CENTRE FOR ADVANCED SPATIAL ANALYSIS

Working Paper Series

Paper 41

SPATIAL CLUSTERING METHOD FOR GEOGRAPHIC DATA

Toshihiro Osaragi

Centre for Advanced Spatial Analysis
University College London
1-19 Torrington Place
Gower Street
London WC1E 6BT

[t] +44 (0) 20 7679 1782
[f] +44 (0) 20 7813 2843
[e] casa@ucl.ac.uk
[w] www.casa.ucl.ac.uk

http//www.casa.ucl.ac.uk/paper41.pdf

Date: January 2002

ISSN: 1467-1298

Toshihiro Osaragi

Toshihiro Osaragi is an Associate Professor in the Graduate School of Information Science and Engineering at Tokyo Institute of Technology. He was an Academic Visitor at the Centre for Advanced Spatial Analysis from March 2001 to January 2002.

Department of Mechanical and Environmental Informatics
Graduate School of Information Science and Engineering
Tokyo Institute of Technology
2-12-1 O-okayama, Meguro-ku, Tokyo 152-8552, JAPAN
Tel:+81-3-5734-3162 Fax:+81-3-5734-2817
email: osaragi@mei.titech.ac.jp

Spatial Clustering Method for Geographic Data

Toshihiro OSARAGI

Abstract: 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 kind of smoothing method that neglects the characteristics of spatial distribution. In order to understand the spatial structure of data, it is also necessary to construct another smoothing method considering the characteristics of the distribution of the spatial data. 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: spatial data, space cluster, Quadtree, AIC (Akaike's Information Criterion), information loss, visualization, classification

1. Introduction

When spatial data are visualized, the attribute values defined numerically have to be classified into some class divisions. In this process, there exists the risk of leading us to miss-judgment or biased understanding, since much information of the original data may be lost, according to the classification method adopted. Therefore, in a previous paper, Osaragi (2001) examined the classification method of spatial data from the viewpoint of information-statistics, and proposed a new classification method based on minimization of information loss. This method is a sort of smoothing technique neglecting the characteristics of spatial data distribution. However, it is necessary to consider the spatial distribution of attributes, to adequately visualize data accompanied with information of "spatial distribution".

In the field of remote sensing, many studies on local smoothing have been carried out. Gilmour (1987) proposed a method that assists in the determination of the optimal neighbourhood size, and Li (1996) proposed a method to integrate GIS so that the shape information, which is frequently used in visual interpretation, can easily be employed to improve the performance of classification. On the other hand, Liebetrau et al. (1977) discussed a classification of spatial distribution based on several cell sizes from the point of view of hypothesis testing. Furthermore, Margules et al. (1985)

presented a numerical method for classifying geographic data in order to incorporate geographic location as an external constraint. Once the matrix of similarity values has been generated and the adjacency coded, a hierarchical agglomerative fusion strategy can be used to construct hierarchical relationships between the objects (Margules et al. 1985). Conversely, Batty (1974, 1976, 1978) discussed the zonal aggregation problem according to a spatial entropy scaled for zone size, and decomposed the information gain into a within-set and a between-set component.

Furthermore, Fotheringham and Wong (1991) has suggested that the sensitivity of analytical results to the definition of units for which data are collected. This stubborn problem related to the use of areal data is commonly referred to as the modifiable areal unit problem (MAUP), which is clearly illustrated through the works of Openshaw and Taylor (1979). Although any 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 with the configuration of the zoning system. Then we have to consider appropriate areal units in this process.

In this paper, we discuss a spatial clustering method considering the characteristics of the local spatial distribution of attributes. Namely, we discuss the question "Which places should be unified as a spatial unit in the sense of a statistical model?" In the following, such a spatial unit is called "space-cluster". Tamagawa (1987) and Higuchi et al. (1988) have proposed a method for deciding the optimum cell-size in which the values of AIC (Akaike's Information Criterion; Akaike 1972, 1974), obtained thorough variously changing the observed range of data, are compared. Furthermore, Nakaya (2000) has also proposed a methodology to select appropriate areal units using AIC and search methods for an informative geographical aggregation in map construction. In this paper, combining these ideas with our spatial classifying and visualizing method, a new spatial clustering method for geographical data is proposed.

2. Definition of Space-cluster

When asking for the appropriate space-cluster, we have two options. The first is to make each space-cluster a uniform size. The second is to change the size of every space-cluster if needed. In this paper, the way of the latter, with higher flexibility than the former, is attempted. That is, we examine how to represent the entire space by a set of space-clusters of various sizes. The fundamental idea is as follows.

First, when the distribution of features is not homogeneous in the study area, it is necessary to divide

the area into some smaller sub-areas. Furthermore, checking the homogeneity of feature distribution in the sub-areas, further division within each sub-area will be done anew, if necessary. The entire study area is divided by repeating this procedure. Thus, if it becomes unnecessary to divide sub-areas any further, i.e., if each sub-area can be *statistically* considered homogeneous, it can be considered that the objective area is filled with appropriate space-clusters at this time. Even if the sub-areas are divided into smaller sub-areas further, we can get only a little information from the data, and the data size will be getting large. That is, we should pay attention to the trade-off relationship between amount of information and amount of data itself.

According to the above discussion, the space-cluster is obtained by a dividing process. However, it is also possible to constitute a space-cluster by unifying smaller sub-areas (see figure 1). According to the author's experience, the latter option is able to constitute a finer space-cluster than the former one. The concrete reason for this will be shown later.

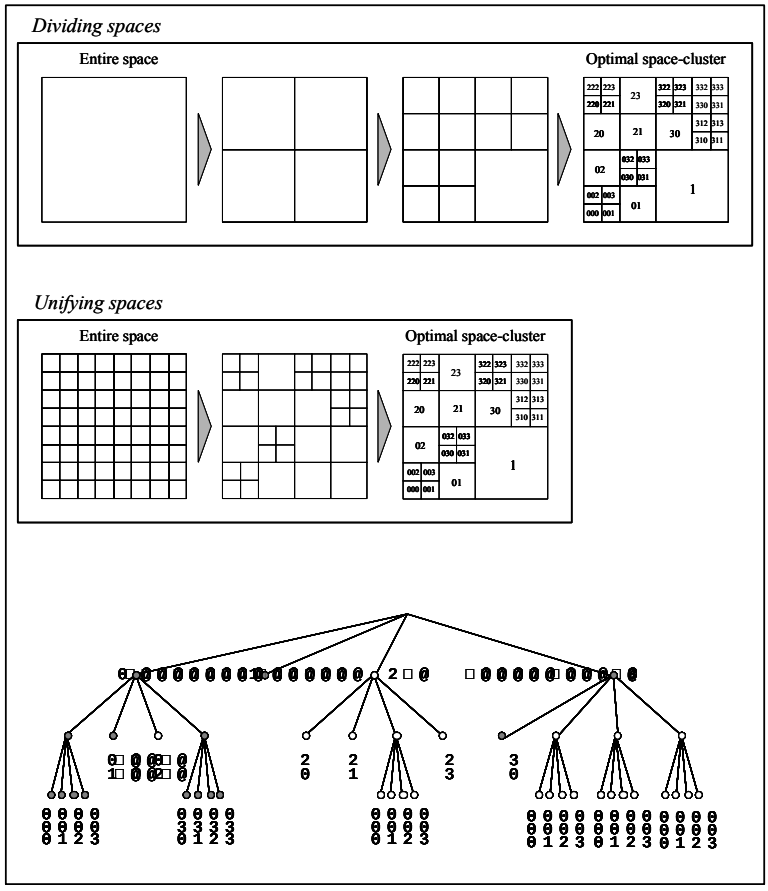


Figure 1: Quadtree data structure and two ways of constructing space-clusters

Margules et al. (1985) tested four agglomerative hierarchical fusion strategies with the adjacency constraint. The choice of classification strategy, which should depend on the type and amount of data and objective of the classification, is an important decision that applies equally to constrained or unconstrained classification. In this research, the Quadtree data structure is used for the process of finding the optimal space-cluster. The applications discussed here are limited to the Quadtree data structure. However, the following method can be applied to any other fusion strategies or data structures. Figure 1 shows an example in which an appropriate space-cluster is expressed using the Quadtree structure. Assuming the top level is the entire study area, the low rank can be considered sub-areas. Furthermore, each leaf can be considered the smallest sub-area, i.e., a space-cluster. That is, an adequate space-cluster can be obtained by traversing the tree using an evaluation function.

3. Space-cluster based on AIC

3.1 Definition of AIC

Tamagawa (1987) and Higuchi et al. (1988) proposed a method based on AIC in order to determine the optimum cell-size. A function of AIC was formulated as follows, transforming the whole area into uniform cell-size (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 ($=\sum_{i=1}^{n} x(i)$). Furthermore, horizontal width and vertical height are represented by a and b respectively when changing cell-size. Furthermore, an attribute value of an axb-cell is denoted by $d(j)$,(j=1, 2, ..., N). As for the data with which the attribute value is defined as a discrete value like point sampling data, the value of AIC can be described as follows:

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

where $d(j)\cdot \log\frac{d(j)}{abX} = 0$ when $d(j)$=0.

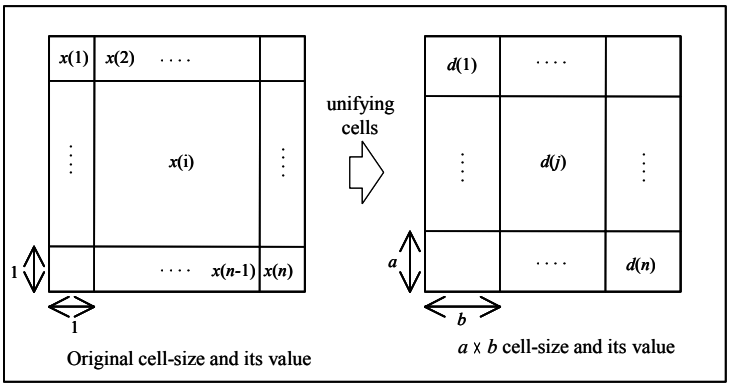


Figure 2: Cell-size and attribute values

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

$$\mathrm{AIC} = n\log 2\pi + n\log\hat{\sigma}^2 + n + 2(N+1), \tag{2}$$

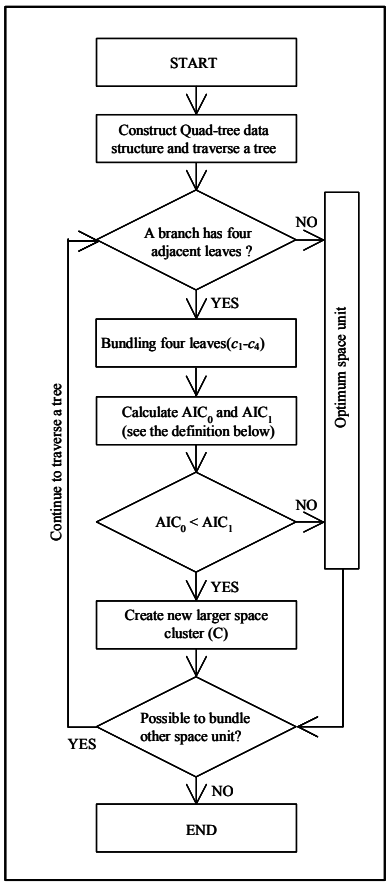
where $\hat{\sigma}^2 = \frac{1}{n}\left(\sum_{i=1}^{n} x(i)^2 - \sum_{j=1}^{N}\frac{d(j)^2}{ab}\right)$.

The cell-size that gives the minimum value of AIC can be regarded as optimal, in a sense of the trade-off relationship between amount of information and amount of data. However, this method is based on the idea of covering the entire area by the same-sized cells.

The author proposes a method for obtaining the optimal space-cluster using the evaluation function of AIC. The fundamental procedure is shown in figure 3. By unifying four sub-areas belonging to the same tree, whose size is 2^k, the new sub-area whose size is 2^{k+1} is formed. Here, the attribute values of smaller sub-areas are expressed as c_1-c_4, and that of larger sub-areas is expressed as C, for convenience. If the larger sub-area, whose size is 2^{k+1}, can be considered as one space-cluster by referring to equation (1), the value of AIC (i.e., the value of AIC_0) can be expressed as follows:

$$\mathrm{AIC}_0 = -2C\cdot\log\frac{C}{2^{2(k+1)}C} + 2(1-1), \tag{3}$$

where the attribute value in this case is a discrete value.



	Four adjacent leaves of level-k	*Leaf of level-k*+1
	c_1 c_2 c_0 c_4 2^k 2^k	2^{k+1} C 2^{k+1}
Discrete variable	$\mathrm{AIC}_1 = -2\sum_{l=1}^{4} c_l \cdot \log \frac{c_l}{2^{2k} C} + 2(4-1)$	$\mathrm{AIC}_0 = -2C \cdot \log \frac{C}{2^{2(k+1)} C} + 2(1-1)$
Continuous variable	$\mathrm{AIC}_1 = 2^{2(k+1)}\left(\log 2\pi + \log \hat{\sigma}^2 + 1\right) + 2(4+1)$ *where* $\hat{\sigma}^2 = \frac{1}{2^{2(k+1)}}\left\{\sum_{l=1}^{4}\sum_{i \in c_l} x_i^{\,2} - \sum_{l=1}^{4} \frac{c_l^{\,2}}{2^{2k}}\right\}$	$\mathrm{AIC}_0 = 2^{2(k+1)}\left(\log 2\pi + \log \hat{\sigma}^2 + 1\right) + 2(1+1)$ *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: Algorithm for obtaining optimal space-cluster using AIC

On the other hand, if the four smaller sub-areas (c_1-c_4), whose size is 2^k, are considered independent, the value of AIC (i.e., the value of AIC_1) can be expressed as follows:

$$\mathrm{AIC}_1 = -2\sum_{l=1}^{4} c_l \cdot \log \frac{c_l}{2^{2k} C} + 2(4-1) \qquad (4)$$

That is, by comparing equation (3) with equation (4), we can say that a model with a small value is adequate when considering the trade-off relationships between amount of information and amount of data. If AIC_0 is less than AIC_1, the sub-area should form the larger sub-area whose size is 2^{k+1}. On the contrary, if AIC_0 is greater than AIC_1, a larger sub-area should not be formed and we should adopt the smaller sub-area whose size is 2^k as the adequate space-cluster. Furthermore, by referring to formula (2), we obtain the following equations when the attribute value is defined as a continuous value. We can then evaluate space-clusters using this equation as follows:

$$\mathrm{AIC}_0 = 2^{2(k+1)}\left(\log 2\pi + \log \hat{\sigma}^2 + 1\right) + 2(1+1), \qquad (5)$$

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

$$\mathrm{AIC}_1 = 2^{2(k+1)}\left(\log 2\pi + \log \hat{\sigma}^2 + 1\right) + 2(4+1) \qquad (6)$$

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

However, $\log \hat{\sigma}^2$ in equation (6) cannot be estimated at the time of k=0 (namely, in case of the smallest unit cell). Therefore, when the attribute value is defined as a continuous value, the sub-area formed of the four smallest unit cells can be the smallest space-cluster.

3.2 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 analysing the data (the number of cells is 64) using two methods is shown in the upper row of figure 4. As a result of adopting a “dividing” method, an optimum is reached when the whole study area is made into one space-cluster, i.e., the local minimum of AIC. We can confirm that the AIC_0 of whole area is smaller than the AIC_1 of divided sub-areas. 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, i.e., 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. That is, 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, the "unifying" method is adopted.

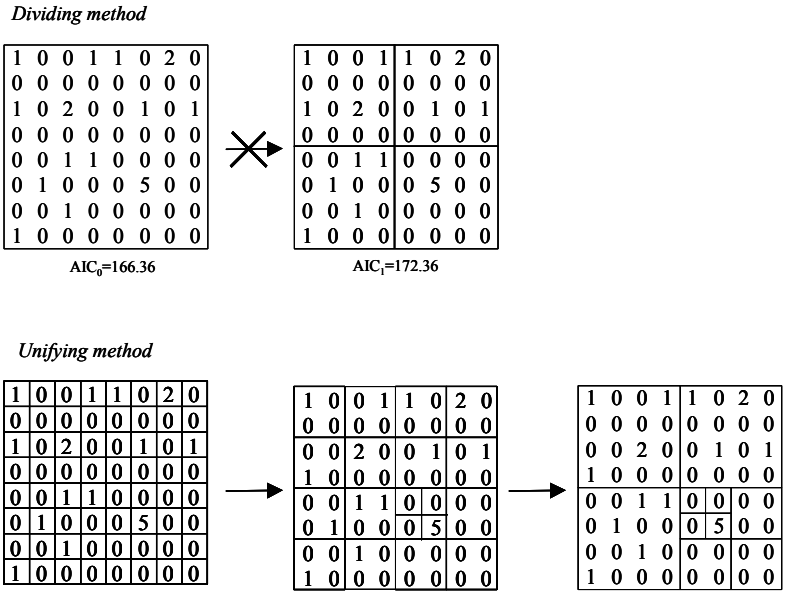


Figure 4: Comparison of dividing method and unifying method

3.3 Application to Actual Spatial Data

Based on the above consideration, the appropriate space-cluster is obtained using actual urban spatial data, and the result is shown in figure 5. The spatial data used here is as follows. (a) "*the number of merchants*", (b) "*the number of fishery workers*", (c) "*the ratio of female workers*", (d) "*nuclear family households*". The source of data is *Digital Mesh Statistics* "1991 *place-of-business statistics*" and "1990 *national-census*". The cell size is about 1 km by 1 km and the number of cells is 256. In addition, the distribution of the attribute values is shown for reference. The data is sorted from the largest to the smallest according to the attribute value. From figure 5, we can see the space-clusters are reduced by our proposed method. For instance, for the data whose attribute value is a discrete value, the cell with the outstanding value is expressed as the smallest space unit, and the other cells are unified into the larger space-clusters. Comparatively, if the attribute value itself has a large value, a small space-cluster is created (see "*the number of merchants*"). On the other hand, as for the data whose attribute value is a continuous value, the cell with the relatively large value is unified together with the other cells. That is, the smallest unit cell is unified with surrounding cells, and is represented as a larger space-cluster that is statistically meaningful.

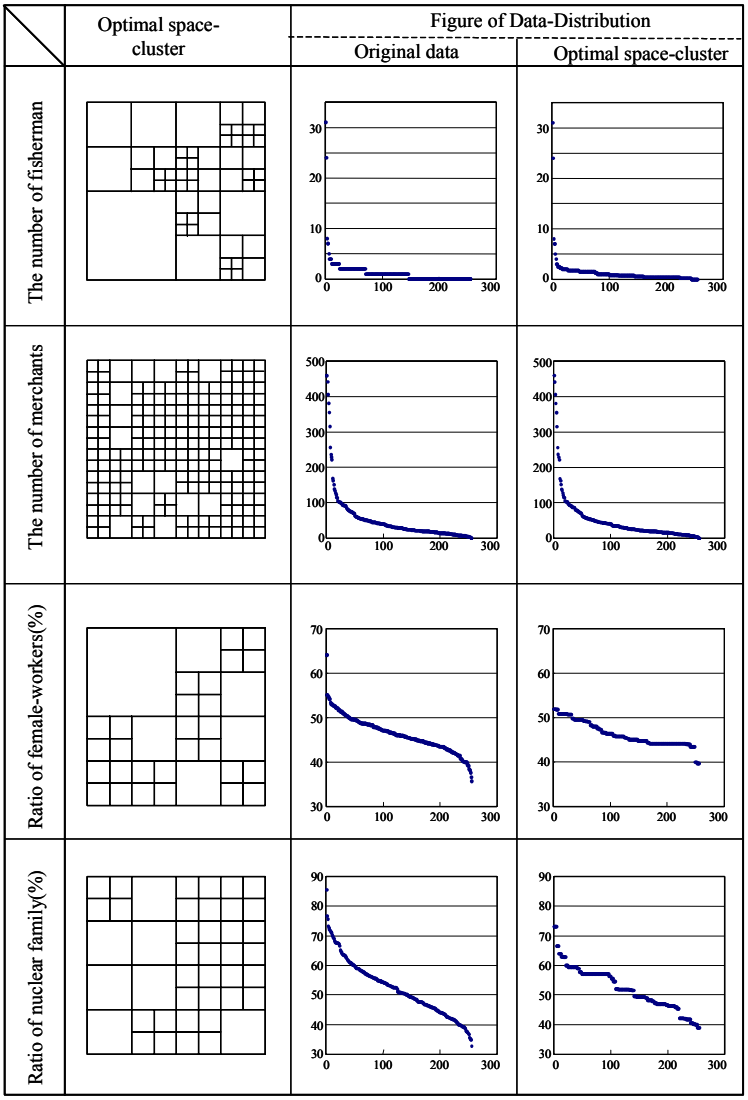


Figure 5: Optimal Space-cluster and Data Distribution

4. Visualization of Space-Cluster

Visualization of spatial data is carried out using the appropriate space-cluster examined in the previous section. The classification of the unified attribute value of space clusters is achieved using the information loss minimization method (Osaragi 2001). The outline of the concept of the information loss minimization method and the study area are shown in figure 6. The results of space-cluster visualization are shown in figure 7. The result in the case where the optimal space-cluster was not asked for is also shown simultaneously in figure 7. Moreover, the ratio (L) of

information loss defined in figure 6 is also shown. The figure 7 shows clearly that if we create the appropriate space-cluster, the space distribution characteristic of the original data can be grasped more easily. The correct results are obtained both in cases where the data is defined as a continuous value and as a discrete value. However, it is necessary to pay attention to the information loss increasing slightly.

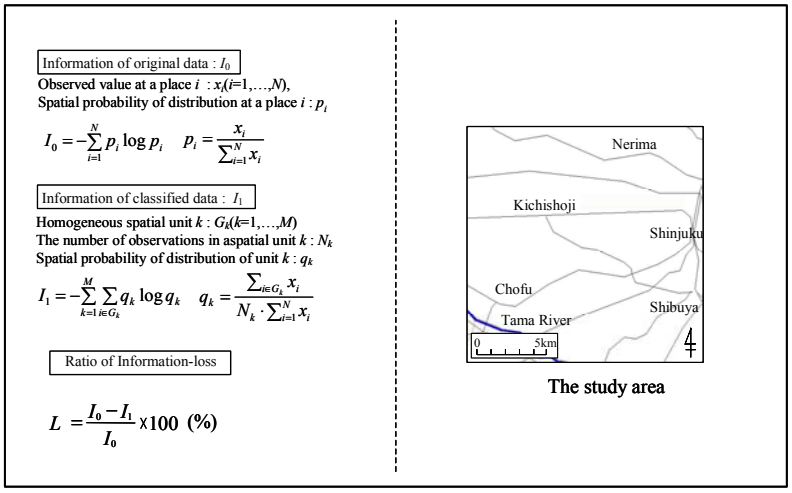


Figure 6: Study area and definition of ratio of information-loss

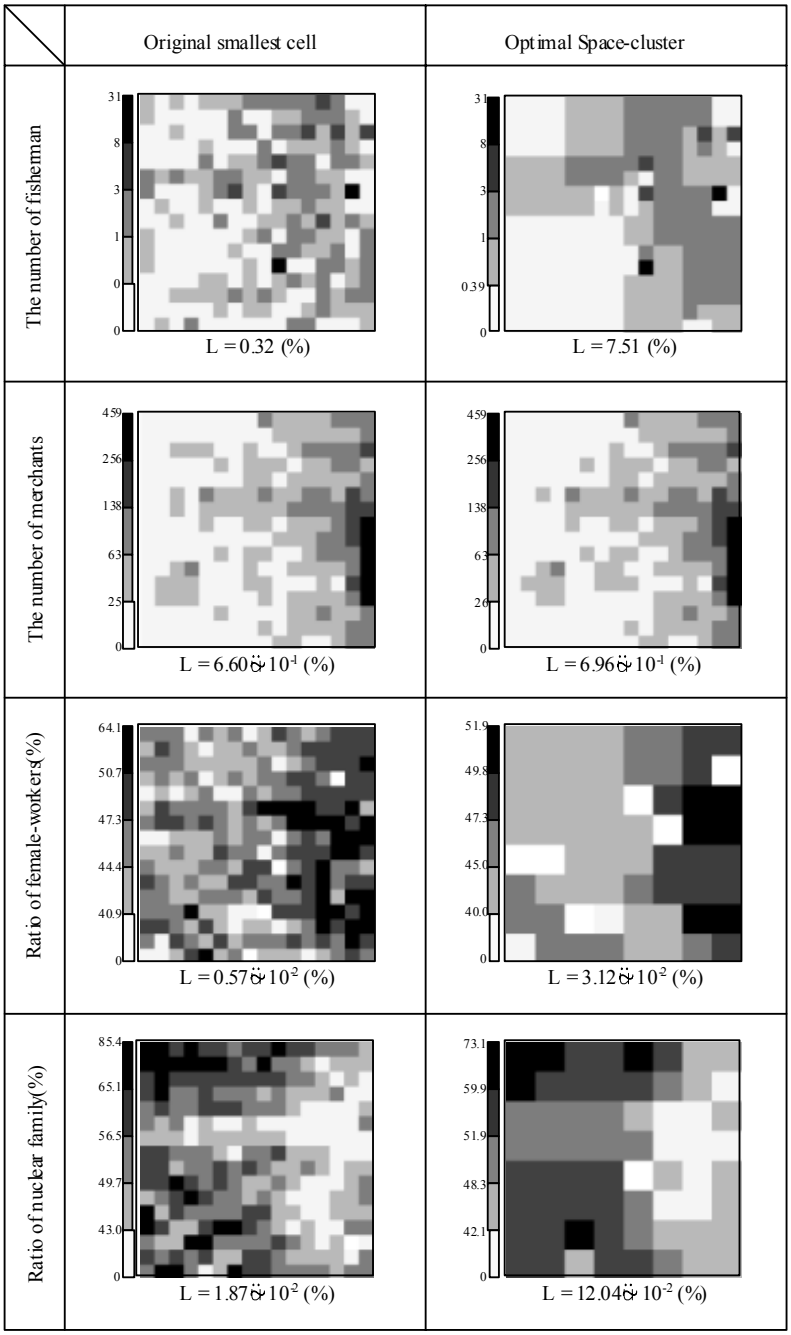


Figure 7: Visualization of homogeneous space-cluster using the classification method of minimizing information-loss

5. Summery and conclusions

The method of obtaining the space-cluster based on the evaluation function of AIC is proposed, with consideration to the distribution characteristic 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.

6. Acknowledgements

The author would like to express his thanks for the valuable comments from Mr. Paul Torrens, Centre for Advanced Spatial Analysis, University College London. Also, the author would like to give his special thanks to Mr. Hiroki Yamanaka, Graduate Student of Tokyo Institute of Technology, for computer-based numerical calculations.

References

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

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

Anselin L, 1995, "Local Indicator of Spatial Association - LISA", Geographical Analysis Vol.27, No.2, pp.93-115.

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

Batty M, 1976, "Entropy in Spatial Aggregation", Geographical Analysis, Vol.8, pp.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 (Martinus Nijhoff, Leiden, The Netherlands), pp.115-147.

Civco D L, 1993, "Artificial neural networks for land-cover classification and mapping", Int. J. Geographical Information Systems, Vol.7, No.2, pp173-186.

Fotheringham A S and Wong D W S. 1991, "The modifiable areal unit problem in multivariate statistical analysis", Environmen and Planning A, Vol.23, pp.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, pp.56-64.

Li-Xia, 1996, “A method to improve classification with shape information”, International Journal of Remote Sensing Vol.17, No.8, pp.1473-1481.

Liebetrau A M and Rothman E D, 1977, “A classification of spatial distributions based upon several cell sizes”, Geographical Analysis, Vol.9, pp.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, Vol.17, pp.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, 2001, “Classification method of spatial data and its information loss”, Working Paper at CASA, University College London.

Roy J R, Batten D F and Lesse P F, 1982, “Minimizing information loss in simple aggregation”, Environment and Planning A, Vol.14, pp.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 Vol.22, pp.229-234. (in Japanese)

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 Vol.23, pp.37-42. (in Japanese)

Wong D W S, Lasus H and Falk R F, 1999, “Exploring the variability of segregation index D with scale and zonal systems: an analysis of thirty US cities”, Environment and Planning A, Vol.32, pp.507-522.