Journal of Information Science
XX (X) pp. 1-13
© The Author(s) 2013
Reprints and Permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/016555150000000
jis.sagepub.com

Geographical Analysis of the Vernacular

John Holliday
Information School, University of Sheffield, United Kingdom

Clive Upton, Ann Thompson
School of English, University of Leeds, United Kingdom

Jonathan Robinson, Jon Herring, Holly Gilbert
British Library, United Kingdom

Paul Norman
School of Geography, University of Leeds, United Kingdom

Abstract
The BBC *Voices* project of 2005 resulted in a large repository of lexical, phonological and grammatical data from the UK, which included geographical references. In order to investigate the relationship between language and geography, various clustering algorithms have been applied to the *BBC Voices* data. Results show a clear spatial relationship, with well-defined, contiguous regions of UK language being identified. In order to prove the clustering methodology, Bayesian models have been generated for each region, and these have been tested using a set of non-standard expressions contributed by a small number of participants. Results of this second stage indicate that the models are, in most cases, able to identify the geographical region of each test participant based on the linguistic items they use.

Keywords
Geolinguistics; BBC Voices; Clustering; GIS; Language analysis; Language mapping

1. Introduction

The BBC *Voices* project of 2005 was an extensive investigation into the use of English language in the UK in the early twenty-first century. The project involved the collection of linguistic data, both spoken and written, through the use of an interactive website and through the BBC's network of reporters and local radio stations. As a result, a wealth of data has been made available for research into many aspects of language variation and dialectology.

Of particular interest is the inclusion of geographical location, as well as further biographical information, with the linguistic data. The geographical data allows for detailed analysis of the relationship between linguistic variation and geography. Analysis of this kind is not new, with related fields such as geolinguistics and dialectology being well established [1,2]. The spatial mapping of variation in dialect is a popular area of research which has led to the production of many maps which aim to define the geographical extent of a wide range of grammatical, phonological and lexical features. These maps and atlases tend, however, to treat these features individually, mapping the variation

Corresponding author:
John Holliday, Information School, The University of Sheffield, Regent Court, 211 Portobello Street, Sheffield, S1 4DP
Email: j.d.holliday@sheffield.ac.uk

in a single linguistic item [3,4].

The possibility of combining many such features allows us to identify regions of natural language in the UK, and to investigate the relationship between features in those regions. Similar regional delineation has been carried out based on a variety of factors, such as economy [5], governance [6], demographics [7], and even human interactions such as internet use [8] and banknote movement [9]. These studies usually seek to partition the country into discrete regions separated by defined boundaries. Whether such defined boundaries exist with regard to English language is highly questionable, a gradual change from one region to a neighbouring region being more likely [4]. The nature of the geographical information in the BBC *Voices* project, however, means that partitioning must be based on existing defined regions, such as postcode areas.

This paper reports on a study that aims to identify regions of natural language in the UK by applying cluster analysis methodologies to the lexical, phonological and grammatical data provided by the BBC *Voices* project. Since it is not initially clear how many such regions exist, the results of the hierarchical clustering are analysed in order to identify the most appropriate number. In order to prove the suitability of the clustering approach, the regions are tested using a machine learning methodology based on Bayesian categorisation. The testing procedure, which uses data provided by a sample of individuals, not only allows the validation of the clusters, but also provides a means for generating a map which indicates the linguistic geography for the test individual.

2. BBC Voices Data

Data for the BBC Voices project came from two main sources, a collection of sound recordings taken by local radio reporters, and an interactive website. The sound recordings, now available on the BBC website [10], are the result of a survey of 300 conversations involving 1201 people from all over the UK. The recordings covered several diverse themes, from fashion and local community to discussions on accent and local dialect. Many recordings concentrated on groups with differing social and ethnic make-up, one being limited to Asian teenagers, another concentrating on a Jamaican family, whilst other groups were more diverse. These recordings were analysed for grammatical and phonological variation by the British Library.

The *Voices* website included the *Language Lab* pages in which the public were asked to contribute their non-standard expressions for a set of 38 standard English concepts. Originally planned for collecting individual words, it soon became apparent that some submissions were expressions or phrases consisting of more than one word; ‘made up’ or ‘over the moon’, for example. For this reason, we use the term ‘lexical item’ to describe the submissions. The response to the website far exceeded expectations, with lexical items continuing to be submitted two years after the project had ended.

2.1. Grammar and Phonology data

A set of 126 grammatical characteristics, or features, and a set of 191 phonological features were identified for use in the following analysis. For each recorded conversation, often involving several members of the public, the occurrence of each feature was recorded in binary format, a one indicating the presence of the characteristic. An example of a grammatical characteristic would be the use of *singular object us*, as in ‘give us a go’ in contrast to standard English ‘give me a go’. Phonological characteristics relate to the linguistic sounds identified in the recordings, such as the vowel in words like ‘ant’ and ‘plant’, which are identical for some speakers but not for others.

Each recording was made at a single location, which was geo-coded in order that it could be represented spatially. These locations could then be represented by two sets of binary values, one for grammar and one for phonology, with each element representing the presence or absence of the respective feature. An example, illustrating the first seven recordings and the first seven phonological features, is shown in Table 1. In this table, ‘KIT’, ‘DRESS’ and ‘TRAP’ refer to lexical sets as defined in Wells [11]; ‘RP’ refers to vowel realisation as in Received Pronunciation; ‘other’ means vowel realisation contrasting with RP; and ‘lex. cond.’ is lexically-conditioned variant. In the recording C11903705, for example, instances of RP KIT, RP DRESS and RP TRAP have been identified. This representation forms the basis for all grammar and phonology analysis described in this paper.

2.2. Lexical data

The *Language Lab* website allowed the public to submit their own terms or expressions for 38 concepts, divided into six themes, as shown in Table 2. Submissions varied from concept to concept, ranging from 9,897 for *to play (a game)* to 29,275 for *drunk*. The submissions included biographical data such as age, gender, postcode and place of birth. The resulting data were sent to the *Whose Voices?* project at the University of Leeds [12] for decoding and analysis.

The result of this extensive analysis was a set of organised spreadsheets, one for each concept, in which analogous lexical terms are grouped appropriately and can be aggregated by postcode area (PCA). The grouping of analogous terms, part of the analysis carried out by the *Whose Voices?* project, allowed, for example, for cases such as ‘skive’, ‘skive off’, ‘skyve’, ‘skiving’, to be represented by one lexical item. For each item, the number of submissions for the respective postcode area was recorded. The top ten submitted lexical items were identified and represented as a percentage of the total for all top ten items. In two cases, a tie in the tenth position meant that the top eleven items were represented. For each PCA we therefore have values representing the proportion of submissions for the ten (or eleven) most frequently submitted lexical items. When taken over all 38 concepts, this produces a non-binary vector of 382 elements for each PCA. An example, illustrating the first seven PCAs and the first ten lexical items, is shown in Table 3. In this table, of those people from postcode AB (representing Aberdeen) whose submissions were found in the top ten terms for the concept ‘hot’, 34.59% used the term ‘Boiling’ or a term analogous to it. This representation forms the basis for all lexical analysis described in this paper.

Table 1. Phonology descriptors

Recording Code	RP KIT	KIT - other	KIT- lex. cond.	RP DRESS	DRESS - other	DRESS – lex. cond.	RP TRAP	TRAP – other
C11900301	1	0	0	1	1	0	1	1
C11900302	1	0	0	1	0	0	0	1
C11900305	1	0	0	1	0	0	1	0
C11903701	1	0	0	1	0	0	1	1
C11903705	1	0	0	1	0	0	1	0
C11903706	1	0	0	1	0	0	1	1
C11900201	1	1	0	1	0	0	1	1

Table 2. Concepts and themes from the Language Lab

Theme	Concept	Theme	Concept
How you feel	hot	What you call them	baby
	cold		mother
	tired		grandmother
	unwell		grandfather
	pleased		friend
	annoyed		male partner
What you do	to play (a game)		female partner
	to play truant		young person in cheap trendy clothes and jewellery
	to throw		
	to hit hard	What they wear	clothes
	to sleep		trousers
Getting personal	drunk		child's soft shoe worn for PE
	pregnant	Inside and out	main room of house (with TV)
	left-handed		long, soft seat in the main room
	lacking money		toilet
	rich		narrow walkway alongside buildings
	insane		to rain lightly
	attractive		to rain heavily
	unattractive		running water, smaller than a river
	moody		

Table 3. Lexical descriptors

PCA	Boiling	Roasting	Hot	Sweltering	Baking	Sweating	Warm	Scorching	Toasty	Melting
AB	34.59	30.83	12.78	3.76	0.75	4.51	6.77	0.75	0.75	4.51
AL	45.61	7.02	5.26	8.77	14.04	8.77	5.26	3.51	1.75	0.00
B	54.63	10.93	4.51	7.13	7.36	6.65	4.28	1.19	1.43	1.90
BA	52.94	6.95	8.56	5.88	8.56	3.74	8.56	1.60	2.14	1.07
BB	48.72	18.80	8.55	3.42	2.56	8.55	3.42	1.71	1.71	2.56
BD	53.33	14.07	1.48	11.1	2.96	4.44	9.63	1.48	0.74	0.74
BH	54.55	8.18	4.55	11.82	7.27	9.09	3.64	0.00	0.91	0.00

3. Cluster analysis

In order to identify regions of natural language in the UK using the data representations described above, a cluster analysis approach has been used. Cluster analysis [13,14], or clustering, is a method which seeks to categorise a set of objects by subdividing them into groups such that the similarity between objects within the same group is maximised, while the similarity between objects in different groups is minimised. Cluster analysis has been applied extensively in many areas, including document collections [15], paintings [16], chemical compound databases [17], and even Scotch whiskies [18].

Many different clustering algorithms have been developed, these being either hierarchical or non-hierarchical in operation. Hierarchical methods seek either to build a hierarchy of clusters, by starting from individual objects and merging similar objects or clusters together (the agglomerative approach), or to divide an initial cluster containing all objects into progressively smaller clusters (the divisive approach). Non-hierarchical clustering uses alternative methods to group the objects. However, this often requires prior knowledge of the desired number of clusters. The objects in this study are the PCAs, since it is our aim to group together those which exhibit similar characteristics, and separate groups which differ in language or dialect.

In order to perform a clustering operation, the degree of similarity between objects, or PCAs, must be quantified. This is facilitated by the use of a similarity (or dissimilarity) measure which takes as its input two vectors, one for each object. The vectors describe features of the object in binary or non-binary form, and the measure quantifies the degree of similarity between the two vectors. Many such measures exist [19]; here we have used the Squared Euclidean distance, the Cosine coefficient, and the Pearson coefficient.

In this study, we applied seven agglomerative hierarchical clustering methods (single linkage, complete linkage, average link between clusters (or UPGMA), average link within clusters, centroid, median, and Ward's minimum variance method [20]) and one non-hierarchical clustering method (*k*-means [13]) to the grammar, phonology and lexical representations described above. Clustering was performed in the SPSS statistical package [21]. These eight clustering methods were applied to the lexical data, Ward's was the only methodology used to cluster the phonology and grammar data. Squared Euclidean distance was applied to all clustering methods, the Cosine and Pearson coefficients were used for single linkage, complete linkage, UPGMA, and average link within clusters.

3.1. Consensus clustering

The clustering methods chosen produced quite different results. In order to assimilate these to produce a definitive solution, a method known as consensus clustering was performed. In standard clustering, the similarity between objects, as described above, is given by some measure of the similarity between the descriptors. The method of consensus clustering chosen here takes, as the similarity value, the number of clusters, from the initial clustering methods, in which two objects co-exist. So, for example, if two PCAs are found in the same cluster in four of the eight clustering methods, then their similarity is 0.5. In order that some methods aren't over represented, only the squared Euclidean analysis for each method was used in the consensus clustering stage. Consensus clustering was carried out using the CLUTO package [22]. The agglomerative clustering methodology available in CLUTO is UPGMA.

3.2. Identifying the number of clusters

Since agglomerative hierarchical clustering merges clusters one-by-one, the procedure can be stopped at any level, from the initial 121 cluster which represent the individual 121 PCAs, right through to the final single cluster representing all PCAs. It is necessary, therefore, to identify the level at which the clustering reflects the natural linguistic regions and output the clustering results for that level. Elmes [23] uses twelve regions of the UK when characterising the BBC *Voices* data. Initial clustering studies would suggest several more. Elmes, for instance, treats Scotland as a whole, whereas our initial investigations indicate several distinct regions within Scotland.

Natural clusters can be identified by applying stopping rules to the hierarchical clustering output [24]. Many such rules have been proposed, some more successfully than others. Many of these rules are based on the inter-cluster variance, others are based on the similarity values at which clusters merge (or *agglomerative schedule* in SPSS). Since we have used a commercial package (SPSS), it is only possible to apply rules based on the later of these. Several rules were applied to the agglomerative schedule with little success, indicating that the merging of clusters is regular. The same result was observed when clustering was repeated using the Clustan Graphics package [25], which incorporates the stopping rule of Mojena [26]. Further investigations into the optimum number of clusters is expected; however, for this study, it was decided to set the number of regions by observation of the cluster maps produced, and their relationship to Elmes's twelve regions.

4. Bayesian modelling

Bayesian modelling is a member of the branch of artificial intelligence known as machine learning. It is a probability-based methodology which seeks to identify those features of an object, or set of objects, which distinguish it as being in one domain or not. Features might, for example, be symptoms and signs; domains might then represent those who are suffering from a specified disease, as opposed to those who are not. The features are quantified by their ability to separate objects between those inside the domain and those outside it. Those features identified as being inside the domain have a positive value, proportional to their discriminating ability. Those which characterise regions outside the domain will have a negative value. Using many such features, and a training set of objects, a model can be built which defines, in probabilistic terms, the features which characterise the domain. Using equivalent features representing a test set of objects, each object can then be assessed against the model for the probability that it lies within the domain.

In this study, the features are the 382 lexical items described above. A model is built for each of the regions identified in the clustering exercise, with each model characterising the features of the PCAs within the region. Model building was carried out using the Pipeline Pilot package, available from Accelrys [27]. For continuous (i.e. non-binary) data, this package bins the values of each feature into a series of ranges and assesses each range for its discriminating power. Given that the descriptors represent the percentage of observations for each lexical item, this is not an appropriate option in this context. Instead, these values have been translated into binary format by the application of a threshold percentage. Those values lying above the threshold are then represented by a one, whilst those below are represented by a zero. In this study, the threshold value has been set at 30%, which equates to an average of 40.56 non-zero bits per PCA, or 1.07 items being set per concept. The models are then generated from this binary vector. This approach is also more appropriate for the testing stage described below.

4.1. Testing the models

A web page has been designed in which candidates are asked to select, for each concept, the term or terms they use most frequently from the top ten terms that are listed for that concept. They are also asked to enter any alternative terms which do not appear in the top ten, although this information is not used at this stage. In addition, biographical data is collected concerning the candidate's age and geographical history. This information produces a 382 element binary vector, equivalent to those used to build the models, which is then tested against all region models. The result is a probability, for each region, that the candidate's terminology matches the terminology of that region. These probabilities can then be displayed on a map, illustrating the candidate's geographical language.

5. Clustering results

Results for all clustering methodologies are available at http://cisrg.shef.ac.uk/gis/voices/. Here we present the results obtained using Ward's method, since this has been found to provide effective clusterings in a wide range of application areas.

5.1. Clustering based on lexical items

The Ward's method produced clusters which appear to be more consistent with our understanding of linguistic variation in the UK. Once identified, the clusters were mapped appropriately using ArcGIS software, available from ESRI [28]. Figure 1 illustrates the clusters identified at the 20 cluster level using Ward's hierarchical clustering. There is a clear relationship between lexical terminology and geography. Most of the clusters are contiguous regions, with the exception of the main Scottish region, which is split into a central and a south-western section, and the Salisbury Plain and Home Counties sections. Appropriate names for these regions have been selected and these, together with their constituent PCAs, are shown in Table 4.

Due to the fact that the CLUTO package uses UPGMA for agglomerative clustering, the consensus clustering study produced results very similar to the initial UPGMA clustering. These consensus clustering results were similar to those of the Ward's method at the 20 cluster level, with the exception that one very large cluster exists which covered the Home Counties, Southern England, East Midlands, North Wales and much of the Avon area of Figure 1. Whether this is nearer to our understanding of the linguistic subdivisions of the UK is unclear. As the number of clusters is reduced, as we move up the clustering hierarchy, this large region extends as far as Yorkshire, accounting for most of the UK. Since the Ward's method, which uses an optimisation routine to ensure minimum variance in the descriptors of each cluster, appears to give more spatially localised clusters, the machine learning exercise was carried out using the results of the Ward's cluster analysis, as shown in Table 4 and Figure 1. One unusual anomaly is the Brecon region, which remains as a single region as far as the 12 cluster level with most clustering methods. Detailed examination does not seem to indicate that this is due to any errors or anomalies in the data and, based on the 38 concepts, evidence seems to suggest that this is indeed an isolated lexical region.

Table 4. Twenty lexical regions identified using Ward's clustering

Region Name	Number (Figure 1)	PCAs
Shetland Islands	1	ZE
Orkneys & Wick	2	KW
Outer Hebrides	3	HS
Scotland Main	4	IV, PH, AB, DD, KY, EH, DG
Argyll	5	PA, FK, G, ML, KA
Berwick	6	TD
Northern Ireland	7	BT
Northern England	8	NE, CA, SR, DH, TS, DL
Yorkshire	9	LA, BD, HG, YO, LS, HX, HD, WF
Humberside	10	HU
North West	11	FY, PR, BB, L, WN, BL, M, OL, SK, WA, CH
East Midlands	12	DN, S, LN, NG, LE, DE
North Wales	13	LL, SY, CW, ST, TF, WS, WV, B, CV, DY, WR
Brecon	14	LD
South Wales	15	SA, CF, NP
Avon	16	HR, GL, BS, SN, BA, TA
Southern England	17	TR, PL, EX, TQ, DT, BH, SO, PO, GU, RG, OX, NN, MK, HP, AL, SG, PE, CB, NR, IP, CO, CM, BN, RH, KT, TW, SW, W, NW, N, E, SE, TN, ME, CT
Home Counties	18	SP, SL, UB, WD, HA, EN, IG, RM, SS, DA, BR, CR, SM, LU
West London	19	WC
East London	20	EC

5.2. Clustering based on grammar and phonology

The grammar and phonology data were clustered using Ward's method with the squared Euclidean distance. Since the recordings were made at single locations, the data points are spatially represented using Thiessen (or Voronoi) polygons [29], in which polygon boundaries are constructed at an equal distance from two neighbouring locations such that all points within a single polygon are always nearest to the location it represents. Neither method produced results with contiguous, identifiable regions, with many of the clusters being disjoint in nature. Similar clustering was observed between the grammar and phonology maps at the same cluster level, but, with the exception of some areas of Scotland and Wales, the geographical relationship was much reduced.

The reason might lie in the method of data collection, which was never intended for analysis of this kind. Most of the recordings included contributions from between three and six members of the public. Although the themes generally relate to opinions about accent, dialect and language, many of the participants exhibit considerable ethnic and geographic diversity, a Jamaican family in one recording, a Polish family in another. In several cases, the participants are not native to the area with many having lived there for fewer than four years.

6. Bayesian modelling results

The machine learning method was tested using a small selection of participants who entered their lexical terms in an online web form. The Bayesian method was not considered for the grammar and phonology data due the fact that the clusters were not as well defined and that testing data would require further recordings and suitable expert analysis.

Participants selected their term or terms for each concept from 38 lists of the top ten terms. An option to enter alternative terms for each concept was included, but not used in the analysis. The data entered was used passed through all 20 Bayesian models, the result of each being a probability score for each model, or region. Figure 2 illustrates the results for an adult female who has lived in Birkenhead, Wirral all her life, indicating the North West region is the best match for this participant's data. Table 5 shows the test results, in terms of probability, for four participants, in which the region with the highest probability score is shown in bold. Negative probability values indicate a poor correlation between the terms of the test subject and those of the region. Test 1 refers to the participant of Figure 2. Test 2 is an adult male whose geographical history covers Clydebank until his mid twenties and Glasgow until very recently. Although the Argyll region is not the best match, it is a close second after the Scotland Main region. Test 3 is an adult male who has lived in northwest London, Hemel Hempstead and, since 1994, Skipton in Yorkshire. Although the Yorkshire region is best match, there seems little evidence of influence from the Home Counties or Southern England regions.

The predictive powers of the methodology are increased when there has been little geographical mobility. Increased mobility does tend to reduce these powers, as can be seen from Test 3, but more obviously from Test 4, whose history includes west London, northwest Birmingham, Leeds (4 years) and, mostly, Buckinghamshire, whose data correlates best with Yorkshire.

One useful feature of Bayesian modelling is the ability to list those features which are most effective at characterising a region, i.e., those which have the highest probability of distinguishing between objects in the region and those outside it. Table 6 lists the top five discriminating lexical items, in terms of this probability, for each region together with their concept (from Table 2) in parentheses. Not surprisingly, some well known terms appear in these lists, such as *ginnel* in the Yorkshire region, *daps* in South Wales and Avon, *scally* in the North West. Interestingly, the subject of Test 4 used the terms *keks*, *ginnel* and *grandma*, three of the top five Yorkshire terms.

7. Conclusions and discussion

The clustering method successfully demonstrates the relationship between lexical variation and geography. In particular, Ward's minimum variance method produces clusters which, with the exception of one region, are contiguous and seem appropriate given our understanding of areas of language in the UK. One region identified which may not reflect this understanding is the Brecon region, representing a single PCA. Whether this is an anomaly due to the data or its collection method, or is in fact a genuine region of lexical variation, is not clear and requires considerable expert analysis at the local level. However, the region is consistent throughout all clustering methods used. Some of the clusters identified are single PCAs, such as Brecon. Ideally, we would use more local geographies, such as postcode district, but this is not possible due to missing data at the postcode district level..

The grammar and phonology data did not produce such well defined clusters, with many be spatially disjoint. This is most likely due to the nature of the data collection method, which was never intended for treatment of this kind. The recordings were theme-based, railway workers in Swindon or skateborders in Milton Keynes, for instance, and covered a diverse ethnic and social mix. In many cases, there is evidence of migration, with many participants originating in a different area or even country. For this reason, and due to the analysis required to test the clustering, it did not seem appropriate to carry out the machine learning strategy for the data from these recordings.

This pilot study does appear to produce Bayesian models which are able to deduce the geographical region of the sample test set studied. This initial study may require refinement, such as change in threshold value for translating the top ten percentage values into binary form. Further improvement may be possible from the use of all lexical items, not just the top ten values, or the generation of models from the raw data rather than the PCA aggregation methodology used here.

Further work on the predictive powers of the technique, with regard to spatial accuracy, is expected to be carried out. The 20 regions identified during the clustering stage are, in most cases, quite large. Further work is expected to concentrate on smaller regions, maybe down to the PCAs themselves, for which similarity search might be a more appropriate methodology.

Table 5. Bayesian test results

Region Name	**Test 1**	**Test 2**	**Test 3**	**Test 4**
Shetland Islands	-20.60	-2.03	-17.21	-19.87
Orkneys & Wick	-15.15	-3.44	-18.59	-21.26
Outer Hebrides	-24.05	-9.68	-22.04	-27.46
Scotland Main	-21.57	**9.69**	-14.16	-22.07
Argyll	-20.72	7.01	-18.14	-26.96
Berwick	-6.29	-0.01	-4.22	-6.89
Northern Ireland	-4.17	-0.67	-6.31	-7.59
Northern England	-6.03	-12.56	-2.39	-9.73
Yorkshire	-4.56	-19.14	**6.37**	**4.35**
Humberside	-1.43	-6.23	-4.95	-0.74
North West	**0.48**	-21.48	-3.48	-3.14
East Midlands	-5.39	-15.35	-4.33	-2.84
North Wales	-6.72	-17.37	-8.12	-4.27
Brecon	-4.80	-4.10	-5.59	-9.62
South Wales	-4.44	-8.16	-4.94	-6.16
Avon	-6.32	-9.04	-5.89	-6.67
Southern England	-20.05	-30.40	-23.57	-22.66
Home Counties	-8.49	-16.83	-18.70	-14.71
West London	-1.47	-2.09	-6.39	-7.67
East London	-6.32	-8.32	-8.45	-2.87

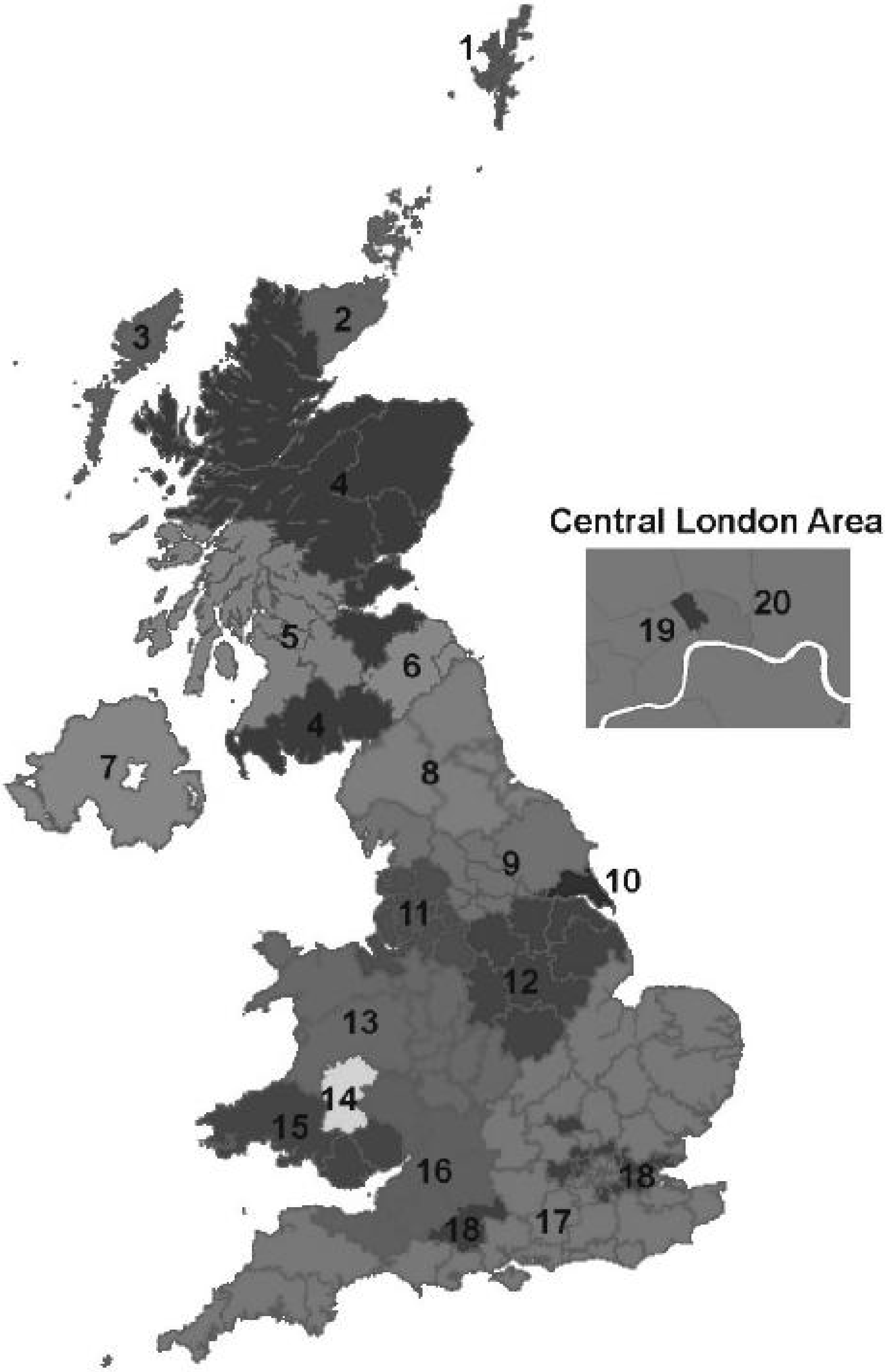

Figure 1. Twenty lexical regions identified using Ward's clustering

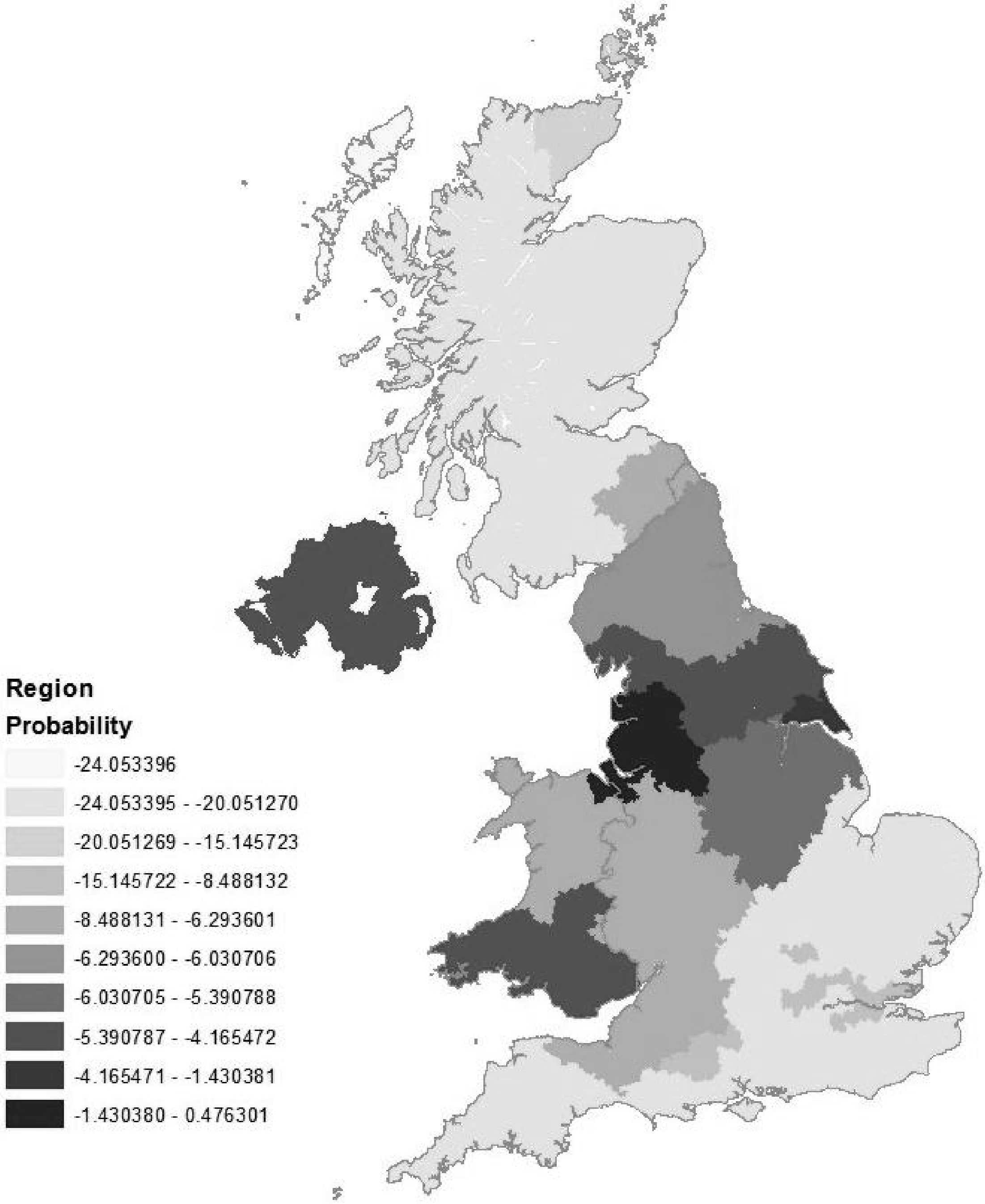

Figure 2. Test results – Adult female from Birkenhead, Wirral

Table 6. Bayesian model - discriminating lexical items

Region Name	Term 1 (Highest)	Term 2	Term 3	Term 4	Term 5
Shetland Islands	Hot (1)	Miserable (20)	Wallop (10)	Rain (36)	Poor (15)
Orkneys & Wick	Passage (35)	Grandfather (24)	Mad (6)	Crazy (17)	Mother (22)
Outer Hebrides	Plastered (12)	Beauty (18)	Angry (6)	Cross (6)	Toilet (34)
Scotland Main	Chav (28)	Roasting (1)	Pal (25)	Ned (28)	Burn (38)
Argyll	Dog (8)	Wean (21)	Bird (27)	Couch (33)	Steamed (12)
Berwick	Minted (16)	Pop (24)	Mental (17)	Wife (27)	Breeks (30)
Northern Ireland	Ma (22)	Thump (10)	Mitch (8)	Broke (15)	Gutties (31)
Northern England	Mam (22)	Lass (27)	Bairn (21)	Hoy (9)	Beck (38)
Yorkshire	Ginnel (35)	Lake (7)	Beck (38)	Grandma (23)	Keks (30)
Humberside	Golly handed (14)	Lark (7)	Townie (28)	Sandshoes (31)	Mam (22)
North West	Scally (28)	Pants (30)	Wag (8)	Ginnel (35)	Trousers (30)
East Midlands	Chuck (37)	Mardy (20)	Poorly (4)	Left (14)	Jitty (35)
North Wales	Mom (22)	Brook (38)	Cag handed (14)	Pumps (31)	Poorly (4)
Brecon	Sulk (20)	Nap (11)	Pretty (18)	Well off (16)	Shattered (3)
South Wales	Daps (31)	Mam (22)	Mitch (8)	Ugly (19)	Skive (8)
Avon	Daps (31)	Minger (19)	Plimsolls (31)	Whack (10)	Pretty (18)
Southern England	Bunk (8)	Plimsoll (31)	Minger (19)	Kip (11)	Happy (5)
Home Counties	Skive (8)	Bunk (8)	Whack (10)	Mad (17)	Stroppy (20)
West London	Sprog (21)	Sitting room (32)	Happy (5)	Brook (38)	Ugly (19)
East London	Pissed off (20)	Zeds (11)	Grandpa (24)	Chilly (2)	Townie (28)

Acknowledgements

BBC Voices lexical data have been provided by the AHRC-funded *Whose Voices?* Project, AH/E509002/1. Phonology and Grammar data have been provided by the British Library's *Voices of the UK* project funded by the Leverhulme Trust (Project F/02 017/C) Copyright © The British Library Board. This work is based on mapping data provided through EDINA UKBORDERS with the support of the ESRC and JISC and used boundary material which is copyright of the Crown and the ED-LINE Consortium.

References

[1] Chambers JK and Trudgill P, *Dialectology*. 2nd ed. Cambridge: Cambridge University Press, 1998.

[2] Kirk J, Sanderson S and Widdowson JDA, *Studies in Linguistic Geography*. London: Croom Helm, 1985.

[3] Upton C, Sanderson S and Widdowson JDA*, Word Maps: A Dialect Atlas of England.* London: Croom Helm, 1987.

[4] Upton C and Widdowson JDA, *An Atlas of English Dialects,* Oxford: Oxford University Press, 1996.

[5] Crone TM, An alternative definition of economic regions in the United States based on similarities in state business cycles, The Review of Economics and Statistics, 2005; 87(4): 617-626.

[6] Lolonis P, Armstrong M P, Location-allocation models as decision aids in delineating administrative regions, *Computers, Environment and Urban Systems,* 1993; 17: 153-174.

[7] Hagood MJ, Statistical methods for the delineation of regions applied to data on agriculture and population, *Social Forces*, 1943, 21 288-297.

[8] Ratti C, Sobolevsky S, Calabrese F, et al. 'Redrawaing the map of Great Britain from a network of human interactions', *PLoS One* 5(12), http://www.plosone.org/article/info:doi/10.1371/journal.pone.0014248 (2010, accessed April 2012).

[9] Thiemann C, Theis F, Grady D, Brune R and Brockmann D. 'The structure of borders in a small world', *PLoS One* 5(11), http://www.plosone.org/article/info:doi/10.1371/journal.pone.0015422 (2010, accessed April 2012).

[10] BBC 'Voices - The Voices Recordings', http://www.bbc.co.uk/voices/recordings/ (2007, accessed April 2012).

[11] Wells JC, *Accents of English,* Cambridge: Cambridge University Press, 1982.

[12] University of Leeds 'Whose Voices?', http://www.leeds.ac.uk/arts/info/125050/whose_voices/ (2012, accessed April 2012).

[13] Everitt, BS, Landau, S and Leese M, *Cluster analysis* 4th ed. London: Edward Arnold, 2001.

[14] Sneath, PH and Sokal, RR, *Numerical taxonomy*. San Fransisco: WH Freeman, 1973.

[15] van Rijsbergen CJ, *Information retrieval.* London: Butterworth, 1979.

[16] Spehr M, Wallraven C and Fleming RW, Image statistics for clustering paintings according to their visual appearance. In: Deussen O and Hall P (eds.) Computational Aesthetics in Graphics, Visualization and Imaging, 2009, pp57-64, https://diglib.eg.org/EG/DL/WS/COMPAESTH/COMPAESTH09/057-064.pdf (2009, accessed April 2012).

[17] Willett P, *Similarity and clustering in chemical information systems*. Letchworth: Research Studies Press, 1987.

[18] Lapointe F-J and Legendre P, A classification of pure malt Scotch whiskies, *Journal of Applied Statistics* 1994; 43(1): 237-257.

[19] Deza MM and Deza E, *Encyclopedia of distances*, Berlin: Springer-Verlag, 2009.

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

[21] IBM 'SPSS software', http://www-01.ibm.com/software/analytics/spss/ (accessed April 2012)

[22] Data Clustering Software – Karypis Lab, http://glaros.dtc.umn.edu/gkhome/views/cluto (2011, accessed April 2012)

[23] Elmes S, *Talking for Britain: A journey through the nation's dialects*, London: Penguin, 2005

[24] Milligan GW and Cooper MC, An examination of procedures for determining the number of clusters in a data set, *Psychometrika* 1985; 50(2): 159-179.

[25] Clustan Graphics, 'Clustan – A Class Act', http://www.clustan.com (1998, accessed April 2012).

[26] Mojena R, Hierarchical grouping methods and stopping rules – an evaluation, *Computer Journal*, 1977; 20: 359-363.

[27] Accelrys 'Scientific Informatics Software for Life Sciences, Materials R&D', http://www.accelrys.com (accessed April 2012).

[28] ESRI 'The GIS Software Leader – Mapping Software and Data', http://www.esri.com (accessed April 2012).

[29] Voronoi G, Nouvelles applications des paramètres continus à la théorie des formes quadratiques. *Journal für die Reine und Angewandte Mathematik* 1908, 133: 97-178.