shift in the Western United States identified by the *Atlas*.[9]

Despite the fact that the *Atlas* does not discuss the possibility of a distinct western shift, the Western Shift (or alternatively the Californian Shift) has been discussed in other studies, which have analyzed the language spoken in several western states including California (Hinton et al, 1987; Fought, 1999), Utah (Di Paolo, 1988), Nevada (Fridland, 2008), Arizona (Hall-Lew, 2004, 2005), Oregon (Conn, 2000; Ward, 2003), and Texas (Koops, 2010). This Western Shift is defined somewhat differently in these various studies but primarily involves the fronting of /uw/ and /ow/ (Fridland, 2008; Hall-Lew, 2004, 2005; Ward, 2003; Hinton et al, 1987), as well as the fronting of /u/ (Fridland, 2008; Hall-Lew, 2004, 2005; Koops, 2010; Fought, 1999), and occasionally the fronting of /o/ (Ward, 2003) and the raising of /ae/ (Hall-Lew, 2004, 2005; although cf. Conn, 2000).

The acoustic data from the *Atlas*, however, tells a somewhat different story. The fronting of /uw/ and /u/ are both identified by Factor 3 as being strong western features, but /ae/, /o/ and /ow/ do not load on Factor 3. In fact, /ae/ is at its lowest average position in the West and /o/ is near its backest average position in the West (see Figure 18). Furthermore, as discussed in the *Atlas*, the relative *stability* of /ow/ in the West is a defining feature of the region, compared to the Southeast, for example, where /uw/ fronting is accompanied by /ow/ fronting (as identified by Factor 1). However, as described above, numerous other vowel formant variables also load on Factor 3, which were neither identified by the *Atlas* as Western features or identified as being part of a Western Shift in previous research. In particular, the fronting of /uh/, /iw/ and /uwf/, the raising of /e/, and the lowering of /ay0/, all are identified as Western features by Factor 3 and all appear to be related to the fronting of /uw/ and /u/. Furthermore, it is possible that all of these changes are triggered by the low back merger (i.e. the fronting and lowering of /oh/), which is also identified by Factor 3 as a Western feature. Given that these vowel shifts appear to form a chain of interrelated changes, it appears that Factor 3 has in fact identified a distinct Western Shift that involves a large number of vowels that has only been partially observed in previous research. This putative Western Shift is diagrammed in Figure 20.

Figure 20 The Western Shift



It is important to note, however, that although the Western Shift identified here spans the West, it appears to be strongest in the Mountain States and the Great Plains, and weakest in the regions where most previous studies have been conducted, including California, Utah, Nevada, Arizona and Oregon. It is therefore possible that there are two related but distinct shifts underway in the West, both perhaps a reaction to the low back merger, with the Californian Shift operating on the West Coast and in the Southwest, and with a Central Shift, as identified by Factor 3, operating in the Mountain States and the Great Plains.

Finally, while chain shifts provide an internal explanation for the patterns identified by this analysis, the standard theory that American dialect regions correspond to historical settlement patterns does provide a satisfactory internal explanation for these results. Settlement patterns cannot account for the distinct Midwestern and Northeastern dialect regions identified in this study because both regions were settled by people from the New England and Midland settlement hearths. On the other hand, the four major dialect regions identified here, including the Northeast and the Midwest, clearly correspond closely to the four major modern American cultural regions. This finding supports a cultural theory of American dialect regions, where American dialect regions correspond to contemporary cultural regions (see Grieve, 2009). This cultural theory of American dialect regions also explains the decline of the Midland dialect region and the emergence of Midwestern and Northeastern dialect regions identified in this study, which follow the apparent decline of the traditional Midland cultural region and emergence of the modern Midwestern and Northern cultural regions.

End Notes

1 The imputation of missing data has very little effect on the results of the analysis because the missing data is minimal (0.02% of the total data) and because the variables with missing data follow the same basic patterns as the other vowel formant variables.

2 Each formant for a vowel is analyzed individually because vowels can pattern differently on the two formants.

3 It is notable that formant 2 variables tend to exhibit higher levels of global spatial autocorrelation than formant 1 variables.

4 A similar technique, known as a principal component analysis, was used in the *Atlas* to cluster informants; however, the analysis was based on the raw variables, the component loadings were not reported, and the component scores were not mapped. A factor analysis was used instead of principal component analysis in this study to focus on the identification of common patterns of regional variation (see Nerbonne, 2006; Grieve et al, 2011). However, a principal component analysis of this dataset would have produced similar results.

5 In all cases vowel measures that load strongly on a particular factor change position in the pair of opposing vowel spaces identified by that factor (see Figure 14). Nevertheless, vowel measures can also change position in the opposing vowel spaces identified by factors upon which they do not load if that change is smaller than the change in the opposing vowel spaces identified by factors upon which they do load.

6 There are several other possible hierarchical clustering algorithms that could have been applied (e.g. see Heeringa, 2004). The main reason why Ward's method is used here (and in many dialectometry studies) is because it tends to identify contiguous dialect regions, which is the goal of the cluster analysis, whereas most other clustering algorithms tend to identify clusters that include a relatively large numbers of geographic outlier locations. In this case, it is completely acceptable to select the clustering algorithm that gives the best results. This is

because a cluster analysis does not test if there are dialect regions in a regional linguistic dataset; dialect regions are assumed to exist (in this case based on the results of the spatial autocorrelation analyses) and the cluster analysis is used to identify their location. Consequently, clustering algorithms that tend to identify contiguous clusters of locations should generally be preferred to clustering algorithms that do not tend to identify contiguous clusters of locations.

7 The analysis presented here is based only on the most complete acoustic vowel data available, whereas the *Atlas* is based on vowel formant variables that were excluded from this analysis due to missing data (see Section 2), as well as additional phonetic and phonological measures.

8 The analysis presented here also aligns closely with the analysis of lexico-grammatical variation in a modern corpus of written American English (see Grieve et al, 2009; Grieve, 2013b).

9 The vowel formant variables loading on Factor 3 were interpreted as identifying a Western Shift as opposed to a Northeastern Shift because the low back merger, which is identified by Factor 3 as characteristic of the Western region, is known to be a change in progress. However, it is also possible that Factor 3 has identified distinct Western and Northeastern features, including perhaps a Northeastern Shift.

References

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

Cliff AD, Ord JK. (1973). *Spatial Autocorrelation*. London: Pion.

Cliff AD, Ord JK. (1981). *Spatial Processes: Models and Applications*. London: Pion.

Conn J (2000). *The Story of /ae/ in Portland*. MA Thesis. Manuscript. Portland State University.

Di Paolo M (1988). Pronunciation and categorization in sound change. In Ferrara K, Brown B, Walters K, Baugh J (eds.), *Linguistic Change and Contact: Proceedings of the Sixteenth Annual Conference on New Ways of Analyzing Variation,* 84-92. Austin, TX: Department of Linguistics, University of Texas.

Fought C. 1999. A majority sound change in a minority community: /u/-fronting in Chicano English. *Journal of Sociolinguistics* 3: 5-23.

Fridland V. (2008). Patterns of /uw/, /upsilon/, and /ow/ fronting in Reno, Nevada. *American Speech* 83: 432- 454.

Goebl H. (1982). *Dialektometrie; Prinzipien und Methoden des Einsatzes der numerischen Taxonomie im Bereich der Dialektgeographie*. Vienna: Verlag der Osterreichischen Akademie der Wissenschaften.

Goebl H. (2006). Recent Advances is Salzburg Dialectometry. *Literary and Linguistic Computing* 21: 411-435.

Gordon MJ. (2002). Investigating chain shifts and mergers. In Chambers J, Trudgill P, Schilling-Estes N (eds.), *The Handbook of Language Variation and Change,* 245-266. London: Blackwell.

Grieve J. (2009). *A Corpus-Based Regional Dialect Survey of Grammatical Variation in Written Standard American English.* Ph.D. Dissertation. Manuscript. 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. (2013a). A comparison of statistical methods for the aggregation of regional linguistic variation. Forthcoming in Szmrecsanyi B., Wälchli B. (eds.), A*ggregating Dialectology and Typology: Linguistic Variation in Text and Speech, Within and Across Languages*. Berlin: Walter de Gruyter.

Grieve J. (2013b). A statistical comparison of regional phonetic and lexical variation in American English. Forthcoming in *Literary and Linguistic Computing*.

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

Hair J, Black B, Babin B, Anderson RE, Tatham RL. (2006). *Multivariate Data Analysis* (6th edition). Englewood Cliffs, NJ: Prentice-Hall. Englewood Cliffs, NJ: Prentice-Hall.

Hall-Lew L. (2004). The western vowel shift in Northern Arizona. Unpublished Manuscript. Available at http://www.lel.ed.ac.uk/~lhlew/index.html

Hall-Lew L. (2005). One shift, two groups: when fronting alone is not enough. In Baranowski M, Horesh U, Evans K, Nguyen G (eds.), *Selected Papers from NWAV 32,* 105-116. Philadelphia PA: University of Pennsylvania Working Papers in Linguistics 10.2.

Heeringa W. (2004). Measuring Dialect Pronunciation Differences using Levenshtein Distance. Ph.D. Dissertation. Manuscript. University of Groningen.

Hinton L, Moonwoman B, Bremner S, Luthin H, Van Clay M, Lerner J, Corcoran H. (1987). It's not just valley girls: a study of California English. In Aske J, Beery N, Michaelis L, Filip H (eds.), *Proceedings of the Thirteenth Annual Meeting of the Berkeley Linguistics Society,* 117-127. Berkeley CA: Berkeley Linguistics Society.

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

Labov W. (2004). *Principles of Linguistic Change, Internal Factors.* Wiley-Blackwell.

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

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

Nerbonne J. (2006). Identifying Linguistic Structure in Aggregate Comparison. *Literary and Linguistic Computing* 21: 463-476.

Nerbonne J, Heeringa W. 2009. Measuring dialect differences. In Schmidt JE, Auer P (editorss), *Language and Space: Theories and Methods*. Berlin: Mouton De Gruyter. 550-567.

Odland JD. (1988). *Spatial Autocorrelation*. Sage Publications.

Ord JK, Getis A. (1995). Local Spatial Autocorrelation Statistics: Distributional Issues and an Application. *Geographical Analysis* 27: 286-306.

Prokic J, Nerbonne J. (2008). Recognizing groups among dialects. *International Journal of Humanities and Arts Computing* 1: 153-172.

Séguy J. (1971). La relation entre la distance spatiale et la distance lexicale. *Revue de linguistique romane,* 35: 335-357.

Séguy J. (1973). La dialectométrie dans l'Atlas linguistique de la Gascogne. *Revue de linguistique romane,* 37: 1-24.

Ward JH. 1963. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association* 58: 236-244.

Ward M. 2003. *Portland Dialect Study: The Fronting of /ow, u, uw/ in Portland, Oregon.* MA Thesis. Manuscript. Portland State University.

Wieling M, Nerbonne J. (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 NLP 16. 33-41.