Sammon's Projection for Clustering Complex Geographical Objects

G. Andrienko, N. Andrienko

Fraunhofer Institute IAIS – Intelligent Analysis and Information Systems
53754, Schloss Birlinghoven, Sankt Augustin, Germany
Email: gennady.andrienko@iais.fraunhofer.de

1. Introduction

Projection methods are often used in information visualization for dealing with multivariate data, which are difficult to visualize. Projection methods place data items in an abstract space with a chosen number of dimensions in such a way that the distances between the positions reflect the differences (dissimilarities) between the data items. For visualization purposes, projections onto one-, two-, or three-dimensional spaces are typically used. Such projections are displayed on the computer screen to enable the exploration of similarities and dissimilarities among data items. When the dataset contains groups of similar data items, they will appear in the projection as clusters of close points. Hence, projection helps the user to discover clusters and to find outliers, i.e. data items that are very distinct from the others.

We have implemented an interactive visual tool that applies projection to different types of geographical objects and thereby enables exploring their similarities and dissimilarities and grouping them into clusters according to relevant properties. There are many projection methods and families of methods, for example, multidimensional scaling (MDS) and principal component analysis (PCA). Self-organizing map (SOM), which has become popular in geographical information science (Agarwal and Skupin 2008), is also a projection method that puts data items in a discrete space (regular grid). In our tool, we use the Sammon's projection algorithm (Sammon 1969). However, it would be possible to use other projection algorithms applicable to a pre-defined *distance matrix*, which consists of numbers expressing item-item dissimilarities; these numbers are called *distances*. The PCA and SOM methods do not satisfy this requirement. The input data for these methods are combinations of values of multiple variables; these combinations are called *feature vectors*.

We use a projection method requiring the input in the form of distance matrix because we want to apply it to complex spatial and spatio-temporal objects whose properties cannot be adequately represented by feature vectors. Distance matrix-oriented projection methods give us the possibility to apply various ad-hoc measures of object dissimilarity. For a given type of objects, a method for assessing the dissimilarity, called *distance function*, is devised taking into account the specifics of the objects. The distance function is used to compute the distance matrix, which is used as the input for the projection algorithm. A distance matrix-oriented projection method can be applied to feature vectors as well. For this purpose, a matrix of e.g. Euclidean distances (or, more generally, Minkowski distances) between the feature vectors is computed and provided as the input to the method.

In our previous research we devised several specific distance functions for two types of spatio-temporal objects: trajectories of moving entities (Rinzivillo et al. 2008) and point events (Andrienko and Andrienko 2009). These functions have been used for clustering the objects by means of the density-based clustering algorithm OPTICS

(Ankerst et al. 1999). A major difficulty that we encountered, especially with trajectories, is visualizing clustering results in a comprehensible way. Since clusters of trajectories are not disjoint in the geographical space, we create a display with multiple small maps each representing a cluster (Andrienko et al. 2009). This display, however, does not convey sufficient information for interpreting clustering results. Regarding a single cluster, it is hard to assess its density, compactness, and internal variation. Regarding the whole set of clusters, it is hard to assess the inter-cluster similarities and differences. Besides, a large part of the input objects usually goes to “noise”, i.e. remains beyond any cluster. It is very hard to estimate how distinct these objects are from the others.

Representing complex objects, such as trajectories, by points in two-dimensional abstract space by means of projection allows the analyst to investigate inter-object and inter-cluster similarities and dissimilarities. For the investigation of clustering results, the points in the projection space are coloured according to the clusters the respective objects belong to. Furthermore, the projection itself can be used for building clusters according to diverse principles. In this paper, we demonstrate our approach by example of exploring a set of 336 car trajectories of one person made during a time period of 316 days.

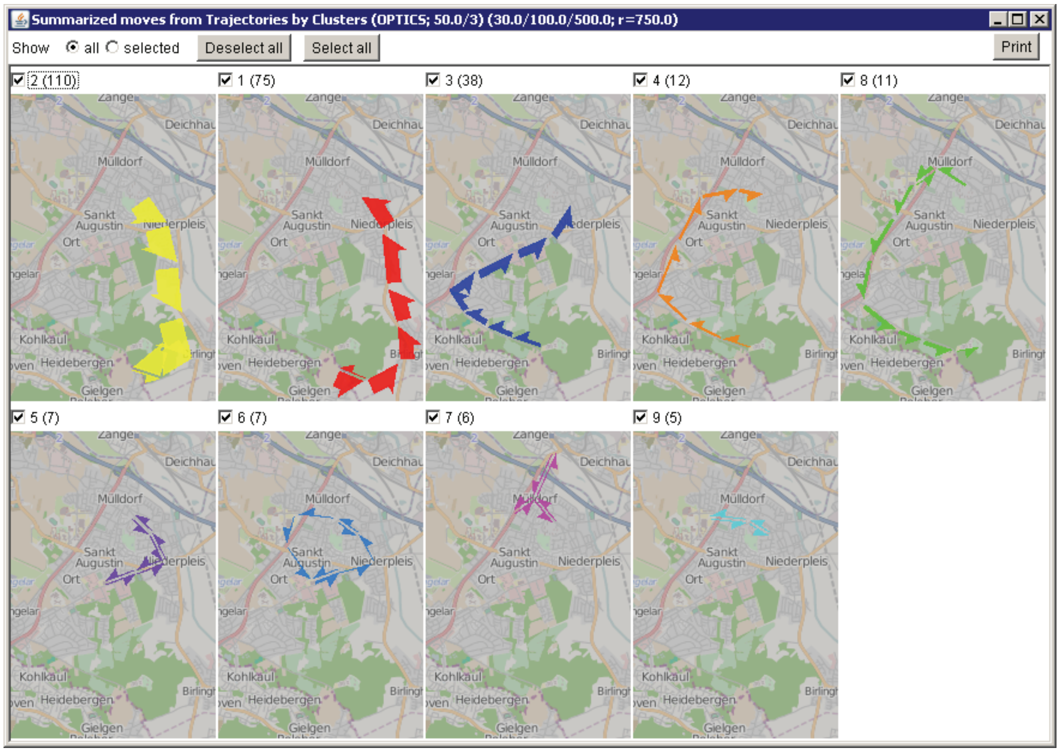
2. Exploration of a Set of Trajectories

2.1 Use of Projection for Exploring Clustering Results

In our first experiment, we use the Sammon’s projection method to explore the outcomes of density-based clustering of the trajectories. We apply the clustering algorithm OPTICS and the distance function ‘route similarity’ (Andrienko et al. 2007), which compares the geometric shapes and spatial positions of trajectories. Figure 1 demonstrates the clusters of trajectories obtained by means of OPTICS. The clusters are presented in a summarized form as flow maps, so that the routes can be easily seen. Not all 336 trajectories have been grouped into clusters by OPTICS; 65 of them have been treated as “noise”. The colours are arbitrarily assigned to the clusters and do not bear any semantics.

To obtain additional information about the clustering results and the relationships between the clustered trajectories and the “noise”, we build a Sammon’s projection of the set of trajectories using the distance matrix computed by means of the same distance function ‘route similarity’ (Figure 2). The trajectories are represented in the projection space by points (small circles), which are coloured according to the OPTICS clusters the respective trajectories belong to. The grey-coloured dots represent the “noise” trajectories. From the relative positions of the dots, we can see that some of the “noise” trajectories are quite unique (their dots are positioned far from the bulk of the dots) while others are close to the clusters. Hence, increasing the distance threshold for defining clusters in OPTICS would, probably, result in grouping the latter trajectories with others into clusters. We can also see that some of the OPTICS clusters, such as cluster 2 (yellow), are very compact, i.e. have low internal variation, while some others, such as cluster 6 (light blue), consist of more diverse objects. The relative distances among the clusters in the projection display show us how distinct the clusters are. Thus, we can see that the trajectories in clusters 6, 8, and 9 (light blue, green and cyan, respectively) do not differ very much in terms of the ‘route similarity’ function.

Hence, the use of projection allowed us to examine more comprehensively the results of OPTICS clustering. The same approach can be applied to clusters obtained by any other clustering algorithm and/or any other distance function.



Summarized moves from Trajectories by Clusters (OPTICS; 50.0/3) (30.0/100.0/500.0; r=750.0)

Show ● all ○ selected

Figure 1. Clusters of trajectories obtained by means of the OPTICS algorithm with the distance function ‘route similarity’.

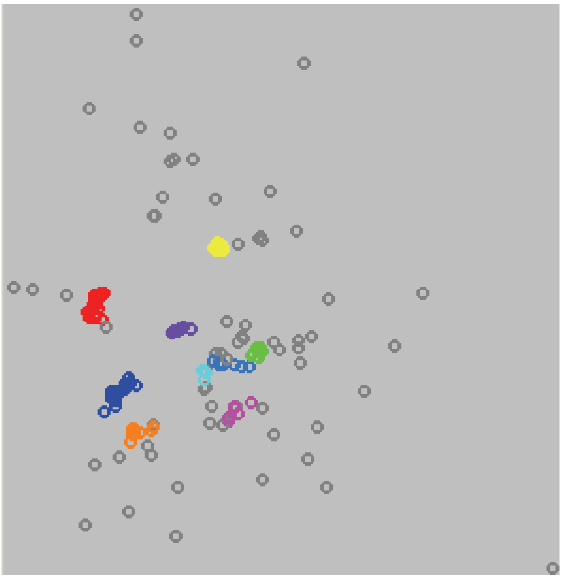


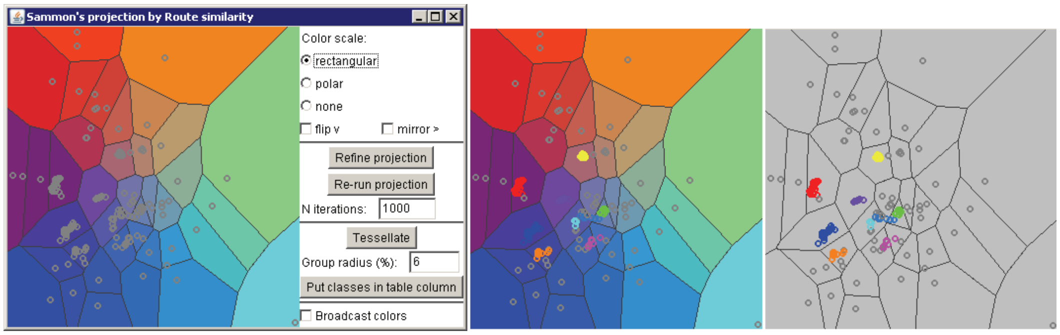
Figure 2. The trajectories are represented by dots in the Sammon’s projection space and coloured according to the OPTICS clusters they belong to.

2.2 Use of Projection for Defining Clusters

There are different approaches to building clusters (Han and Kamber 2006). Partition-based methods such as k-means divide the set of objects into a user-specified number of subsets such that the dissimilarities within the subsets are minimized and the dissimilarities between the subsets are maximized. Hierarchical approaches work through an iterative hierarchical decomposition of the dataset. Density-based clustering methods (Ester et al 1996) rely on the concept of density: for each object inside a cluster, the neighborhood of a given radius has to contain at least a given number of objects, i.e. the density of the cluster has to be not less than the density threshold.

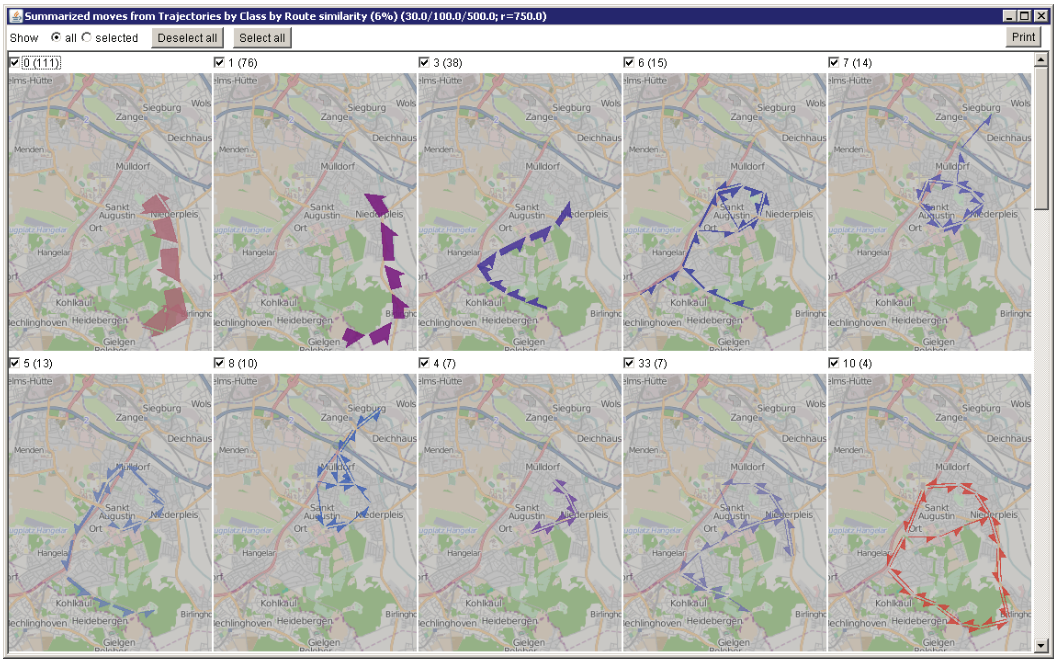
In (Andrienko and Andrienko 2010) we suggested another possible principle of defining clusters, which was applied to points in geographical space: points are organized in groups such that the radii of the groups do not exceed the specified parameter R. The parameter regulates the accepted degree of internal variation within a cluster: the distance between any two points in a cluster must not exceed 2*R. We have devised a clustering algorithm that tends to preserve existing concentrations of close points: when the radius of a concentration of points is below the threshold, the points are included in the same cluster. The same clustering principle can be applied to points in a projection space. In this way, our radius-based clustering algorithm can be applied to arbitrary objects and arbitrary measures of object dissimilarity.

In our second experiment, we apply the radius-based clustering algorithm to the points representing the trajectories in the Sammon's projection space (Figure 2). In the case of geographical clustering, the desired radius is specified in metres. In the case of an abstract projection space, there are no meaningful units for the distances. Therefore, the radius is specified in percents to the width of the projection space. After the points are organized in groups, we use the medoids of the groups (i.e. the points with the minimal sum of distances to all other group members) as the generating points for the Voronoi tessellation of the projection space. Each Voronoi cell defines a cluster. Colours are assigned to the clusters according to the positions of their generating points in the projection space. For this purpose, we use two variants of two-dimensional colour scales, rectangular and polar. In the rectangular scheme, four distinct primary colours are put in the corners and the colours for all other positions are obtained by mixing the primary colours proportionally to the distances from the corners. In the polar scheme, the positions are expressed in polar coordinates with respect to the centre. The angle defines the colour hue and the distance to the centre defines the saturation and brightness. With this approach to assigning colours to clusters, similarity of colours means similarity of the clusters.


Sammon's projection by Route similarity
Refine projection
Re-run projection
Tessellate
Put classes in table column

Color scale:
● rectangular
○ polar
○ none
□ flip v □ mirror >
N iterations: 1000
Group radius (%): 6
□ Broadcast colors

Figure 3. Building clusters on the basis of Sammon's projection (left) and comparison with the OPTICS clusters (centre and right).



Summarized moves from Trajectories by Class by Route similarity (6%) (30.0/100.0/500.0; r=750.0)

Show ● all ○ selected Deselect all Select all Print

☑ 0 (111) ☑ 1 (76) ☑ 3 (38) ☑ 6 (15) ☑ 7 (14)

☑ 5 (13) ☑ 8 (10) ☑ 4 (7) ☑ 33 (7) ☑ 10 (4)

Figure 4. 10 biggest clusters of trajectories (out of 38) obtained on the basis of the Sammon's projection.

Figure 3 demonstrates the division of the projection space into Voronoi polygons on the basis of point clusters with the radius 6% of the space width. In this way, we have obtained 38 clusters; 10 biggest clusters are shown in a summarized form in Figure 4. 19 clusters consist of singular objects, 6 clusters are made of pairs of objects, and there are two clusters with three objects and two clusters with four objects. In terms of a density-based algorithm, these would be not clusters but "noise". The screenshots in the centre and on the right of Figure 3 allow the reader to compare the radius-based clusters with the OPTICS clusters, which are represented by the colours of the points as in Figure 2.

Rinzivillo et al. (2008) describe the idea of progressive clustering, in which clustering is applied to results of previous clustering. Thus, the user may select one or more of previously obtained clusters, or "noise" in case of density-based clustering, and apply clustering with another distance function and/or with other parameters to the selected subset. In this way, clusters can be gradually refined. This idea is applicable in an obvious way also to projection-based clustering. Progressive clustering may be particularly useful when the dataset contains a few outliers that are very distinct from the other objects. In this case, the bulk of the objects occupies a small area within the projection space and is hard to divide into clusters. Hence, the first step of clustering should separate the outliers from the other objects, and the second step should be applied to the bulk of the objects.

3. Conclusion

Building a projection of complex geographical objects according to their pair-wise dissimilarities is useful for several purposes. First, the user can explore the distribution of the objects, discover groupings of similar objects and detect outliers. Second, the user can examine and interpret clusters produced by any clustering method (we demonstrated this by example of OPTICS) and investigate the sensitivity of the clustering results to the parameters. Third, the projection itself can be used for building clusters since any generic clustering algorithm such as OPTICS or k-means can be applied to the points representing objects in the projection space. In this way, generic algorithms become applicable to complex objects, such as trajectories, requiring specialized distance functions. Furthermore, clusters so obtained can be represented by colours that reflect their similarities and differences.

We have suggested a new approach to clustering, in which clusters with the specified maximum radius are built. The method was initially designed for points in geographical space. The use of projection allows us to apply this method to any objects for which an appropriate distance function exists. The results of this way of clustering are comparable to results obtained from the SOM method. In both cases, there is a grid with cells including similar objects. In case of SOM, the grid is predefined; in our method, the grid is built after positioning the objects in the projection space and can be dynamically refined or coarsened; hence, the number of clusters and the degree of variation within the clusters can be directly controlled by the user. Besides, the SOM is applicable only to feature vectors while our approach can be used with any distance function.

Acknowledgements

This work is supported by EU in project ESS and by DFG in project ViAMoD.

References

Agarwal P and Skupin A (eds), 2008, *Self-Organising Maps: Applications in Geographic Information Science*. Wiley.

Andrienko G, Andrienko N and Wrobel S, 2007, Visual Analytics Tools for Analysis of Movement Data. *ACM SIGKDD Explorations*, 9(2): 38-46.

Andrienko G and Andrienko N, 2009, Interactive Cluster Analysis of Diverse Types of Spatiotemporal Data, *ACM SIGKDD Explorations* (in press)

Andrienko G, Andrienko G, Rinzivillo S, Nanni M, Pedreschi D and Giannotti F, 2009, Interactive Visual Clustering of Large Collections of Trajectories, In: *Proceedings of IEEE Visual Analytics Science and Technology (VAST 2009)*, 3-10

Andrienko N and Andrienko G, 2010, Spatial Generalization and Aggregation of Massive Movement Data. *IEEE Transactions on Visualization and Computer Graphics*, IEEE computer Society Digital Library. IEEE Computer Society, http://doi.ieeecomputersociety.org/10.1109/TVCG.2010.44

Ankerst M, Breunig M, Kriegel H-P and Sander J, 1999, OPTICS: Ordering points to identify the clustering structure. In: *Proceedings of ACM SIGMOD 1999*, 49–60.

Ester M, Kriegel H-P, Sander J and Xu X, 1996, A density-based algorithm for discovering clusters in large spatial databases with noise. In: *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*, Portland, Oregon, USA, 226-231.

Han J and Kamber M, 2006, *Data Mining. Concepts and Techniques*. Morgan Kaufmann, San Francisco, CA

Rinzivillo S, Pedreschi D, Nanni M, Giannotti F, Andrienko N and Andrienko G, 2008, Visually–driven analysis of movement data by progressive clustering, *Information Visualization*, 7(3/4): 225-239.

Sammon JW, 1969, A nonlinear mapping for data structure analysis. *IEEE Transactions on Computers*, 18: 401–409.