

Analyzing Phonetic Variation in the Traditional English Dialects: Simultaneously Clustering Dialect and Phonetic Features

^aMartijn Wieling*, ^bRobert G. Shackleton, Jr. and ^aJohn Nerbonne

^aUniversity of Groningen, The Netherlands; ^bCongressional Budget Office, United States of America

*Corresponding author: Martijn Wieling, Postbus 716, 9700 AS Groningen, The Netherlands, m.b.wieling@rug.nl

Abstract

This study explores the linguistic application of bipartite spectral graph partitioning, a graph-theoretic technique that simultaneously identifies clusters of similar localities as well as clusters of features characteristic of those localities. We compare the results using this approach to previously published results on the same dataset using cluster and principal component analysis (Shackleton, 2007). While the results of the spectral partitioning method and Shackleton's approach overlap to a broad extent, the analyses offer complementary insights into the data. The traditional cluster analysis detects some clusters which are not identified by the spectral partitioning analysis, while the reverse also occurs. Similarly, the principal component analysis and the spectral partitioning analysis detect many overlapping, but also some different linguistic variants. The main benefit of the bipartite spectral graph partitioning method over the alternative approaches remains its ability to *simultaneously* identify sensible geographical clusters of localities with their corresponding linguistic features.

Introduction

A great deal of language variation is conditioned geographically, giving rise to geographic dialects, which have been studied in dialectology for well over a century (Chambers and Trudgill, 1998). Dissatisfaction with dialectology's tendency to focus on details gave rise in the 1970's to dialectometry (Séguy, 1973; Goebel, 1984), which systematizes procedures and obviates the need for feature selection, at least to some extent. Nerbonne (2009) argues that dialectometry has been successful because of its emphasis on measuring aggregate levels of differentiation (or similarity), strengthening the geographic signals in the linguistic data, which are often complex and at times even contradictory. The professional reception of dialectometry has been polite but less than enthusiastic, as some scholars express concern that its focus on aggregate levels of variation ignores the kind of linguistic detail that may help uncover the linguistic structure in variation. For this reason there have been several recent attempts to supplement (aggregate) dialectometric techniques with, on the one hand, techniques to identify linguistic variables which tend to be strongly associated throughout geographic regions and, on the other hand, techniques to extract prominent linguistic features that are especially indicative of aggregate differentiation.

Grieve, Speelman and Geeraerts (2011) analyzed a large dataset of written English with respect to lexical variation. They used spatial autocorrelation to detect significant geographical patterns in 40 individual lexical alternation variables, and subsequently applied factor analysis to obtain the importance of individual lexical alternation variables in every factor (which can globally be seen as representing a geographical area). In the following step, they applied cluster analysis to the factor scores in order to obtain a geographical clustering.

Shackleton (2007) used cluster analysis and principal component analysis (PCA) to identify linguistic variables which tend to correlate when compared across many localities. We illustrate the basic idea with an example: if the localities in which a standard /æ/ is raised to [ɛ] tend to be the same as those in which /e/ is also raised (to [eɪ]), then a good cluster analysis should identify a cluster of localities that share those variables, while PCA should identify a principal component which is common to the two linguistic variables. Shackleton (2007) identified several interesting clusters and components, which we discuss below at greater length.

Wieling and Nerbonne (2010) use bipartite spectral graph partitioning (BiSGP), a graph-theoretic technique, which clusters localities on the basis of the features they share and features on the basis of the localities in which they occur. To continue with the example in the last paragraph, a good BiSGP would identify the two variables as associated and also the sites in which this (and other) associations are evident. From a dialectometric point of view BiSGP is attractive in attributing a special status to features as well as to localities, but like all procedures for seeking natural groups in data, it needs to be evaluated empirically.

In this study, we apply BiSGP to Shackleton's (2007) data. We compare these results to those on the basis of cluster analysis and PCA reported by Shackleton (2007).

Dataset

In this study we use the dataset described by Shackleton (2007), derived mainly from Anderson's (1987) *A Structural Atlas of the English Dialects* (henceforth *SAED*). The *SAED* contains more than 100 maps showing the geographic distribution and frequency of occurrence of different phonetic variants in groups of words found in the *Survey of English Dialects* (Orton and Dieth, 1962; henceforth *SED*), the best broad sample of the most traditional dialect forms that were still

in use in 313 rural localities throughout England the mid-20th century. The dataset assembled from the *SAED* maps classifies over 400 responses from the *SED* by assigning each to one of 39 groups. All of the words in a given group include a segment or combination of segments that is believed to have taken a single uniform pronunciation in the ‘standard’ Middle English dialect of the Home Counties of southeastern England. The segments include all of the Middle English short and long vowels, diphthongs, and most of the relatively few consonants that exhibit any variation in the English dialects. For each idealized Middle English pronunciation, in turn, the responses may take any of several 20th-century pronunciations – and, in any given location, may take different pronunciations for different words in the group. The dataset thus tabulates frequencies of use of a total of 209 different variant pronunciations of the 39 idealized phonemes. For example, one group includes a number of words, such as *root* and *tooth*, all of which included a segment pronounced /o:/ in Middle English. Several maps are associated with that group, one for each modern variant: one of the maps shows the frequency with which that segment is pronounced as [u:] (that is, the percentage share of the words with the vowel articulated as [u:]) in each locality in the *SED*; another shows the frequency with which the segment is pronounced as [y:], and so on. (Throughout the presentation, we write the Middle English form considered common to the group as /x/ and the variants recorded in the *SED* as [x].) A complete list of variants is given by Shackleton (2010: 180-186).

In a few cases, Anderson classified localities from geographically separate regions as having ‘different’ variants even though the variants are actually the same, on the grounds that the variant is likely to have arisen independently in the two regions. Moreover, many maps actually show a range of distinguishable pronunciations that Anderson somewhat arbitrarily took to be similar enough to be classified into a single variant. Although it tends to understate the true range

of variation in the speech it characterizes, the dataset summarizes a large body of phonetic information in a tractable form that enables straightforward quantitative analyses of phonetic variation in the traditional English dialects.

Most variants have a relatively unique distribution among and frequency of use within localities, and very few large geographic correlations with others. Variants with a large number of high geographic correlations with each other are found either in the far Southwest or in the far North of England, suggesting that those regions tend to have relatively distinctive speech forms with several features that regularly co-occur in them (exemplified by the very similar geographic distributions of voiced fricatives in the Southwest). The comparative lack of geographic correlation raises challenges for analytic techniques, such as the bipartite spectral graph partitioning presented here, that seek to identify groups of linguistic features that can be said to characterize regional dialects.

Methods

Clustering Varieties and their Variants Simultaneously

We use hierarchical spectral partitioning of bipartite graphs (Wieling and Nerbonne, 2010) to *simultaneously* identify the geographical clusters in the dataset as well as their characteristic linguistic variants. A bipartite graph is a graph which has two sets of vertices (one representing geographical varieties and the other linguistic variants) and a set of edges connecting vertices from one set to the other set (each edge represents the occurrence of the variant in a variety). Note that no other edges, e.g. between varieties, are allowed. Hierarchical spectral partitioning refers to the hierarchical clustering method, which is based on calculating the singular value decomposition of the input matrix. For an extensive mathematical explanation of the bipartite

spectral graph partitioning method, we refer to Wieling and Nerbonne (2010) and Wieling and Nerbonne (2011).

In this study, the bipartite graph is represented by a geographic variety \times linguistic variant matrix where every position in the table marks the relative variant frequency as used by Shackleton (2007) in his principal component analysis. To ensure every variant carries comparable weight in the analysis, we scaled all individual columns of the matrix (relative variant frequency) between zero and one; that is, for each variant, all of the relative frequencies are divided by the highest relative frequency for that variant.¹ This approach potentially places greater emphasis than other approaches on regionally distinctive but comparatively uncommon variants.²

Determining the Most Important Variants for Every Cluster

As every cluster will contain many varieties and corresponding variants, and we are interested only in the most important linguistic variants for every geographical cluster, we need a method to distinguish the most variants. Following Wieling and Nerbonne (2011), we define the importance of a variant in a cluster as a linear combination of two measures, distinctiveness and representativeness. The representativeness of a variant measures how frequently the variant occurs within the cluster. E.g. if there are ten varieties in the cluster and the variant occurs only in four varieties, the representativeness equals 0.4 (four divided by ten). The distinctiveness of a variant measures how frequently the variant occurs within a cluster as opposed to outside of the cluster, taking the relative size of the cluster into account. For example, if the variant does not occur outside of the cluster, the distinctiveness is one (the variant perfectly distinguishes the cluster from the rest), no matter how large the cluster. Alternatively, if a cluster contains 50% of

the geographic varieties and 50% (or less) of the total variant frequency, the distinctiveness is zero (the variant does not distinguish the cluster at all). The values of distinctiveness (ignoring uninteresting cases) and representativeness both range between zero and one.

Normally representativeness and distinctiveness are averaged to obtain the importance score for every variant (higher is better), but it is also possible to assign different weights to representativeness and distinctiveness. When the input matrix contains many variants which occur almost everywhere, representativeness will be very high for these (non-informative) sound correspondences. In that case it makes sense to weight distinctiveness more heavily than representativeness. Alternatively, if there are many variants occurring only in a few varieties, the distinctiveness of these (non-informative) variants will be very high. In that case, it makes sense to weight representativeness more heavily than distinctiveness. As our matrix contained many frequent variants, we weighted distinctiveness twice as heavily as representativeness.

Results

Applied to the data from Shackleton (2007), the BiSGP analysis initially distinguishes northern and southern English dialects along a line that roughly traces the border separating Northern from Midlands and Southern dialects in Shackleton's (2007) cluster analysis. Figure 1 shows this division. The southern region includes 198 (63%) of the localities and 123 (62%) of the variants, while the northern region includes the remaining 115 (37%) localities and 76 (38%) variants. A few of the roughly 70 high-scoring southern variants are widely found throughout the region, including those reflecting movements or lengthening of the Middle English short vowels, but most are members of several groups that, on closer inspection, tend to be restricted to sub-regions of the south; these include upgliding diphthongization of the Middle English long vowels

(e.g. [ɛin] or [læin] for *lane*) characteristic of the Great Vowel Shift and occurring mainly in the southeast, fricative voicing and retention of rhotics ([vɑrm] for *farm*) largely in the southwest, and fronting of many vowels ([ny:n] for *noon*) in Devonshire. The roughly 50 high-scoring northern variants are similar in that some reflect widely distributed conservative retentions of the Middle English short vowels (e.g. [man] for *man*) along with more restricted ingliding diphthongs for some Middle English long vowels ([liən] for *lane*), limited retention of rhotics, and fronting of some vowels ([bø:n] for *bone*) in the far north.

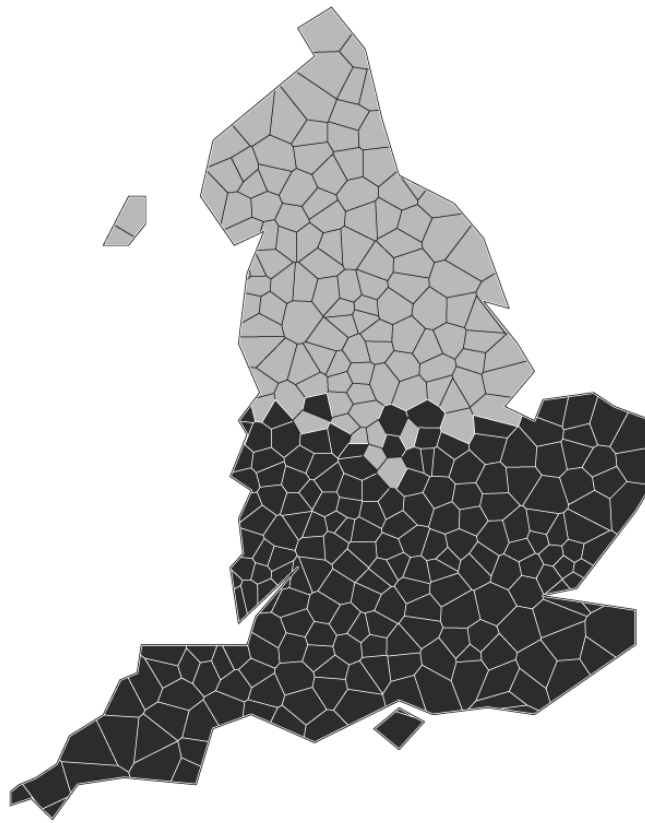


Fig. 1 Bipartite spectral graph partitioning in two groups

The second round of partition of the north and south divides England into four separate regions, somewhat more clearly reflecting regionally coherent distributions of variants. Figure 2

shows this division. A small far northern region emerges restricted mainly to Northumbria, with 11 localities and 21 variants including retained rhotics ([r] and, in one location in Cumberland, [r̥]) and aspirates as well as fronting or ingliding of Middle English long vowels ([liən] for *lane*), while the rest of the north – 104 localities with 55 variants – includes a number of other features irregularly distributed through that region. The southwest – 51 localities with 42 variants – includes the voiced fricatives characteristic of the entire region (e.g. [vɑrm] for *farm*) as well as the fronted vowels characteristic only of Devonshire ([ny:n] for *noon*), while the southeast – 147 localities with 81 variants – includes the upgliding diphthongization of the Middle English long vowels characteristic of much of the southeast ([leɪn] or [lain] for *lane*) as well as a number of more sporadically occurring variants.

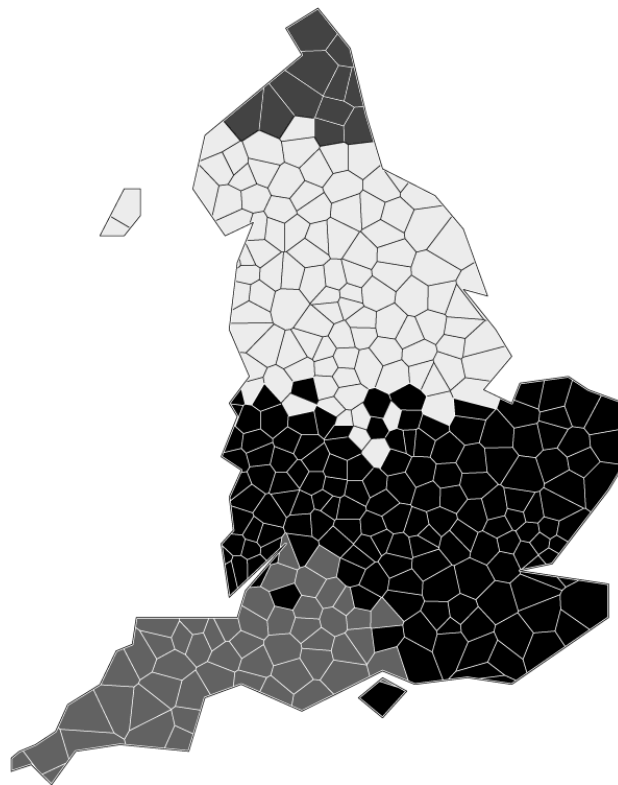


Fig. 2 Bipartite spectral graph partitioning in four groups

A further round of partition into eight regions (shown in Fig. 3) yields yet more coherent distributions in the north and south. In the far north, a single locality in Cumberland is distinguished by its alveolar trill [r] (marked with number 1 in Fig. 3), with the rest of the far north (marked with number 2 in Fig. 3) characterized by the ‘Northumbrian burr’ (a uvular trill [ʀ]), retained aspirates, and fronting or ingliding of Middle English long vowels. Most of the remaining northerly localities – 82 localities with 44 variants (marked with number 3 in Fig. 3) – have irregular distributions of variants, but an irregularly shaped region of 22 localities (marked with number 4 in Fig. 3) centered on Staffordshire and Derbyshire is associated with 9 unusual variants that, on closer examination, include 5 of the 8 variants that Trudgill (1999) associates together as characteristics of a regional ‘Potteries’ dialect: [ti:l] for *tail*, [bɛɔt] for *boot*, [dain] for *down*, [ʃeip] for *sheep*, and [koɔt] for *caught*. In the southwest, 13 localities in Devonshire and Cornwall (marked with number 8 in Fig. 3) are associated with 14 variants, mainly fronting of Middle English back vowels ([ny:n] for *noon*) and the development of a low monophthong for Middle English /i:/ ([na:f] for *knife*), while the remaining 38 localities (two regions marked with number 7 in Fig. 3) are associated with 28 other variants, the highest scoring of which nearly all involve the voicing of fricatives and the retention of a retroflex rhotic ([vaɾm] for *farm*). Much of the southeast (marked with number 6 in Fig. 3) – 69 localities with 38 variants – is associated with the upgliding diphthongization of the Middle English long vowels ([leɪn] or [lain] for *lane*) and particularly strong movements of Middle English short vowels (e.g. [mɛn] for *man*), as well as a number of less extensively distributed variants such as those restricted mainly to East Anglia. The rest of the south, including most of the West Midlands (marked with number 5 in Fig. 3) – 78 localities with 43 variants – is associated only quite loosely with a wide variety of variants that are generally distributed throughout a much wider region or are found only in more

isolated regions. The highest-scoring variants in this region, for example, include the development (mainly in the Severn Valley) of a back unrounded vowel [ɑ:] in *daughter*, *law*, and *cough*, [faiv] for *five* mainly in Shropshire, and the palatalization of Middle English /ɛ:/ ([bjʌnz] for *beans*) in the Southwest Midlands.

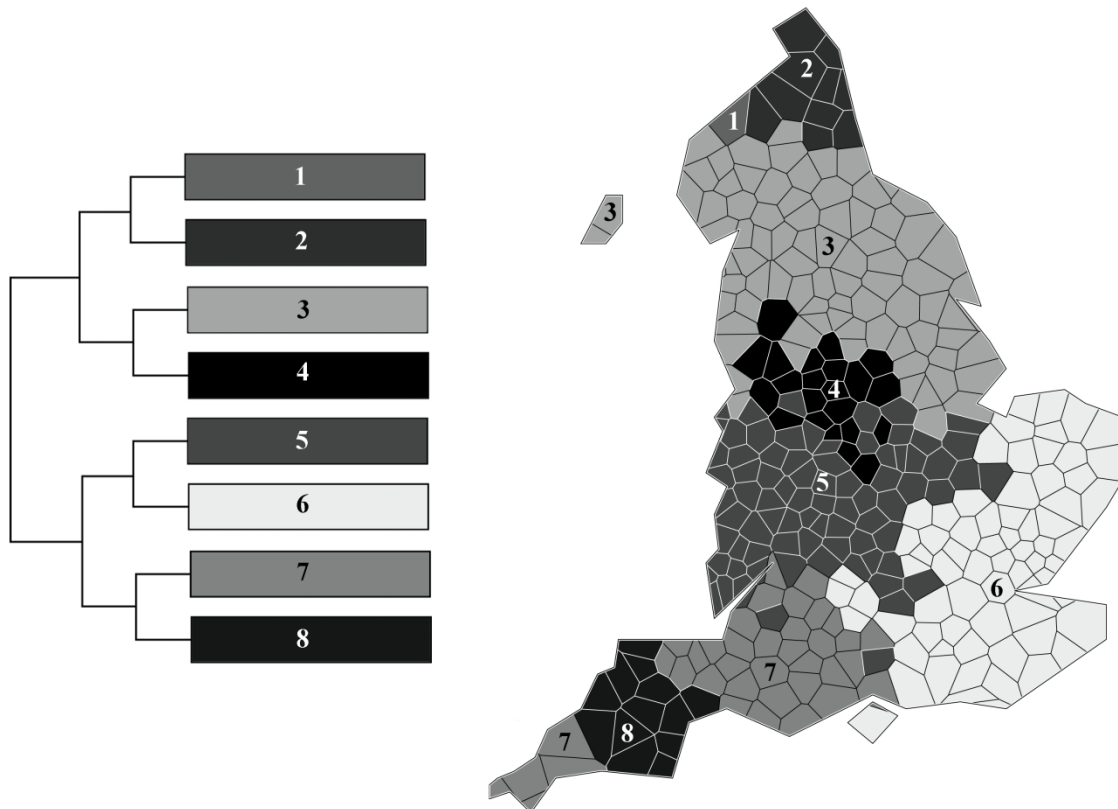


Fig. 3 Bipartite spectral graph partitioning in eight groups including the hierarchy

Comparison to Traditional Cluster Analysis

The results from the BiSGP analysis can be usefully compared with those that emerge from Shackleton's (2007) cluster analysis of the same data, thus illustrating the comparative strengths of the two approaches. In contrast to the bipartite spectral graph partitioning approach described here, cluster analysis may use a variety of techniques to group localities on the basis of some

measure of the aggregate similarity of the localities' patterns of usage rather than optimizing over a balance of representativeness and distinctiveness. Shackleton (2007) applied several different clustering techniques to the English data set and combined them into a single site \times site table of mean cophenetic differences (i.e. distances in dendrograms). He then used multidimensional scaling to reduce the variation in the results to a relatively small, arbitrary number of dimensions that summarize fundamental relationships in the data. For visualization purposes the variation is reduced to three dimensions, which can be mapped onto the RGB color spectrum. The resulting pattern, shown in Fig. 4, shows many similarities to the eight regions resulting from the BiSGP analysis.

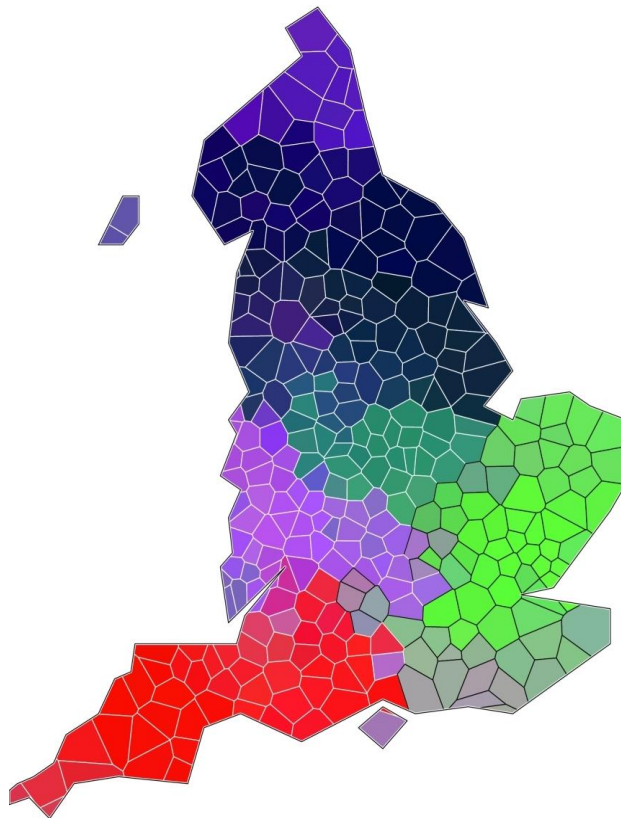


Fig. 4 Multidimensional scaling of cluster analyses. The original image was taken from Shackleton (2010).

As mentioned above, the demarcation of northern and southern dialect regions is similar to Shackleton's delineation of Northern dialects from Midlands and Southern dialects except that the BiSGP analysis classifies a few localities in Shackleton's transitional Central Midlands region into the north. The BiSGP analysis distinguishes almost exactly the same southeastern and southwestern regions as Shackleton on the basis of very similar sets of dialect features, and does the same for the Northumberland region, except that the BiSGP analysis isolates the single locality in Cumberland by its rhotic trill [r]. Those peripheral regions of the English dialect landscape tend to be distinguished by distinct sets of variants that have comparatively coherent geographic distributions – rhotics and aspirates in the far north, fricative voicing in the southwest, fronting in Devonshire, and the particularly strong upgliding diphthongization found in the southeast – and that are therefore relatively straightforward to identify.

Differences arise in the two analyses' delineation of dialect regions in the lower north and much of the Midlands, where the various traditional dialect developments tend to be less coherently or much more locally distributed. For example, the BiSGP analysis groups together Shackleton's Upper Southwest with most of his Central Midlands region and consequently does not detect a region corresponding with the Central dialect region as identified by Trudgill (1999), whereas the cluster analysis (partly) does (Shackleton, 2007).

Interestingly, however, the BiSGP analysis identifies a region in the northwest Midlands, centered on Staffordshire and Derbyshire, on the basis of a number of variants associated by Trudgill (1999) with the 'Potteries' region that Shackleton's analysis consistently fails to distinguish.

Comparison to Principal Component Analysis

The regions resulting from the BiSGP analysis can also be usefully compared to those isolated by Shackleton's varimax principal component analysis of the data, illustrating the comparative strengths of those approaches. In contrast to the BiSGP's focus on the representativeness and distinctiveness of variant usage in localities, principal component analysis identifies groups of variants that are strongly positively or negatively correlated – that is, that tend to occur together or that always occur separately – and combines them into principal components that are essentially linear combinations of the correlated variables. A principal component typically has two 'poles', one involving large positive values for a group of variables that tend to be found together, and another involving large negative values for a different group of variables that are also found together but never with the first group. (Varimax rotation tends to sharpen the focus and concentration of each component by increasing the loading on its most highly correlated variants, and when applied to linguistic data, tends to yield groups of variants that are more readily interpretable in linguistic terms.)

Localities can be assigned component scores that indicate the extent to which the variants in a given principal component appear in that particular locality, and in many cases a group of localities may have sufficient geographic cohesion to suggest a dialect region identified by the variants with high scores in that component. Indeed, principal component analysis of our data set identifies groups of identifying variants for about a dozen regions of England, accounting in the process for roughly half of the variation in the data set. In some cases, the principal components appear to provide a fairly objective method for characterizing traditional English dialect regions on a quantitative basis. However, unlike the BiSGP analysis, principal component analysis does not comprehensively divide England into regions; moreover, it often isolates variants that are

unique to fairly small regions or include variants that are not unique to the relevant region; and few localities in an identified region even use most of the variants identified by the relevant principal component.

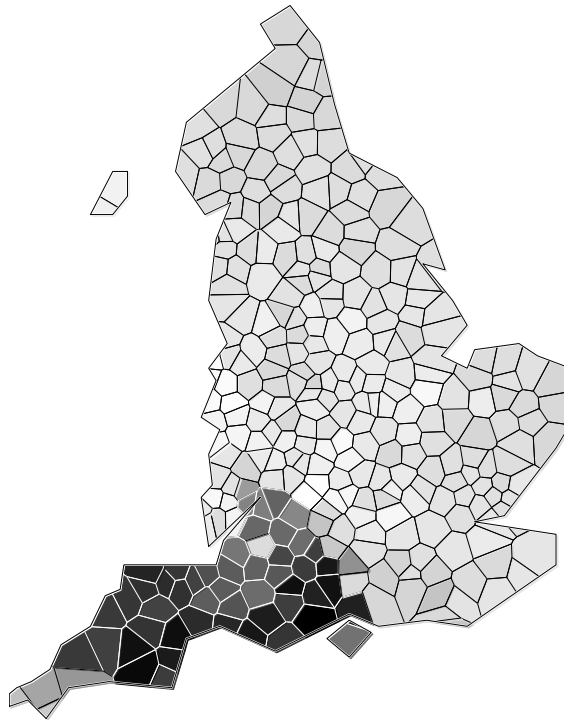


Fig. 5 Component scores for the first varimax principal component. Darker shades of grey indicate higher component scores. The original image was taken from Shackleton (2010).

For example, as illustrated in Fig. 5, the high-scoring localities of the first component largely overlap with the broad Southwest dialect region identified by the BiSGP analysis (regions 7 plus 8 in Fig. 3), while the loadings indicate that the features most closely associated with the principal component are the voicing of fricatives (also linked to this region by the BiSGP analysis) and occasionally the voicing of medial dentals (e.g. [vist] for *fist* and [bʌder] for *butter*), the plausibly related voicing and dentalizing of medial fricative [s] ([ɪrɪt] for *isn't*), and

lowering and unrounding of /u/ ([bʌt] for *but*). (The principal component also assigns comparatively high loadings to strong rhoticity as well as to a set of vocalic features that nearly fully describe a nonstandard regional dialect system of vowels, but the rhotic features are not unique to the Southwest while the vocalic features appear only sporadically.) Nonetheless, the variants associated with the principal component are never found all together in any single Southwestern locality and can only rather loosely be thought of as representing a Southwestern dialect – a limitation that appears to be inherent in the principal component analysis of linguistic variation. Instead of strictly delineating a dialect region in the manner of the BiSGP analysis, the principal component analysis is at best suggestive of where the region’s boundaries might lie.

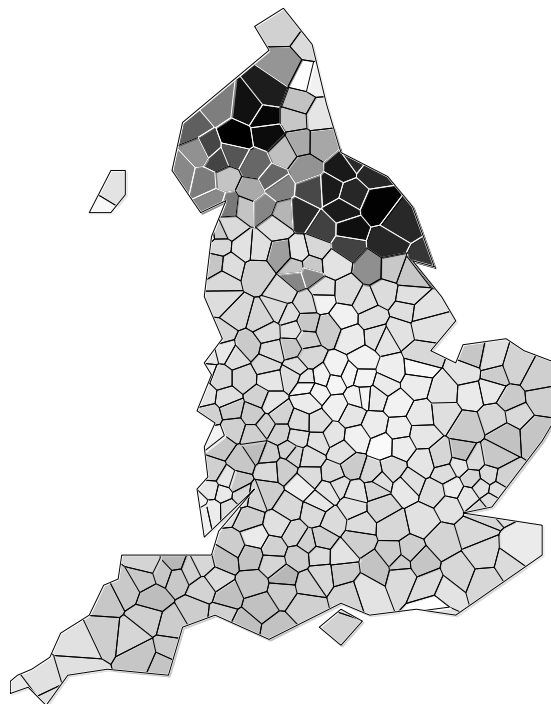


Fig. 6 Component scores for the second varimax principal component. Darker shades of grey indicate higher component scores. The original image was taken from Shackleton (2010).

The second rotated principal component (shown in Fig. 6) appears to be strongly associated with a large region of the Upper North that is not identified by the BiSGP analysis. The defining variants in this principal component all involve the development of ingliding from a high front onset for low long vowels (e.g. [lɛən] for *lane* and [koəl] for *coal*). The component also includes nearly all of the variants that are the most common regional pronunciations of Middle English long vowels. Nevertheless, as with the first component, the high-scoring variants may only rather loosely be said to define a dialect or group of dialects. In contrast to the first principal component, however, the second principal component cannot be said to strongly delineate a dialect region; it is more suggestive of than explicit about the region's boundaries.

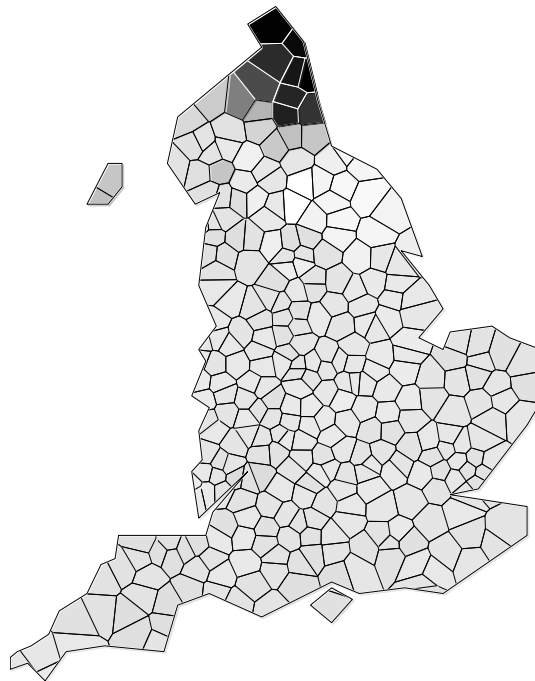


Fig. 7 Component scores for the third varimax principal component. Darker shades of grey indicate higher component scores. The original image was taken from Shackleton (2010).

The third rotated principal component (shown in Fig. 7) assigns high component scores to localities in the Far North, and assigns high positive loadings to the same variants associated with that region by the BiSGP analysis. In this case, the variants are sufficiently highly correlated with each other and also sufficiently unique to the region to allow both approaches to arrive at essentially the same classification – although, again, the principal component does not delineate the region as distinctly as the BiSGP analysis does. Also note that the area clearly overlaps with that of the second principal component (see Fig. 6).

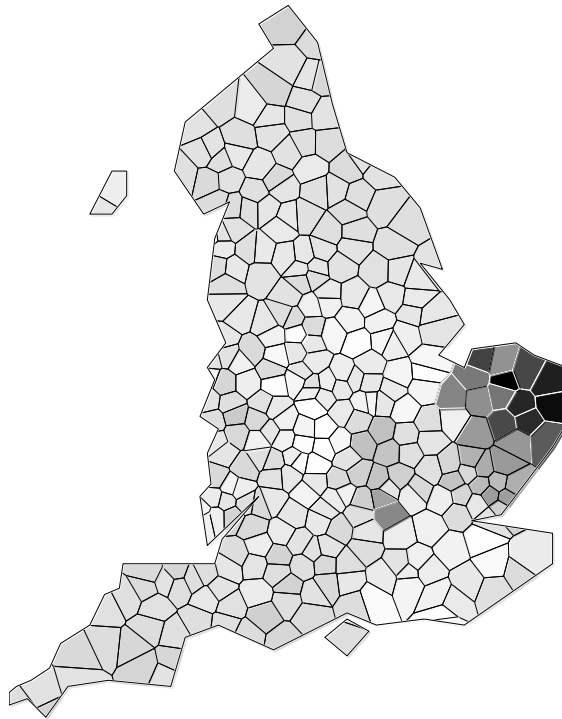


Fig. 8 Component scores for the fourth varimax principal component. Darker shades of grey indicate higher component scores. The original image was taken from Shackleton (2010).

The fourth principal component (shown in Fig. 8) rather weakly delineates most of East Anglia on the basis of the development of [w] for /v/ (e.g. [wɪnɛgr] for *vinegar*) and the development of a centered, unrounded onset in /i:/ (e.g. [ʌɪs] for *ice*). This classification has no counterpart in the BiSGP analysis, but it does appear (and is also similar in terms of characteristic variants) when the number of regions distinguished by the BiSGP approach is increased. We did not discuss this finer-grained division in this study, as this also resulted in many additional (uninformative) regions consisting of only a single locality.

Several other principal components (not shown) match dialect regions identified in the BiSGP analysis. For instance, the sixth principal component distinguishes Devonshire (region 8 in Fig. 3) from the rest of the Southwest by its unique fronting of back vowels, the same features associated with that region by the BiSGP analysis. Somewhat similarly, the seventh principal component, to some extent, distinguishes the ‘Potteries’ zone (part of region 4 in Fig. 3) by the use of [i:] and [u:] for /a:/ and /ɔ:/, respectively (e.g. [gi:t] for *gate* and [gu:t] for *goat*). The seventeenth principal component isolates the single locality (region 1 in Fig. 3) on the Scottish border in Cumberland that uses the alveolar trill [r]. Except for the last example, however, principal component scores provide suggestive evidence for regional boundaries, but do not strictly delineate regions in the manner of the BiSGP analysis.

Discussion

Hierarchical bipartite spectral graph partitioning complements other exact approaches to dialectology by simultaneously identifying groups of localities that are linguistically similar and groups of linguistic variants that tend to co-occur. This introduces an inductive bias in which the linguistic and the geographic dimensions reinforce one another. In a *post-hoc* step we identified

the important variants associated with a dialect region, by examining a linear combination of the variant's distinctiveness (usage frequency in the region as opposed to outside the region) and representativeness (comparative frequency within the region). That approach contrasts with and complements one-dimensional clustering techniques, which identify regions as groups of localities with similar aggregate patterns of variant frequencies, and principal component techniques, which identify correlated groups of variants without reference to patterns of distinctiveness or representativeness (and without reference to the sites where the variants are found).

Applied to the English dialect data used in this chapter, the BiSGP analysis identifies dialect regions that are broadly similar to those identified by clustering and principal component techniques, and isolates sets of variants distinctive for those regions that are also broadly similar to many of the sets identified by principal component analysis. In some cases, however, the BiSGP analysis failed to identify such well-accepted clusters as the Central dialect region (Trudgill, 1999), which was detected (in part) using cluster analysis (Shackleton, 2007); but, in other cases – most notably in the ‘Potteries’ region --- the BiSGP analysis distinguishes regionally distinctive combinations of variants that the other methods largely fail to identify.

Principal component analysis applied to linguistic material identifies groups of variables whose values tend to co-occur with one another. It applies primarily to numerical values, but also works well with frequency counts, once these attain a substantial size. PCA attaches no special value to solutions which privilege finding coherent groups of sites, i.e. finding groups of sites which tend to share strong values for one or more principal components. It is remarkable that (rotated) principal components normally do identify regions, i.e. geographically coherent groups

of sites where variables tend to co-vary. A proponent of PCA might therefore question the need for hierarchical bipartite spectral graph partitioning.

BiSGP seeks partitions of an input matrix that simultaneously identify co-varying linguistic variants (just as PCA does) and also co-varying sites, i.e. sites which share linguistic variants. It is more broadly applicable than PCA, even supporting the analysis of binary data (see Wieling and Nerbonne, 2011). Since, as dialectologists, we are interested in identifying common structure in both the linguistic and the geographic dimensions, hierarchical bipartite spectral graph partitioning is intuitively appealing. This intuitive appeal motivated our empirical examination.

Notes

¹ Wieling and Nerbonne (2011) used a binary matrix (with a threshold), but here we opted to use the scaled values as the *SAED* input matrix already included an aggregation step by having grouped several words, e.g. *root* and *tooth*, and we did not wish to add another aggregation step.

² Note that when using the raw frequencies, results were generally similar to those using the scaled frequencies (as most columns already had a maximum value of one).

References

Anderson, P. M. (1987). *A Structural Atlas of the English Dialects*. London: Croom Helm Ltd.

Chambers, J.K. and Trudgill, P. (1998). *Dialectology*, 2nd edn. Cambridge: Cambridge University Press.

Goebel, H. (1984). *Dialektometrische Studien: Anhand italoromanischer, rätoromanischer und galloromanischer Sprachmaterialien aus AIS und ALF.*, 3 Vol. Tübingen: Max Niemeyer.

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.

Nerbonne, J. (2009). Data-Driven Dialectology, *Language and Linguistics Compass* 3: 175-198.

Orton, H. and Dieth, E. (1962). *Survey of English Dialects*. Leeds: E.J. Arnold.

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

Shackleton, R. G., Jr. (2007). Phonetic variation in the traditional English dialects: a computational analysis, *Journal of English Linguistics*, 33: 99-160.

Shackleton, R. G., Jr. (2010). *Quantitative assessment of English-American speech relationships*. PhD thesis, University of Groningen. Groningen.

Trudgill, P. (1999). *The Dialects of England*. 2nd edn. Oxford: Blackwell

Wieling, M. and Nerbonne, J. (2010). Hierarchical spectral partitioning of bipartite graphs to cluster dialects and identify distinguishing features. In Banea, C., Moschitti, A., Somasundaran, S. and Zanzotto, F.M. (eds.), *Proceedings of the 2010 Workshop on Graph-based Methods for Natural Language Processing*. Stroudsburg, PA: Association for Computational Linguistics, pp. 33-41.

Wieling, M. and Nerbonne, J. (2011). Bipartite spectral graph partitioning for clustering dialect varieties and detecting their linguistic features, *Computer Speech and Language*, 25: 700-715.