Spatial Clustering for Mining Knowledge in Support of Generalization Processes in GIS

Bin Jiang

Division of Geomatics, Dept. of Technology and Built Environment
University of Gävle, SE-801 76 Gävle, Sweden
Email: bin.jiang@hig.se

Abstract: This paper proposes spatial clustering to be used for mining knowledge in support of generalization processes in GIS. Three aspects are investigated in the study. First, spatial clustering facilitates the detection of hidden structures and patterns (with the source map or database) that should be retained in the course of generalization, in particular when multiple attributes from geometric, topological and semantic perspectives are involved. Second, hierarchical spatial clustering supported by a dendrogram provides an efficient and effective tool for visual interrogation and exploration of clusters for multi-representation. Third, spatial clustering provides some objective criteria in quality assessment in terms of how derived clusters represent the initial patterns or structures of the dataset. All these aspects are illustrated with the conducted case studies in this paper.

Keywords: Cluster analysis, hierarchical clustering, dendrogram, visualization and generalization

1. Introduction

Generalization processes require maintaining the overall structures and patterns presented with the source map or database. Thus the recognition of the structure, mainly implicit and previously unknown, has been an important task to achieve a better generalization outcome. Over past two decades, tremendous research efforts have been spent in GIS community on the knowledge acquisition for automatic generalization based on the rule-based systems (Buttenfield and McMaster 1991). The knowledge involved in the rule-based systems includes geometry, topological and semantic relationships of the cartographic objects, as well as procedural knowledge in terms of what operations, algorithms and parameter settings to perform a generalization task (Armstrong 1991). However, knowledge is difficult to formalize, and knowledge acquisition has been a kind of bottleneck for rule-based cartographic generalization systems (Weibel et al. 1995).

Parallel to cartographic knowledge, geographic knowledge or geographic knowledge discovery has been emerging as a similar concept in the context of spatial data mining with large databases (Miller and Han 2001, Shekhar et al. 2001). Such a geographic knowledge discovery process appears to have potential applications to generalization processes, for instance using spatial clustering techniques to find clusters in a multi-dimensional dataset. On the other hand, interactive visual tools integrated with clustering techniques provide tremendous powerful tools to conduct knowledge discovery tasks (e.g. Seo and Shneiderman 2002). It is our belief that geographic knowledge discovery can be viewed as a pre-requisite for generalization processes, and some of data mining techniques like spatial clustering can be directly applied into generalization processes.

Spatial clustering is a way of detecting groups of spatial objects, or clusters, so that the objects in the same cluster are more similar than those in different clusters. Spatial clustering has been an important technique for spatial data mining and knowledge discovery (Han et al. 2001). It was mainly developed from spatial statistics and has gained intensive studies in machine learning as an unsupervised classification method. Spatial clustering applies to the situation where there is no prior knowledge about the possible classes, so it is an unsupervised classification mainly for multi-dimensional datasets.

This paper aims to explore spatial clustering techniques for generalization processes in GIS. Spatial clustering can facilitate finding the similarity or dissimilarity between cartographic objects, thus group them into clusters in terms of a variety of geometric, topological and semantic properties. The cluster detection is an important stage for the subsequent generalization operations such as selection, aggregation, and merging. Importantly hierarchical clustering accompanied by a tree like diagram (dendrogram) provides an intuitive visualization means for different levels of detail for a cluster hierarchy. The benefit of using spatial clustering in generalization processes relies also on the fact that it provides an objective criterion for assessing the quality as to how the clusters or selected objects reflect the initial patterns of the dataset.

The remainder of this paper is structured as follows. In the next section, we review the related work on knowledge acquisition and clustering techniques for generalization. Section 3 introduces some basic concepts and principles of spatial clustering with a brief introduction of two clustering algorithms. Section 4 develops some principles of spatial clustering in support for generalization operations. Section 5 reports two case studies on how spatial clustering is used in mining knowledge for generalization processes. Finally section 6 draws conclusion.

2. Related work

In this section, we briefly review related work linking to knowledge discovery via spatial clustering for generalization purposes. More comprehensive overview on the domain of geographic knowledge discovery can be found for instance in Han et al. (2001). This survey focuses on knowledge discovery from large spatial databases, not particularly linked to generalization. GIS database is huge and complex in term of the number of objects, spatial and nonspatial attributes involved. In a GIS database, it is often integrated with a variety of layers characterised as point, line and polygon objects. The relationships between objects in a same layer and objects between different layers are very complex to deal with. Furthermore attributes involving geometric, topological and semantic properties can be up to dozens of variables. The complexity of database also refers to the diversity of data types involved with the attributes, such as numerical, nominal, ordinal, binary and missing data records.

Given the complex nature of spatial database, it is beyond of human's capability to detect the knowledge hidden in the large database. Therefore automatic derivation of useful information (or knowledge) has emerged as a research field, so called knowledge discovery in database (KDD) (Fayyad et al. 1996). It sets a clear difference from relevant data mining developments such as artificial intelligence, pattern recognition and spatial statistics. KDD is a combination of several sub-processes such as data extraction, data cleaning, feature selection, algorithm design and tuning, involving a close interaction between domain experts and data mining analyst (Shekhar et al. 2001). Thus visualization is an important component for knowledge discovery processes (e.g. Seo and Shneiderman 2002).

Related work has been done in the generalization field to extract necessary knowledge for generalization purpose. Knowledge acquisition, another similar term to knowledge discovery, is a prerequisite for a rule-based generalization system, and it mainly relies on human cartographers, although sources of the knowledge can be maps and textbooks (Weibel et al. 1995). However most of cartographic knowledge is implicit and previously unknown, thus the conventional methods are not always valid. To overcome so-called knowledge acquisition bottleneck, Weibel et al. (1995) have attempted to use ccomputational intelligence methods to derive knowledge for generalization with interactive environments.

Some sporadic studies have been carried out in using spatial clustering as a major mechanism for generalization purpose. Worth noting in this respect is the application of Kohonen's self-organized map (Kohonen 2001) in generalization processes. In essence it is an unsupervised (the meaning of self-organization) clustering algorithm combined with two-dimensional map (the meaning of map) of a set of neurons. It helps to project a higher dimensional dataset into a low dimension, but maintain the initial patterns of the dataset. Højholt (1995) applied Kohonen map into a settlement pattern and derived a small set of representative houses for the purpose of typification. More recently, Jiang and his colleagues have applied the techniques to the selection of spatial objects, e.g. the selection of street from a network (Jiang and Harrie 2003), and filtering points for line simplification (Jiang and Nakos 2003). All these studies are based on the clustering capability of the self-organizing map algorithm, and its visualization for detecting and showing clusters.

Graph-based hierarchical clustering has been used in generalization processes in GIS. For instance, Jiang and Claramunt (2002) proposed a filtering model based on the graph hierarchy for the selection of streets from a network. Based on the concept of subgraphs, Anders (2003) developed an approach called hierarchical parameter-free graph clustering for selection and typification processes in GIS. Despite the efforts mentioned above, we believe that there is still a need for further development, because of numerous clustering algorithms, sensitive settings with the algorithms and diversity of data types with the datasets. We propose that spatial clustering can be used in mining knowledge for generalization rather than directly for generalization operations. In the context of this paper, knowledge is referred to hidden structures and patterns through spatial clustering. It could be knowledge on parameter settings for clustering. It could also be something of value for generalization emerged from interactive clustering processes. This proposal focuses on discovering patterns, validation of structure, and determining right parameter settings for generalization processes. Our study does not constrain to any particular clustering algorithms, although we adopt two algorithms for example in this paper.

3. Spatial clustering: concepts and algorithms

3.1 Basic concepts on spatial clustering

In the past, many clustering methods and algorithms have been developed (see Han et al. 2001 for an overview). It is worthwhile to add a remark on what is special for spatial clustering. The distinction between spatial clustering and general clustering can be said to be that between spatial statistics and general statistics and it lies in the fact that spatial processes are not a pure random process. The spatial process is likely to show clusters, which is exactly what spatial clustering aims to. There are different ways to classify the clustering methods. For instance, clustering processes can be deterministic or probabilistic, clustering learning could be incremental or batch, and the produced clusters could be exclusive or overlapping each other. In this paper, we adopt a dichotomy: hierarchical and non-hierarchical clustering techniques. The hierarchical method puts different objects into appropriate clusters iteratively and thus ends up hierarchically organized clusters. Non-hierarchical method partitions objects into different clusters, thus only one pattern with the requested number of clusters will be formed. Interested readers can refer to relevant literature (e.g. Kaufman and Rousseeuw 1990, Berkhin 2004) for an introduction and overview of the techniques. In this section, we briefly present some basic concepts on spatial clustering.

For a set of m spatial objects $X = \{x_1, x_2, ...x_m\}$, where $x_i \in \Re^n$ is an n dimensional vector in Euclidean space and $i = 1,2,...m$. The dataset can be represented a m-by-n matrix as follows:

$$\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \cdots & & & \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix} \quad [1]$$

Where each row represents an object with n attributes. The attributes for spatial objects could involve spatial variables, temporal variables and other socio-economic variables depending on applications.

Similarity is a basic notion involved in spatial clustering processes. It is a fundamental aspect to determine which objects are clustered together. Based on the notion of similarity, similar objects are clustered together as clusters. A cluster c_j is a subset of input *m* objects, i.e. $c_j \subseteq X$. The *m* objects are likely partitioned into *k* different clusters $C = \{c_1, c_2, ...c_k\}$, where $c_j \in \Re^n$ and $j = 1,2,...k$ based on some similarity measure. All the clusters constitute the entire set of the input objects, i.e. $c_1 \cup c_2 \cup ... \cup c_k = X$, noting that there may be overlaps among the clusters.

For the hierarchical clustering, it is usually supported by a dendrogram that shows a tree-like hierarchy about cluster structure of objects. With the dendrogram, all the objects $X = \{x_1, x_2, ...x_m\}$ constitute leaves of the tree. The root of the hierarchical tree is the set of all the objects. The nodes between root and leaves represent intermediate clusters at a certain level of detail. All nodes have at least one parent node and one child node. The links represent a kind of subset relationship. A dendrogram is usually represented by a U-shaped branch (see the following figure 3 as an illustration).

3.2 K-means clustering algorithm

The k-means algorithm is a classical clustering method of grouping *m* objects into *k* mutually exclusive clusters (*m*>*k*). The algorithm is often considered to be a partitioning clustering method, and it works as follows. First arbitrarily choose *k* centres and compute the distances from all the objects to the cluster centres, and then update the centres according to the new memberships of the objects. Repeat the procedure till the sum of distances from all objects in the individual clusters are minimised. Formally an object x_i is assigned to the cluster c_j iff $d(x_i, c_j) = \min(x_i)$. Note the convergence condition is with respect to distance measures one specifies, e.g., squared Euclidean distance or city block (sum of absolute differences).

Silhouette value (Kaufman and Rousseeuw 1990), also known as silhouette width, gives a sort of compactness of a cluster with respect to the other clusters. Let $a(i)$ be the average distance between the *i*th object and all of the objects in a given cluster *c*, $b(i)$ be the minimum average distance between the *i*th object and all of the objects in other clusters, the silhouette value $S(i)$ can be formally defined as:

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad [2]$$

This value gives an index to what extent the ith object belongs to the given cluster *c*. A value close to 1 suggests the object is rather centrally located; a value close to 0 indicates that the object is on the edge of two neighbouring clusters; and a value close to –1 shows that the object is probably clustered into the wrong cluster. An average value of S(i) for all the objects within a cluster describes the compactness of the cluster, so it is a useful measure for validating a clustering process.

3.3 Agglomerative hierarchical clustering algorithm

Distinct from k-means algorithm with one partition (p_1) of k clusters, a hierarchical clustering ends up with multiple partitions at the different levels of similarity, i.e. $(p_m,...p_2, p_1)$ where p_m consists of *m* original objects as clusters (most similar level), while p_1 consists of a single cluster containing all the *m* objects (most dissimilar level) (Kaufman and Rousseeuw 1990). Such a cluster hierarchy is represented as a tree-like diagram, so called dendrogram. An agglomerative hierarchical clustering is a bottom up approach. The approach joins together the two clusters that are most similar, and goes successively to bigger and bigger clusters till all the objects group as one single cluster.

In the course of the hierarchical clustering process, various agglomerative techniques, also known as linkage methods, are developed namely single linkage, complete linkage, average linkage, centroid linkage and ward linkage. The difference between the linkages is up to how one defines the distance between two clusters c_1 and c_2 with respect to all pairs of objects (note that only pairs consisting of one object from each cluster are considered). The single, complete and average linkages respectively define the minimum, maximum, average distance of all the pairs of objects as the distance of two clusters, whereas centriod linkage considers the distance between two cluster centroids. The ward linkage is based on the notion of information loss that is measured by an error sum-of-squares within a cluster. It is important to note that different linkages have different fits to different datasets and applications.

To verify how a cluster tree (or dendrogram) reflects the similarities of individual objects and clusters, a quantitative measure so called cophenetic correlation coefficient is developed for estimation of possible distortion (Kaufman and Rousseeuw 1990). The measure basically compares distance information from among all pairs of objects (for *m* objects, there are $m \times (m-1)/2$ pairs of distance), and distance information generated by a linkage method during the clustering process. Let y_{ij} be the Euclidean distance between objects *i* and *j*, z_{ij} the distance between objects *i* and *j* given by linkage, and $\bar{y}$ and $\bar{z}$ the average distance for y_{ij} and z_{ij} respectively, then cophenetic correlation coefficient (γ) is formally defined by,

$$\gamma = \frac{\sum_{i<j}(y_{ij} - \bar{y})(z_{ij} - \bar{z})}{\sqrt{\sum_{i<j}(y_{ij} - \bar{y})^2 \sum_{i<j}(z_{ij} - \bar{z})^2}} \qquad [3]$$

Note the coefficient value is drawn from the unit interval [0, 1]. The closer to unity the value is, the less distortion of the cluster tree is.

4. Generalization supported by spatial clustering: principles

The principles introduced in this section are not necessary constrained to the above two algorithms, and they are generally applicable to all clustering techniques.

4.1 Typification and selection based on clustering processes

The clustering process can be considered to be a transformation from the set of objects X to the set of clusters C, i.e. $f: X \to C$. The mean values of all attributes of the objects within a same cluster can be used to represent the cluster, i.e. the centroid of the cluster. Thus with $m \geq k$, the transformation can be regarded as a morphism between the two sets. Put it differently, the set of centres (or centroids) of *k* clusters represent to some extent the structure of the initial objects. This can be considered as a typification process for generalization. Given a subset of objects (X_i) forming a cluster, the objects can be represented (or typified) by the cluster centre (c_j), denoted by

$$X_i \Leftrightarrow c_j \qquad [4]$$

Another generalization operation selection can be achieved by picking up closet objects to the centres of individual clusters to represent individual clusters or the structure of the initial objects. The selection process can be formalised by

$$\min_i \{| x_i - c_j |\} \qquad [5]$$

where | . | is the similarity between input objects (x_i) and their respective cluster centres (c_j). Those with the best matching objects to the centres of clusters are selected.

4.2 Dendrogram as a visual support for multi-representation

Hierarchical clustering is supported by dendrogram, a tree like diagram that shows the cluster hierarchy. Its root (p_1) represent the whole set of objects, and each leaf (p_m) represents a single object of the whole data set. Such a cluster hierarchy is represented by a tree, with which a U-shaped branch represents the union of similar objects and the height of the branch represents the distance or similarity between the respective clusters. Cutting the dendrogram somewhere along a line perpendicular its branch leads to different cluster division that corresponds to the different levels of detail for a multi-representation. It provides an intuitive visualization means of showing the hierarchical structure of nested clusters. It is apparent via dendrogram that how a set of objects is organised hierarchically, and it supports a kind of generalization from leaves to root. With the above-defined dendrogram, similarity between objects or clusters can be shown by the height of the dendrogram (or distance far from leaves). The larger the height is, the less the similarity among clusters is.

$$(LOD_m, ...LOD_2, LOD_1) \Leftarrow (p_m, ...p_2, p_1) \qquad [6]$$

4.3 Cluster validity techniques for quality assessment of generalization

Cluster validity is one of important issues for cluster analysis, i.e. how can we ensure the clusters detected are valid? Cluster validity techniques aim to develop various quantitative measures for assessing the quality of clustering. For instance, we have introduced two measures respectively for k-means algorithm and the hierarchical clustering in the above section. It has been an important research in clustering techniques (Halkidi et al. 2001). The same issue is valid for generalization as well, i.e. how can we assess the quality of a generalization task? If a generalization task is carried out using clustering techniques, then we can use the relevant cluster validity technique for assessing the generalization quality.

We believe it provides another important support for generalization processes. The comparison and assessment can be performed with different cluster methods or with a same method but different parameter settings. It is on the one hand to assess the sensibility of parameter settings and on the other hand to examine the validation of cluster techniques for a particular dataset.

5. Case studies

Based on the above-introduced principles, we carried out two case studies in order to illustrate the usefulness of clustering in mining knowledge for generalization processes. The case studies focused on three important aspects: (1) clustering in support of selection and typification, (2) hierarchical clustering and dendrogram for multi-representation, and (3) cluster validity techniques for quality assessment of generalizations. The spatial clustering was done based on Statistics Toolbox with MATLAB, and outcomes were imported into ArcView GIS for selection and visualization. Let's start with a simple case study.

5.1 The first case study

This simple case study takes coordinates (x, y) of individual point objects as attributes, thus only geometric properties are involved in the generalization processes. This is a set of 179 objects that represent a real residential pattern (figure 1a). Using k-means clustering method, we derive 100 clusters from the initial 179 point objects. Figure 1b and 1c illustrates the two derived patterns representing typification and selection operations described in the section 4.1.

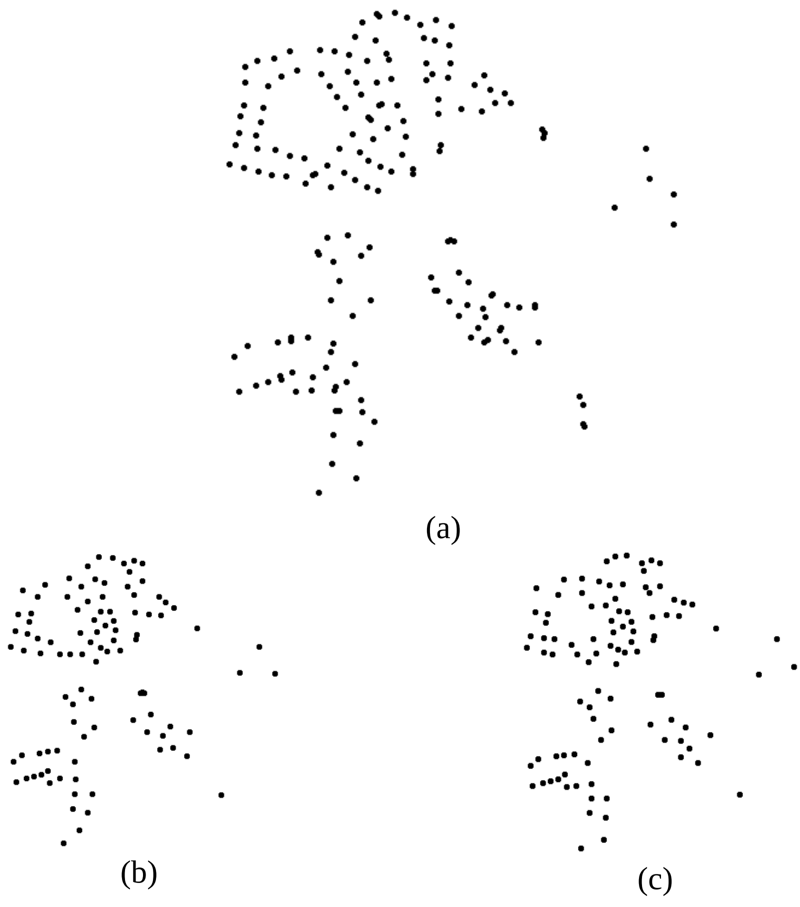


Figure 1: Original point pattern (a), its typification (b) and selection (c)

We can note from figure 1 that how the derived set of points retains the pattern of the initial data set with visual inspection. However it is important to understand the fitness of the patterns, i.e. how good the derived points through typification and selection represent the respective clusters. For illustration purpose, we computed the silhouette value for individual points with the patterns of the typification and selection. Figure 2 illustrates that the large dots means more valid as representatives of individual clusters according to the computed silhouette values.

(a) (b)

Figure 2: Silhouette value of the points with respect to the typification and selection

We applied the agglomerative hierarchical clustering algorithm and experimented all the distance measures and linkage methods. Eventually, we found that with respect to this particular dataset, Euclidean distance and average linkage are best options for distance measure and linkage method, respectively. Cophenetic correlation coefficient for the cluster tree tends to be 0.78. Figure 3 show the tree that reflects the similarities of individual objects and clusters. This can be viewed as an important knowledge discovery process, which supports the subsequent generalization operations.

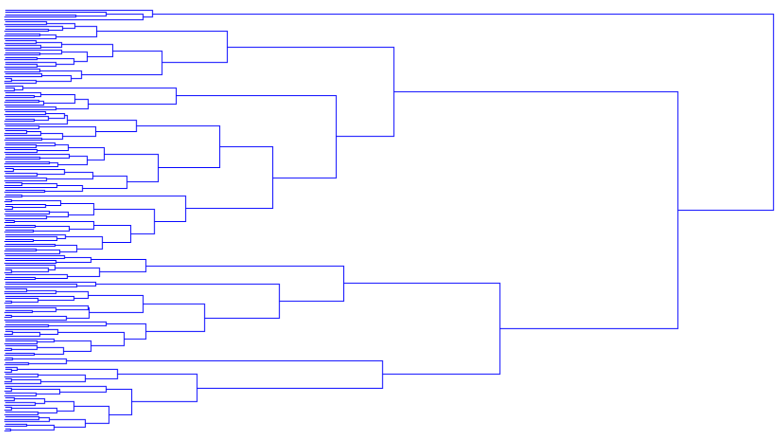


Figure 3: A cluster tree that reflects the similarities of individual objects and clusters

With the best scenario, we can make different cutoffs of the cluster tree in order to derive generalized point patterns at the different levels of detail. For illustration purpose, a series of point patterns based on the typification operation is represented in figure 4. These patterns to a great extent represent natural pattern with the initial dataset. However it is important to note that k-means algorithm is based on the spherical distance, thus it is difficult to find arbitrarily shaped clusters. Also the outcome can be very much dependent on the parameter k for the algorithm (Kaufman and Rousseeuw 1990). We can note from patterns figure 4c and 4d are a bit distorted, in particular when compared to the initial pattern. We will further see in the next case study that the k-means algorithm does not fit to some particular dataset for clustering processes.

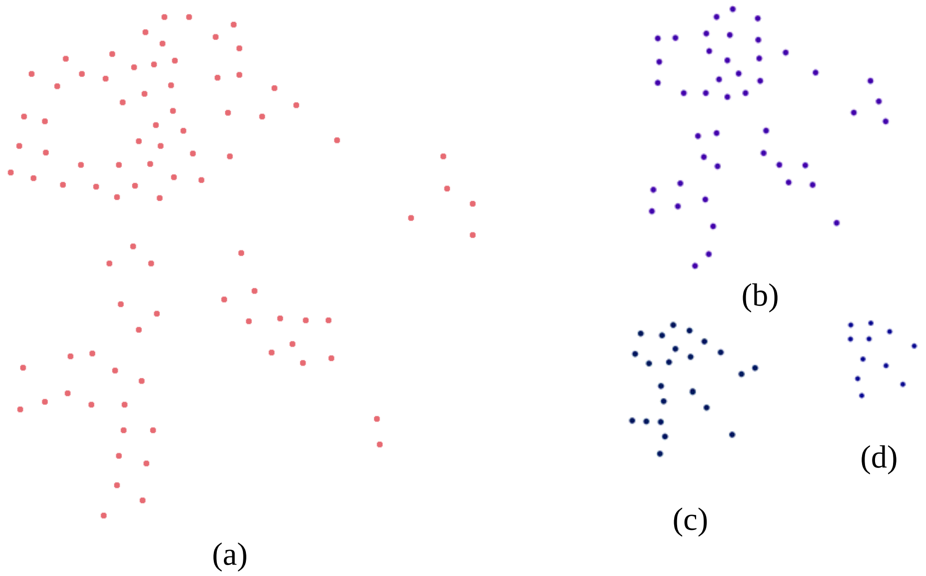


Figure 4: Typification based on hierarchical clustering at different levels of detail with 50% (a), 25% (b), 12.5% (c) and 6.25% (d) reduction

5.2 The second case study

We must consider, in the practical generalization activities, many other geometric attributes such as size (distance, area, perimeter, and volume) and orientation for line and area objects. Furthermore many other attributes from topological and semantic perspectives should be considered in the course of generalization processes. Topological properties in terms of neighbouring, adjacency, and integration (how an object is integrated to all others) are important for showing relationships of various objects. For instance well-connected objects and well-integrated objects tend to be more important. Both geometric and topological properties can be directly or indirectly seen, whereas the semantic property in terms of meaning may not be seen from its appearance. For instance some historical streets may not be significant in terms of both geographic and topological attributes, but it is often considered as landmarks that should be retained in reduced scales of map. Semantic properties could be also referred to the different class levels of a spatial object, e.g. highway, motorway and normal streets.

The second case study involves the selection of streets from the Munich network that consists of initially 785 streets. Totally 7 attributes are defined respectively from geometric, topological and semantic perspectives. More specifically, these attributes are degree, closeness, betweenness centralities (topology), length, lanes (geometry), speed and class (semantics). Details on the definition of these attributes can be found in Jiang and Harrie (2003), in which a street oriented representation is adopted for the selection of streets. Our clustering processes were based on the range of attributes with the weight [1, 1, 1, 2, 2, 2, 3] for the seven attributes in the orders of [degree, closeness, betweenness, length, lanes, speed, class].

Firstly we adopt k-means algorithm and attempt to derive three levels of detail of the network. We first used squared Euclidean distance measure for similarity computation and derived two levels of selection as shown in figure 5. This result appears to suggest that the algorithm or the parameter settings do not end up a satisfied solution to the selection. Importantly it appears that the structure of the network is not successfully discovered with the algorithm. The partial reason we believe could be the limitations of k-means algorithm we remarked before with the first case study.

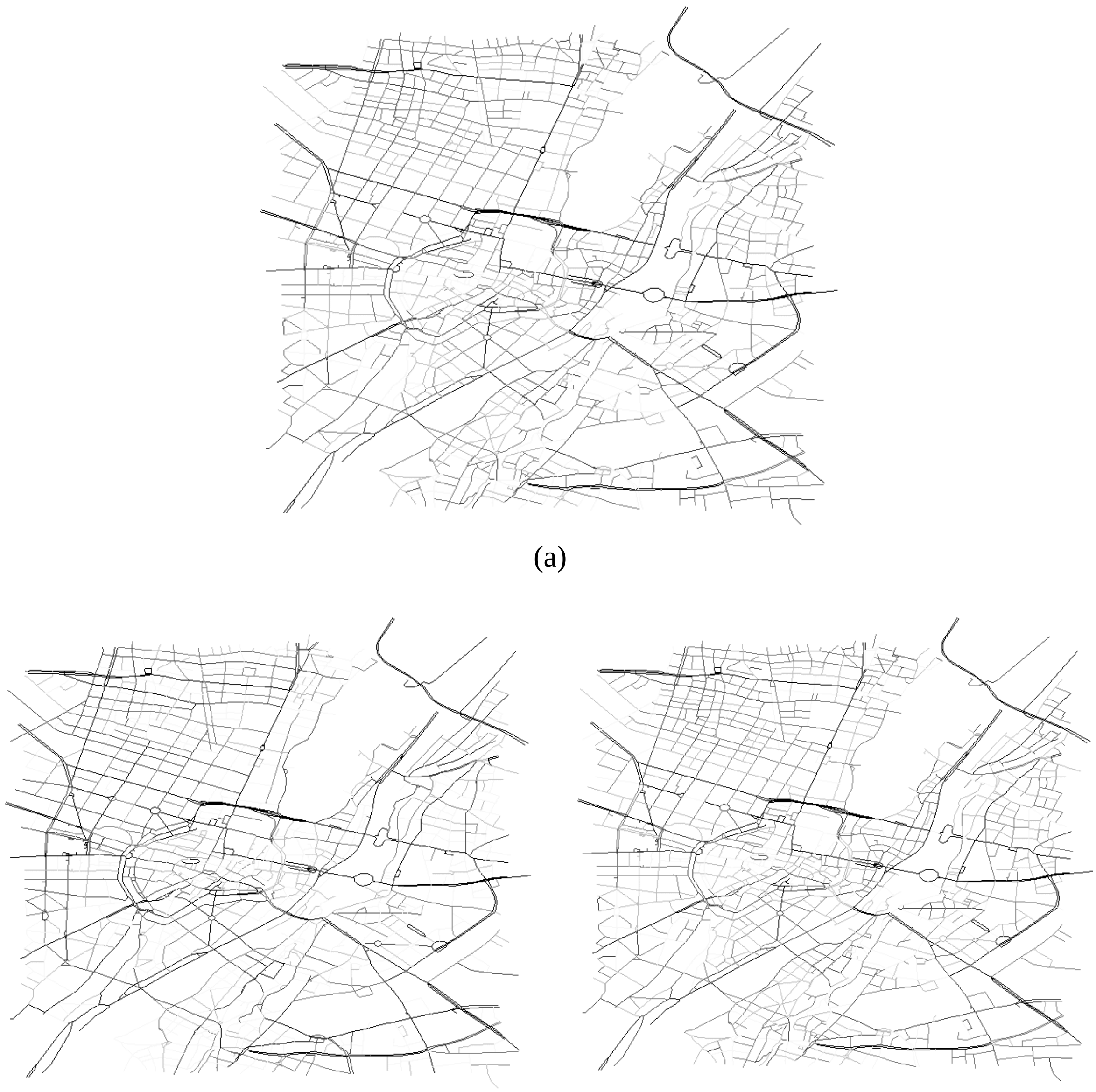
(a)

(b)

(c)

Figure 5: The original network (a) and the selected streets at the level of 50% (b) and 25% (c) based on k-means algorithm

Secondly we tried all options for distance measure and linkage methods for hierarchical clustering. It ends up a very high cophenetic correlation coefficient of 0.926 with Euclidean distance and average linkage respectively for distance measure and linkage methods. Based on the settings, we determined to cluster the streets into 589, 393 and 196 categories with respect to the levels of detail of 75%, 50% and 25%. Then we make a second classification in terms of the categories. Interestingly, we found a sort of network skeletons emerged from the network with the levels of detail as shown in figure 6.



(a)

(b) (c)

Figure 6: Network skeletons emerged from the hierarchical clustering process

We have shown with the case studies that how spatial clustering helps in generalization. However the focus of this paper is rather on mining knowledge related to generalisation processes using spatial clustering techniques. An interactive environment that integrates all kinds of spatial clustering techniques, parameter settings, and validity techniques should be developed for knowledge discovery and generalization processes. This can be done via some open sources for knowledge mining, e.g. WEKA (Witten and Frank 2000). With such a computing environment, domain experts (geographers and cartographers) can be involved together with spatial mining experts to achieve better outcomes.

5. Conclusion

This paper examined how spatial clustering techniques can be used for mining knowledge in support of generalization processes in GIS, with focus on two types of spatial clustering techniques, namely hierarchical and non-hierarchical clustering. It is demonstrated that spatial clustering can be important support for generalization processes in terms of pattern and knowledge discovery. It appears to extend the applications of other spatial data mining techniques such as classification, association rule, outliers' detection and so on in generalization processes. Apart from structure detection with spatial clustering, cluster validity provides another important support of generalization processes in terms of quality assessment. The two case studies have illustrated some aspects of knowledge discovery to do with generalization processes based spatial clustering.

Although we used the two algorithms in this paper, the principles we suggested should not be constrained to these two particular algorithms, but rather applicable to all kinds of clustering algorithms. Also no intention or evidence whatsoever is intended to show these two algorithms are best for generalization processes. On the contrary, an ideal scenario would be that for a generalization task, all possible clustering algorithms should be experimented in order to have better outcome with respect to pattern recognition and knowledge discovery. Our future work will target to such developments.

References:

Anders K. (2003), A hierarchical graph-clustering approach to find groups of objects, working paper for ICA generalization commission workshop, available at http://www.geo.unizh.ch/ICA/docs/paris2003/papers03.html

Armstrong M. P. (1991), Knowledge classification and organization, in: Buttenfield B. P. and McMaster R. B. (eds.), *Map generalization – making rules for knowledge representation*, Longman: London, pp. 86 – 102.

Berkhin P. (2004), Survey of clustering data mining techniques, *Datanautics, Inc. Research Papers*, available at http://www.accrue.com/products/rp_cluster_review.pdf

Buttenfield B. and McMaster R. (1991, eds.), *Map generalization: Making rules for knowledge representation*, Harlow: Longman.

Fayyad U. Piatetsky-Shapiro G., and Smyth P. (1996), From data mining to knowledge discovery in databases, *AI Magazine*, 17(3), pp. 37 – 54.

Halkidi M., Batistakis Y. and Vazirgiannis M. (2001), On clustering validation techniques, *Journal of Intelligent Information Systems*, Kluwer Academic Publishers. 17: 2/3, pp. 107 – 145.

Han J, Kamber M. and Tung A. K. H. (2001), Spatial clustering methods in data mining, in: Miller H. J. and Han J. (2001, eds.), *Geographic data mining and knowledge discovery*, Taylor & Francis, London and New York, pp. 188 – 217.

Højholt P. (1995). Generalization of build-up areas using Kohonen-networks, *Proceedings of Eurocarto XIII*, 2-4 October, Ispra, Italy.

Jiang B. and Claramunt C. (2002), A Structural Approach to Model Generalisation of an Urban Street Network, *presented at the 4th AGILE Conference on Geographic Information Science*, 24-27 April 2002, Mallorca, Spain (a revised version of this paper has been published in *GeoInformatica*, **8** (2): 157-171, June 2004).

Jiang B. and Harrie L. (2003), Selection of Streets from a Network Using Self-Organizing Maps, ICA Generalization Workshop, Paris, April 28 - 30, 2003 (a revised version of this paper has published in *Transactions in GIS*, Blackwell Publishers, Vol. 8, No.3, pp. 335 - 350.)

Jiang B. and Nakos B. (2003), Line Simplification Based on Shape Detection Using Self-Organizing Maps, *Proceedings of ISPRS workshop on spatial analysis and decision making (SADM 2003),* 3-5 December Hong Kong.

Kaufman L. and Rousseeuw P. 1990. *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley and Sons, New York, NY.

Kohonen T. (2001), *Self-Organizing Maps* (third edition), Springer, Berlin, Heidelberg, New York.

Miller H. and Han J. (2001, eds.), *Geographic data mining and knowledge discovery*, Taylor and Francis, London.

Seo J. and Shneiderman B. (2002), Interactively Exploring Hierarchical Clustering Results, *IEEE Computer*, Vol. 35, No. 7, pp. 80 – 86.

Shekhar S., Huang Y., Wu W., and Lu C.T. (2001), What's Special about Spatial Data Mining: Three Case Studies, in: V. Kumar, R. Grossman, C. Kamath, R. Namburu (eds.), *Data mining for scientific and engineering applications*, Kluwer Academic Pub.

Weibel R. Keller S. And Reichenbacher T. (1995), Overcoming the knowledge acquisition bottleneck in map generalization: the role of interactive systems and computational intelligence, in: A. U. Frank and W. Kuhn (eds.), *Spatial information theory, int. conference COSIT'95*, Springer, pp. 139 – 156.

Witten I. H. and Frank E. (2000), *Data mining: practical machine learning tools and techniques with Java implementations*, Morgan Kaufmann Publishers.