Voices dialectometry at the University of Groningen

Martijn Wieling

In order to allow experimentation with the data of the *Voices* project, a website has been developed at the University of Groningen where dialectological analyses using the Voices data can be readily conducted. This website is a tailored version of the online dialect analysis application *Gabmap* developed by the University of Groningen, and is available at http://www.gabmap.nl/voices.

Five datasets (i.e. projects) are available on the website for consultation and download: (1) the complete dataset which includes the results of men and women and older and younger people; (2) a subset including only the responses of women; (3) a subset including only the responses of men; (4) a subset including only the responses of people younger than 30; and (5) a subset including only the responses of people older than 30. For each dataset to be of manageable size, data are grouped by postcode area and include only the ten most popular variants per variable. The relative frequency of each variant is stored for each postcode area.

Within each project there are six analyses possible. The purpose of this technical appendix is to explain how these analyses work (on the basis of the complete dataset). In addition to the information presented here, the website includes a tutorial to help you run the various analyses.

VALUE MAPS

Value maps can be used to explore the raw data as they assign a shade of grey to every postcode area on the basis of the relative frequency of occurrence of a certain variant in the postcode area. Higher relative frequencies are assigned a darker shade of grey. Table 13.1 shows the relative frequencies for a subset of ten postcode areas for the top ten variants of the variable YOUNG TRENDY PERSON. The numbers in each row sum to 1, as variants not in the top ten are excluded from the calculations. It is clear from the table that people in Aberdeen have a preference for the variant *ned* as the corresponding value is 0.624. In contrast, the other places listed (except for Bolton) are seen to prefer *chav*, as it has the highest value in each row.

TABLE 13.1 NEAR HERE

Figure 13.1 shows the value maps on the basis of *ned* and *chav*, which clearly show a separation between Scotland and the rest of the United Kingdom.

FIGURE 13.1 NEAR HERE

DIFFERENCE MAPS

It is also possible to determine how much two sites (i.e. postcode areas) differ in total. This can be done by averaging the differences between the relative frequencies for each variant. For example, the difference between Aberdeen and St. Albans for variant *chav* is 0.385 (0.633 – 0.247; see Table 13.1). After calculating the differences for all other variants (for all variables) and averaging these, the aggregate linguistic difference between the two sites is obtained. In similar fashion, the linguistic difference between every pair of sites (e.g. Aberdeen and St. Albans, Aberdeen and Birmingham, St. Albans and Birmingham) can be calculated. Table 13.2 shows the resulting aggregate linguistic differences for a sample of

four postcode areas. Note that the table on the basis of all 121 postcode areas would consist of 121 rows and 121 columns.

TABLE 13.2 NEAR HERE

A difference map visualizes these differences by drawing a (greyscale) line between pairs of sites. The darker the connecting lines, the more similar the postcode areas (i.e. the lower the aggregate linguistic difference). Figure 13.2 (left) visualizes the differences between neighbouring postcode areas only, while Figure 13.2 (right) visualizes the differences between all pairs of postcode areas. We can clearly identify the separation of the Scottish sites from the more strongly linked sites in England.

FIGURE 13.2 NEAR HERE

REFERENCE POINT MAPS

An alternative to the difference map is the reference point map, which visualizes the aggregate linguistic difference between a single reference site and all other sites. A shade of grey is assigned to each site based on its aggregate linguistic difference from the reference site. Sites which are more similar (i.e. which have a lower aggregate linguistic difference) to the reference site have a darker shade of grey than those which are less similar. This approach allows one to assess how different sites are when taking the perspective of the reference site. Figures 13.3 left and right show the reference point maps from ‘PH – Perth’ and ‘LS – Leeds’ respectively, these sites being indicated by a star. The sites linguistically most similar to Perth are located close to Perth, and are generally located in Scotland. Similarly, the sites most similar to Leeds are generally located close to Leeds.

FIGURE 13.3 NEAR HERE

MDS MAPS

While we are able to visualize the complete set of aggregate linguistic differences (see Table 13.2) using difference maps, these maps can be hard to interpret due to the overlapping of lines. A better approach to visualize the differences between all postcode areas is by applying multidimensional scaling (MDS).

Multidimensional scaling takes advantage of the fact that the aggregate differences in the original table are not independent. For example, in Table 13.2 we see that the difference between Bath and St. Albans is relatively small. Consequently, the difference between Bath and other sites will be similar to the difference between St. Albans and those sites. Indeed, we can observe that the difference between Aberdeen and both Bath and St. Albans is very similar. Based on these dependencies, multidimensional scaling converts the original difference table to a new, reduced table with the same number of rows, but only a few (e.g., three) columns or dimensions. When the values in these columns are similar, this indicates that the sites are similar. Table 13.3 shows the first three dimensions of the MDS result for the same four postcode areas as shown in Table 13.2. Clearly, Bath and St. Albans are highly similar.

TABLE 13.3 NEAR HERE

The first dimension (i.e. column) of the reduced table is the most important, as it captures most of the variation of the original difference table. Adding more dimensions increases the fit between the original difference table and the new table, but three dimensions are generally enough to provide an excellent fit (i.e. more than 80% of the variation of the original difference table is represented by the new table).

Having just three dimensions (i.e. columns) also enables easy visualization. Colour being available, when mapping the first column to red, the second column to the green, and the third column to blue, the values in the three columns can be combined to a single colour. For example, if one site has a 1 in the first and second columns and a zero in the third column, the colour of this site would be set to yellow (red mixed with green). In similar vein, every site can be assigned its own colour. The interpretation of the resulting map is then straightforward: sites which are similar in the relative frequency of their variants will have a similar colour, while the colour will be very different when sites have very different relative frequencies.

FIGURE 13.4 NEAR HERE

Lacking colour here we do not show the combined map (which is included in the online tutorial), but three separate maps in Figure 13.4, one for each dimension. While the first dimension clearly contrasts Scotland from Wales and England (Northern Ireland lies in between), the second dimension separates the central part of the United Kingdom from the rest. The third MDS dimension clearly separates ‘KW – Kirkwall’, ‘ZE – Lerwick’, ‘HS – Outer Hebrides’, and ‘LD - Llandrindod Wells’ from the remaining sites. The first dimension explains 74% of the variation of the original difference table. Adding the second dimension increases this number to 79%, while also adding the third dimension increases it to 85%. In conclusion, just three dimensions are able to capture the information present in the original difference table (of 121 columns) to a great extent, in turn enabling straightforward visualization.

CLUSTER MAPS AND DENDROGRAMS

While the MDS approach is suggestive of where dialect borders might lie, it does not distinguish clear dialect groups. Using a clustering approach yields a pre-specified number of dialect areas. There are various clustering approaches, each using a different method to determine a group of related sites. We distinguish three approaches. Each iteratively merges the two nearest sites on the basis of their aggregate linguistic difference, but they differ in how they determine the difference between a group of merged sites and the other (groups of) sites:

1. 'Complete Link' sets the difference between a group of newly merged sites and another group of sites, to the maximum of all differences between pairs of sites in one group and pairs of sites in the other group.
2. 'Group Average', also known as unweighted pair-group method using arithmetic averages (UPGMA), sets the difference between a group of newly merged sites and another group of sites to the average of all differences between pairs of sites in one group and pairs of sites in the other group.
3. 'Ward's Method' iteratively merges sites in a way that minimizes the variance in each group. This method tends to create clusters of similar size.

Figure 13.5 shows the clustering results of the three algorithms when clustering in three groups. It is clear that no algorithm yields the same result. Note that the shades of grey are arbitrary and do not convey a measure of similarity: they simply distinguish three different groups.

FIGURE 13.5 NEAR HERE

As mentioned earlier, all three clustering approaches are based on the iterative merger of sites. This iterative procedure is visualized using a dendrogram in which the farther to the right two groups are connected, the more different these are. For example, Figure 13.6 shows the dendrogram corresponding to Figure 13.5 (right).

FIGURE 13.6 NEAR HERE

It is clear that different clustering methods yield different clustering results, a drawback being that it is unclear which method is 'right'. Another disadvantage of clustering is that it is very unstable. If there are small errors in the data (and this is generally quite likely), clustering results might vary significantly when these errors are corrected. A further disadvantage of clustering is that the number of desired clusters needs to be specified in advance, and clustering will always yield that number of clusters, even though there may be fewer significant clusters in the data. For example, imagine a draughts board in the initial setting (black on one side and white on the other). It is clear that there are only two clusters of pieces (black and white), but not more. Any clustering algorithm, however, will yield more than two clusters when a higher number of required clusters is specified. Of course, determining what the real clusters are when the data are much more complex (as in our case) is not trivial. In the following section, we describe a robust approach aimed at alleviating most of the problems associated with standard clustering.

PROBABILISTIC DENDROGRAMS

'Noisy' (or 'fuzzy') clustering is a robust clustering method. Instead of using the single difference table, this generates 100 new difference tables by adding noise. That is, small random values are added to or subtracted from some of the values in the original difference

table. Consequently, each difference table differs slightly from the original difference table, and also from the other 99 tables. Each table is then used to obtain a separate clustering. Using the resulting 100 clustering results, we may count how frequently a group of sites is clustered together. A robust cluster will be characterized by sites which cluster together very frequently (e.g., in more than 90 cases), while less robust clusters will not group as frequently. To make the procedure even more robust, we can repeat the approach using different clustering approaches to obtain several hundred clustering results.

We may visualize the results of noisy clustering by using a dendrogram as before. In this case numbers are added which denote the percentage of times the sites (located to the left of the vertical bar) were grouped together. In contrast to the standard dendrogram, clusters detected in fewer than 50 percent of all cases are not shown, to allow a focus on more robust clusters. Figure 13.7 shows the probabilistic dendrogram on the basis of the complete dataset, and illustrates the locations of the dendrogram by placing them on a map.

FIGURE 13.7 NEAR HERE

The results show the lightest cluster (representing most of the Scottish sites) to be very robust, as the contained sites group in all cases (i.e. the number next to the top-most cluster equals 100). Similarly, ‘KW – Kirkwall’, ‘ZE – Lerwick’ and ‘HS – Outer Hebrides’ (the darkest cluster) also grouped together in 100 percent of cases. Finally, most of the remaining sites also group together every time. Within this large cluster several small groups of sites clustered together. Examples include ‘B – Birmingham’, ‘CV – Coventry’ and ‘ST – Stoke on Trent’ (grouped in all cases), and also ‘DH – Durham’, ‘NE – Newcastle upon Tyne’, ‘DL – Darlington’, ‘TS – Cleveland’, ‘CA – Carlisle’, and ‘HU – Hull’ (grouped in 93 percent of the cases).

CONCLUSION

Most of the methods described above can be adequately used to obtain a comprehensive view of the regional patterns in linguistic variation. Standard clustering, however, is not wholly suitable for this purpose unless different algorithms are simultaneously compared and the number of predefined clusters is not overly large (to prevent reporting clusters which might not be robust at all). Noisy clustering is a considerable improvement over standard clustering, as it succeeds well in identifying robust clusters and is able to combine the results of many clustering algorithms simultaneously to increase objectivity. The other methods are also useful. A line map yields an objective view of the differences, but can be cluttered, thus limiting interpretability. Reference point maps also offer an objective view, but only from a predetermined reference site, so not offering the complete picture.

We might regard MDS maps as most suitable for the obtaining of a comprehensive view of the regional patterns in linguistic variation. MDS maps do not presuppose a distinct number of groups, but nevertheless detect distinct groups if there are any (see for example, Figure 13.4 (left)). More importantly, however, MDS maps allow one to observe the ever-present continuum of regional variation.

Table 12.1. Sample of the dataset for the variable YOUNG TRENDY PERSON. Numbers represent the relative frequencies of each variant in the various postcode areas.

	Chav	Townie	Scally	Ned	Pikey	Tart	Kev	Slapper	Trendy	Teenagers
AB - Aberdeen	0.247	0.011	0.032	0.624	0.011	0.022	0.000	0.043	0.000	0.011
AL - St. Albans	0.633	0.122	0.041	0.020	0.041	0.020	0.020	0.020	0.000	0.082
B - Birmingham	0.502	0.107	0.077	0.015	0.021	0.009	0.223	0.009	0.024	0.015
BA - Bath	0.588	0.175	0.026	0.018	0.044	0.070	0.026	0.000	0.026	0.026
BB - Blackburn	0.476	0.342	0.134	0.012	0.000	0.012	0.000	0.024	0.000	0.000
BD - Bradford	0.614	0.168	0.119	0.010	0.030	0.030	0.000	0.010	0.000	0.020
BH - Bournemouth	0.509	0.208	0.009	0.009	0.170	0.009	0.019	0.019	0.028	0.019
BL - Bolton	0.309	0.136	0.444	0.012	0.025	0.025	0.000	0.012	0.037	0.000
BN - Brighton	0.609	0.135	0.032	0.026	0.122	0.039	0.000	0.026	0.000	0.013
BR - Bromley	0.607	0.036	0.018	0.000	0.321	0.000	0.000	0.000	0.000	0.018

Figure 12.1. Value maps of two variants of the variable YOUNG TRENDY PERSON: *ned* (left) and *chav* (right). A darker shade indicates a greater relative frequency of use.

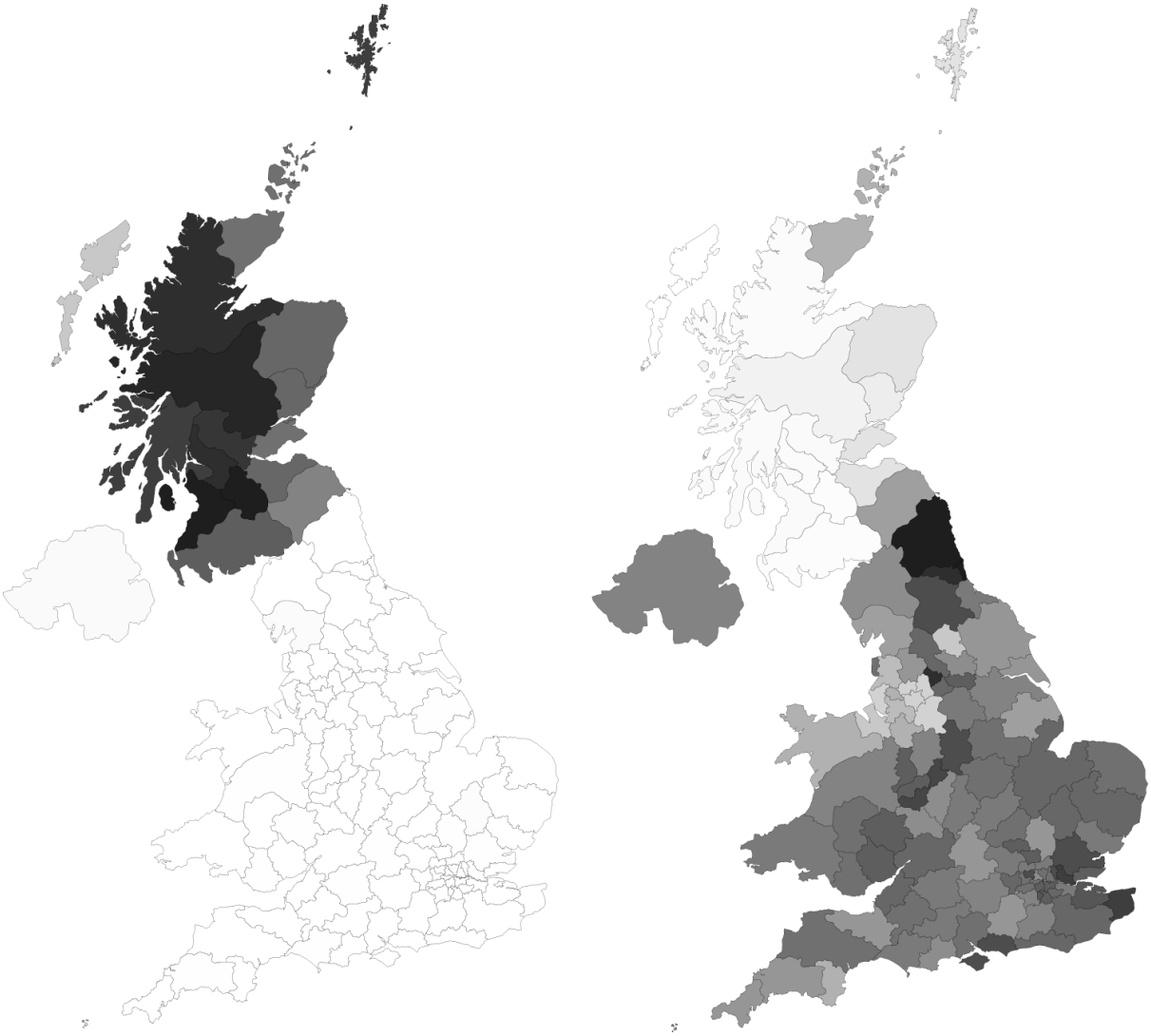

Table 12.1. Aggregate linguistic differences between a sample of four postcode areas.

	AB - Aberdeen	AL - St. Albans	B - Birmingham	BA - Bath
AB - Aberdeen	0	0.05994	0.05719	0.05856
AL - St. Albans	0.05994	0	0.03758	0.03363
B - Birmingham	0.05719	0.03758	0	0.03244
BA - Bath	0.05856	0.03363	0.03244	0

Figure 12.2. Difference maps. The left map connects neighbouring postcode areas, while the right map connects all pairs of postcode areas. Darker lines connect more similar sites.

Figure 12.3. Reference point maps. The left map shows the visualization when using ‘PH – Perth’ as the reference point, while the right map has ‘LS – Leeds’ as the reference point.

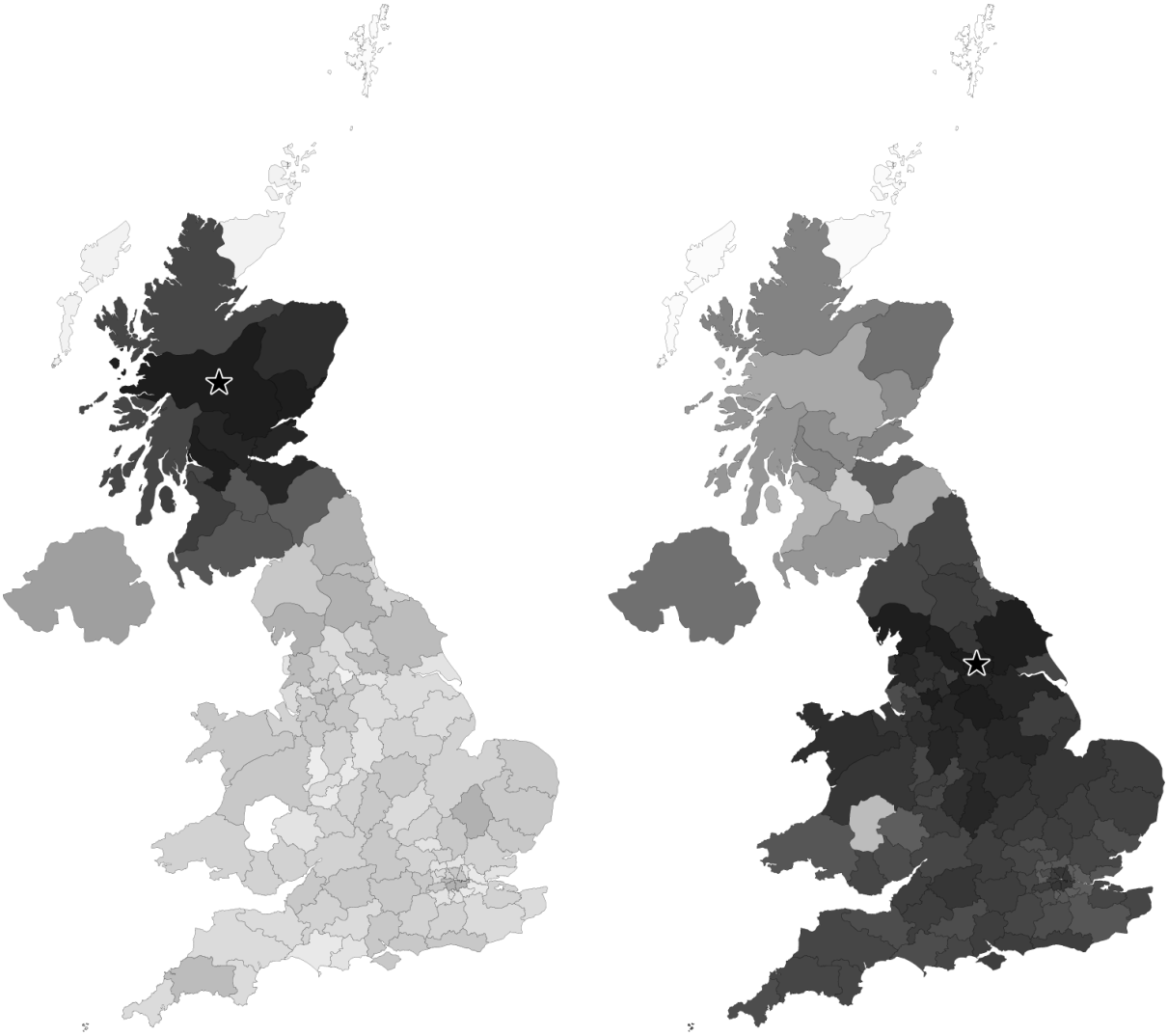

Table 12.2. Subset of MDS results on the basis of the complete aggregate linguistic difference table.

	Dimension 1	Dimension 2	Dimension 3
AB - Aberdeen	0.616	0.592	0.725
AL - St. Albans	0.161	0.780	0.721
B - Birmingham	0.110	0.545	0.690
BA - Bath	0.149	0.765	0.722

Figure 12.4. Visualization of the first (left), second (centre) and third (right) MDS dimensions from Table 12.3. Similar shades of grey indicate similar sites.

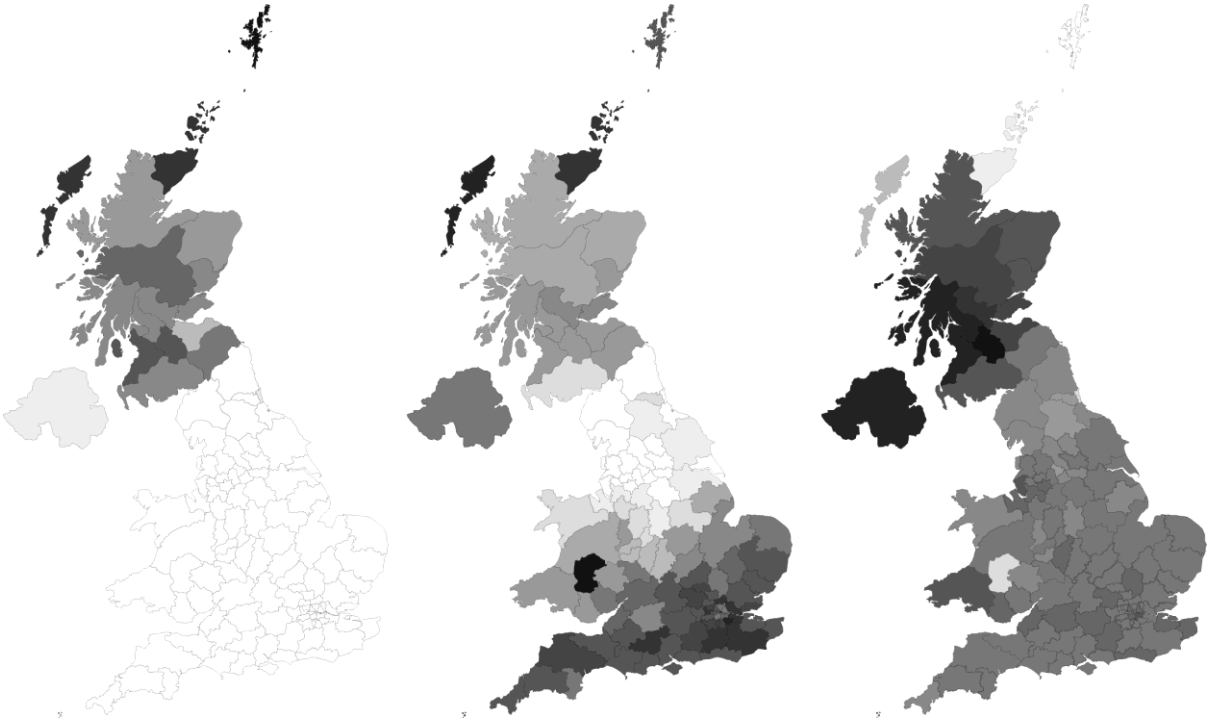

Figure 12.5. Clustering in 3 groups on the basis of ‘Complete Link’ (left), ‘Group Average’ (centre) and ‘Ward’s Method’ (right).

Figure 12.6. Dendrogram on the basis of ‘Ward’s Method’. The corresponding map distinguishing three groups is shown in Figure 12.5 (right).

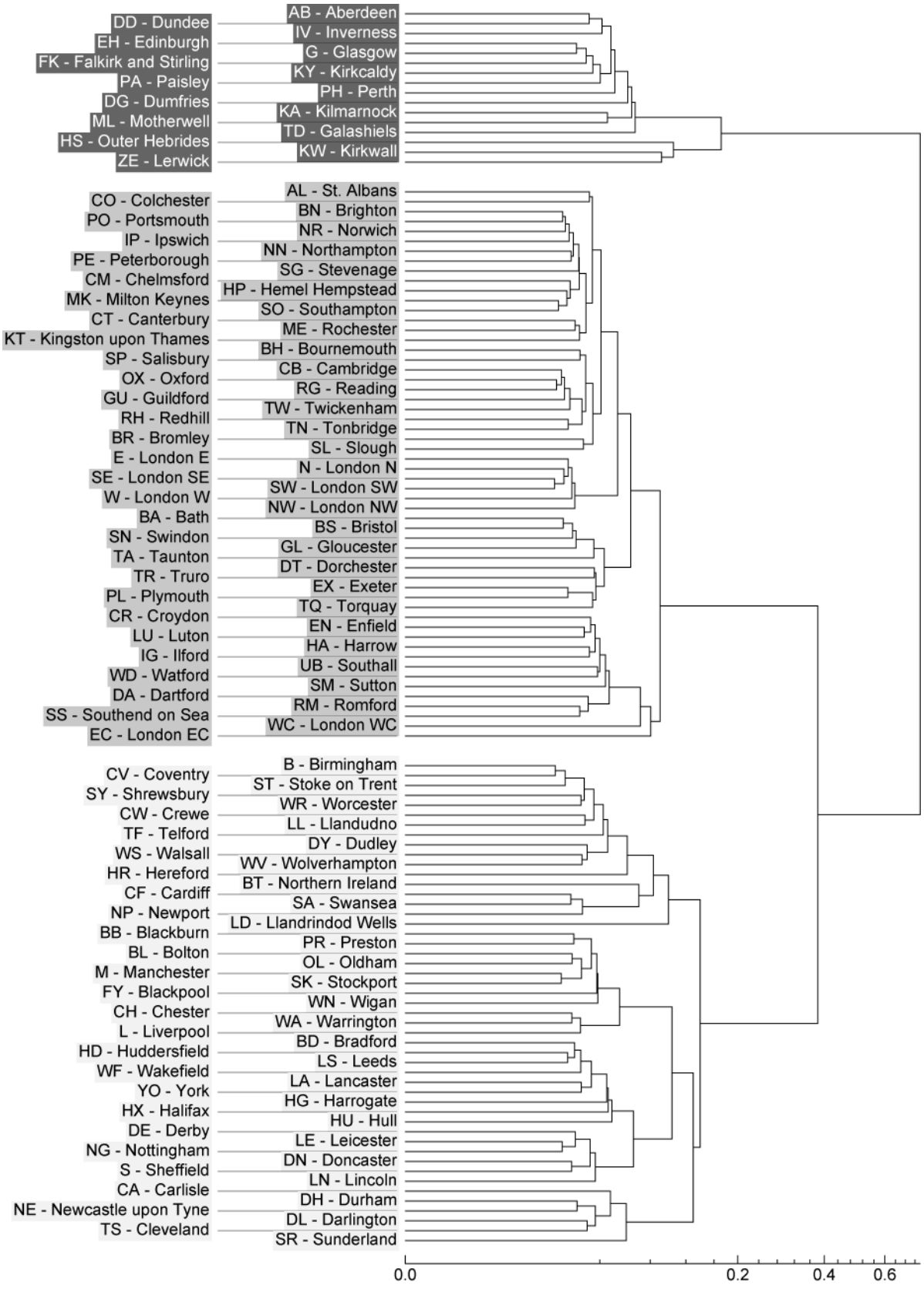

Figure 12.7. Probabilistic dendrogram (left). Higher numbers indicate more robust clusters. Clusters in less than 50% of cases are not shown. For illustration purposes, the shades of grey of the labels are mapped onto the map (right).