UCL CENTRE FOR ADVANCED SPATIAL ANALYSIS

UCL

casa

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

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

Margules et al. (1985) tested four agglomerative hierarchical fusion strategies with the adjacency

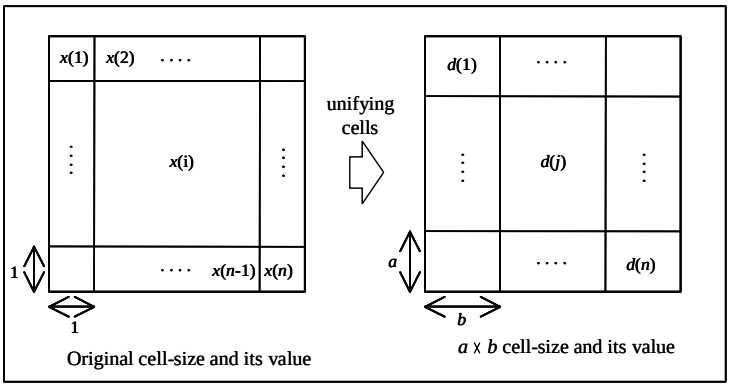


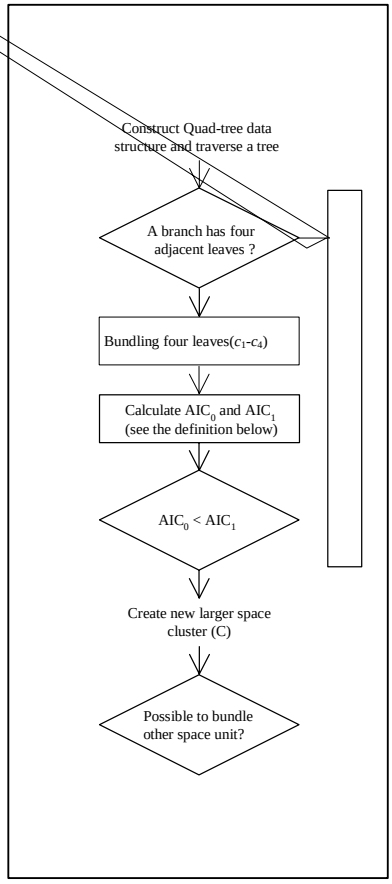
Figure 2: Cell-size and attribute values

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:

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

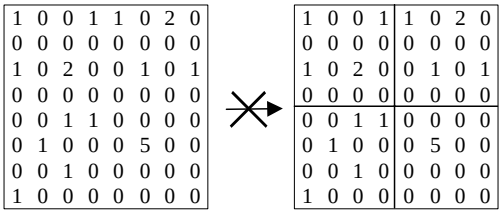
where $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} x(i)^2 - \sum_{j=1}^{N} \frac{d(j)^2}{ab}$.

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. However4620.fn [(on ant)7(ofnt)7(6er46206.1(a s


Construct Quad-tree data structure and traverse a tree
A branch has four adjacent leaves ?
Bundling four leaves(c_1-c_4)
Calculate AIC_0 and AIC_1 (see the definition below)
$AIC_0 < AIC_1$
Create new larger space cluster (C)
Possible to bundle other space unit?

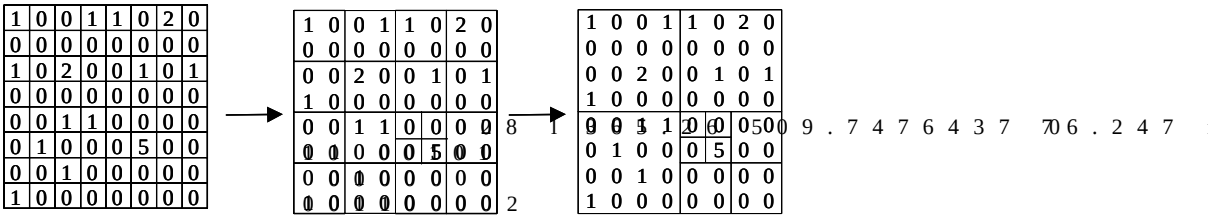
$$AIC_1 \quad 2 \sum_{1}^{4} \log \frac{}{2^2} + 2(4-1)$$

Dividing method

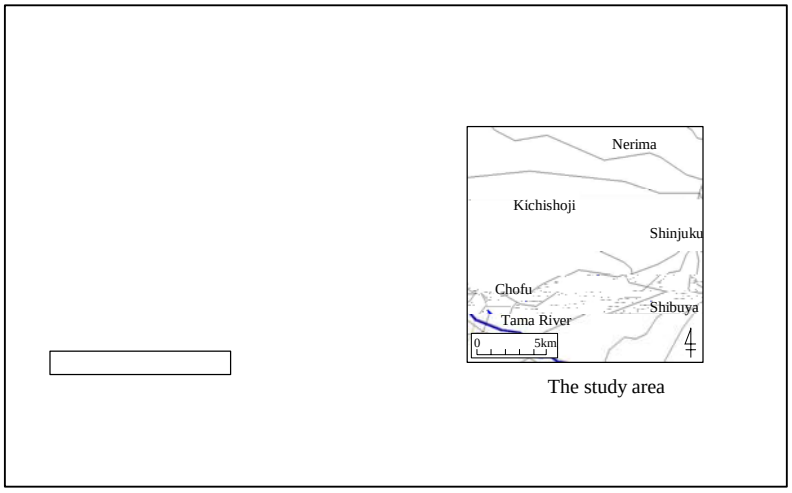


AIC_0=166.36 AIC_1=172.36

Unifying method



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.



The study area

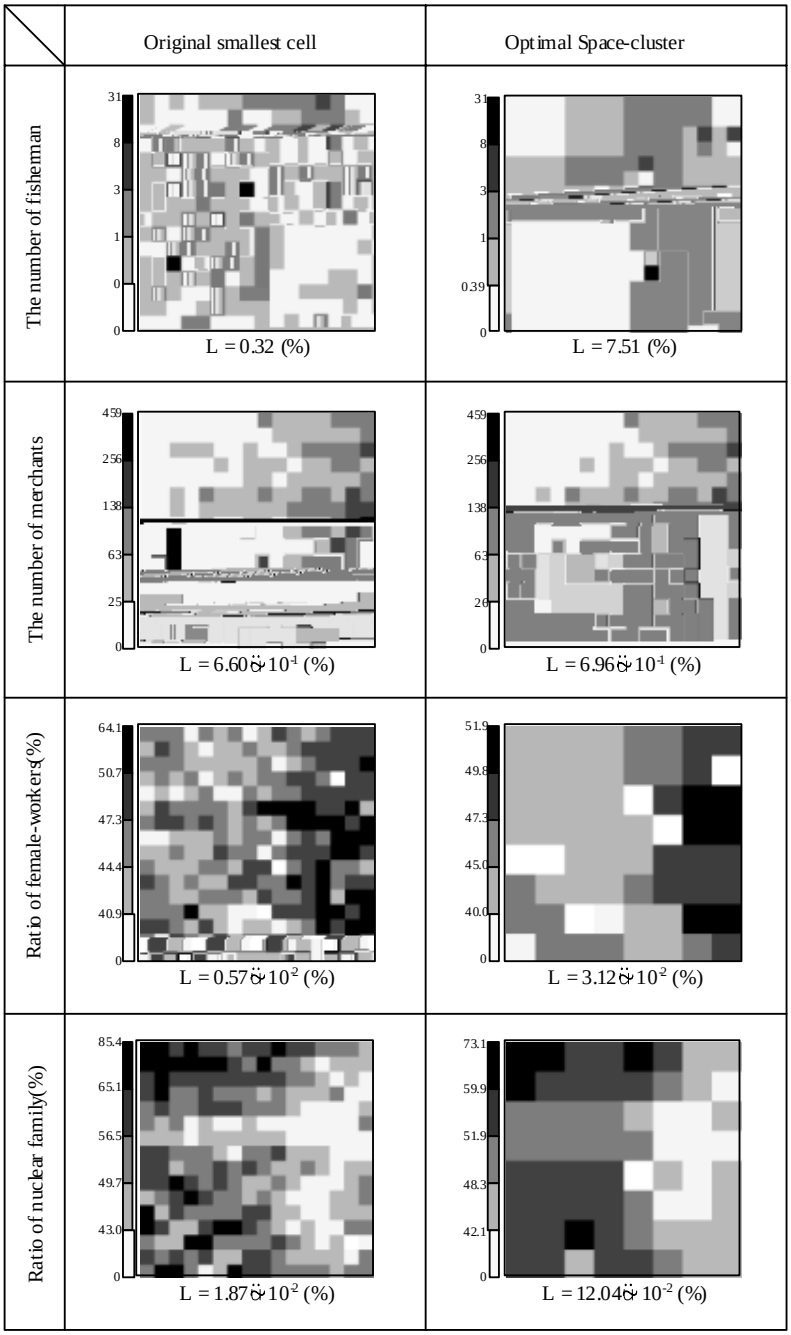


Figure 7: Visualization of homogeneous space-cluster using the classification method of minimizing information-loss

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 lsper3.5(i(sp)4ni4T)11.4i-g(“7ty)5

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.