A Method for Detecting Space Cluster Using Geographic Raster Data

Toshihiro Osaragi
Department of Mechanical and Environmental Informatics,
Graduate School of Information Science and Engineering,
Tokyo Institute of Technology
2-12-1-W8-10 O-okayama, Meguro-ku, Tokyo, JAPAN 152-8552
osaragi@mei.titech.ac.jp

SUMMARY

In the process of visualizing quantitative spatial data, it is necessary to classify attribute values into some class divisions. In a previous paper, the author proposed a classification method for minimizing the loss of information contained in original data. This method can be considered as a type of smoothing technique that neglects the characteristics of spatial distribution. In order to understand the spatial structure of data, it is necessary to formulate another smoothing method that considers the characteristics of spatial data distribution. In this paper, a spatial clustering method based on Akaike's Information Criterion is proposed. Furthermore, numerical examples of its application are shown using actual spatial data for the Tokyo metropolitan area.

KEYWORDS: *space cluster, quadtree, AIC, visualization, classification*

INTRODUCTION

When spatial data are visualized, the attribute values defined numerically have to be classified into some class divisions. In this process, there is a risk of incorrect judgment or biased understanding since considerable information of the original data may be lost, depending on the classification method adopted. Therefore, in a previous paper, Osaragi (2003) examined the classification method of spatial data from the viewpoint of information statistics and proposed a new classification method based on the minimization of information loss. This method is a type of smoothing technique that neglects the characteristics of spatial data distribution. However, it is necessary to consider the spatial distribution of attributes in order to adequately visualize data accompanied with "spatial distribution" information.

Many studies on local smoothing have been carried out in the field of remote sensing. Gilmour (1987) proposed a method that facilitates the determination of the optimal neighborhood size. Further, Li (1996) proposed a method to integrate GIS so that the shape information, which is frequently used in visual interpretation, can easily be employed in order to improve the performance of classification. On the other hand, Liebetrau et al. (1977) discussed a classification method of spatial distribution based on several cell sizes as a hypothesis test. Furthermore, Margules et al. (1985) presented a numerical method for classifying geographic data in order to incorporate geographic location as external constraint. When a matrix of similarity values is generated and the adjacency matrix is coded, a hierarchical agglomerative fusion strategy can be used to construct hierarchical relationships between the objects (Margules et al. 1985). On the other hand, Batty (1974, 1976, 1978) discussed the zonal aggregation problem with regard to spatial entropy scaled according to zone size and decomposed the information gain into a within-set and a between-set component.

Furthermore, Fotheringham and Wong (1991) have suggested the sensitivity of analytical results to the definition of units for which data are collected. This crucial problem that is related to the use of areal data is commonly referred to as the modifiable areal unit problem (MAUP), which is clearly illustrated in the works of Openshaw (1977). Although specific statistical analysis is usually not employed in the process of visualizing spatial data, the results are likely to vary with the level of

aggregation and the configuration of the zoning system. We then have to consider appropriate areal units in this process.

In this paper, we discuss a method for detecting space clusters, taking into account the characteristics of the local spatial distribution of attributes. We primarily discuss the places that should be unified as a spatial unit with regard to a statistical model. In the following study, such a spatial unit is referred to as "space cluster." Tamagawa (1987) and Higuchi et al. (1988) have proposed a method for deciding the optimum cell size where the values of AIC (Akaike's Information Criterion; Akaike 1972, 1974) obtained by varying the observed range of data are compared. Furthermore, Nakaya (2000) has also proposed a methodology for the selection of appropriate areal units using AIC and search methods for informative geographical aggregation in map construction. In this paper, by combining these methods with our spatial classification and visualization method, we propose a new spatial clustering method for geographical data.

DEFINITION OF SPACE CLUSTER

There are two approaches for finding an appropriate space cluster. The first one is to consider each space cluster as uniformly sized. The second is to adjust the size of each space cluster as needed. In this paper, we examine/study the latter approach that has higher flexibility than the former. In other words, we examine the representation of the entire space by a set of space clusters of various sizes. The fundamental idea is as follows:

First, if the features are not homogenously distributed in the study area, it is necessary to divide this area into some smaller subareas. Further, the homogeneity of the feature distribution in the subareas will be checked, and each subarea will be further divided anew, if necessary. The entire study area is divided by repeating this procedure recursively. Thus, if it becomes unnecessary to divide subareas any further, i.e., if each subarea can be *statistically* considered homogeneous, it can be considered that the objective area is filled with appropriate space clusters at this time. When the subareas are further divided, although the amount of data increases, we obtain only limited information from the data. Therefore, we should take into account the trade-off between the amount of information and data.

According to the above discussion, the space cluster is obtained by a dividing process. However, it is also possible to form a space cluster by unifying smaller subareas (see figure 1). According to the author's experience, the latter approach produces a finer space cluster than the former one. The concrete reason for this will be discussed later.

Margules et al. (1985) tested four agglomerative hierarchical fusion strategies with the adjacency constraint. The choice of the classification strategy is dependent on the type and amount of data and the objective of the classification; it is an important decision that is equally applicable to constrained and unconstrained classification. Figure 1 shows an example in which an appropriate space cluster is represented using the quadtree structure. By assuming the top level to be the entire study area, the lower ranks can be considered as the subareas. Furthermore, each leaf can be considered to be the smallest subarea, i.e., a space cluster. In this research, the quadtree data structure is used in the process of finding the optimal space cluster. The applications discussed here are with regard to the quadtree data structure mainly. However, the following method is applicable to other fusion strategies or data structures.

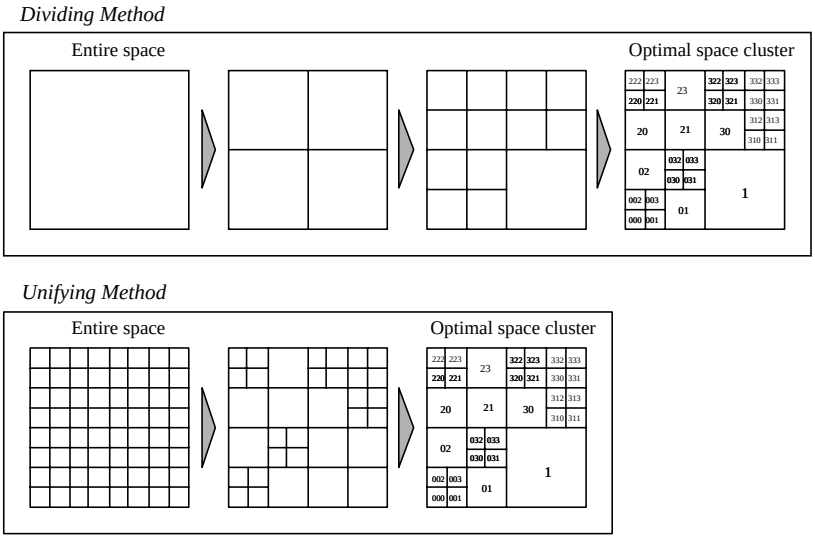


Figure 1: Quadtree data structure and the two approaches for constructing space clusters.

SPACE CLUSTER BASED ON AIC

Definition of AIC

Tamagawa (1987) and Higuchi et al. (1988) proposed a method based on AIC (Akaike 1972, 1974) in order to determine the optimum cell size. By using the values of AIC, we can evaluate models from the synthetic viewpoint of the fitness of the model and the simplicity of the model simultaneously. A model, which shows smaller AIC value, is superior to others, if the data used in models are the same. The value of AIC is given by the following equation:

$$\text{AIC} = -2\ [\textit{Maximum Likelihood}] + 2\ [\textit{the number of free parameters}]. \qquad (1)$$

A function of AIC that transformed the entire area using uniform cell size was formulated as follows (see figure 2). The attribute value of a unit cell is denoted by $x(i)$, (i=1, 2,…, n), and the sum of values in the entire area is denoted by X ($= \sum_{i=1}^{n} x(i)$). The horizontal width and vertical height are represented by a and b, respectively, when changing the cell size. Further, an attribute value of an $a \times b$ cell is denoted by $d(j)$, (j=1, 2, ..., N). In cases where the data with which the attribute values are discrete like point sampling data, the value of AIC can be expressed as follows:

$$\text{AIC} = -2\sum_{j=1}^{N} d(j)\ln\frac{d(j)}{abX} + 2(N-1), \qquad (2)$$

$$\text{where } d(j)\ln\frac{d(j)}{abX} = 0 \quad \text{when } d(j) = 0.$$

Further, in the case of data with which attribute values are continuous like a ratio, the value of AIC is expressed as follows:

$$\text{AIC} = n\ln 2\pi + n\ln\hat{\sigma}^2 + n + 2(N+1), \qquad (3)$$

$$\text{where } \hat{\sigma}^2 = \frac{1}{n}\left(\sum_{i=1}^{n} x(i)^2 - \sum_{j=1}^{N}\frac{d(j)^2}{ab}\right).$$

The cell size that produces a minimum value of AIC is considered optimal with regard to the trade-off between the fitness (the amount of information) and the simplicity (the amount of data) of the model. However, this method is based on the idea of covering the entire area with uniformly sized cells.

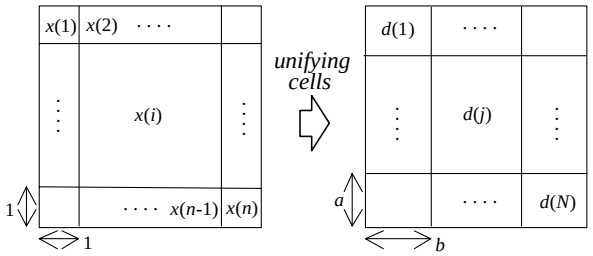


Figure 2: Cell size and attribute values.

Method for Detecting Optimal Space Cluster

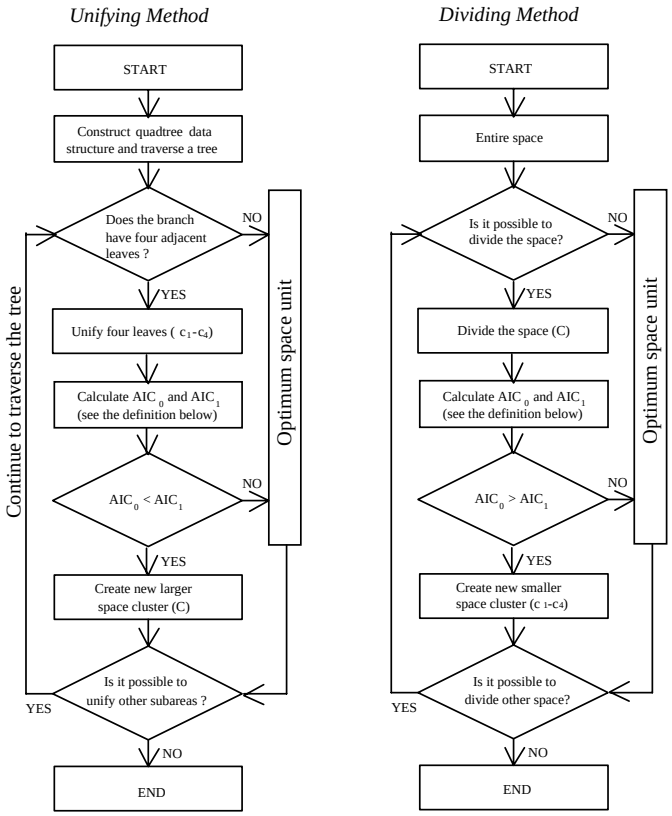
A method for obtaining an optimal space cluster using the evaluation function of AIC is proposed. The fundamental procedures of unifying and dividing subareas are shown in figure 3. By unifying four subareas of size 2*k*, belonging to the same branch, a new subarea whose size is 2*k*+1 is formed. Here, the attribute values of smaller subareas are expressed as $c_1,..,c_4$, while those of larger subareas are expressed as *C*, for convenience. If the larger subarea, whose size is 2*k*+1, is considered as one space cluster by referring to equation (2), the value of AIC (i.e., AIC_0) can be expressed as follows:

$$AIC_0 = -2C \ln \frac{1}{2^{2(k+1)}}, \tag{4}$$

where the attribute value in this case is discrete. On the other hand, if the four smaller subareas ($c_1,..,c_4$), whose size is 2*k*, are considered independent, the value of AIC (i.e., AIC_1) can be expressed as follows:

$$AIC_1 = -2\sum_{l=1}^{4} c_l \ln \frac{c_l}{2^{2k} C} + 6 \tag{5}$$

Therefore, by comparing equations (4) and (5), we can state that a model with a small value of AIC is adequate when considering the trade-off between the fitness (the amount of information) and the simplicity (the amount of data) of the model. If AIC_0 is less than AIC_1, the subareas should form the larger subarea with size 2*k*+1. On the contrary, if AIC_0 is greater than AIC_1, a larger subarea should not be formed and we must adopt the smaller subarea with size 2*k*, as the adequate space cluster. Furthermore, by referring to the equation (3), we obtain the following equations when the attribute value is continuous.



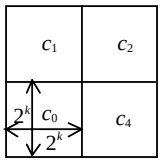
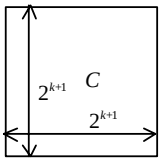
	Four adjacent leaves of level k	*Leaf of level k* +1
		
Discrete variable	$\mathrm{AIC}_1 = -2\sum_{l=1}^{4} c_l \ln \frac{c_l}{2^{2k} C} + 6$	$\mathrm{AIC}_0 = -2C \ln \frac{1}{2^{2(k+1)}}$
Continuous variable	$\mathrm{AIC}_1 = 2^{2(k+1)} \left(\ln 2\pi + \ln \hat{\sigma}^2 + 1\right) + 10$ where $\hat{\sigma}^2 = \frac{1}{2^{2(k+1)}} \left\{ \sum_{l=1}^{4} \sum_{i \in c_l} x_i^2 - \sum_{l=1}^{4} \frac{c_l^2}{2^{2k}} \right\}$	$\mathrm{AIC}_0 = 2^{2(k+1)} \left(\ln 2\pi + \ln \hat{\sigma}^2 + 1\right) + 4$ where $\hat{\sigma}^2 = \frac{1}{2^{2(k+1)}} \left\{ \sum_{i \in C} x_i^2 - \frac{C^2}{2^{2(k+1)}} \right\}$

Figure 3: Flow chart of the algorithm for obtaining optimal space cluster using AIC.

$$\mathrm{AIC}_0 = 2^{2(k+1)}\left(\ln 2\pi + \ln \hat{\sigma}^2 + 1\right) + 4, \qquad (6)$$

$$\text{where} \quad \hat{\sigma}^2 = \frac{1}{2^{2(k+1)}}\left\{\sum_{i \in C} x_i{}^2 - \frac{C^2}{2^{2(k+1)}}\right\}.$$

$$\mathrm{AIC}_1 = 2^{2(k+1)}\left(\ln 2\pi + \ln \hat{\sigma}^2 + 1\right) + 10,$$

$$\text{where} \quad \hat{\sigma}^2 = \frac{1}{2^{2(k+1)}}\left\{\sum_{l=1}^{4}\sum_{i \in c_l} x_i{}^2 - \sum_{l=1}^{4}\frac{c_l{}^2}{2^{2k}}\right\}.$$

However, in equation (7) AIC cannot be estimated when $k = 0$ (namely, in case of the smallest unit cell). In the above calculations, it is possible to weight attribute by multiplying the values by the weight.

Comparison of Methods: "Dividing" and "Unifying"

The difference between the "dividing method" and the "unifying method" is examined using artificial spatial data. The result of the analysis of the data (the number of cells is 64) using these methods is shown in the figure 4. When a dividing method is adopted, an optimum is reached only when the whole study area is unified into one space cluster, i.e., the local minimum of AIC. We can confirm that the AIC_0 of the entire area is smaller than the value of AIC_1 of the divided subareas. On the other hand, if the unifying method is applied to the same artificial data in order to form the space cluster, we can avoid the above problem—a local minimum of AIC. In addition, considering that our research is aimed at decreasing information loss (Roy et al. 1982), the unifying method is preferable. Therefore, using the dividing method we risk losing vital information regarding the original data, as is clearly shown by this simple example. Thus, in the following sections, the unifying method is adopted.

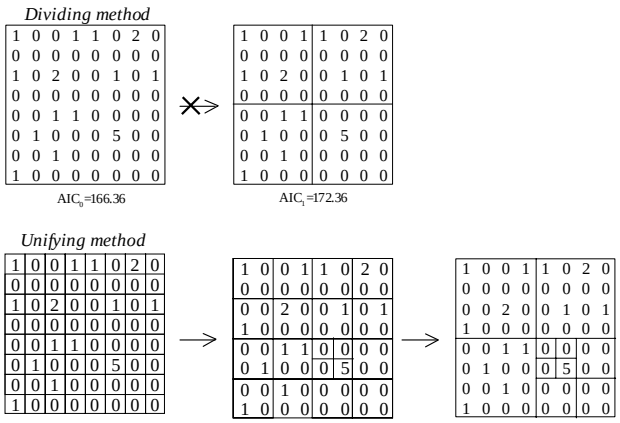


Figure 4: Comparison of the dividing method and unifying method.

A Method for Unifying Surrounding Subareas

Since the quadtree is a efficient data structure for GIS, this technique can be set naturally in GIS as a visualization tool. However, the extracted space clusters are extremely rigid, i.e. it implies only "square" clusters. Therefore, a method for obtaining more flexible space clusters is proposed (see figure 5). First, consider a subarea, denoted by C_0, and another subarea surrounding C_0. Calculate the AIC values for the cases of unifying this pair of two subareas into one larger subarea (AIC_0), and of leaving them as two subareas (AIC_1). If AIC_0 is less than AIC_1, consider this pair of cells satisfies the "*condition of unification*." After performing this operation using all surrounding subareas (C_1,..,C_8),

and count the number of pairs, denoted by *S*, which satisfies the condition of unification. If *S* is equal to a threshold (θ:θ=1,2,..,8) or more, these pairs are unified. According to the above method, the space cluster, which is not based on the quadtree data structure, can also be constituted. The threshold θ is called the "*intensity of unification.*" The greater θ forms less and larger space clusters. On the contrary, the smaller θ forms many and smaller space clusters. Since the method for unifying surrounding subareas does not have restrictions in the contiguity relation of subareas like quadtree, it can constitute the space clusters of more flexible figures. Furthermore, the abstraction level of visualization can be controlled by changing the intensity of unification θ.

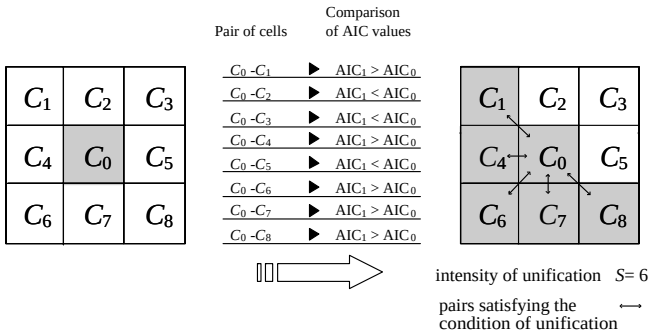


Figure 5: A method of unifying surrounding subareas and concept of intensity of unification.

APPLICATION TO ACTUAL SPATIAL DATA

Based on the above consideration, an appropriate space cluster is obtained using actual urban spatial data. The study area is shown in figure 6. The spatial data used here is as follows: (a) the ratio of nuclear families, (b) the ratio of female workers, (c) the number of commuters whose travel time is more than 1.5 hours, and (d) the number of fishery workers. The source of data is Digital Mesh Statistics—"1991 place-of-business statistics" and "1990 national census." The cell size is approximately 1 km x 1 km and the number of cells is 256.

The results of space cluster visualization are shown in figure 7. The result in the case where the optimal space cluster was neglected is also shown simultaneously. From figure 7, we observe that the size of the space clusters is reduced by our proposed method. For instance, for the data whose attribute value is discrete, the cell with an outstanding value is expressed as the smallest space unit, and the other cells are unified into larger space clusters. Comparatively, if the attribute value itself is large, a small space cluster is created. On the other hand, for the data whose attribute value is continuous, the cell with a relatively large value is combined with the other cells. In other words, the smallest unit cell is unified with the surrounding cells and is represented as a larger space cluster that is statistically meaningful. Hence, the figure 7 clearly shows that if we create the appropriate space cluster, the space distribution characteristic of the original data can be understood more easily. Comparing the methods based on quadtree and regular data structure, the former can represent the efficient space clusters from the point of view of the data structure, while the latter can achieve more flexible shapes.

Figure 6: Map of the study area.

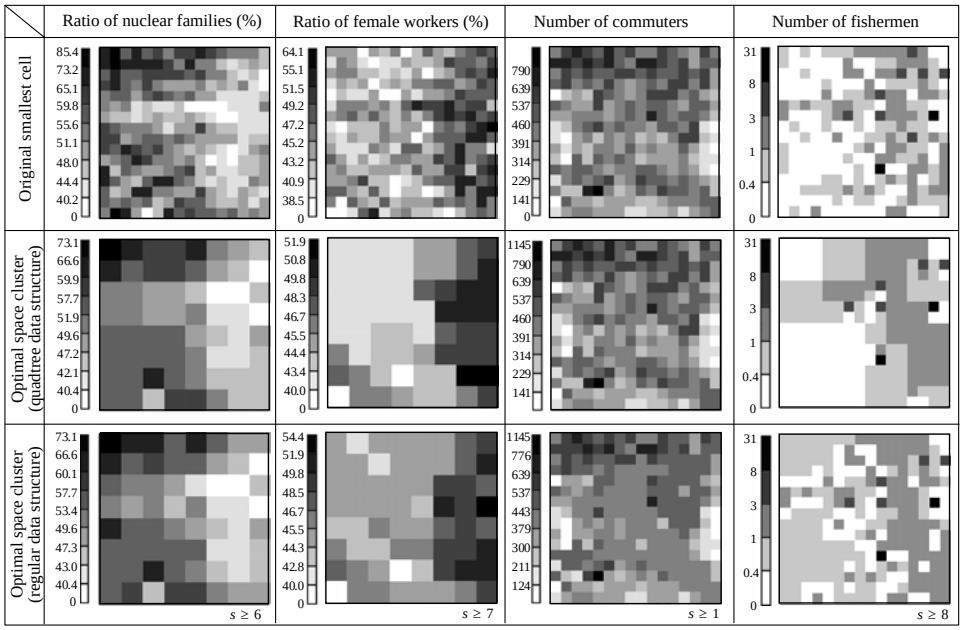


Figure 7: Visualization of space cluster using the classification method based on the minimization of information loss (Osaragi, 2003).

SUMMARY AND CONCLUSIONS

A method of obtaining a space cluster using the evaluation function of AIC is proposed by considering the distribution characteristics of spatial data. Moreover, the appropriate space cluster is visualized by the information loss minimization method. Using the proposed method, the information contained in the original spatial data can be visualized, and we can grasp and understand the statistical characteristics of geographical data.

ACKNOWLEDGMENTS

The author would like to specially thank Mr. Hiroki Yamanaka, Graduate Student of Tokyo Institute of Technology, for the computer-based numerical calculations. Further, the author would like to express his gratitude for the valuable comments provided by the anonymous referees and the researchers at the Centre for Advanced Spatial Analysis, University College London.

BIBLIOGRAPHY

Akaike H, 1972 Information theory and an extension of the maximum likelihood principle, Proceedings of the 2nd International Symposium on Information Theory, B N Petron, F Csak (ed.), 267-281.

Akaike H, 1974 A new look at the statistical model identification, IEEE Transactions on Automatic Control AC19, 716-723.

Batty M, 1974 Spatial Entropy, Geographical Analysis, 6, 1-31.

Batty M, 1976 Entropy in Spatial Aggregation, Geographical Analysis, 8, 1-21.

Batty M, 1978 Speculations on an information theoretic approach to spatial representation, in Studies in Applied Regional Science 10: Spatial Representation and Spatial Interaction, Edt I Masser, P Brown (ed.), 115-147.

Fotheringham A S and Wong D W S, 1991 The modifiable areal unit problem in multivariate statistical analysis, Environment and Planning A, 23, 1025-1044.

Gilmour T, 1987 Image smoothing as an aid to classification, in Advances in Digital Image Processing, Proc. Remote Sensing Society 13th annual conference, Nottingham, 56-64.

Higuchi T, Tamagawa H and Ishak A B P, 1988 A study on the optimum mesh size for continuous variables: An example by using a mental map, Papers on City Planning 23, 37-42 (in Japanese).

Li-Xia, 1996 A method to improve classification with shape information, International Journal of Remote Sensing, 17 (8), 1473-1481.

Liebetrau A M and Rothman E D, 1977 A classification of spatial distributions based upon several cell sizes, Geographical Analysis, 9, 14-28.

Margules C R, Faith D P and Belbin L, 1985 An adjacency constraint in agglomerative hierarchical classifications of geographic data, Environment and Planning A, 17, 397-412.

Nakaya T, 2000 An information statistical approach to the modifiable areal unit problem in incidence rate maps", Environment and Planning A, Vol.32, pp.91-109.

Openshaw S, 1977, "Optimal zoning systems for spatial interaction models", Environment and Planning A, Vol.9, pp.169-184.

Osaragi, T., 2003 Information Loss Minimization for Spatial Data Representation, Journal of Architectural Planning and Engineering, AIJ, 574, 71-76 (in Japanese).

Roy J R, Batten D F and Lesse P F, 1982 Minimizing information loss in simple aggregation, Environment and Planning A, 14, 973-980.

Tamagawa H, 1987 A study on the optimum mesh size in view of the homogeneity of land use ratio, Papers on City Planning 22, 229-234 (in Japanese).