Journal of Linguistic Geography, page 1 of 20.
doi:10.1017/jlg.2018.6

ORIGINAL RESEARCH

Probabilistic corpus-based dialectometry

Christoph Wolk,[1]* and Benedikt Szmrecsanyi[2]

[1] *University of Giessen*
[2] *KU Leuven*

Researchers in dialectometry have begun to explore measurements based on fundamentally quantitative metrics, often sourced from dialect corpora, as an alternative to the traditional signals derived from dialect atlases. This change of data type amplifies an existing issue in the classical paradigm, namely that locations may vary in coverage and that this affects the distance measurements: pairs involving a location with lower coverage suffer from greater noise and therefore imprecision. We propose a method for increasing robustness using generalized additive modeling, a statistical technique that allows leveraging the spatial arrangement of the data. The technique is applied to data from the British English dialect corpus FRED; the results are evaluated regarding their interpretability and according to several quantitative metrics. We conclude that data availability is an influential covariate in corpus-based dialectometry and beyond, and recommend that researchers be aware of this issue and of methods to alleviate it.

1. INTRODUCTION

In this paper, we discuss new technologies to measure dialect distances on the basis of dialect corpora that sample authentic, naturalistic usage data. More specifically, we show how probabilistic modeling techniques can be used to address a thorny challenge in corpus-based dialectometry: the amount of text and speech that is available to cover particular dialect locations is typically not constant but variable, and this variability can seriously confound unless it is neutralized by using the mathematics of uncertainty.

Let us fix some basic concepts and methodological preliminaries at the outset. Practitioners of traditional dialectology study "interesting" dialect phenomena, one feature at a time, often only in a handful of dialects; cross-feature comparisons remain rather impressionistic through the bundling and/or grading of isoglosses and the resulting areal classifications. In contrast, DIALECTOMETRY is the branch of geolinguistics dedicated to measuring, visualizing, and analyzing aggregate dialect similarities or distances as a function of properties of geographic space (see Goebl 1984, 2007; Heeringa, 2004; Nerbonne, 2009; Nerbonne, Heeringa & Kleiweg, 1999; Séguy, 1971 for foundational work).

As for DATA SOURCES, traditional dialectometry draws on dialect atlases as its data source. Take, for example, Goebl (1982), who investigates similarities between Italian dialects on the basis of 696 linguistic features that are mapped in the *Sprach- und Sachatlas Italiens und der Südschweiz* (AIS), an atlas that covers Italy and southern Switzerland. It is important to bear in mind that the data that (most) linguistic atlases provide speak primarily to the issue of what informants KNOW about their dialects. To study LANGUAGE USAGE analysts typically turn not to surveys and dialect atlases but to DIALECT CORPORA. Linguistic corpora are principled and broadly representative collections of naturalistic texts or speech that sample usage data—a data type that is increasingly popular in dialectology (Anderwald & Szmrecsanyi, 2009; Grieve, 2016) and beyond (see the papers in Szmrecsanyi & Wälchli, 2014).

CORPUS-BASED DIALECTOMETRY (henceforth: CBDM), then, combines the study of dialectometric research questions with corpus-linguistic methodologies. CBDM utilizes aggregation methodologies to explore quantitative and distributional usage patterns extracted from dialect corpora (see Szmrecsanyi, 2008, 2011, 2013; Szmrecsanyi & Wolk, 2011; Wolk, 2014; Wolk & Szmrecsanyi, 2016). Turning to corpora enables analysts to address questions about usage versus knowledge, production/comprehension versus intuition, chaos versus orderliness, and so on.

In this contribution, we discuss a recent methodological advance in CBDM. To do justice to the fact that textual coverage of individual dialects may (and typically does) vary in dialect corpora (e.g. dialect A may be represented by 20 interviews, but dialect B by only 5 interviews), the first wave of CBDM approaches merely normalized text frequencies prior to aggregation: so instead of saying that a particular linguistic variant occurred, e.g., 100 times in total in material from some particular dialect, a normalized measure ("the linguistic variant occurs 30 times per 10,000 words of running text") was used to calculate dialect distances. The

*Address for correspondence: Otto-Behaghel-Straße 10 B 403, 35394 Giessen, Germany, +49 641 - 99 301 53, christoph.b.wolk@anglistik.uni-giessen.de

innovation we discuss in the present paper introduces statistical modeling into the CBDM pipeline: subcorpus size turns out to be a crucial covariate of aggregate distances, and so probabilistic CBDM draws on stochastic reasoning to take this covariate more seriously than normalization-based CBDM does. The outcome, as we shall see, is a less noisy and arguably more accurate portrayal of aggregate dialect relationships.

This paper is structured as follows: Section 2 motivates our approach by discussing the role of data availability in dialectometric analyses, with a particular focus on frequency- and corpus-based dialectometry. We will also introduce some of the technicalities of our particular solution to the challenges this factor provides. The remainder of the paper will work in detail through a case study concerning British dialects. Section 3 will introduce the data set, while sections 4 and 5 will present two related techniques, straightforward frequency-based CBDM and the probabilistically enhanced version, as well as their results on the present dataset. The final section will conclude by reviewing and discussing these outcomes in light of the discussion in section 2.

2. ON THE INFLUENCE OF DATA AVAILABILITY IN DIALECTOMETRY

The principal question underlying the CBDM approach is the following: How can we derive accurate representations of linguistic divergence from naturally occurring discourse?[2] This is, in principle, not too different from questions regularly asked in atlas-based dialectometry, which has greatly benefited from a long methodological discussion and a varied set of techniques (e.g., Heeringa, 2004). In general, a dialectometric analysis proceeds as follows:

1. Establish a feature set (lexical items, pronunciations etc) by which the dialects are to be compared.
2. Determine the realizations of these features in each location.
3. Compare all features in all locations and derive a numerical value indicating the degree of dissimilarity.
4. Aggregate over all features to yield a composite score, or distance, for each pair of locations.
5. Analyze the resulting scores using exploratory and/or confirmatory data analysis.

Steps 3–5 in particular have seen considerable advancement and extension. Steps 1 and 2 typically rely on the results of large survey projects, compiled into dialect atlases. Atlas data are in many ways well-suited for such analyses; for instance, the network of locations tends to have a rather fine mesh, which allows a geographically high resolution. Particularly crucial for present purposes is that the amount of data per location tends to be quite homogeneous: sites usually have similar amounts of informants, and ideally each informant has a complete set of responses to the survey items. In practice, atlas data are not always complete. The problems that missing entries can cause, and how to mitigate them, have been recognized as important issues almost from the inception of dialectometry. Goebl (1977: 46) discusses the issue and employs a method that has since become the standard treatment (see also, e.g., Nerbonne & Kleiweg, 2007: 159): features which are missing for one or both locations in a pair are to be completely left out of the analysis for that pair, continuing as if they were never in the feature set. While this introduces noise into the measurements—Goebl (1993: 286) terms this the *missing data effect*—in general, it has not led to major problems for dialectometry. The well-documented increase in robustness that dialectometry achieves through aggregation seems sufficient to cancel out minor noise resulting from this (Nerbonne, 2009). Nevertheless, other approaches have been suggested by Viereck (1988: 546): missing entries could be included, estimated based on the closest neighbor, or an aggregate of neighbors. Taking geographic information into account like this can reduce noise, and therefore diminish the missing data effect. The cost of this is, of course, that the additional assumption may not be warranted—that a location may well be unlike its neighbors with regard to the missing feature, and the resulting analysis is biased. This matches a central trade-off in statistical learning, the "bias-variance trade-off" (James, Witten, Hastie & Tibshirani, 2013). Statistical learning—such as learning the distances between dialects from samples of individual informants' judgments and productions—depends both on the data set itself and on the specific characteristics and implicit assumptions inherent in the method. As one increases the flexibility of a method, incidental properties of the data (such as missing entries at certain locations) will have greater weight; lower flexibility, however, leads to greater dominance of the assumptions that the method makes. Both yield the danger of distorted and poorly generalizable results. For categorical atlas-based signals, increased flexibility may often be more desirable, as incidental distortions should even out in the long run.

One recent extension of the dialectometric tool box involves adapting the methods to data of a different nature: instead of relying on (typically questionnaire-based) categorical surveys filtered through dialect atlases, authentic speech is tapped directly. This is made possible through the emergence of specialized dialect corpora, i.e., naturally occurring linguistic material, typically whole interviews, collected from dialectologically appropriate speakers. We believe this to be a central locus for the advancement of the dialectometric

project: it opens new possibilities for frequency-based analyses that integrate well with the kind of usage-based approaches that are gaining ground in many branches of linguistics, and allows tackling research questions that crucially depend on frequency. Nevertheless, this change of data source necessitates a retooling of the usual approaches in dialectometry. As we shall demonstrate, for typical dialect corpora, the importance of the *data availability* factor greatly increases, and relying only on aggregation may lead to wrong inferences.

How can one establish comparability of frequency measurements between corpora, and therefore locations? In a realistic scenario, it is highly likely that the two corpora are unequal in size, be it due to availability of raw material (because, e.g., more interviews happen to have been conducted in location A than in location B) or corpus design. But when one corpus is substantially larger than the other, similar counts do not imply similar usage rates. The usual solution in corpus linguistics involves *normalization*, i.e., transforming raw counts into occurrence rates. The process is straightforward:

$$(\text{raw counts} / \text{text size}) \times \text{normalization constant}$$

Hence, the total number of occurrences in a (sub)corpus is divided by the number of words, and, to make the numbers more easily interpretable, the resulting figure is scaled by a fixed number, yielding the number of occurrences per, say, ten thousand words (*pttw*). As these normalized values have a common basis, they can be compared with one another and their difference can be quantified.

There are, as we shall show, cases in which the normalization procedure may lead to biased results, and these bear a strong resemblance to the missing data effect. While the process will always yield a numerical value of the difference between the corpora, the accuracy of this value crucially depends on the text size prior to normalization, and in particular on that of the smaller corpus. To illustrate this, consider a hypothetical feature with a (population) frequency value of 1 *pttw* in two communities; from the first, we sample a corpus of one hundred thousand words (C1), from the second a corpus of only ten thousand words (C2). We should expect the normalized value for C1 to approximate the population value relatively well. C2, however, may well be far from the true value—it would not be surprising to see the feature completely absent, or have a normalized frequency two to three times as high as in C1. The distance between the corpora is quite likely to be substantially higher than the difference between the populations—exactly zero. On the other hand, C1 will likely be more similar to a population where there are actual frequency differences; higher or lower depending on the accident of chance. Note that this is a property of the corpora and their sizes; when moving toward combining such individual measurements into a multi-feature dialectometric analysis, this will apply to all of them individually. It follows that the distances between similar points are likely to be too high, and the distances between dissimilar points may be too high or even too low. Goebl (1993) reports that the missing data effect for categorical data, with missing entries left out of the analysis, entails "measurement results which are too high [i.e., similar, as Goebl uses similarity instead of distance] in comparison to the general trend" (286). In frequency-based analyses, absence corresponds to a frequency of zero and the feature is not left out of the analysis. This suggests that measurements affected by such issues would be prone to be more *dissimilar* than the trend suggests at close locations, but may be too similar at more distant locations.

As we shall show, this is not just a hypothetical issue, but has direct implications for dialectometric practice. While we will be focusing on frequencies, similar considerations apply to proportions of categorical alternants and possibly even categorical atlas realizations. Streck (2014), for example, reports that his corpus-oriented analysis of phonological variation in southwestern German yielded a substantially stronger relationship between geographic distance and linguistic similarity after removing the locations with the least amount of data. In principle, only datasets that are complete and large are fully safe, although minor amounts of size discrepancies or empty cells need not display any adverse effect.

How can the effect that the amount of data has on the result be mitigated in corpus-based analyses? Similar solutions apply as for the missing data approaches described above, but with a crucial difference: instead of a categorical presence signal we have a gradient quality signal applying to all features at the same time, which makes the consensus solution for atlas-based dialectometry unavailable—at least without dropping the location completely. One way around the problem is to just accept it, to not "fill in missing data artificially" (Goebl, 1991: 283). The benefit here is that every step of the analysis is purely based on actually observed data, and few additional assumptions are necessary. This comes at a cost: the usual visualizations and maps hide the fact that some measurements are more imprecise than others, and various statistical results become difficult to interpret. A second possibility involves restricting the analysis to those locations where there is ample data. This seems statistically reliable, but will typically result in substantial reductions in geographic coverage. The third possibility is to use a larger corpus. Our data suggest that effects of differences in the amount of text is still detectable even as the number of words in both corpora increases (see section 6 below), but it is clear that having more data leads to more precise measurements even

with simple normalization. This is clearly the best solution, but is in general not feasible for spoken dialect data, due to difficulties in acquiring material and the costs of transcription. For studying geographic variation in written material, such as letters to the editor in regional newspapers (Grieve, 2016) or Twitter data (Eisenstein, 2018; Huang, Guo, Kasakoff & Grieve, 2016), where corpus sizes can reach dozens of millions or even billions of words, normalized values alone may suffice fully. The final possibility, and one that is applicable to the relatively small spoken dialect corpora, is to use a method that is less sensitive to the incidental variance in the data through the use of additional assumptions. Given the geographic nature of dialectal data, the best candidates for such analyses are those that build on the fundamental dialectological postulate that "geographically proximate varieties tend to be more similar than distant ones" (Nerbonne & Kleiweg, 2007; see also Tobler's (1970) first law of geography: "Everything is related to everything else, but near things are more related than distant things.").[3] The observed values should be considered in their spatial context, for example by discounting large differences between close locations if they are based on little data. Of course, this correction should not be too strong, and the method should be able to still find differences between close locations if this is warranted.

Within the frequency- and corpus-based dialectological literature, several methods have been proposed that can achieve this (e.g., Grieve, 2009; Pickl, Spettl, Pröll, Elspaß, König & Schmidt, 2014). We believe that *generalized additive models* (GAMs) have particularly nice properties for this purpose (Wood, 2006). GAMs share conceptual similarities with generalized linear models, which are familiar to many linguists (e.g., in the form of VARBRUL models, cf. Sankoff, 1987). GAMs extend this by including smooth terms, smooth functions whose shape is determined from the data. They can be two-dimensional, and are therefore suitable to represent dialectal "frequency landscapes" that represent the geographic distribution of features as mountains and valleys of high and low usage. GAMs are not new in geolinguistics; they were previously used successfully for dialectometric purposes in Wieling, Nerbonne & Baayen (2011) and Wieling, Montemagni, Nerbonne & Baayen (2013). Our use differs quite substantially from theirs: instead of building a single model based on the linguistic distances of all points to a reference variety, we build individual models for each feature. This model represents an improved estimate of how frequent that feature is in each subcorpus, and is itself suitable for visualization and interpretation. An example can be seen in Figure 1, which displays the smooth term for one of the features that we include in our analysis: multiple negation, as in (1)

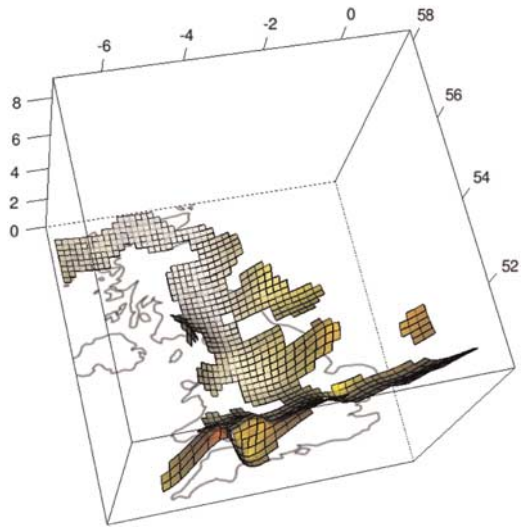


Figure 1. Frequency landscape for feature [33], multiple negation.

(1) 'cause you dare not say nothing ... < LND_001 >

The GAM estimates a frequency of around 8 *pttw* in southern England, which drops as one moves north. The Isle of Man also shows a higher rate of usage.

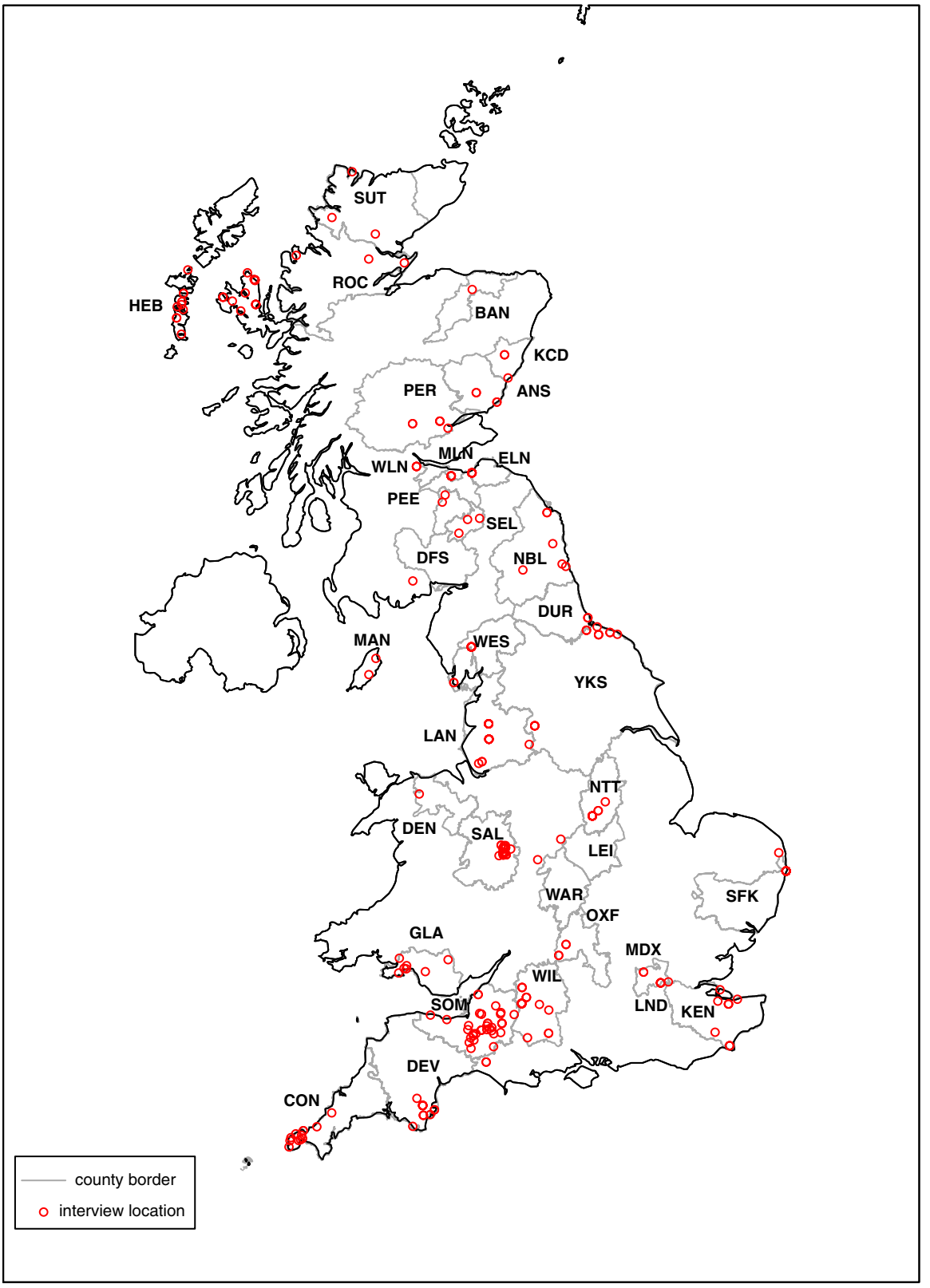
As is apparent from the plot, such landscapes can be quite complex. Nevertheless, the GAM implementation used here limits this complexity by means of generalized cross-validation: the effect of leaving out individual data points is determined by analyzing subsets of the data, and the final result is chosen such that single points do not have undue influence. This yields a landscape that can be steeply sloped when this is warranted based on the data, but where the pattern tends to be flat when it is not.

It is important to note that such models have the capabilities of generalized linear models, and can therefore include speaker- and/or text-oriented covariates, such as speaker age or some measure of conversational interactivity. We make only limited use of this capability in the present paper, but believe that it is a central advantage of this method. What is crucial, however, is that both normalization and modeling both yield numerical values representing the same thing: estimates of frequency. Therefore, both can be analyzed in exactly the same way, and are easily compared and contrasted. It is to this task that we turn next.

3. DATA

3.1. Corpus Database

This case study taps into FRED, the *Freiburg Corpus of English Dialects* (Hernández, 2006; Szmrecsanyi & Hernández, 2007), a major dialect corpus that covers traditional dialect speech (mainly transcribed so-called "oral



Map 1. Counties and interview locations that contributed data to the corpus as used in the analyses to follow.

Table 1. *N = 34 objects (i.e. FRED counties/dialects) considered in the present study: map labels, membership in a-priori dialect areas roughly following Trudgill's dialect division on pronunciational grounds (Trudgill 1999: Map 9), textual coverage (running words) in FRED.*

map label	county	*a-priori* dialect area	no. words sampled in FRED
ANS	Angus	Sc Lowlands	19,899
BAN	Banffshire	Sc Lowlands	5,655
CON	Cornwall	Southwest of E	107,072
DEN	Denbighshire	Wales	5,794
DEV	Devon	Southwest of E	97,080
DFS	Dumfriesshire	Sc Lowlands	9997
DUR	Durham	North of E	28,069
ELN	East Lothian	Sc Lowlands	40,190
GLA	Glamorganshire	Wales	53,110
HEB	Hebrides	Hebrides	72,761
MAN	Isle of Man	Isle of Man	10,930
KCD	Kincardineshire	Sc Lowlands	7,509
KEN	Kent	Southeast of E	176,908
LAN	Lancashire	North of E	205,342
LEI	Leicestershire	E Midlands	5,864
LND	London	Southeast of E	110,802
MDX	Middlesex	Southeast of E	31,794
MLN	Midlothian	Sc Lowlands	32,040
NBL	Northumberland	North of E	30,644
NTT	Nottinghamshire	E Midlands	150,810
OXF	Oxfordshire	Southwest of E	15,109
PEE	Peebleshire	Sc Lowlands	14,955
PER	Perthshire	Sc Lowlands	20,896
ROC	Ross and Cromarty	Sc Highlands	10,475
SAL	Shropshire	E Midlands	168,572
SEL	Selkirkshire	Sc Lowlands	9,325
SFK	Suffolk	Southeast of E	312,560
SOM	Somerset	Southwest of E	207,502
SUT	Sutherland	Sc Highlands	10,967
WAR	Warwickshire	E Midlands	8,269
WES	Westmorland	North of E	157,562
WIL	Wiltshire	Southwest of E	185,928
WLN	West Lothian	Sc Lowlands	18,418
YKS	Yorkshire	North of E	90,816

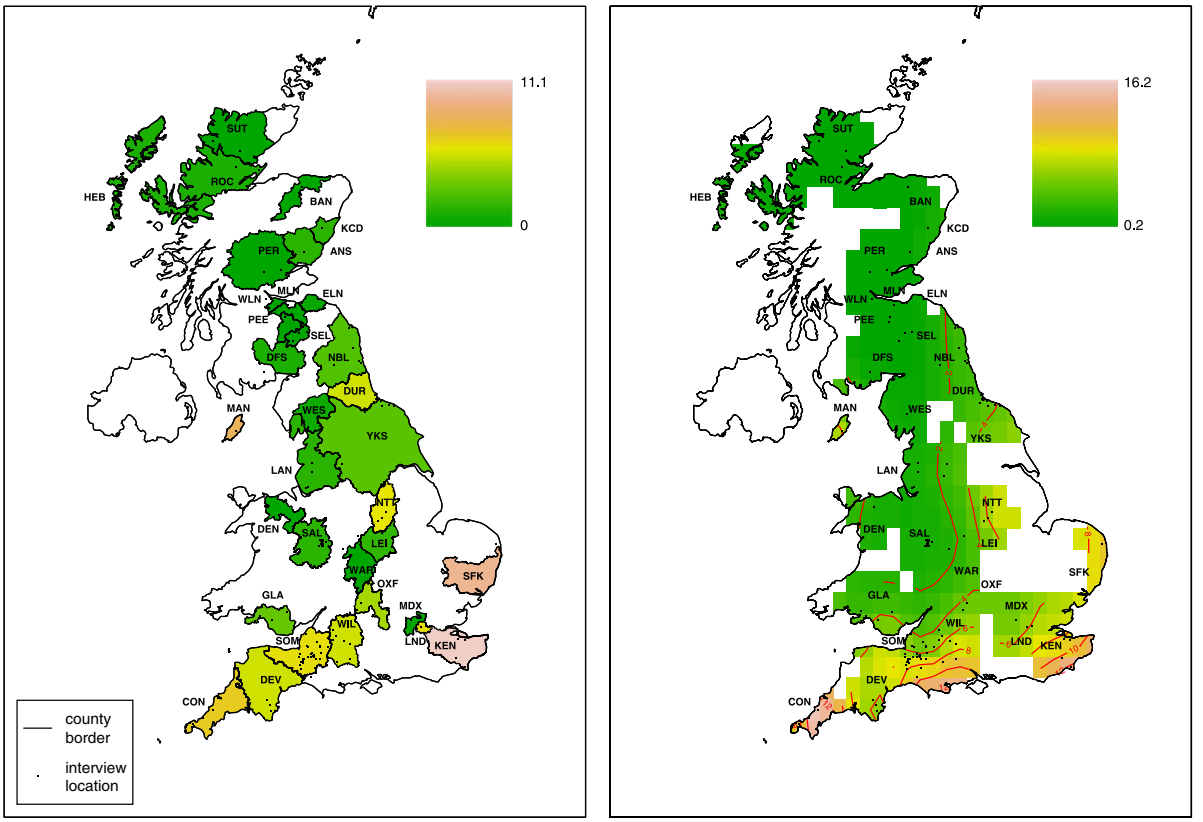
history" material) all over England, Scotland, and Wales.[5] The vast majority of the interviews were recorded between 1970 and 1990. In most cases, a fieldworker interviewed an informant about life, work, etc., in former days. The informants sampled in the corpus are typically elderly people with a working-class background, so-called 'non-mobile old rural males' (Chambers & Trudgill, 1998, 29).

The version of FRED we use here consists of interviews with 376 informants and spans approximately 2.4 million words of running text. The interviews were conducted in 34 pre-1974 counties in Great Britain, including the Isle of Man and the Hebrides. To mitigate data sparsity, the level of areal granularity investigated in the present study will be the level of individual counties. Map 1 displays the county boundaries and interview locations. See Wolk (2014, chapter 3) for an in-depth discussion of the dataset. (Table 1) Q6

3.2. *Features*

The analysis in this paper is based on the feature set used in Szmrecsanyi (2013). The set comprises 57 features, and overlaps with, but is not identical to, the comparative morphosyntax survey in Kortmann & Szmrecsanyi (2004) and the battery of morphosyntax features covered in the Survey of English Dialects (Orton, Sanderson & Widdowson 1978; Viereck, Ramisch, Händler, Hoffmann & Putschke, 1991). The features in the catalog fall into eleven major grammatical domains: (i) pronouns and determiners (e.g., non-standard reflexives, as in (2)), (ii) the noun phrase (e.g., the *s*-genitive, as in (3)), (iii) primary verbs (e.g., the verb *to do*, as in (4)), (iv) tense & aspect (e.g., the present perfect with auxiliary *be*, as in (5)), (v) modality (e.g., epistemic/deontic *must*, as in (6)), (vi) verb morphology (e.g., non-standard weak past tense and past participle forms, as in (7)), (vii) negation (e.g., *never* as a preverbal past tense negator, as in (8)), (viii) agreement (e.g., non-standard *was* as in (9)), (ix) relativization (e.g., the relative particle *what*, as in (10)), (x) complementation (e.g., unsplit *for to*, as in (11)), and (xi) word order and discourse phenomena (e.g., lack of auxiliaries in *yes/no* questions, as in (12)). We cannot discuss the features in much detail here, but the Appendix provides the complete list of features. See Szmrecsanyi (2013, chapter 3) for guidelines regarding feature selection and Szmrecsanyi (2010) for a detailed description of the feature extraction procedure.

(2) But old Silvain, he used to look after hisself, really <CON_003>
(3) But his wife is dead my brother Quentin's wife is dead two years ago <GLA_002>
(4) I don't know <CON_001>
(5) Joe, if you weigh them up, and you 're got an odd britch, I could do with a pair o' them <SFK_038>
(6) […] we must have been tough nuts you know, really. <LAN_009>
(7) Oh, but you wouldnae be telled the wages. <BAN_001>
(8) […] and they never moved no more, neither one of them, never tried to <CON_005>
(9) they was half gypsies you see? <OXF_001>
(10) See that up on the top there, the stamp what you hammer in... <WIL_024>
(11) For to screw down the cover on the churn <CON_002>



Map 2. Normalized frequency (left) and model-based frequency prediction (right) for multiple negation [33]. Yellow colors indicate areas where high frequencies are observed or predicted, green colors indicates low (observed or predicted) frequencies.

(12) They didn't want him prosecuted? <DEV_001>

4. THE NORMALIZATION-BASED CBDM APPROACH

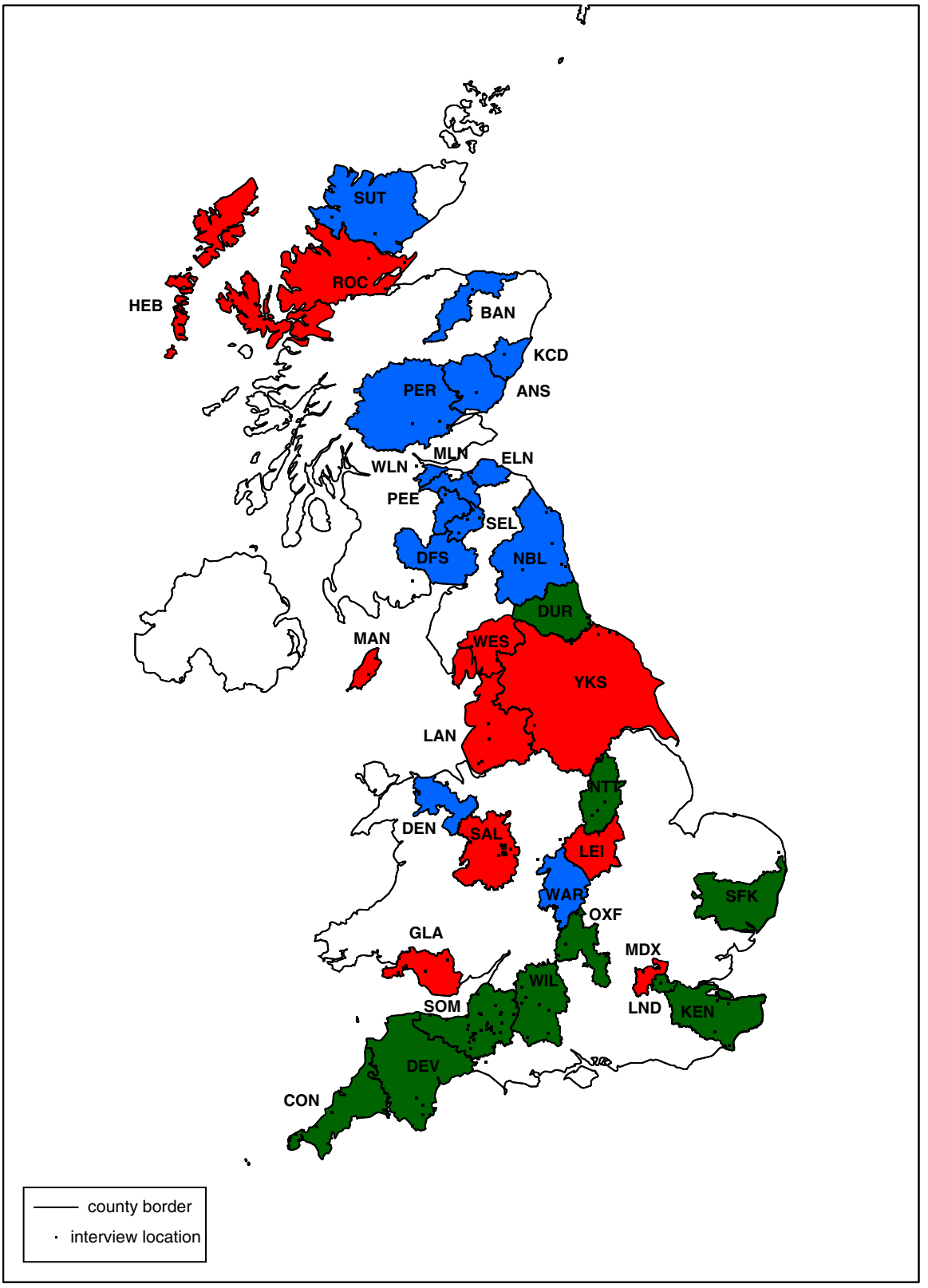
How does the normalization-based CBDM approach without probabilistic enhancements following Szmrecsanyi (2013) study dialect relationships as a function of geographic space? We first determine the text frequency of the features in the corpus material: how often do we find particular features—say, multiple negation—in interviews from particular locations. Next, we normalize text frequencies to frequency per 10,000 words because textual coverage of individual dialects varies, and round the result to whole numbers. At this stage, we also perform a log-transformation, which is a customary method to de-emphasize large frequency differentials and to alleviate the effect of frequency outliers (Shackleton, 2007, 43), thus increasing reliability of the measurements. For features that are absent from a corpus, the value is set to -1, corresponding to a frequency of 0.1 *pttw*, as the logarithmic transformation requires its input to be larger than 0. Let us illustrate the procedure: in FRED, the county Cornwall has a textual coverage of 12 interviews totaling about 107,000 words of running text (interviewer utterances excluded). In this material, feature [34] (negative contraction, as in (13)) occurs 326 times, which translates into a normalized text frequency of $326 \times 10{,}000/107{,}000 \approx 30$ occurrences per ten thousand words.

(13) They won't do anything. <WES_011>

A *log*-transformation of this frequency yields a value of $log10(30) \approx 1.5$. This is the measurement that characterizes this specific measuring point (Cornwall) in regard to feature [34].

In the next step, we create an $N \times p$ frequency matrix, in which the $N = 34$ objects (that is, dialects) are arranged in rows and the $p = 57$ features in columns, such that each cell in the matrix specifies a particular (normalized and log-transformed) feature frequency. Our case study thus yields a 34×57 frequency matrix: 34 British English dialects, each characterized by a vector of 57 text frequencies. To illustrate, Map 2 (left) projects feature frequencies of feature [33] (multiple negation) to geography. (Note: parallel maps for the other 56 features in the catalog are available in the online appendix.)

Frequency matrices can serve as input to a number of multivariate analysis techniques, such as Principal Component analysis or Factor Analysis (see, e.g.,



Map 3. Cluster map based on the normalization-based CBDM approach. Displayed: 3-cluster solution.

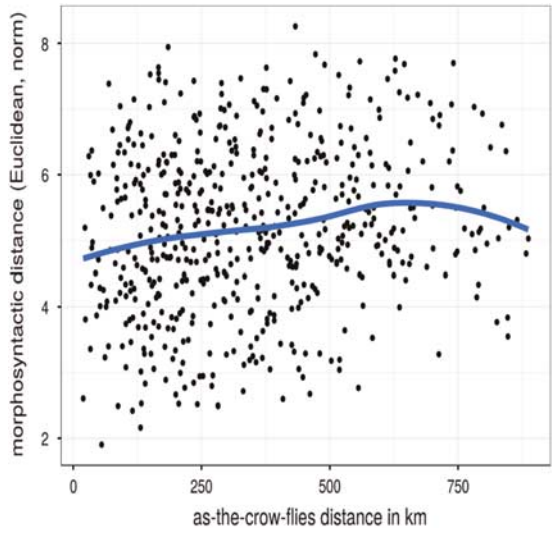


Figure 2. Correlating normalization-based morphosyntactic distances with as-the-crow-flies distances ($r = .19$, $p < .001$, logarithmic $R^2 = 3.6\%$).

Grieve, 2014). Most aggregational procedures customary in dialectometry, however, are empirically based on so-called distance matrices, which are obtained by transforming an $N \times p$ frequency matrix into an $N \times N$ distance matrix. This transformation abstracts away from individual feature frequencies and instead provides pairwise distances between the dialect objects considered. To create a distance matrix, we relied on the Euclidean distance measure (Aldenderfer & Blashfield, 1984, 25), which defines the distance between two dialect objects as the square root of the sum of all p squared frequency differentials.

In this paper, we analyze distance matrices in two ways: via cluster maps[1] and by correlating linguistic distances with geographic distances[4]. Cluster maps are a staple analysis technique in dialectometry (see, e.g., Goebl, 2007, Map 18; Heeringa, 2004, Figure 9.6)—the $N \times N$ distance matrix is subjected to hierarchical agglomerative cluster analysis (Jain, Murty & Flynn, 1999), a statistical technique used to group a number of objects (in this study, dialects) into a smaller number of discrete clusters.[6] Each of the clusters is assigned a distinct color, and the clusters are subsequently visually depicted on a map.

Thus Map 3 projects a 3-cluster categorization based on the normalization-based distances to geography. We can see that there clearly is a geographic signal in the dataset: Scottish counties are colored in blue, Northern English dialects tend to belong to the red cluster, and Southern English dialects tend to be assigned to the green cluster. That being said, it is clear that there is also a good deal of geographic incoherence and noise: there are blue counties in Wales and England, red spots in Southern Wales, the Scottish Highlands, and the Hebrides, and Durham in the North of England is mysteriously green.

We move on to correlating linguistic distances with geographic distances, for the sake of precisely quantifying the extent to which normalization based dialect distances are predictable from geographic distance (specifically: pair wise as-the-crow-flies distances, which can be easily calculated from longitude/latitude information) between dialect locations. The relationship is visually depicted in Figure 2. There is a significant relationship, but as-the-crow flies distance accounts for only 3.4% of the normalization-based morphosyntactic variance; a sub-linear logarithmic relationshp only fares marginally better at 3.6%. This is not a big share: in the realm of syntax-focused atlas-based dialectometry, analysts have reported R^2 values of up to 45% (Spruit, Heeringa & Nerbonne, 2009). Compared to that, the geolinguistic signal in our normalization-based dataset is quite weak.

5. THE MODEL-BASED APPROACH

As we have argued in section 2, the results of the method discussed in the previous section may be influenced by imprecise measurements in some features and/or locations. We now move on to a method that can alleviate this, namely regression modeling using GAMs.

Regression models are essentially statistical models, in which the values of one variable are represented as combinations of the effects of other variables, the so-called predictors. The effects of the individual predictors are determined from the data during the fitting process. Generalized additive models, in particular, allow the estimation of complex non-linear predictor behavior in one or more dimensions, and thus allow the representation in maps. When building regression models for linguistic phenomena, the analyst faces a bewildering amount of choices, ranging from the basic representation of the data over the precise model specification to the details of the fitting process. The principles guiding our selection processes for present purposes are the following: First, the models and their results should be as straightforwardly comparable to the normalization based results as possible; the fewer deviations from the process outlined in the previous section, the better. This enables comparative analysis that clearly shows where the methods differ. Second, where possible, we should choose the methods such that the result is still responsive to local conditions and the frequency patterns at individual locations; after all, simply parroting geography for its own sake would be dialectologically meaningless. The aggregation process can alleviate the impact of overfitting to a degree, but may struggle on severely underfit data.

The first choice to be made pertains to the representation of the outcome. Many linguistic phenomena can be analyzed and represented in several ways; consider, for example, feature [2], non-standard reflexive pronouns such as *hisself* in (2) where Standard British English would use *himself*. This feature could be studied in terms of its frequency (e.g., how often are such non-standard forms used per ten thousand words) or its share of all constructions fulfilling similar functions (e.g., what is the share of non-standard forms among all reflexives?).

Different operationalizations allow different aspects of the data to shine through: proportions are more robust toward variation in the base frequency of the feature—a generally higher frequency of reflexive use in some areas may or may not be linguistically relevant. Even where it is relevant, frequency appears to mix two different aspects of the phenomenon. To give a hypothetical extreme example, a county with ten reflexive pronouns *pttw*, all of which are non-standard, is intuitively different with regard to non-standardness from one where ten of 100 reflexives *pttw* are non-standard, even if the normalized frequency of the non-standard variant is the same. Nevertheless, there is considerable debate whether pure frequencies or relative metrics are what is ultimately of relevance to cognition and linguistic theory (Bybee, 2010; Gries, 2012). For present purposes, we decided to model only frequencies, as this minimizes the differences between both normalization-based and probabilistic analyses, and removes the selection of relevant contexts as a source of errors. There are several model types that allow modeling such frequencies, of which perhaps the best-known is the Poisson regression. One assumption of this method, however, is that mean and variance of the dependent variable are the same. Linguistic material often violates these assumptions, especially content words and other "bursty" features, i.e., those that stray from even dispersion throughout a text/corpus (Manning & Schütze, 2000: 547). This is particularly troubling as grammatical features have been shown to reliably occur more often after they have already appeared (Branigan, 2007; Szmrecsanyi, 2006) and are therefore likely to be bursty. An alternative is negative binomial regression, which includes an additional parameter (*theta*), allowing the shape (and therefore variance) of the distribution to vary. This parameter can be pre-specified, or determined from the data. This makes the negative binomial distribution more appropriate for word and/or grammatical feature distributions. Note, however, that there are still potential issues—these models may still suffer from overdispersion, and especially from zero inflation, i.e., more observations of zero than the distribution allows (Hilbe, 2007). Models specifically designed for such situations exist, but are difficult to operate, are not available in standard tools, and may lead to model fitting issues. For these reasons, overdispersed models are sometimes recommended against,[7] and we will employ regular negative binomial regression. The software package we use, *mgcv* version 1.8–10 (Wood, 2006), allows two major ways of determining the additional coefficient for the negative binomial distribution. One makes use of restricted maximum likelihood (REML), the other of generalized cross-validation (GCV); both also affect the other estimations in the model, and in particular the general shapes of the frequency landscapes (as in Figure 1) that are our primary interest. In our sample, the GCV approach seems to lead to more varied, hilly landscapes, whereas the REML-based models are flatter, and may remove too much of the geographic specificity in the data. We therefore use GCV where possible, with the search space for the theta parameter rather wide (ranging from 0.01 to 50). However, there is a small number of features where the GCV-based model either does not converge ([4], [9], [29], and [39]), or leads to degenerate estimates for the family parameter ([27], [31], [40] and [43]). In those cases,[8] the REML-based model was used, which is more robust in convergence (*mgcv* documentation: *negbin*). To include geographic location in the models, we follow the practice recommended by Wieling et al. (2014: 679) and use *thin plate regression splines*, which are a "highly suitable approach to model the influence of geography in dialectology." Each interview is coded with the geolocation of the informant's village or town, and we use this information directly for modeling instead of first aggregating on the county level.

One of the big advantages of the GAM-based approach is that the models can easily be combined with further information, whether it pertains to the sociolinguistic situation (such as speaker characteristics) or the immediate linguistic context (which may often be feature-specific, such as subject type). Taking such information into account can increase the reliability of the analysis by reducing the effect of non-geographic variation in the data, and can serve as corroborating evidence for the analysis—if non-geographic variables have the expected effects, this should raise our confidence in both data and models—and is interesting from the single-feature perspective. Nevertheless, there are also significant costs to this. Feature-based annotation can require tremendous effort, which is not feasible for holistic analyses that cover a large number of features. Information that is trivial to code, such as a speaker's age, may not be available in the corpus metadata. This applies in our case: while most speakers in the dialect corpus we are using have metadata containing the sociolinguistically relevant predictors gender and age, or have information that allows relatively accurate estimation (such as birth and interview decades), some do not. This forces us to choose between

either excluding these factors in the models or excluding the speakers where the information is not available (which would remove some counties completely). Finally, adding such predictors would also add another difference to the normalization-based approach in the previous section. We therefore restrict ourselves mostly to geography-only models, but do present a brief summary of the sociolinguistic patterns observed in the next section. Previous research suggests that, at least for this dataset, age and gender are not substantially significant factors, as their effect on the aggregate level is very restricted (Wolk, 2014). For the models included in this brief summary, age, gender, and their interaction are always included. To make the models more easily interpretable, the age variable was centered around the mean, so that frequency differences between genders are evaluated toward the center of the observed data, and not at a hypothetical value for speakers that are zero years old.

5.1. Sociolinguistic Predictors

In this section, we report on the models that contain the sociolinguistic predictors age and gender. The purpose of this is twofold: First, to demonstrate that our approach can integrate sociolinguistic aspects and thus link dialectometry proper to social dialectology. Second, from sociolinguistic and dialectometric research, we have clear ideas on the kinds of patterns that should be expected. If our modeling results violate these expectations, it would be curious and potentially troubling; if, on the other hand, they largely match the expectations, our confidence in the results can be raised. To make our expectations more concrete, a large body of literature (e.g., Labov's (2003: 266) principle 2 and the evidence presented in support of it) suggests that women tend to use more standard forms, particularly in linguistically stable conditions. Similarly, many traditional dialectal forms are believed to be receding; we therefore assume, as per the apparent time hypothesis, that older speakers will tend to have higher frequencies for these features. We apply statistical significance at the customary threshold of .05 as a crude filter to keep out particularly unreliable signals.

For female speakers, we find that there are four clearly non-standard features where the usage frequency is clearly lower than for male speakers: [43] auxiliary deletion, as in (14), [47] relativizer *what*, as in (10), [49] *as what/than what*, as in (15), and [50] unsplit *for to*, as in (11).

(14) They gettin' too much. < DEV_007 >

(15) [...] but years ago they were a lot harder than what they are today [SAL_013]

In contrast, there is only one feature where women use a non-standard feature more often: [36] *never* as a past tense negator, as in (8) above. This feature is an atypical case: Cheshire, Edwards & Whittle (1989) note that although a range of authors consider it non-standard, they also attest its widespread use even in formal written English. It is not implausible that this feature may behave differently from more clearly stigmatized features. There are some features that cover both standard and non-standard usages. Feature [37], *wasn't* as in (16), is one such case. The frequency of this form is clearly going to depend on the prevalence of the *was - weren't* split: if the preferred negated form is always *weren't*, we should expect a lower frequency of *wasn't*. Female speakers, however, use *wasn't* more often than men. The remaining features with significant differences are features that also exist in Standard English and either have non-standard extensions or are features undergoing language change. These include pronominal forms, primary and modal verbs, and features involved in classic grammatical alternations such as *that*/zero complementizers.

(16) There wasn't a great deal. < LND_004 >

For speaker age, we find a much clearer picture. We find that, as expected, older speakers have a clear tendency to use more archaic and non-standard features. These include [28] non-standard weak verb forms, as in (7) above, [33] multiple negation, as in (1) above, and [50] unsplit for to, as in (11) above, as well as the following:

(17) [27] *a*-prefixing: And the other week she was a-telling me, she said, [..] < KEN_003 >

(18) [30] non-standard *come:* [...] he come home on a Saturday afternoon [...] < LND_006 >

(19) [32] *ain't:* He says, You ain't had your rotten teeth out. < NTT_012 >

(20) [39] non-standard verbal *-s:* [...]so I goes round see, and hits the belt like that with mi hand < WIL_001 >

Again, we also find standard grammatical features, often ones that are undergoing language change. Our results here often match those reported in the literature; to exemplify, [16] possessive *have got*, as in (21), and [26] *(have) got to* as a marker of epistemic or deontic modality, as in (22), are both used less often by older speakers. Both match previous results by Tagliamonte (2004), where, for the birth dates covered in our corpus, older speakers showed a dispreference for the variants involving *got* while younger speakers used them more often. Interactions turn out not to matter too much: only four features exhibit a significant interaction between age and gender.

(21) Each class have got their own form captain [...] <HEB_023>

(22) Ehr, you 've got to bit them up proper […] <SOM_005>

Overall, the results of the models including sociolinguistic information often seem to match our expectations. There are, however, some effects that would be expected based on the literature but fail to show up. Feature [24], *must* as a marker of epistemic or deontic modality as in (6), for instance, is a feature that is widely considered an "obsolescing form" (Jankowski, 2004: 101). However, there is no general trend for lower frequencies with younger speakers, only for female speakers as an interaction. Feature [17], the *going to* future, which has been shown to still grammaticalize and expand in British dialects (Tagliamonte, Durham & Smith, 2014), exhibits no pattern based on speaker age. Nevertheless, we consider these results to be reassuring —what we do find is quite plausible, and largely in line with our expectations.

5.2. Geography

We now turn to the models that contain all speakers, but do not include any predictors beyond geography and, as an offset, text size. We first report summary information on the individual feature models, then take the aggregational perspective. To generate distances from the individual models, we follow the method outlined earlier for the normalization-based CBDM approach as closely as possible. The major difference is that, as input to the distance calculation, we use the model predictions *pttw* instead of the normalized counts. There is another minor technicality, which involves exceedingly rare phenomena. For the normalized counts, absences were coded as −1, corresponding to 0.1 *pttw*, instead of taking the logarithm, as the logarithm of zero is undefined. The models will in general not predict exactly zero tokens, but numbers that are arbitrarily close to zero, and therefore without lower bound under logarithmic transformation. This means that, in contrast to the process on normalized values, rare features would have an undue influence. Instead, we enforce a lower bound of again -1 for the logarithmically transformed frequency, keeping the resulting values in a similar range for both processes.

Looking at the models individually, we find that for the majority of features geographic information is significant. Only eleven of the 57 features have a geographic smoother with a non-significant *p*-value.[10] The GAM solution used here also reports the proportion of the total deviance in the data that the model explains; in some cases, a non-significant (or marginally significant one) can still have considerable explanatory power. The major example here is [9] (the s-genitive), where the geographic smoother is only marginally significant, but the model nevertheless explains almost 46.6 percent of the deviance. Similar cases are [43] auxiliary deletion, as in (14) above, at 37.3 percent and, to a lower degree, [49] *as what/than what*, as in (15) above, and [52] gerundial complementation, as in (23), where the model explains about 10 percent.

(23) Oh, it was before I started working [...] <GLA_001>

Most of the other non-significant features hover between 1 and about 7 percent of the deviance. This pattern holds across the full dataset: the *p*-value correlates negatively (at around r = -0.41) with explanatory power. In other words, the lower the *p*-value, the better geography can explain the observed distribution of features. Restricting our analysis to those cases where the smoother is significant, we find a left-skewed distribution, with the peak at around 20 percent of explained deviance. The median is at around 28 percent, the mean slightly higher at 32 ($SD = 20$). There is a small number of outlier features where the simple model accounts for over 80 percent, namely [22] the present perfect with *be,* as in (5) above, [27] *a-prefixing*, as in (24), and the negator *nae* [31], as in (25).

(24) [...] What 's he 're a-jumpin' at? <SFK_006>

(25) Ach, it s- s- sounds good, but it wasnae really. <MLN_006>

All of these are features with a very marked geographic distribution, with high peaks in individual counties or regions (the Scottish Lowlands for [31], Suffolk for the others), and almost complete absence elsewhere. On the other side of the spectrum, there are a few features where geography is relevant, but not very informative. Looking at the sixteen cases where geography accounts for less than 20 percent of the deviance, we find mostly features of Standard English: the pronouns *us* and *them*, relativizers, the *of*-genitive, zero complementation and so on. There are only three clearly non-standard features in this list: [1] non-standard reflexives, as in (2) above , [43] *there is/was* with plural subjects, as in (26), and [44] non-standard *was*, as in (9) above. All of these are among the most widespread features in Britain or even worldwide, as the relevant surveys attest (Britain, 2010; Kortmann, 2004, 2013). In short, the features that are particularly weakly influenced by geography are those that are available in most locations —a quite plausible result. The remaining features, about half of the total feature set, lie in the region ranging from 20 to about 60 percent.

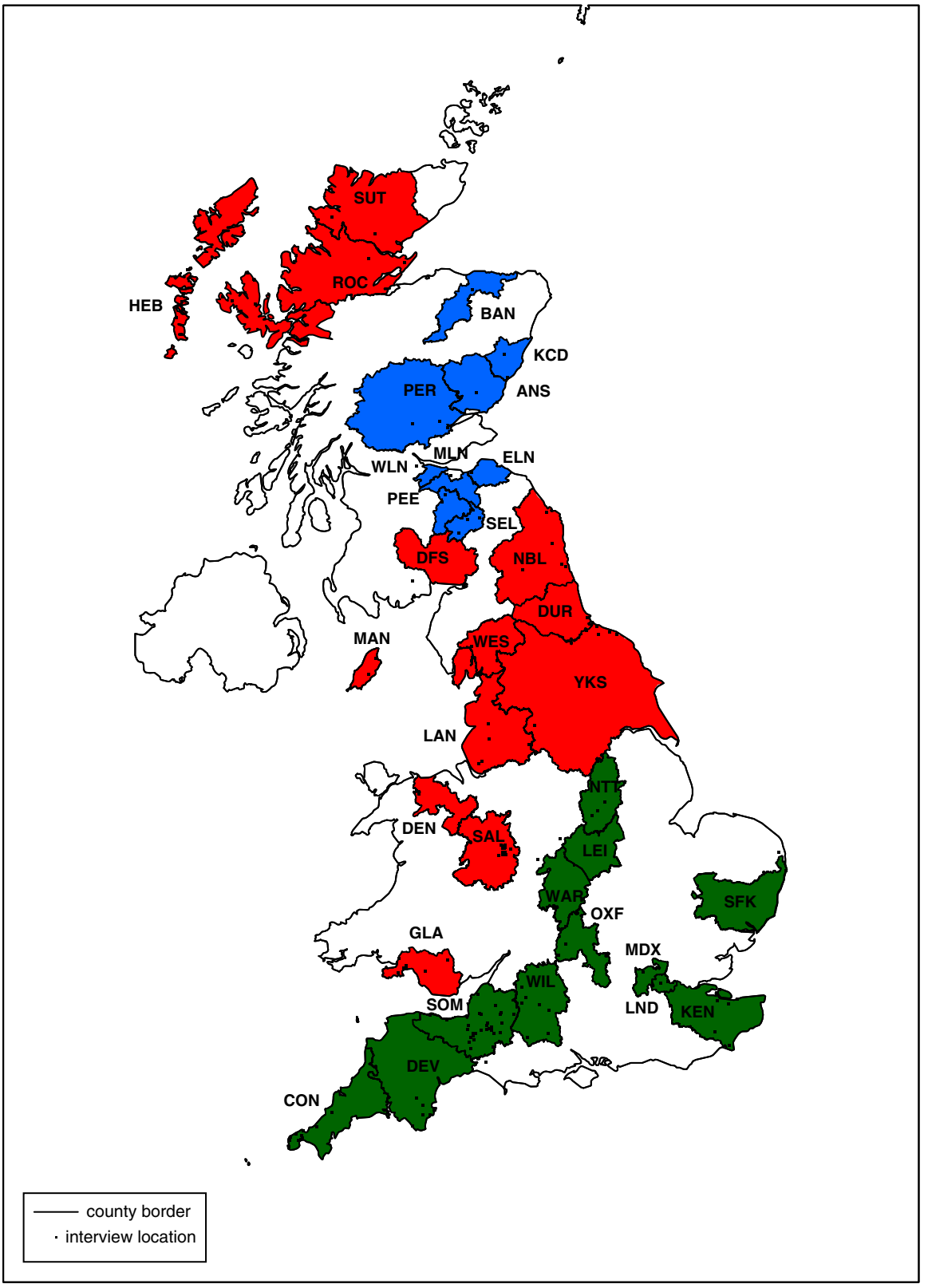
(26) There was all kinds of bits of quirks to keep the job as quiet as ever they possible could. <YKS_008>

Map 2 (right) shows the resulting geographic pattern for one feature, namely multiple negation [33], as in (1). The geographic pattern displayed in the background (i.e., the probabilistic pattern) is highly significant ($p < 0.01$) and very informative at almost 48 percent of explained deviance. Lighter yellow/orange tones indicate higher frequencies, while the dark green indicates lower frequencies, in this case virtual absence. The small black dots in the background show the interview locations; the display of the smoother is constrained to the area surrounding these points. The left-hand side of Map 2 displays the normalization-based values for the same feature. The legends in the top right corners illustrate the color gradients used, and highlight the endpoints of the scale, in observations *pttw*. To give an example, in the county with the most tokens, Kent, we have 11 tokens *pttw*, and the GAM predicts a higher value of 16.2 as the highest value for any individual location. Observe that, in this case, the highest prediction is even higher than the observed value; this results from within-county variation patterns.[6] The lowest observed county value is 0, complete absence, while the lowest predicted value is 0.2. For these areas, there is not enough variation to assert a lower value, even if no token was observed. Even in Scotland, which seems uniformly green in the normalized values, there are seven observations in around 200,000 words, which suggests an overall frequency of 0.34. Of course, there is variation within Scotland, and Angus alone accounts for almost half of the Scottish tokens. Nevertheless, it is plausible that our best guess even for the lowest-frequency areas is a value that is very close to zero, but slightly higher. The red lines indicate the overall shape of the frequency landscape: the feature is most prevalent along the southern and, to a lesser degree, eastern coasts of England, and decreases as one moves north or west from there.

Using these models, we can calculate the predictions for the counties, more precisely their centers, and proceed as previously described. To recapitulate briefly, the resulting per-county frequencies *pttw* are collected into an $N \times p$ frequency matrix. This frequency matrix is then used to calculate an $N \times N$ distance matrix using the Euclidean distance measure. The distance measure we then subject to hierarchical clustering with noise using Ward's method. The result is displayed in Map 4. Despite the substantial differences between the methods—one working with straightforward normalized frequencies, the other with an elaborate post-processing of the raw data using generalized additive modeling—the large-scale distribution is a remarkably similar tripartite division into a southern English area, a northern English area, and a Scottish Lowlands area. Gone, however, are the many outliers that were present in Map 2; all clusters are now geographically contiguous, with the exception of the Scottish Highlands and the Hebrides, which again show the largest similarity to the northern English cluster. Looking at the results in slightly greater detail, we find that there are smaller differences at the borders of these areas: Dumfriesshire and Northumberland have moved from the Scottish Lowlands cluster to the northern English cluster, and the Midlands counties east of Shropshire now exclusively group with the south. This increase in areal cohesion is also reflected in the relationship between geographic and GAM-derived linguistic distances, displayed in Figure 3. It bears noting that the curve has the sublinear curve predicted by Seguy's law (Nerbonne, 2010), in contrast to the one resulting from the normalization-based procedure. Furthermore, the explanatory power of (logarithmically transformed) geography increases greatly, from less than 4 percent to 58.1 percent. This is hardly surprising—the assumption of geographic coherence is central to the model formulation. Nevertheless, we can interpret this as an upper boundary: it is quite likely that having more data for particularly sparse counties, where the model assumptions carry greater weight, would increase the variability somewhat, but less likely to lead to a reduction. This also means that there is significant linguistic information left in the aggregated model predictions, as 58 percent is very high, but much closer to other dialectometrical estimates (e.g., the 45 percent reported in Spruit et al., 2009, for syntactic distances in Dutch dialects) than to (a dialectologically almost meaningless) 100 percent.

6. DISCUSSION AND CONCLUSION

Let us briefly recapitulate our approach and main results. We began by discussing data availability in dialectometry, the *missing data effect*, and why it may be particularly troubling for frequency-based dialectometry, including corpus-based variants. After outlining the data used in this paper, we presented the normalization-based approach to CBDM and showed that on the present dataset, it yields plausible results, with some peculiarities: the normalization-based solution suffered from outliers that were hard to explain, and the relationship between geographic and linguistic distances had both a much lower explanatory power than one would expect based on the dialectometrical literature and a shape that is linear rather than sublinear. In the previous section, we showed that a technically more sophisticated solution based on generalized additive models yields large-scale dialect areas consistent with the first method. There were, however, major differences with regard to the peculiarities, as we had hypothesized. The presence of outlying locations is a characteristic that Goebl (1991) reports for



Map 4. Cluster map based on the model-based CBDM approach. Displayed: 3-cluster solution.

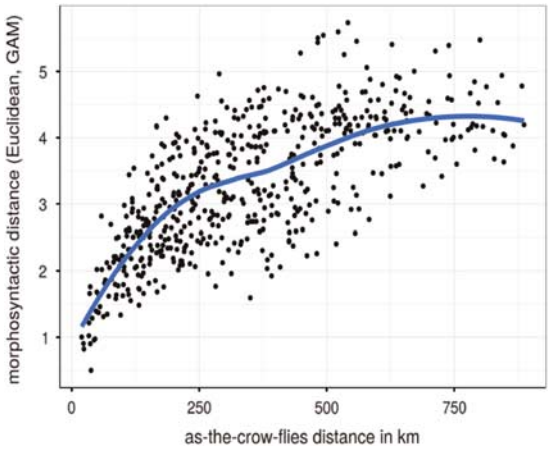


Figure 3. Correlating model-based morphosyntactic distances with as-the-crow-flies distances (logarithmic $r = .76$, $p < .001$, $R^2 = 58.1\%$).

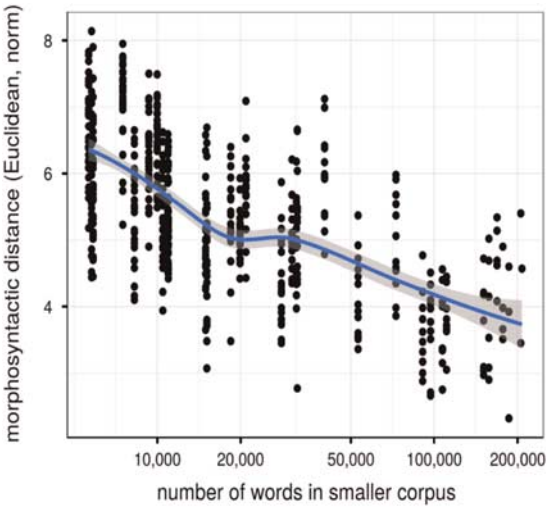


Figure 4. Number of words in smaller corpus of county pair plotted against their morphosyntactic normalization-based distance.

locations in atlas-based dialectometry where data is sparse. The second issue has been linked to such sparsity as well: Streck (2014) reports that removing the 40 percent of location pairs that involve those locations with the least amount of data almost doubled the percentage of variance that as-the-crow-flies distance can explain. We argued earlier that processing the raw data with GAMs can alleviate these symptoms, and indeed this has turned out to be the case. The model-derived distances correlate more strongly, and sublinearly, with the spatial configuration of their locations, and the outliers have vanished. We will now discuss the two methods by comparing some of their properties directly, and argue that such modeling is an appropriate choice for dialectometric purposes.

If our hypothesis that the outliers and low fit result from sparsity is correct, we would expect the counties that have reduced coverage to behave systematically differently from those with ample text. The distances, however, apply to pairs of locations, while the number of words is a property of an individual location. As our hypothesis predicts that smaller sizes should have the greatest impact, it is sensible to associate each distance with the minimum size of either subcorpus involved. This way, a pairing that is assigned the value of 50,000 ensures that both corpora at least reach that level of coverage, and the higher this number is, the more confidence we can place in the distance measurement of this pair. A small downside to this is that the right-hand side of the scale thins out—the county with the lowest amount of running text (Banffshire), contributes 33 individual points to the analysis, as it is the smallest corpus in all its 33 pairings. This makes this county particularly prominent visually. As the number of words for a county increases, the number of points on the corresponding spot on the x-axis decreases, as there are fewer and fewer counties that have at least as much text. Figure 4 displays the minimum number of words on the x-axis on a logarithmic scale, and the y-axis shows the normalization-based distance. The blue line is a LOESS smoother that indicates the overall trend. The data follow an almost perfectly straight downward-sloping line, with a small plateau of stability after 20,000 words. In other words, the relationship is a logarithmic decay: as the number of words in a county subcorpus increases, its distance to larger subcorpora decreases, but the rate at which it does so decreases as well. This relationship is visually strong, and this is confirmed by examining the correlation numerically: the log-transformed minimum number of words can account for 44 percent of the variance in the normalization-based distances. Note also that there is no discernible relationship between minimum size and geographic distance ($r = -0.08$, and no visible pattern when plotted)—any such relationship is therefore not due to the spatial distribution of locations. In the model-derived distances, this pattern disappears almost completely, with only 2 percent of the variance in linguistic distances attributable to minimum text size. All of this is consistent with the hypothesis. While we do expect the logarithmic decay to cease at some point (after all, it is unlikely that the actual difference between dialects is arbitrarily close to zero), none is apparent so far—the influence of corpus size on the normalized distances diminishes, but does not vanish.

How good, then is our solution to this problem—geographic smoothing using generalized additive models? It seems to pass a gauntlet of consistency checks—strength and shape of the relationship of linguistic and geographic distances, influence of geography, and coherence and interpretability of the resulting areal

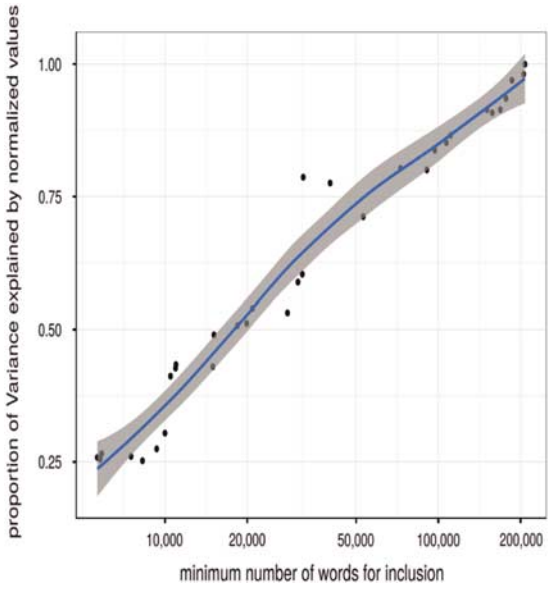


Figure 5. Percent of explained variance between normalization-based and model-based distance as a function of the minimum corpus size required for inclusion.

classification. Nevertheless, this comes at a cost: the Fundamental Dialectological Postulate has to be accepted beforehand, limiting the inferences one can make from the data. The smoothing may also be too aggressive, presenting a limited picture of the actually existing variation. Given that there is no external measure for morphosyntactic differences between these locations, we cannot directly evaluate exactly how accurate this strategy is. What we can do, however, is compare it to another strategy—a tactical retreat to the counties that are most plentiful in terms of the textual coverage, where normalization should work best. Taking all the counties, we find that there is a clear but limited relationship between normalization- and model-based metrics, with a linear R^2 of 0.26. We can now successively drop the county with the lowest amount of text and see how this changes the result (without recalculating the models). The development is displayed in Figure 5, with the minimum number of words required for inclusion on the x-axis (again on a log scale) and the linear R^2 on the y-axis. There is a clear relationship: the more one restricts attention to well-covered areas, the more the results of the two methods approximate one another. The R^2 for the logarithmic relationship is 0.85. In words, the two analyses increasingly resemble one another. This means that there are few downsides to modeling compared to exclusion; including the low-data counties yields largely the same results for the high-data counties, but modeling also leads to plausible results for the rest, and allows them to contribute to the analysis.

We wish to make a final point: Dialectometry has heralded the beneficial properties of aggregation for noise reduction and pattern identification (see e.g., Nerbonne, 2009: 129, who considers it "at the heart of the benefits of dialectometry"). Our analysis confirms this, but also shows important limitations: there are biases that are impervious to aggregation; for noise that persists across features, or limits which features can be compared for individual location pairs, aggregation may simply not be enough. Our example was related to frequency-based dialectometry, but similar concerns should apply for categorical data. We urge scholars to be mindful of this, and include analyses based on data availability in cases where the data basis is not exactly equivalent for all locations.

Acknowledgements

The research presented in this article is partially based on work done for the first author's PhD dissertation as well as the second author's book *Grammatical variation in British English dialects*, but has been substantially reworked and expanded. We are grateful to Bernd Kortmann, Guido Seiler, Peter Auer, John Nerbonne, Joan Bresnan, and to the audience at Methods in Dialectology XV where an earlier version of this research was presented. We use cartographic material provided by Natural Earth, GADM, the Scottish Government Spatial Data Infrastructure, and the Great Britain Historical GIS Project; licensing terms can be found in the appendix. Funding from the Freiburg Institute for Advanced Studies (FRIAS) is gratefully acknowledged.

Supplementary material

To view supplementary material for this article, please visit http://dx.doi.org/10.1017/jlg.2018.6

Notes

[1] For the purposes of this paper, we consider frequency differences of linguistic phenomena as observed in naturalistic corpora a (potentially noisy) proxy of the underlying linguistic differences between locations. Some readers may prefer methods that are explicitly based on linguistic variables, showing likelihood of use as opposed to surface frequency. Our method can also incorporate this approach; examples can be found in Wolk (2014). The advantage of surface frequencies for present purposes is that they can be directly applied to all the features under study equally.

[2] The somewhat coarsely meshed network of interview locations makes it hard to read continuum maps, as subtle color comparisons over areas separated by white space are visually challenging. Interested readers can find the continuum maps for both approaches in the online appendix.

[3] An anonymous reviewer points out that the Fundamental Dialectological Postulate need not hold, as linguistic

behavior may be influenced by factors other than distance, including migration or political borders. We agree in general, but consider it applicable in situations with a long settlement history; also note that both the principle and our application of it are intended as empirical generalizations and allow for deviations given sufficient evidence.

[4] The use of measures of spatial autocorrelation, namely Moran's I, instead of correlation coefficients was recommended to us by an anonymous reviewer. We agree that there are clear advantages to using Moran's I to investigate the distribution of individual features, but see no clear way to extend this method to patterns of aggregate distances.

[5] The interview locations are unfortunately not balanced throughout the regions, and may be more concentrated in certain counties and spread out elsewhere. We will not make claims about places not covered in our data.

[6] We specifically used Ward's minimum variance method (Ward, 1963), an algorithm that tends to create small and even-sized clusters. Note that simple clustering can be unstable, which is why we clustered with noise (Nerbonne, Kleiweg, Manni & Heeringa, 2008): The original distance matrix was clustered repeatedly, adding some random amount of noise ($c = \sigma/2$) in each run.

[7] For example, Paul Allison advises against overdispersed models at http://statisticalhorizons.com/zero-inflated-models.

[8] For the set of models that include social information on the speakers, the list of features fit using REML because of convergence problems is [19], [39], [55], while the set with too large family parameters remains the same.

[9] The full list is: [7] synthetic comparison, [9] the s-genitive, [10] preposition stranding, [17] the *going to* future, [36] preverbal *never*, [43] auxiliary deletion, [49] *as what/than what*, [51] infinitival and [52] gerundial complementation , [55] lack of inversion, and [57] the prepositional dative.

[10] In other cases, especially those involving exceptionally high values in low-data counties, the GAM prediction may be much lower. See for example feature [31], *-nae,* where the highest observed value is 144 *pttw*, but the highest prediction only 79.

References

Aldenderfer, Mark S. & Roger K. Blashfield. 1984. *Cluster analysis*. London: Sage Publications.

Anderwald, Lieselotte & Benedikt Szmrecsanyi. 2009. Corpus linguistics and dialectology. In Lüdeling, Anke & Merja Kytö (eds.), *Corpus linguistics: An international handbook*, Handbücher Zur Sprach- Und Kommunikationswissenschaft /Handbooks of Linguistics and Communication Science 29(1), 1126–39. Berlin: Mouton de Gruyter.

Branigan, Holly. 2007. Syntactic priming. *Language and Linguistics Compass* 1(1/2). 1-16. doi: 10.1111/j.1749-818X.2006.00001.x.

Britain, David. 2010. Grammatical variation in the contemporary spoken English of England. In Kirkpatrick, Andy (ed.), *The handbook of world Englishes*, 37-58. London: Routledge.

Bybee, Joan L. 2010. *Language, usage and cognition*. Cambridge: Cambridge University Press.

Chambers, Jack K. & Peter Trudgill. 1998. *Dialectology*, 2nd edn. Cambridge: Cambridge University Press.

Cheshire, Jenny, Viv Edwards & Pamela Whittle. 1989. Urban British dialect grammar: the question of dialect levelling. *English World-Wide* 10. 185-225.

Eisenstein, Jacob. 2018. Identifying regional dialects in on-line social media. In Boberg, Charles, John Nerbonne & Dominic Watts (eds.), *The handbook of dialectology*. New York: Wiley-Blackwell.

Goebl, Hans. Rätoromanisch versus Hochitalienisch versus Oberitalienisch. *Ladinia* 1. 39-71.

Goebl, Hans. 1982. *Dialektometrie: Prinzipien und methoden des einsatzes der numerischen taxonomie im bereich der dialektgeographie*. Wien: Österreichische Akademie der Wissenschaften.

Goebl, Hans. 1984. *Dialektometrische studien: Anhand Italoromanischer, Rätoromanischer Und Galloromanischer sprachmaterialien aus AIS und ALF*. Tübingen: Niemeyer.

Goebl, Hans. 1993. Dialectometry: A short overview of the principles and practice of quantitive classification of linguistic atlas data. In Köhler, Reinhard & Burghard Rieger (eds.), *Contributions to Quantitative Linguistics: Proceedings of the First International Conference on Quantitative Linguistics, QUALICO, Trier, 1991*, 277–315. Dordrecht: Kluwer.

Goebl, Hans. 2007. A bunch of dialectometric flowers: A brief introduction to dialectometry. In Smit, Ute, Stefan Dollinger, Julia Hüttner, Gunter Kaltenböck & Ursula Lutzky (eds.), *Tracing English through time: Explorations in language variation*, 133-172. Wien: Braumüller.

Great Britain Historical GIS Project. 2004. Great Britain Historical GIS. University of Portsmouth.

Gries, Stefan Th. 2012. Frequencies, probabilities, and association measures in usage-/exemplar-based linguistics: Some necessary clarifications. *Studies in Language* 36(3). 477-510.

Grieve, Jack. 2014. A comparison of statistical methods for the aggregation of regional linguistic variation. In Szmrecsanyi, Benedikt & Bernhard Wälchli (eds.), *Aggregating dialectology, typology, and register analysis: Linguistic variation in text and speech*, Lingua & Litterae 2853-2888. Berlin: Walter de Gruyter.

Grieve, Jack. 2016. *Regional variation in written American English*. Cambridge: Cambridge University Press.

Heeringa, Wilbert. 2004. *Measuring dialect pronunciation differences using Levenshtein distance*. Groningen, Netherlands: University of Groningen dissertation.

Hilbe, Joseph M. 2007. *Negative binomial regression*. Cambridge: Cambridge University Press.

Hernández, Nuria. 2006. *User's guide to FRED*. Freiburg: University of Freiburg.

Huang, Yuan, Diansheng Guo, Alice Kasakoff & Jack Grieve. 2016. Understanding US regional linguistic variation with Twitter data analysis. *Computers, Environment and Urban Systems* 59. 244-255.

Jain, Anil K., M. Narasimha Murty & Patrick J. Flynn. 1999. Data clustering: A review. *ACM Computing Surveys* 31(3). 264-323.

James, Gareth, Daniela Witten, Trevor Hastie & Robert Tibshirani. 2013. *An introduction to statistical learning: With applications in R*. New York: Springer.

Jankowski, Bridget. 2004. A transatlantic perspective of variation and change in English deontic modality. *Toronto Working Papers in Linguistics* 23(2). 85-113.

Kortmann, Bernd & Benedikt Szmrecsanyi. 2004. Global synopsis: Morphological and syntactic variation in English. In Bernd Kortmann, Edgar Schneider, Kate Burridge, Rajend Mesthrie & Clive Upton (eds.), *A Handbook of varieties of English* vol. 2. 1142-1202. Berlin: Mouton de Gruyter.

Manning, Chris & Hinrich Schütze. 1999. *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.

Nerbonne, John. 2009. Data-driven dialectology. *Language and Linguistics Compass* 3(1). 175-198.

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

Nerbonne, John & Peter Kleiweg. 2007. Toward a dialectological yardstick. *Journal of Quantitative Linguistics* 14 (2). 148-166.

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

Nerbonne, John, Peter Kleiweg, Franz Manni & Wilbert Heeringa. 2008. Projecting dialect differences to geography: Bootstrapping clustering vs. clustering with noise. In Christine Preisach, Lars Schmidt-Thieme, Hans Burkhardt & Reinhold Decker (eds.), *Data analysis, machine learning, and applications. Proceedings of the 31st Annual Meeting of the German Classification Society*, 647-654. Berlin: Springer.

Orton, Harold, Stewart Sanderson & J. D. A. Widdowson. 1978. *The Linguistic Atlas of England*. London: Croom Helm.

Pickl, Simon, Aaron Spettl, Simon Pröll, Stephan Elspaß, Werner König & Volker Schmidt. 2014. Linguistic distances in dialectometric intensity estimation. *Journal of Linguistic Geography* 2(1). 25-40.

Sankoff, David. 1987. Variable Rules. In Ulrich Ammon, Norbert Dittmar & Klaus J. Mattheier (eds.), *Sociolinguistics / Soziolinguistik: An International Handbook / Ein internationales Handbuch, Vol.2*, 984-997. Berlin: de Gruyter.

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

Shackleton, R. G. 2007. Phonetic variation in the traditional English dialects: A computational analysis. *Journal of English Linguistics* 35(1). 30-102. doi: 10.1177/0075424206297857.

Spruit, Marco René. 2006. Measuring syntactic variation in Dutch dialects. *Literary and Linguistic Computing* 21(4). 493-506.

Spruit, Marco René, Wilbert Heeringa & John Nerbonne. 2009. Associations among linguistic levels. *Lingua* 119(11). 1624-1642. doi: 10.1016/j.lingua.2009.02.001.

Streck, Tobias. 2014. Alemannisch quantitativ. Zur Erklärungskraft der Geografie für aggregierte Dialektunterschiede. In Pia Bergmann, Karin Birkner, Peter Gilles, Helmut Spiekermann & Tobias Streck (eds.), *Sprache im Gebrauch: Räumlich, zeitlich, interaktional. Festschrift für Peter Auer*, 157-173. Heidelberg: Winter (OraLingua 9).

Szmrecsanyi, Benedikt. 2006. *Morphosyntactic persistence in spoken English: A corpus study at the intersection of variationist sociolinguistics, psycholinguistics, and discourse analysis*. Berlin: Mouton de Gruyter.

Szmrecsanyi, Benedikt. 2008. Corpus-based dialectometry: Aggregate morphosyntactic variability in British English dialects. *International Journal of Humanities and Arts Computing* 2(1/2). 279-296.

Szmrecsanyi, Benedikt. 2010. *The morphosyntax of BrE dialects in a corpus-based dialectometrical perspective: Feature extraction, coding protocols, projections to geography, summary statistics*. Freiburg: University of Freiburg. URN: urn:nbn:de:bsz:25-Opus-73209, URL: http://www.freidok.uni-freiburg.de/volltexte/7320/.

Szmrecsanyi, Benedikt. 2011. Corpus-based dialectometry: A methodological sketch. *Corpora* 6(1). 45-76.

Szmrecsanyi, Benedikt. 2013. *Grammatical variation in British English dialects: A study in corpus-based dialectometry*. Cambridge: Cambridge University Press.

Szmrecsanyi, Benedikt & Nuria Hernández. 2007. *Manual of information to accompany the Freiburg Corpus of English Dialects Sampler ("FRED-S")*. Freiburg: University of Freiburg. http://www.freidok.uni-freiburg.de/volltexte/2859/.

Szmrecsanyi, Benedikt & Bernhard Wälchli. eds. 2014. *Aggregating dialectology, typology, and register analysis: Linguistic variation in text and speech. Lingua & Litterae*. 28 Berlin: Walter de Gruyter.

Szmrecsanyi, Benedikt & Christoph Wolk. 2011. Holistic corpus-based dialectology. *Brazilian Journal of Applied Linguistics/Revista Brasileira de Linguística Aplicada* 11(2). 561-592.

Tagliamonte, Sali. 2004. *Have to, gotta, must*: Grammaticalisation, variation and specialization in English deontic modality. In Christian Mair & Hans Lindquist (eds.), *Corpus approaches to grammaticalization in English*, 33-55. Amsterdam: Benjamins.

Tagliamonte, Sali A., Mercedes Durham & Jennifer Smith. 2014. Grammaticalization at an early stage: future *be going to* in conservative British dialects. *English Language and Linguistics* 18(1). 75-108.

Tobler, Waldo. 1970. A computer movie simulating urban growth in the Detroit region. *Economic Geography* 46(2). 234-240.

Viereck, Wolfgang. The Computerisation and quantification of linguistic data: Dialectometrical methods. In Alan R. Thomas (ed.), *Methods in Dialectology : Proceedings of the Sixth International Conference Held At the University College of North Wales, 3rd-7th August 1987*. 523-550. Clevedon: Multilingual Matters.

Viereck, Wolfgang, Heinrich Ramisch, Harald Händler, Petra Hoffmann & Wolfgang Putschke. 1991. *The Computer Developed Linguistic Atlas of England*. Tübingen: Niemeyer.

Ward, Joe H. Jr. 1963. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association* 58. 236-244.

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

Wieling, Martijn, John Nerbonne & R. Harald Baayen. 2011. Quantitative social dialectology: Explaining linguistic variation geographically and socially. *PLoS ONE* 6 (9).

Wolk, Christoph. 2014. *Integrating aggregational and probabilistic approaches to language variation*. Freiburg: University of Freiburg dissertation.

Wolk, Christoph & Benedikt Szmrecsanyi. 2016. Top-down and bottom-up advances in corpus-based dialectometry. In Marie-Hélène Côté, Remco Knooihuizen & John Nerbonne (eds.), *The future of dialects*, 225-244. Berlin: Language Science Press.

Wood, Simon N. 2006. *Generalized additive models: an introduction with R*. Boca Raton, FL: Chapman & Hall/CRC.

Appendix: the feature catalogue

A. Pronouns and determiners

[1] non-standard reflexives (e.g. they didn't go theirself)
[2] standard reflexives (e.g. they didn't go themselves)
[3] archaic thee/thou/thy (e.g. I tell thee a bit more)
[4] archaic ye (e.g. ye'd dancing every week)
[5] us (e.g. us couldn't get back, there was no train)
[6] them (e.g. I wonder if they'd do any of them things today)

B. The noun phrase

[7] synthetic adjective comparison (e.g. he was always keener on farming)
[8] the of-genitive (e.g. the presence of my father)
[9] the s-genitive (e.g. my father's presence)
[10] preposition stranding (e.g. the very house which it was in)
[11] cardinal number + years (e.g. I was there about three years)
[12] cardinal number + year-Ø (e.g. she were three year old)

C. Primary verbs

[13] the primary verb TO DO (e.g. why did you not wait?)
[14] the primary verb TO BE (e.g. I was took straight into this pitting job)
[15] the primary verb TO HAVE (e.g. we thought somebody had brought them)
[16] marking of possession – HAVE GOT (e.g. I have got the photographs)

D. Tense and aspect

[17] the future marker BE GOING TO (e.g. I'm going to let you into a secret)
[18] the future markers WILL/SHALL (e.g. I will let you into a secret)
[19] WOULD as marker of habitual past (e.g. he would go around killing pigs)
[20] used to as marker of habitual past (e.g. he used to go around killing pigs)
[21] progressive verb forms (e.g. the rest are going to Portree School)
[22] the present perfect with auxiliary BE (e.g. I'm come down to pay the rent)
[23] the present perfect with auxiliary HAVE (e.g. they've killed the skipper)

E. Modality

[24] marking of epistemic and deontic modality: MUST (e.g. I must pick up the book)
[25] marking of epistemic and deontic modality: HAVE TO (e.g. I have to pick up the book)
[26] marking of epistemic and deontic modality: GOT TO (e.g. I gotta pick up the book)

F. Verb morphology

[27] a-prefixing on -ing-forms (e.g. he was a-waiting)
[28] non-standard weak past tense and past participle forms (e.g. they knowed all about these things)
[29] non-standard past tense done (e.g. you came home and done the home fishing)
[30] non-standard past tense come (e.g. he come down the road one day)

G. Negation

[31] the negative suffix -nae (e.g. I cannae do it)
[32] the negator ain't (e.g. people ain't got no money)
[33] multiple negation (e.g. don't you make no damn mistake)
[34] negative contraction (e.g. they won't do anything)
[35] auxiliary contraction (e.g. they'll not do anything)
[36] never as past tense negator (e.g. and they never moved no more)
[37] WASN'T (e.g. they wasn't hungry)
[38] WEREN'T (e.g. they weren't hungry)

H. Agreement

[39] non-standard verbal -s (e.g. so I says, What have you to do?)
[40] don't with 3rd person singular subjects (e.g. if this man don't come up to it)

[41] standard doesn't with 3rd person singular subjects (e.g. if this man doesn't come up to it)
[42] existential/presentational there is/was with plural subjects (e.g. there was children involved)
[43] absence of auxiliary BE in progressive constructions (e.g. I said, How ø you doing?)
[44] non-standard WAS (e.g. three of them was killed)
[45] non-standard WERE (e.g. he were a young lad)

I. Relativization

[46] wh-relativization (e.g. the man who read the book)
[47] the relative particle what (e.g. the man what read the book)
[48] the relative particle that (e.g. the man that read the book)

J. Complementation

[49] as what or than what in comparative clauses (e.g. we done no more than what other kids used to do)
[50] unsplit for to (e.g. it was ready for to go away with the order)
[51] infinitival complementation after BEGIN, START, CONTINUE, HATE, and LOVE (e.g. I began to take an interest)
[52] gerundial complementation after BEGIN, START, CONTINUE, HATE, and LOVE (e.g. I began taking an interest)
[53] zero complementation after THINK, SAY, and KNOW (e.g. they just thought ∅ it isn't for girls)
[54] that complementation after THINK, SAY, and KNOW (e.g. they just thought that it isn't for girls)

K. Word order and discourse phenomena

[55] lack of inversion and/or of auxiliaries in wh-questions and in main clause yes/no-questions (e.g. where you put the shovel?)
[56] the prepositional dative after the verb GIVE (e.g. she gave [a job] to [my brother])
[57] double object structures after the verb GIVE (e.g. she gave [my brother] [a job])