A Hierarchical Graph-Clustering Approach to find Groups of Objects

Karl-Heinrich Anders
Institute of Cartography and Geoinformatics
University of Hannover
Appelstraße 9a, 30167 Hannover, Germany
karl-heinrich.anders@ikg.uni-hannover.de

KEY WORDS: Map Generalization, Unsupervised Clustering, Neighbourhood Graphs

1 Introduction

Nowadays, the necessity of automatic interpretation and analysis of spatial data is getting more and more important, because the amount of digital spatial data continuously increases. On the one hand, there are raster data sets, on the other hand vector data that are predominantly based on different landscape models. Differences between these landscape models are, e.g., the object type, the degree of generalization, or the geometric accuracy of the captured landscape objects. The pure interactive processing and analysis of large spatial databases is very time-consuming and expensive. Especially the manual analysis of spatial data for the purpose of data revision will reach the limit of technical feasibility in the near future, because modern requirements on the up-to-dateness of data lead to ever shorter update cycles.

The automatic interpretation of digital landscape models needs the integration of methods of the field of spatial data mining or knowledge discovery in spatial databases into geographical information systems (GIS). In general, the automatic interpretation of a digital landscape model (DLM) can be divided into the interpretation based on a specific model of the DLM, the interpretation based on a generic model of the basic elements of the DLM, and the unsupervised interpretation of the DLM. Clustering methods can be divided into supervised and unsupervised methods. Unsupervised clustering or learning methods can be used for the third case of DLM interpretation. Especially unsupervised clustering methods are well suited for the model generalization and the cartographic generalization of DLM data if these methods can recognize clusters of arbitrary shape. There are a lot of different clustering approaches, but most of them need certain prerequisites, like the distribution function of the data, or thresholds for similarity tests and terminating conditions. In many cases, clustering methods can only find clusters with a convex shape and without holes (e.g., maximum-likelihood).

In this paper a new unsupervised clustering approach called *Hierarchical Parameter-free Graph CLustering* (HPGCL) for the automatic interpretation of spatial data is described (Anders 2003). The HPGCL algorithm can find clusters of arbitrary shape and needs neither parameters like thresholds nor an assumption about the distribution of the data or number of clusters. The novelty of the HPGCL algorithm lies on the one hand in the application of the hierarchy of neighbourhood graphs (also called proximity graphs) to define the neighbourhood of a single object and object clusters in a natural and common way and on the other hand in the definition of a median based, threshold free decision criteria for the similarity of clusters. In the HPGCL algorithm the Nearest-Neighbour-Graph, the Minimum-Spanning-Tree, the Relative-Neighbourhood-Graph, the Gabriel-Graph, and the Delaunay-Triangulation are used. It will be shown that the hierarchical relationship of these proximity graphs can be used for a natural generalization process in the sense of a coarse-to-fine segmentation of a data set. One additional feature of the HPGCL algorithm is that in general a limiting number of clusters greater than one will be found. In contrast, general hierarchical cluster algorithms require the minimal number of clusters as a parameter, otherwise they will always group all objects of a data set in one big cluster.

2 Related Work

In the context of data aggregation, there are many approaches in GIS and in digital cartography, namely in model or database generalization. (Richardson 1996) and (van Smaalen 1996) present approaches to come from one detailed scale to the next based on a set of rules. If such rules are known or models of the situation are available, good results can be achieved (cf. (Sester, Anders & Walter 1998)). However, the main problem being the definition of the rules and the control strategy to infer new data from it (Ruas & Lagrange 1995). Current concepts try to integrate learning techniques for the derivation of the necessary knowledge (Plazanet, Bigolin & Ruas 1998), (Sester 1999).

Clustering is a well established technique for data interpretation. It usually requires prior information, e.g. about the statistical distribution of the data or the number of clusters to detect. Existing clustering algorithms, such as k-means (Jain & Dubes 1988), PAM (Kaufman & Rousseeuw 1990), CLARANS (Ng & Han 1994), DBSCAN (Ester, Kriegel, Sander & Xu 1996), CURE (Guha, Rastogi & Shim 1998), and ROCK (Guha, Rastogi & Shim 1999) are designed to find clusters that fit some static models. For example, k-means, PAM, and CLARANS assume that clusters are hyper-ellipsoidal or hyper-spherical and are of similar sizes. The DBSCAN algorithm assumes that all points of a cluster are *density reachable* (Ester et al. 1996) and points belonging to different clusters are not. All these algorithms can breakdown if the choice of parameters in the static model is incorrect with regarding to the data set being clustered, or the model did not capture the characteristics of the clusters (e.g. shapes, sizes, densities). In the following, we give a brief overview of existing clustering algorithms.

2.1 Non-hierarchical Schemes

Non-hierarchical clustering techniques are also called partitioning clustering techniques. These approaches attempt to construct a simple partitioning of a data set into a set of k non-overlapping clusters such that the partitions optimize a given criterion. Each cluster must contain at least one data element, and each data element must belong to exactly one group. In most of the partitioning methods an initial partitioning is chosen and then the cluster membership is changed in order to obtain a better partitioning. *Centroid based* methods like the k-means method (MacQueen 1967), (Jain & Dubes 1988) and the ISODATA (Ball & Hall 1965) method try to assign data elements to clusters such that the mean square distance of data elements to the centroid of the assigned cluster is minimized. These techniques are suitable only for data in metric spaces, because they have to compute a centroid of a given set of data elements. *Medoid based* approaches as CLARANS (Ng & Han 1994) and PAM (Kaufman & Rousseeuw 1990) try to find a so called medoid which is a representative data element that minimize the sum of the distances between the medoid and the data elements assigned to this medoid. One disadvantage of centroid and medoid based methods is that not all values of k lead to natural cluster so it is useful to run the algorithm several times with different values for k to select the best partition. With a given optimization criterion this decision can be automated. The main drawback of both methods is that they will fail for data sets in which data elements belonging to a cluster are closer to the representative of another cluster than to the representative of their own cluster. This case is typical for many natural clusters if the cluster shapes are concave or their sizes vary largely.

2.2 Hierarchical Schemes

Hierarchical cluster schemes constructs a dendrogram is a tree structure which represents a sequence of nested clusters. This sequence represents multiple levels of partitioning. On the top is a single cluster which includes all other clusters. At the bottom are the data elements representing single element clusters. Dendrograms can be constructed top-down or bottom-up. The bottom-up method is known as the agglomerative approach, where each data element starts out as a separate cluster. In each step of an agglomerative algorithm the two most similar clusters are grouped together based on similarity measures in subsequent steps and the total number of clusters is decreased by one. These steps can be repeated until one large cluster remain or a given number of clusters is obtained or the distance between two closest clusters is above a certain threshold. The top-down method known as the divisive approach works in the reverse direction. Agglomerative methods seems to be the most popular in the literature. In the literature one can find many different variations of hierarchical algorithms. Basically, these algorithms can be distinguished by their definition of similarity and how they update the similarity between existing clusters and the merged clusters. In general, the approaches described are alternative formulations or minor variations of the following three concepts: *centroid or medoid based methods, linkage based methods, variance or error sum of squares error*. The centroid or medoid based approaches also fail on clusters of arbitrary shapes and different sizes like non-hierarchical methods, such as k-means and k-medoid. The oldest linkage based method is the *single linkage* algorithm, sometimes referred to as the nearest neighbor approach. In the single linkage method, no representative exists. The cluster is represented by all data elements in the cluster and the similarity between two clusters is the distance between the closest pair of data elements belonging to different clusters. The single linkage method is able to find clusters of arbitrary shape and different sizes, but it will fail at poorly separated clusters and is susceptible to noise and outliers. In order to avoid these drawbacks algorithms like the *shared near neighbors method* (Jarvis & Patrick 1973), *CURE* (Guha et al. 1998) or *ROCK* (Guha et al. 1999) were proposed. Instead of using a single centroid to represent a cluster, CURE choose a constant number of representative points to describe a cluster. The ROCK algorithm operates on a derived similarity graph and scales the aggregate *inter-connectivity* with respect to a predefined inter-connectivity model. The shared near neighbors method use a k-nearest-neighbour graph to determine the similarity between two clusters. The advantage of this clustering method over most other alternatives is that it is independent of absolute scale. A major limitation of existing agglomerative hierarchical schemes such as the *Group Averaging Method* (Jain & Dubes 1988), CURE, and ROCK is that the merging decisions are based on static modeling of the clusters to be merged. More information about the limitations of existing hierarchical methods can be found in (Karypis, Han & Kumar 1999).

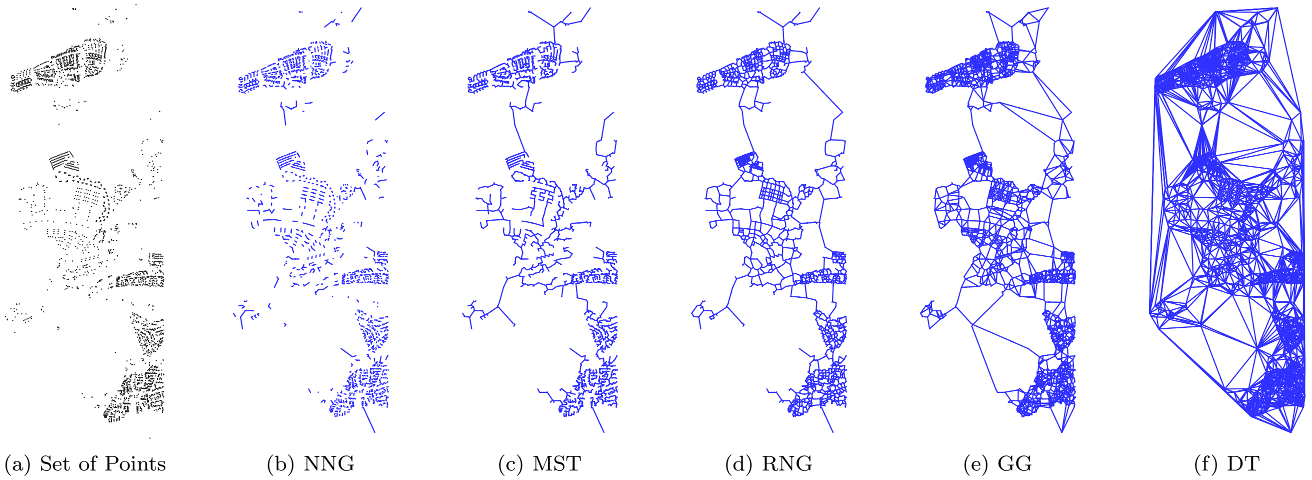
3 Graph-based Clustering

The most powerful methods of clustering in difficult problems, which give results having the best agreement with human performance, are the graph-based methods (Jaromczyk & Toussaint 1992). The idea is extremely simple: Compute a neighborhood graph (such as the minimal spanning tree) of the original points, then delete any edge in the graph that is much longer (according to some criterion) than its neighbors. The result is a forest and each tree in the forest represents a cluster. In general, hierarchical cluster algorithms work implicitly or explicitly on a similarity matrix such that every element of the matrix represents the similarity between two elements. In each step of the algorithm the similarity matrix is updated to reflect the revised similarities. Basically, all these algorithms can be distinguished based on their definition of similarity and how they update the similarity matrix. In spatial clustering algorithms one can discriminate between *spatial similarity* and *semantic similarity* which means the similarity of non-spatial attributes. Spatial Similarity implies the definition of a neighborhood concept which can be defined on geometric attributes, such as coordinate, distance, density, and shape. The computation of a spatial similarity matrix can be seen as the construction of a weighted graph, so called *neighborhood graph*, where each element is represented by a node and each neighborhood relationship (similarity) is an edge.

3.1 Neighbourhood Graphs

A general introduction to the subject of Neighbourhood Graphs is given in (Jaromczyk & Toussaint 1992). Neighbourhood graphs capture proximity between points by connecting nearby points with a graph edge. The many possible notions of *nearby* lead to a variety of related graphs. It is easiest to view the graphs as connecting points only when a certain region of space is empty. In our approach we use the following neighbourhood graphs: The Nearest Neighbour Graph (NNG) (Eppstein, Paterson & Yao 1997, Nakano & Olariu 1997, Jarvis & Patrick 1973), the Minimum Spanning Tree (MST) (Yao 1982, Supowit 1983, King 1995), the Relative Neighbourhood Graph (RNG) (Toussaint 1980), the Gabriel Graph (GG) (Gabriel & Sokal 1969), and the Delaunay Triangulation (DT) (Lee 1980, O'Rourke 1982, Preparata & Shamos 1988). Figure 1 shows all these graphs for an example point set. The important relationship between these neighbourhood graphs is that they build a hierarchy:

$$\mathsf{NNG} \subseteq \mathsf{MST} \subseteq \mathsf{RNG} \subseteq \mathsf{GG} \subseteq \mathsf{DT}.$$



(a) Set of Points (b) NNG (c) MST (d) RNG (e) GG (f) DT

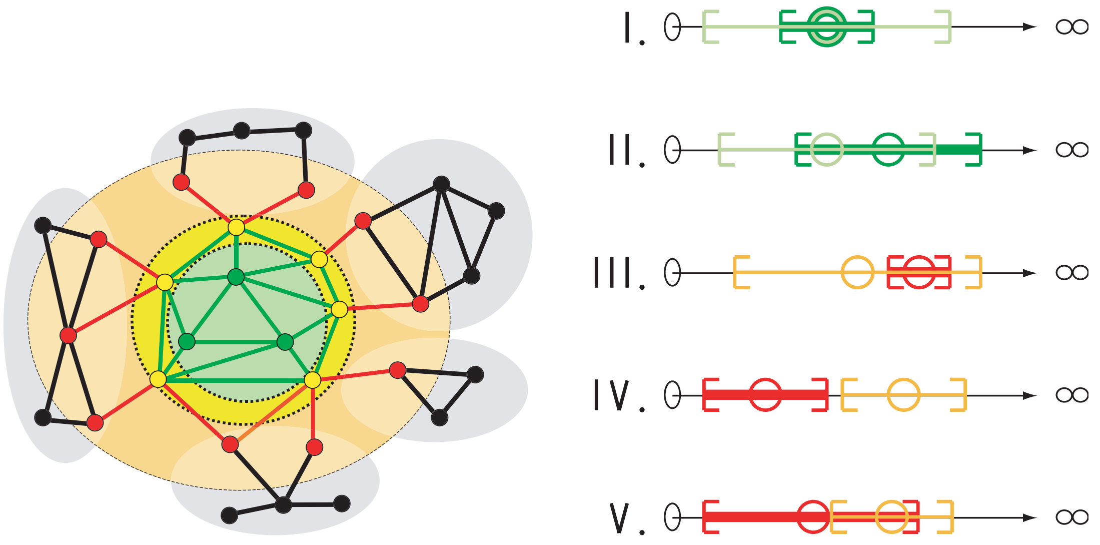
Figure 1: Modelling local to global neighbourhood. From left to right: Set of points, Nearest Neighbour Graph (NNG), Minimum Spanning Tree (MST), Relative Neighbourhood Graph (RNG), Gabriel Graph (GG), and Delaunay Triangulation (DT).

4 HPGCL-Algorithm

In our approach we use the hierarchical relationship between proximity graphs to represent a local to a global neighbourhood model. The first step in our approach is the computation of all used neighbourhood graphs (DT, GG, RNG, MST, and NNG). Then we activate the edges of the NNG to start the most local neighbourhood. Then all given nodes (data that should be clustered) are initialized as a single cluster. In our model every cluster contains a set of *inner* edges and a set of *outer* edges. The inner edges connect nodes which belongs to the same cluster and the outer edges connect nodes which belongs to different clusters (figure 2(a)). Every cluster is characterized by the median of the inner edge sizes (*cluster density*) and the *cluster variance.* the cluster variance is the absolute median deviation of all inner and outer edge sizes from the cluster density, which introduce an uncertainty model (tolarance interval)to our clustering approach. At the beginning every initial cluster has no inner edges and therefore a density of zero, but the variance will be none zero, because every node in the NNG belongs at least to one edge. All initial clusters are put into a priority queue, ordered by their density and variance values. The first cluster in the priority queue is selected and merged with all of his *valid* neighbour clusters. Valid neighbour clusters X and Y are clusters which are connected by at least one outer edge and meet the following three constraints: **Density compatibility** (see figure 2(b)), **Distance compatibility**, which means that that the median distance between X and Y belongs to the tolerance intervals of X and Y, and **Variance compatibility**, which means that the variance of the merged cluster XY is at most the maximum variance of X and Y. From the priority queue all valid neighbours are removed and the new merged cluster is inserted. Then repeat the selecting and merging step until no more clusters with valid neighbours can be found. The result is the set of clusters based on the NNG. In the next step the edges of the MST are activated and the same selecting and merging procedure as for the NNG is repeated. Tis procedure is repeated for the edges of the RNG, GG, and the DT, which represents the most global neighbourhood. Figure 3 shows the clustering result for an artificial test set with and without noise and for a 3D laser scan data set. The data set shown in figure 1(a) are the centroids of building groundplans. In figure 4 the cluster results are shown if one is using only a single graph of the hierarchie and figure 5 shows the different results if one is using the different subhierarchies.

5 Conclusion

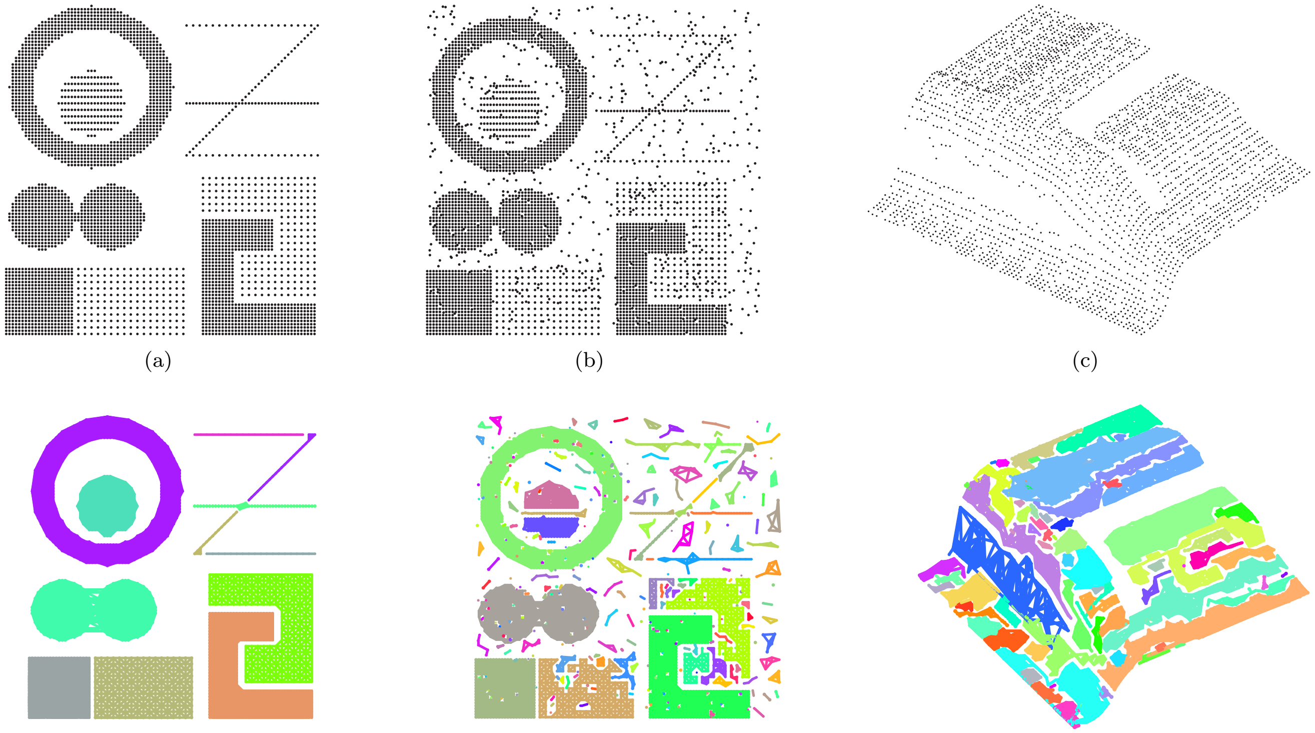
One important operation for the cartographic generalization is the search for groups of neighboured objects. We have shown that neighbourhood graphs are a very good tool to find such objects groups in a natural way without



(a) Clusterdefinition

(b) Density compatibility

Figure 2: Figure (a): Green inner edges, red outer edges and yellow cluster border. Figure (b): A circle is the cluster density and an interval is the cluster variance. Cases I and II are compatible and the others are incompatible.



(a)

(b)

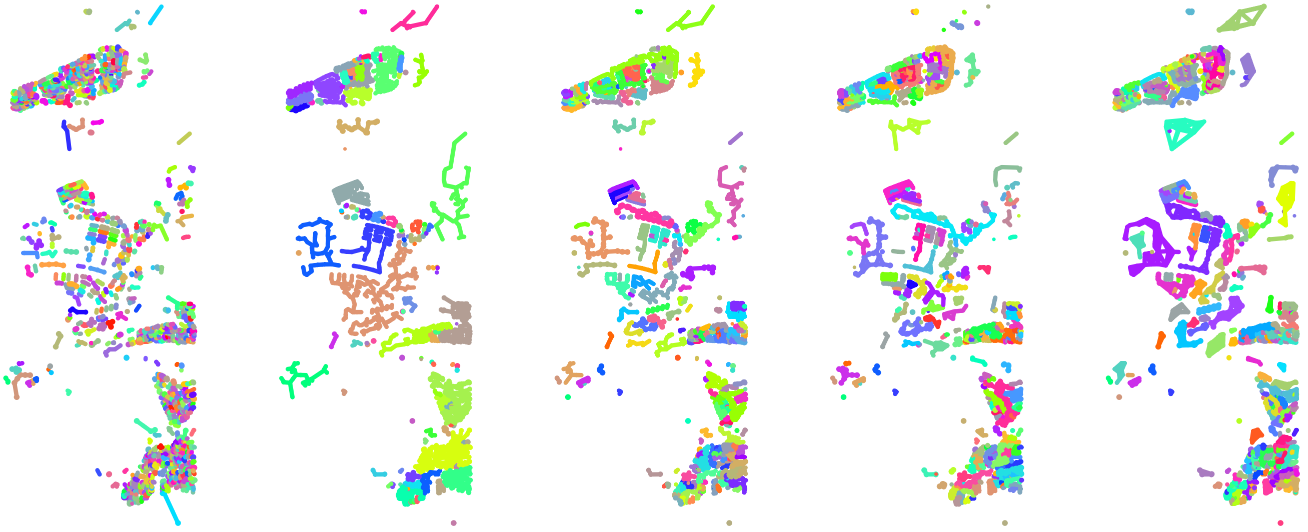
(c)

(d)

(e)

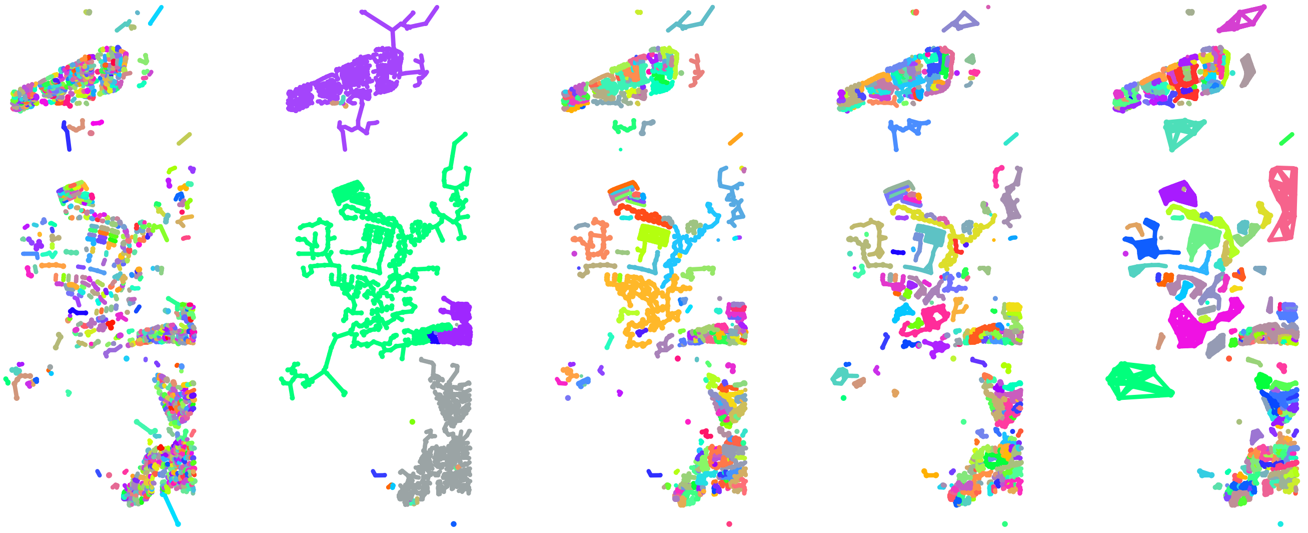
(f)

Figure 3: (a) artificial test set, (b) artificial test set with noise, (c) 3D laser scan data. (d)(e)(f) results of the clustering method.



(a) NNG (b) MST (c) RNG (d) GG (e) DT

Figure 4: Cluster result by using only a single graph.



(a) NNG (b) NNG-MST (c) NNG-RNG (d) NNG-GG (e) NNG-DT

Figure 5: Cluster result by using the different graph hierarchies.

need of any parameter. We think that more research should be investigated on the usage of neighbourhood graphs for map generalization, because neighbourhood graphs, like the Relative Neighbourhood Graph, can give us a good structural represantion of spatial objects. This structural information can be used to find regular structures, which provides us with more cartographic meta information.

References

Anders, K.-H. (2003), Parameterfreies hierarchisches Graph-Clustering Verfahren zur Interpretation raumbezogener Daten, PhD thesis, Universität Stuttgart.

Ball, G. & Hall, D. (1965), 'Isodata: a novel method of data analysis and pattern classification', Stanford Research Institute AD 699616.

Eppstein, Paterson & Yao (1997), 'On nearest neighbor graphs', *GEOMETRY: Discrete & Computational Geometry* **17**.
*citeseer.nj.nec.com/eppstein97nearestneighbor.html

Ester, M., Kriegel, H.-P., Sander, J. & Xu, X. (1996), A density-based algorithm for discovering clusters in large spatial databases with noise, *in* 'Proceedings of 2nd. International Conference on Knowledge Discovery and Data Mining (KDD-96)'.

Gabriel, K. & Sokal, R. (1969), 'A new statistical approach to geographic variation analysis', *Systematic Zoology* **18**, 259–278.

Guha, S., Rastogi, R. & Shim, K. (1998), Cure: An efficient clustering algorithm for large databases, *in* 'Proc. of 1998 ACM-SIGMOD International Conference on Management of Data'.

Guha, S., Rastogi, R. & Shim, K. (1999), Rock: A robust clustering algorithm for categorical attributes, *in* 'Proc. of the 15th International Conference on Data Engineering'.

Jain, A. & Dubes, R. (1988), *Algorithms for Clustering Data*, Prentice Hall.

Jaromczyk, J. & Toussaint, G. (1992), Relative neighborhood graphs and their relatives, *in* 'Proceedings IEEE', Vol. 80(9), pp. 1502–1517.

Jarvis, R. & Patrick, E. (1973), 'Clustering using a similarity measure based on shared near neighbours', *IEEE Transactions on Computers* **22**(11), 1025–1034.

Karypis, G., Han, E.-H. S. & Kumar, V. (1999), Chameleon: A hierarchical clustering algorithm using dynamical modeling. To appear in the IEEE Computer or via internet at http://winter.cs.umn.edu/~karypis/publications/data-mining.html.

Kaufman, L. & Rousseeuw, P. (1990), *Finding Groups in Data: An Introduction to Cluster Analysis*, John Wiley & Sons.

King, V. (1995), A simpler minimum spanning tree verification algorithm, *in* 'Workshop on Algorithms and Data Structures', pp. 440–448.
*citeseer.nj.nec.com/king95simpler.html

Lee, D. (1980), 'Two dimensional voronoi diagram in the l_p metric', *Journal of ACM* (27), 604–618.

MacQueen, J. (1967), Some methods for classification and analysis of multivariate observations, *in* 'Proc. of the Fifth Berkeley Symposium on Mathematical Statistics and Probability', Vol. 1, pp. 281–297.

Nakano, K. & Olariu, S. (1997), 'An optimal algorithm for the angle-restricted all nearest neighbor problem on the reconfigurable mesh, with applications:', *IEEE Transactions on Parallel and Distributed Systems* **8**(9), 983–990.
*citeseer.nj.nec.com/129316.html

Ng, R. & Han, J. (1994), Efficient and effective clustering method for spatial data mining, *in* 'Proc. of 1994 Int. Conf. on Very Large Data Bases (VLDB'94)', Santiago, Chile, pp. 144–155.

O'Rourke, J. (1982), 'Computing the relative neighborhood graph in the l_1 and l_∞ metrics', *Pattern Recognition* pp. 45–55.

Plazanet, C., Bigolin, N. & Ruas, A. (1998), 'Experiments with learning techniques for spatial model enrichment and line generalization', *GeoInformatica* **2**(4), 315–334.

Preparata, F. P. & Shamos, M. I. (1988), *Computational Geometry*, Springer-Verlag, New York.

Richardson, D. (1996), 'Automatic processes in database building and subsequent automatic abstractions', *Cartographica, Monograph 47* **33**(1), 41–54.

Ruas, A. & Lagrange, J. (1995), Data and knowledge modelling for generalization, *in* J.-C. M"uller, J.-P. Lagrange & R. Weibel, eds, 'GIS and Generalization - Methodology and Practice', Taylor & Francis, pp. 73–90.

Sester, M. (1999), 'Knowledge acquisition for the automatic interpretation of spatial data', *Accepted for Publication in: International Journal of Geographical Information Science* .

Sester, M., Anders, K.-H. & Walter, V. (1998), 'Linking objects of different spatial data sets by integration and aggregation', *GeoInformatica* **2**(4), 335–358.

Supowit, K. (1983), 'The relative neighborhood graph, with an application to minimum spanning trees', *J.Assoc.Comput.Mach.* **30**, 428–448.

Toussaint, G. (1980), 'The relative neighborhood graph of a finite planar set', *Pattern Recognition* **12**, 261–268.

van Smaalen, J. (1996), 'Spatial abstraction based on hierarchical re-classification', *Cartographica, Monograph 47* **33**(1), 65–74.

Yao, A.-C. (1982), 'On constructing minimum spanning trees in k-dimensional spaces and related problems', *SIAM J. Comput.* (11), 721–736.