Regionalisation as Spatial Data Mining Problem: A Comparative Study

P V V S Srinivas, Susanta K Satpathy, Lokesh K Sharma, Ajaya K Akasapu

Department of Computer Science and Engineering

Rungta College of Engineering and Technology, Bhilai (CG) -India

***Abstract*— Regionalisation, an important problem from socio-geography. It could be solved by a classification algorithm for grouping spatial objects. A typical task is to find spatially compact and dense regions of arbitrary shape with a homogeneous internal distribution of social variables. Grouping a set of homogeneous spatial units to compose a larger region can be useful for sampling procedures as well as many applications such as customer segmentation. It would be helpful to have specific purpose regions, depending on the kind of homogeneity one is interested in. In this paper we perform comparative study on various regionalisation techniques available in literatures.**

***Keywords*— Spatial Data Mining, Regionalisation, Spatial Cluster.**

I. INTRODUCTION

Geographical information systems are becoming rich deposits of spatial data in many application areas (i.e., geology, meteorology, traffic planning, emergency aids). Moreover, the GISs provide the user with the possibility of querying a territory for extracting areas that exhibit certain properties, i.e., given combinations of values of the attributes. Just as it is intuitive to extend standard database query language to embody inductive queries, researchers believe that an analogous approach can be explored for geographical information systems, and, in general, for spatial databases. This explosively growing spatial data creates the necessity of knowledge/information discovery from spatial data, which leads to a promising emerging field, called spatial data mining or knowledge discovery in spatial databases [9]. Knowledge discovery in spatial databases can be defined as the discovery of interesting, implicit, and previously unknown knowledge from spatial databases. Spatial clustering is an important task in spatial data mining. It aims group similar spatial objects into classes or clusters so that objects within a cluster have high similarity in comparison to one another but are dissimilar to objects in other clusters [1]. An important application area for spatial clustering algorithms is social and economic geography. In the scope a classical methodical problem of social geography, "regionalisation" can be considered. Regionalisation is a classification procedure applied to spatial objects with an area representation, which groups them into homogeneous contiguous regions [2][3]. It would be helpful for many applications, e.g. for direct mailing, to have specific purpose regions, depending on the kind of homogeneity one is interested in [1].

In this paper comparative study of various regionalisation techniques are reported. Regionalisation techniques can be divided into four parts: Conventional clustering method, maximization of regional compactness approach, an explicit spatial contiguity constraint approach, and density based approach. Rest of this paper discusses these four approaches in briefly.

II. CONVENTIONAL CLUSTERING FOR REGIONALIZATION

This is probably the simplest regionalization method. Regionalization via conventional clustering algorithms was proposed by Openshaw [5] as a methodological approach for regionalizing large datasets, comprising two stages. The first stage applies any conventional partitioning, or hierarchical, clustering algorithm to aggregate areas that are similar in terms of a set of variables. In the second stage, each cluster is revised in terms of spatial contiguity by applying the following rule: If the areas included in the same cluster are geographically disconnected, then each subset of contiguous areas assigned to the same cluster is defined as a different region. Openshaw and Wymer [5] formalized this method on a step-by-step basis for classifying and regionalizing census data.

Note that the number of clusters defined in the first stage is always smaller than or equal to the number of contiguous regions resulting in the second stage. Thus, adjustments in the number of clusters are required in order to obtain the number of regions desired. In some cases, this is not possible; for example, an increment (reduction) of one unit in the number of clusters in the first stage can generate an increment (reduction) greater than one in the number of regions in the second stage.

Openshaw and Wymer [6] stressed the fact that regional homogeneity is guaranteed in the first stage. Moreover, this strategy may also help in providing evidence of spatial dependence between the areas. Thus, when the clusters in the first stage tend to be spatially contiguous, this may imply that the classification variables have some spatial pattern.

Another characteristic of this methodology is that it does not impose regional compactness. In this case; the regional shape depends heavily on the spatial distribution of the classification variables and on the clustering algorithm chosen for the analysis. The selection of the clustering algorithm is very important for identifying certain spatial patterns is pointed out. For example, the centroid and Ward's algorithms

can easily identify circular and dense spatial patterns, whereas single linkage algorithm is useful to identify elongated spatial patterns.

Finally, when optimizing the aggregation criterion, conventional clustering algorithms like the k-means approach only allow improving moves [2]. This makes the algorithm converge quickly, mainly because it can easily be trapped in suboptimal solutions. It is also known that the final solution is very sensitive to changes in the initial centroids. One approach to this problem consists of solving it with different initial centroids and then selects the best solution. Openshaw et al [7] proposed a simulated annealing variant to this algorithm which allows non-improving moves as a way to force the algorithm to explore more solutions and avoid suboptimal ones.

III. MAXIMIZATION OF REGIONAL COMPACTNESS APPROACH

Another way to obtain spatially contiguous regions is to force the regions to be as compact as possible. This strategy was introduced in the early 1960s as a methodological approach to design political districts. The authors saw the opportunity to adapt the mathematical formulation for solving the warehouse location-allocation problem to the political districting problem. The aim is to select a subset of areas to be region centers (warehouses) to which the other areas (customers) are assigned.

The aggregation criterion consists of maximizing regional compactness by minimizing the sum of the "moments of inertia," defined as the product of the population per area and the squared distance from the centroid of each area to the region center it was assigned to. It is important to note that region centers are not a decision variable but a parameter in the formulation. The only decision variable in the models is the assignation of areas to the predefined region centers.

The formulation also requires exactly equal populations in the regions. In order to satisfy these constraint fractional assignations of one or more areas to more than one region must be allowed. An iterative procedure fixes those fractional assignments and re-calculates the new regional centers. When the solution without fractional assignments leads to a change of the regional centers, the warehouse location-allocation model is solved again with this new set of centers. The process stops when no change of centers is needed after solving fractional assignments.

The satisfaction of the spatial contiguity constraint is not always guaranteed, since the assignation of the areas to their closest regional center is based on a weighted distance measure (population * distance). Thus, a final inspection of the final solution is required to correct spatial disconnections.

Hess et al. [10] made a more formal presentation of this method. The fragmentation of areas is theoretically solved by relaxing the equal population constraint to a "nearly equal" population requirement that allows the regional population to be between a lower and an upper bound. This relaxation makes it possible to formulate the problem as an integer programming model with a decision variable xij = 1 if the population of area j is assigned to the center i, and xij = 0 otherwise. xij =1 with i=j, means that area i is selected as region center. A final revision for spatial contiguity is still required.

Kaiser [11] proposed an aggregation criterion based on a weighted combination of two components. The first component is a measure of population equality, in which the population of each region should be as close as possible to the ratio between the total population and the number of regions to be designed. The second component is a measure of relative geometric compactness, where the shape of each region should be as close as possible to a circle. This relative compactness is calculated as the proportion of the geometric moment of inertia for region j and the moment of inertia of a circle with the same geometric area. The minimum value of this quotient is one, signifying that region j is a perfect circle. For a given solution, the global compactness is measured as the average of the relative compactness for all regions. Kaiser's regionalization procedure starts from an initial feasible solution that is improved, in terms of the aggregation criterion, by moving areas between regions. Two types of moves are allowed: first, moving an area from its region to every other region, and second, exchanging every pair of areas belonging to different regions. Only improving moves are accepted, which means that the process may well be trapped in local optimal solutions and be sensitive to the starting solution. The iteration process stops when no improving moves are possible. Finally, feasibility in terms of contiguity constraint depends on the weight given to the population equality component with respect to the compactness component.

Mills [12] extends the Hess et al. [10] location-allocation approach by taking into account natural boundaries in the regionalization process9 in such a way that a region is not split by these types of boundaries. This condition is achieved by performing what he called "permanent assignations," which consist of assigning an area to a particular center in order to avoid this area being assigned to another center located on the opposite side of a given natural boundary.

Bacao et al. [13] proposed a methodology that applies genetic algorithms to define the location of the region centers. The algorithm starts by creating an initial set of solutions. Each solution comprises a set of region centers, the assignation of each area to the closest region centers, and a value for the aggregation criterion. With this initial set of solutions, new solutions are created by applying selection, crossover and mutation operators in order to improve the aggregation criteria. The algorithm stops when a predefined number of solutions are generated without improvements in the aggregation criteria.

IV. AN EXPLICIT SPATIAL CONTIGUITY CONSTRAINT APPROACH

The methods covered in this section include, within their solution process, additional instruments that ensure the spatial

contiguity of each region. This implies that these models require information about the spatial neighbouring relationships between areas.

These methods can be categorized into three main categories: exact optimization models, heuristic models, and hybrid or mixed heuristic models.

Duque [15] formulated the regionalization problem as a Mixed Integer Programming (MIP) model. As in the method we just saw, Duque borrows concepts from graph theory in order to deal with the spatial contiguity constraint. Thus, the areas and their neighboring relationships are represented as a connected graph with nodes representing areas and links representing first order spatial connectivity between areas.

V. THE DENSITY BASED APPROACH FOR REGIONALISATION

The fourth approach is density based algorithm [9]. Sharma et al proposed efficient clustering technique for regionalisation of a spatial database (RCSDB) [1]. This algorithm combines the 'spatial density' and a covariance based method to inductively find spatially dense and non-spatially homogeneous clusters of arbitrary shape. RCSDB tackled above mention problems by a special clustering method which takes into account spatial point distributions as well as the distribution of several non-spatial characteristics. RCSDB classify a database of geographical locations into homogeneous, planar and density-connected subsets called "regions". It finds internal density connected sets (that is density-connected sets which allow to "touching" other clusters, but do not allow for "overlapping"). Furthermore, these sets have to own a certain minimal homogeneity. This can be measured by a normalized variance-covariance based parameter which takes into account local and global variances as well as the "extravagance" of a cluster. Furthermore, homogeneity has outlier robust and in order to increase the clustering quality, RCSDB follows avoid and reconsider noise and merge cluster heuristics.

The notion of regionalisation clustering in [1] is defined as follows:

Suppose a spatial database D of geo-referenced addresses (point data) is given. Let $X=\{X_1, \ldots, X_j, \ldots, X_m\}$ be a set of variables associated with D, so that each address $o_i \in D$ has got the m-tuple $(x_{1i}, \ldots, x_{ji}, \ldots, x_{mi})$ of values.

Definition 1: Let $CL = \{C_1, \ldots, C_k\}$ be a (not necessarily maximal) mutually exclusive set of nonempty subsets of D, denoting the result of a regionalisation clustering, so that each cluster is defined to be a regionalisation cluster, but not each possible regionalisation cluster is part of CL.

The noise can be defined in D with respect to a given clustering CL as the set of objects in D not belonging to any cluster in CL, noise = $D\setminus(C_1 \cup \ldots \cup C_k)$. Let Npred be a reflexive and symmetric binary predicate on D meaning that two points are spatial neighbours. Let Card be a function returning the cardinality of a subset of D, and MinC be a minimum cardinality.

Definition 2: Internally directly density reachable iddr(): An object p is internally directly density reachable from an object q with respect to Npred, MinC, and CL, iddr(p, q), if

$Npred(p, q)$ (neighbourhood condition)

$Card(\{o \in D | Npred(o, q)\}) > MinC$ (core object condition)

$\forall o \in D: Npred(o, q) \Rightarrow \exists C_i \in CL: o \in C_i \wedge q \in C_i$ (planarity condition)

This binary predicate is not symmetric and means that p is part of an inhabited and dense neighbourhood of q which entirely belongs to one cluster. Based on this predicate, we define "internally density reachable" idr() and "internally density connected" idc() accordingly. These definitions imply the ones in Sander et al, but they do not follow from them, so idc(p,q) implies dc(p,q), but dc(p,q) does not imply idc(p,q). Furthermore, let H be a homogeneity predicate, meaning that a subset of D is homogeneous with respect to a variable X_j and a minimum homogeneity MinH.

Definition 3. A regionalisation cluster C_i in D with respect to a set of variables $X=\{X_1, \ldots, X_m\}$ is a nonempty subset of D, satisfying the following formal requirements:

- For all addresses p, q from C_i, $p \in C_i \wedge q \in C_i$: p is internally density connected to q (internal density connectivity)
- The addresses of C_i are homogeneous with respect to each variable in X, so: $\forall X_j \in X: H(\{o \in D | o \in C_i\}, X_j)$ (homogeneity).

Outlier robust homogeneity can be measured by following formula.

$$H_{comb}(C,X) = (H_{NLC}(C,X))(H_{NLC}(C,X)) + (1 - H_{NLC}(C,X))(H_{NLV}(C,X))$$

$$H_{comb}(C,X) = \left(1 - \frac{Var_{local-local}}{Var_{local-global}}\right)\left(1 - \frac{Var_{local-local}}{Var_{local-global}}\right) + \left(\frac{Var_{local-local}}{Var_{local-global}}\right)\left(1 - \frac{Var_{local-local}}{Var_{global}}\right)$$

Definition 4: Let $MinH \in [0...1]$ be a fixed normalized homogeneity minimum, for example 0.7. Let Q_u and Q_l be an upper and lower quartile of a variable Xj in cluster C, for example $Q_{80\%}$ and $Q_{20\%}$. Then we consider the cluster C of D to be homogeneous with respect to X_j, H(C, X_j, MinH), if:

$Hcomb(C', X_j) > MinH$, with $C' = \{o_i \in C \mid Q_l - 1.5*(Q_u - Q_l) < x_{ji} < Q_u + 1.5*(Q_u - Q_l)\}$,

which means the predicate is true, if the cluster shows a minimal homogeneity for an outlier-free subset of its values.

TABLE I: REGIONALISATION METHODS AND THEIR MAIN CHARACTERISTICS

REGINALSATION METHOD / CHARACTERISTICS	CONVENATIONAL CLUSTERING	MAXIMIZATION OF REGIONAL COMPACTNESS APPROACH	AN EXPLICIT SPATIAL CONTIGUITY CONSTRAINT APPROACH	DENSITY BASED APPROACH
Number of regions is required	√	√	√	X

Information about all pairwise relationships can be considered	√	√	√	√
Neighboring relationships must be provided	X	X	X	X
Regional shape is not constrained	√	√	√	√
May be used to solve large regionalization problem	X	X	√	√
Homogeneity	X	X	X	√

VI. CONCLUSIONS

Regionalisation is an important task in spatial data mining. In this paper we summarized the various regionalisation techniques those are reported on literatures. We found first three techniques do not solve the MAUP problem. Density based approach for regionalisation provides significant process to avoid this problem. The drawback of density based approach is that it is parameter sensitive. When we do not use proper parameter for density measure, it produces more noises. In our opinion it can be solved by using evolutionary approaches.

Regionalisation offers various application areas in real life. A possible application of Regionalisation is customer segmentation. Customer segment based metric points are being developed in order to affect the shift from globalization to regionalisation in the marketing strategy. All businesses need to know where their best customers are located, how much Buying power they posses and how far any given customer must travel to nearest sales or service point. Regionalisation can be effective use for customer segmentation to indentify the target customers. Table 1 shows the regionalisation methods and it characteristics. Density based approach we do no need to provide the number of clusters. It identifies number of clusters itself. It also provides the method to calculate the homogeneity among social variables.

REFERENCES

[1] L. K. Sharma, S. Scheider, W. Kloesgen and O. P. Vyas, "Efficient clustering technique for regionalisation of a spatial database", Int. J. Business Intelligence and Data Mining, Vol. 3, No. 1, pp. 66-81, 2008.

[2] C. D. Juan, R. Raul, S. Jordi, "Supervised Regionalization Methods: a Survey", Research Institute of Applied Economics, 2006.

[3] R. M. Assuncao, M. C. Neves, G. Câmara, and C. C. Freitas, "Efficient regionalization techniques for socio-economic geographical units using minimum spanning trees," Int. J. of Geographical Information Science, vol. 20, no. 7, pp. 797-811, August 2006.

[4] S. Wise, R. Haining and J. Ma, "Providing spatial statistical data analysis functionality for the GIS user: the SAGE project", Int. J. of Geographical Information Science, vol. 15, no. 3, pp. 239–254, 2001.

[5] S. Openshaw and L. Rao, "Algorithms for reengineering 1991 census geography", Environment and Planning, vol. 27, no. 3, pp. 425–446, 1995.

[6] S. Openshaw and C. Wymer. Census Users Handbook, chapter Classifying and regionalizing census data, pages 239–270. Cambridge, UK, GeoInformation International, 1995.

[7] S. Alvanides, S. Openshaw and P. Rees, "Designing your own geographies. In The Census Data System", P. Rees, D. Martin and P. Williamson (Ed.), pp. 47-65, 2002.

[8] R. Xu and D. Wunsch, "Survey of Clustering Algorithms", IEEE Tran. on Neural Networks, vol. 16, no. 3, May 2005, pp. 645-678.

[9] M. Ester, H. P. Kriegel, J. Sander and X. Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise". In: Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining (KDD'96), Portland, AAAI Press, pp. 291-316, 1996.

[10] S. W. Hess, J. B. Weaver, H. J. Siegfeld, J. N., Whelan and P. A. Zitlau, Nonpartisan political redistricting by computer. Operations Research, 13(6):998–1006, 1965.

[11] H. F. Kaiser, "An objective method for establishing legislative districts", Midwest Journal of Political Science, 10(2):200–213, 1966.

[12] G. Mills, "The determination of local government electoral boundaries", Operational Research Quarterly, 18(3):243–255, 1967.

[13] F. Bacao, V. Lobo, and M. Painho, "Applying genetic algorithms to zone design. Soft Computing", 9(5):341–348, 2005.

[14] D. GUO, "Regionalization with dynamically constrained agglomerative clustering and partitioning (REDCAP)", Int. J. of Geographical Information Science Vol. 22, No. 7, July 2008, 801–823.

[15] J. C. Duque, Design of homogeneous territorial units. A Methodological Proposal and Applications. PhD thesis, University of Barcelona, 2004.