# Geographic Projection of Cluster Composites

Peter Kleiweg, John Nerbonne, and Leonie Bosveld

Rijksuniversiteit Groningen, 9700 AS Groningen, The Netherlands
{kleiweg,nerbonne,bosveld}@let.rug.nl
WWW home page: http://www.let.rug.nl/~kleiweg/ccmap/

**Abstract.** A composite cluster map displays a fuzzy categorisation of geographic areas. It combines information from several sources to provide a visualisation of the significance of cluster borders. The basic technique renders the chance that two neighbouring locations are members of different clusters as the darkness of the border that is drawn between those two locations. Adding noise to the clustering process is one way to obtain an estimate about how fixed a border is. We verify the reliability of our technique by comparing a composite cluster map with results obtained using multi-dimensional scaling.

## Projecting Classifications Geographically

A large variety of applications (ranging from image segmentation to data mining) have made use of clustering techniques [1]. Clusters may be visualised as an aid in identifying similar attributes, as well as to identify significant classes of individuals, the task we focus on here. Visualisation of geographic information is extensively studied by Bertin [2].

Iterative clustering produces a hierarchical categorisation that can be represented by a DENDROGRAM, i.e. a tree showing the history of the clustering process. Each time two elements are fused, a new node is introduced with branches to the fused elements. The length of a branch reflects the COPHENETIC DISTANCE, the distance between the elements when they fuse.

Cutting the dendrogram anywhere along a line perpendicular through its branches gives you a clean cluster division: each element is stored into one of several groups (see Fig. 1). To inspect for geographic influences in the data, we project this classification onto a map, making use of standard tiling techniques (see Fig. 1). This is useful, but to a limited extent, because the map shows a clear division in a number of equal groups (a rather arbitrarily chosen number), that may not reflect reality, and at best reflects a small fraction of the information in the dendrogram. Fig. 2 notes a second problem with the standard projection, namely that there is no reflection of how significant the borders are.

The COMPOSITE CLUSTER MAP is obtained by collecting chances that pairs of neighbouring elements are part of different clusters. The order in which hierarchical clustering proceeds gives one estimate: the later two elements are joined, the larger the chance they belong to different clusters. The cophenetic differences provide another estimate.

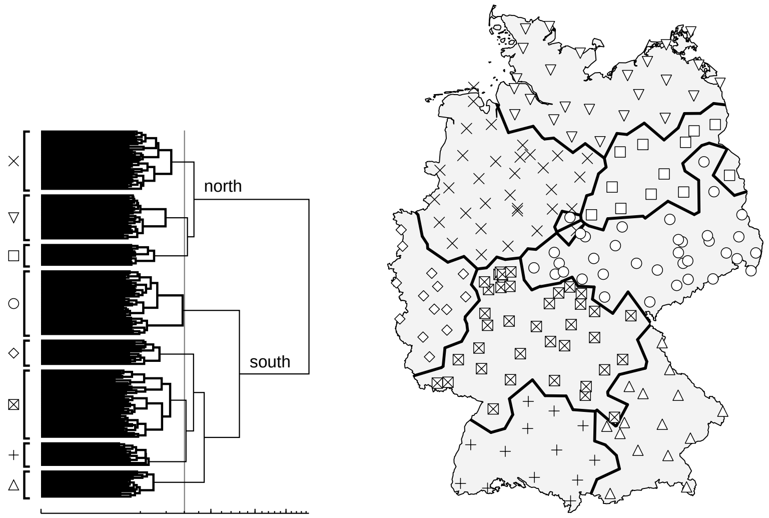

**Fig. 1.** A classification of pronunciation via edit distance [3] when subjected to clustering yields a dendrogram (left). We can project any arbitrary level of the dendrogram to create a categorial map (right), which however, loses a great deal of the information in the original classification.

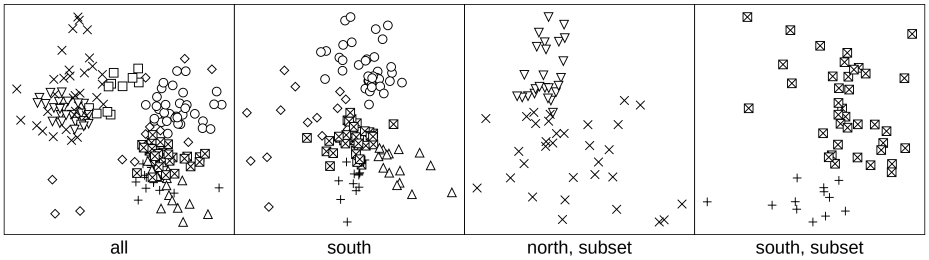

**Fig. 2.** We examine the distinctness of the groups classified in Fig. 1 by applying multidimensional scaling (MDS): items are located in the plane so that the relative distances between them correlate optimally with their mutual differences. The icons used in this display correspond to those in the dendrogram and map in Fig. 1. The dendrogram in Fig. 1 suggests that the groups ought to be distinguished well, as there is a reasonable horizontal distance between the groups and the nodes that subsume them. The MDS plot (all, left) demonstrates that the north-south break in the data is indeed robust, as are some southern distinctions (2nd) but the details (3rd and 4th plots) are less encouraging about the degree to which the data clusters naturally. We will look for this structure in the composite cluster map as well.

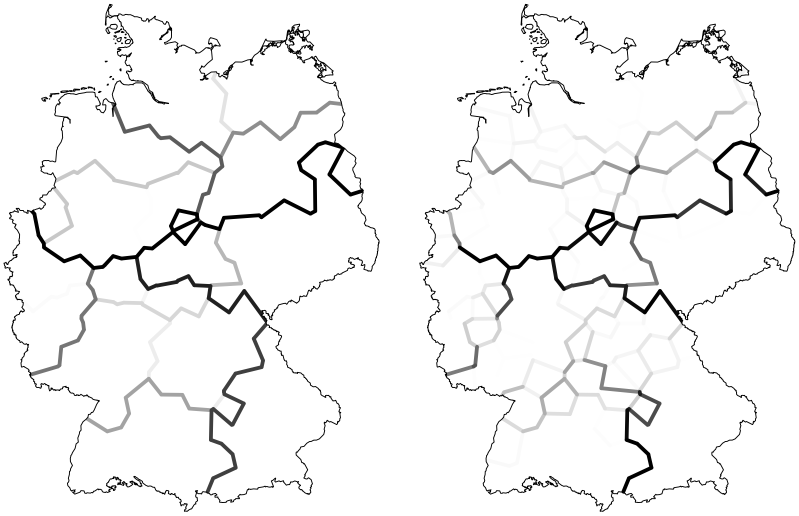

**Fig. 3.** A composite cluster map is obtained by drawing higher levels of clustering as borders of increasing darkness (left). Alternatively we repeatedly cluster the same data with variable amounts of noise (right). Note that the North-South division prominent in the MDS analysis (Fig. 2) emerges clearly.

Because hierarchical clustering is inherently unstable [1], we add noise to the clustering, and combine the results of many clustering runs. In fact, we exploit the instability, which is usually considered to be a weakness, to distinguish naturally sharp borders, shown as dark lines, from transition areas, shown as nets of light lines (Fig. 3, right).

### The next steps...

Our composite cluster maps give a more differentiated picture than a simple cluster division. The next step will be to inspect this ‘granularity’, comparing several clustering algorithms with results of other techniques, such as multidimensional scaling. A major test will be to have experts in the field of study evaluate these maps.

## References

1. Jain, A.K., Murty, M.N., Flynn, P.J.: Data clustering: A review. ACM Computing Surveys **31** (1999) 264–323
2. Bertin, J.: Semiologie graphique. Editions Gauthier-Villars, Paris (1967)
3. Nerbonne, J., Heeringa, W., Kleiweg, P.: Edit distance and dialect proximity. In Sankoff, D., Kruskal, J., eds.: Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison. 2nd edn. CSLI, Stanford, CA (1999) v–xv