CENTRE FOR ADVANCED SPATIAL ANALYSIS
Working Paper Series

Paper 40

CLASSIFICATION METHODS FOR SPATIAL DATA REPRESENTATION

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/paper40.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

Classification methods for spatial data representation

Toshihiro OSARAGI

Abstract: It is necessary to classify numerical values of spatial data when representing them on a map and visually understanding it. In consequence, loss of information from original data is inevitable in the process of this classification. A gate loss of information might lead to a misunderstanding of the nature of original data. In this study, a classification method of spatial data is proposed, in which the loss of information is minimized. Comparing our method with other existing classification methods, some new findings are shown.

Keywords: spatial data, visualization, classification, information loss, AIC (Akaike's Information Criterion)

1. Introduction

Our natural interpretation capabilities, originally endowed to human being, are excellent. Thematic maps are therefore very effective to understand spatial distribution of geographical features, since we can use our natural interpretation capabilities to understand colors, patterns, and spatial relevance. This fact simultaneously suggests the importance of expression methods, i.e., how to represent spatial data on a map. That is, according to how creation of the thematic map on a geographic information system is carried out, the characteristics of the original data might be overlooked, or there might be a risk of mistaking judgment about the characteristic which original data has.

The error problem in geographical information systems was briefly summarized by Goodchild et al. (1992). In order to estimate the uncertainty of a product, we have to discuss not only the uncertainty existing in the database, but also the propagation of uncertainty through the operations performed on the data by the systems. Focusing on the so-called 'area class map', an error model for categorical data was proposed (Goodchild et al. 1992). In the visualization process of the bi-dimensional spatial data defined quantitatively, it is necessary to classify the data (i.e., places) into several class-divisions. Namely, the places that have values in a certain range are classified into the same class and represented using the same display-colour. The aim of this paper is to discuss the uncertainty performed in the process of classifying numerical data.

Generally, if we employ a number of classes, the distribution-characteristic of original data can be expressed faithfully. However, if there are too many classes, its legends will become complicated and the map will be difficult to understand; we cannot distinguish delicate color differences. On the other hand, when we only have a few classes, the information such as small vibrating factors or local peaks might be ignored; namely, much information of original data will be lost. Hence, a classification problem has to be discussed from the following viewpoint.

(1) How many classes are necessary to represent the spatial data?

Automatic classification methods are being incorporated in existing geographic information systems. However, the characteristics of the original data might be overlooked, or there might be a risk of mistaking judgment, if we do not have enough knowledge about the classification method as well as the distribution characteristics of the original data. Even if we are using the same number of classes and the same spatial data, we might obtain the quite different maps. A typical example is shown in figure 6. Hence, the following viewpoint is also important for a classification problem.

(2) How the boundary value between each class should be set?

It is true that the classification method to be used depends on the nature of data, and what we want to show about the data. However, more flexible, simple and easy methods are necessary for the non-expert end-users of GIS. In this research, we discuss this primitive and fundamental problem. Hence, the existing classification methods are examined from the viewpoint of information statistics, and we attempt to propose a new classification method for the visualization of spatial data.

As for the former question (1), Umesh (1988) developed an algorithm for achieving efficient classification of data with no a-priori information available about the number of groups. From this a performance index was defined so that minimizing it results in appropriate clustering of the given data.

As for the latter question (2), various methods have been considered and already built in existing geographical information systems. For instance, we have options called *Natural-Breaks*, *Quantile*, *Equal-Area*, *Equal-Interval*, *Standard-Deviation* in a popular GIS software, Arc View. According to the *Natural-Brakes*, the so-called "*Jenks' optimization method*" is employed and realized (Jenks, 1967). This is the method of determining boundary value so that the average of a squared deviation in each class will be minimized. The capabilities of each method will be shown through

application to actual spatial data in section 4.

Considering the above two fundamental questions, one study that consideres about the cell-size of raster data should be cited. Tamagawa (1987) has analyzed point sampling data by changing the observed spatial-range variously, and proposed a method based on AIC (An Information Criterion) by Akaike (1972 and 1974) to obtain the optimum cell-size. The AIC is a synthetic-measurement considering "*model's fitness*" and "*model's simplicity*". This idea is employed to our classification problem.

In this study, the classification method using the evaluation function based on AIC is examined first, and is applied to actual spatial data. Next, based on the consideration about its result, a new classification method based on the minimization of information loss will be proposed. Furthermore, verification of this method is achieved through the examination comparing it with the existing classification methods.

2. Classification Method based on AIC

2.1 Formulation of Classification Method

2.1.1 Discrete variables

Firstly, the spatial data such as point sampling data is discussed. Namely, we focus on the spatial data obtained by counting an attribute value within a certain spatial unit.

First, we denote the variable x_i (i= 1, 2, .., n) as an observed value within a certain spatial unit. It is assumed that its value is obtained by distributing the total number of observation within the whole objective-space, denoted by $X(=\sum_{i=1}^{n} x_i)$, into n space units. That is, the multinomial probability distribution is assumed here. Next, if the objects are distributed into some spatial units according to the same distributing probability q_k (k= 1, 2, .., m), these spatial units should be classified into the same class, denoted by G_k. Therefore, under the condition that the values x_i (i= 1, 2, .., n) are observed, the logarithm of Maximum Likelihood Estimates can be written as follows:

$$\ln L(q_k) = C + \sum_{k=1}^{m} \sum_{i \in Gk} x_i \log q_k , \qquad (1)$$

where C is a constant value. Distribution probability q_k has a constraint of $\sum_{k=1}^{m} N_k q_k = 1$,

where N_k is the number of spatial units in a class of G_k. The Maximum Likelihood Estimates of q_k can be estimated, by using the Lagrange's method of undetermined multiplier considering the constraint of q_k, as follows:

$$\hat{q}_k = \frac{\sum_{i \in Gk} x_i}{X\, N_k}, \tag{2}$$

The number of free parameters q_k of this model is m-1. Then, the value of AIC is given by the following equation (Akaike 1972 and 1974), when we classify the original data into m classes:

$$\begin{aligned} \text{AIC} &= -2\,[\textit{Maximum Likelihood}] + 2\,[\textit{the number of free parameters}] \\ &= -2\sum_{k=1}^{m}\sum_{i \in Gk} x_i \log \hat{q}_k + 2(m-1), \end{aligned} \tag{3}$$

where the constant term is omitted.

2.1.2 Continuous variables

Spatial data composed of continuous variables can be discussed in the same way. First, the observation value at each place i (i= 1, 2, .., n) is denoted as x_i. Next, the parameter common to all observation value is denoted as θ_0, and a parameter peculiar to a observation value is expressed by θ_k $(k = 1,2,\cdots,m,\ \sum_{k=1}^{m} \theta_k = 0)$.

The number of places which have the same parameter θ_k is denoted as N_k $(\sum_{k=1}^{m} N_k = n)$, and it is assumed that these places are contained in the same class G_k. That is, the observation value included in class G_k is assumed to follow the normal distribution $(\theta_0 + \theta_k, \sigma^2)$. Then, the logarithm likelihood of observed data can be written as follows:

$$\ln L\{\theta_0, \theta_k\} = C - \frac{n}{2}\ln 2\pi - n \ln \sigma - \frac{\sum_{k=1}^{m}\sum_{i \in Gk}[x_i - \theta_0 - \theta_k]^2}{2\sigma^2}, \tag{4}$$

where, C is a constant.

Next, the maximum likelihood estimator which makes equation (4) the maximum can be estimated

as follows, if the undetermined multiplier method of Lagrange is used considering the constraint conditions about θ_0 and θ_k:

$$\hat{\theta}_0 = \frac{X - \sum_{k=1}^{m} \theta_k N_k}{n}, \tag{5}$$

$$\hat{\theta}_k = \frac{\sum_{i \in Gk} x_i}{N_k} - \hat{\theta}_0, \tag{6}$$

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{k=1}^{m} \sum_{i \in Gk} \left[x_i - \frac{\sum_{i \in Gk} x_i}{N_k} \right]^2. \tag{7}$$

Since there is constraint condition about θ_k, the number of free parameters of a model is (m+1) in all. That is, the value of AIC when classifying the data into n classes is given by the following formula, where the constant term is omitted:

$$\mathrm{AIC} = n \ln \hat{\sigma}^2 + 2(m+1). \tag{8}$$

2.2 Method of boundary setting for classification

We can evaluate each model by comparing values of AIC, which is a synthetic-measurement considering "*model's fitness*" and "*model's simplicity*". This is the principal difference between the method we propose and the method proposed by Umesh (1988). Hence, the optimum classification under the condition of the data currently observed can be achieved by minimizing the value of AIC given in the equation (3). Namely, optimum classification can be achieved by the following two steps:

(i) Fix the number of classes m, and search the boundary value of optimum classification, which minimize the value of AIC.

(ii) Repeat the above process by changing the number of classes m one by one, and search the number of classes m that gives the minimal value of AIC.

Although the step (ii) seems to be simple and easy, the step (i) is not so clear. In the following, we discuss in detail how we can achieve the step (i).

The problem of setting the boundary values of classification will become equivalent to the problem of setting the ranks of boundary values, if the data is sorted in the order of its attribute values (see figure 1). The rank of boundary value is hereafter called "boundary-rank", and a possible procedure of setting the boundary-rank is shown in figure 2. In the following discussion, we consider the condition that the number of class *m* is fixed for simplicity. The procedure given by figure 2 is described as the following four processes:

(1) Make a group of the spatial units, if their values show a tie.

(2) Set up the initial boundary-rank of (*m*-1) classes, and consider them as initial values.

(3) Calculate a value of AIC by moving a boundary-rank up or down, and set the boundary-rank so that the value of AIC becomes minimum.

(4) Repeat the above operation until the value of AIC does not decrease.

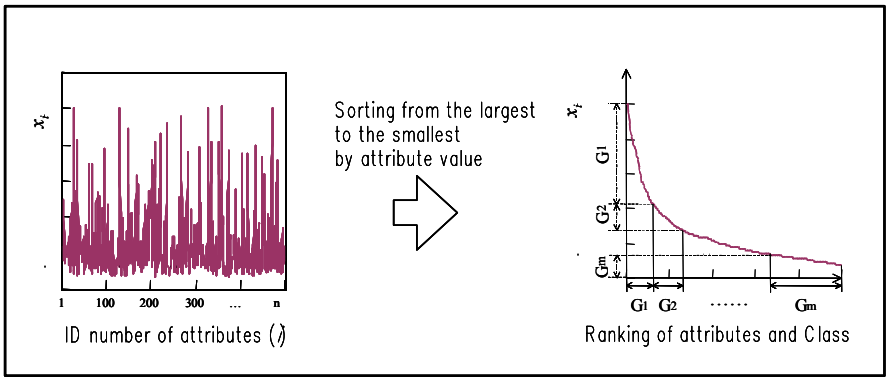

Figure 1: Basic concept of data classification

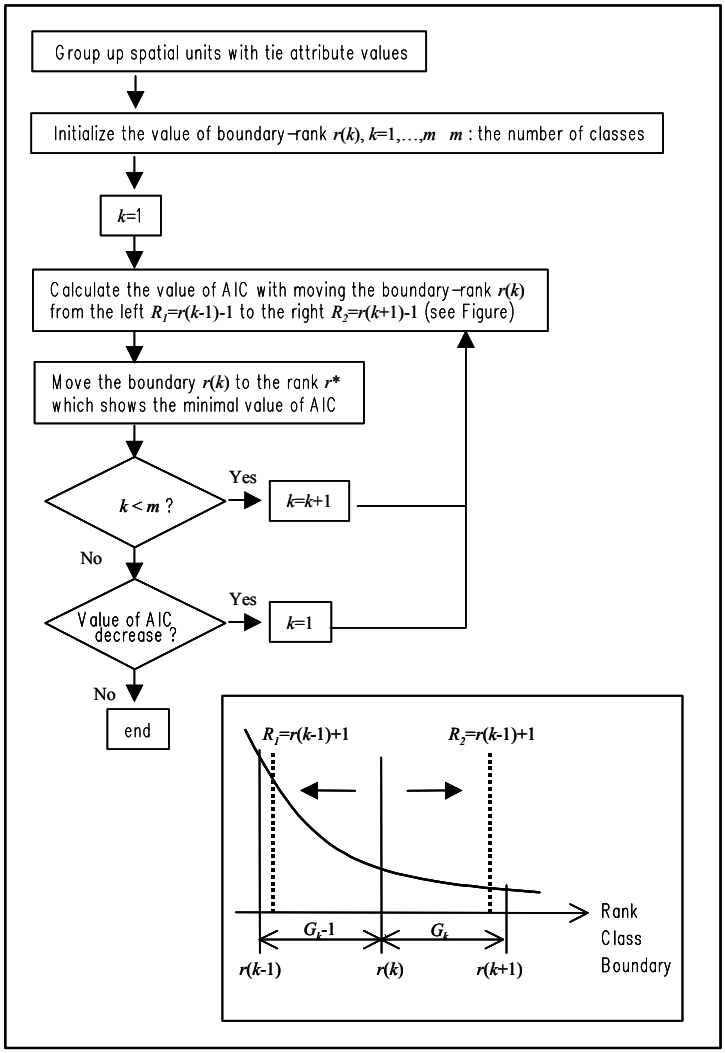

Figure 2: Algorithm for minimizing the value of AIC

The above method can correspond flexibly in such a case. For instance, it can be achieved by setting the original data into a round number at the end, in advance of the process (1), to obtain the boundary value with a round number at the end. However, this procedure might lead us to a risk that the value of AIC is a local minimum. Therefore, it is more desirable to introduce the probabilistic method in spite of the above deterministic method, in the process (3) setting of boundary-rank. Namely, set the boundary-rank according to the size of probability calculated from the value of AIC. The technique called "*annealing*" of the Neural Network Theory can be utilized here (see figure 3).

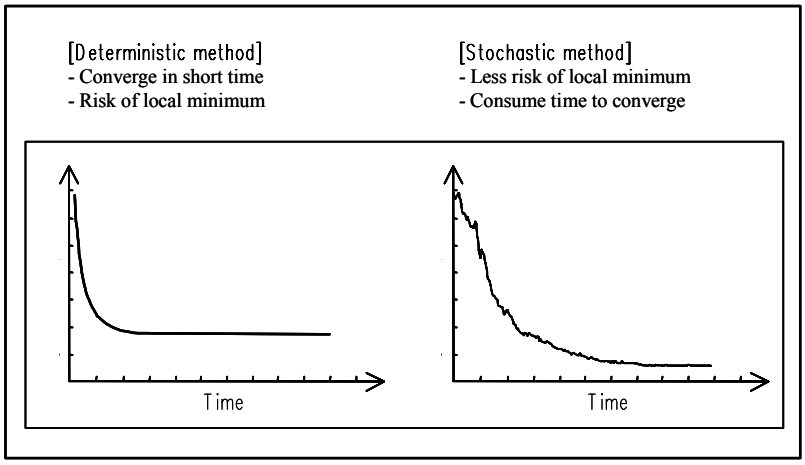

Figure 3: Comparison of deterministic method and stochastic method in the minimization process of AIC

According to our experiments using actual data, we can obtain quasi-optimum boundary-rank that gives almost the minimum value of AIC, if the process of minimization is performed several times by changing the initial boundary-rank. The above method (the deterministic method) does not produce any major practically problem, from the author's experience.

2.3 Application to Actual Data

The above method is first applied to the raster data, *Digital Mesh Statistic* compiled by *Statistic Bureau & Statistic Center of Japan*, and the result is shown in figure 4-a. The data source is "*agriculture/forestry/fishery worker households, 1989 Population Census*". The cell-size is about 1 km by 1 km, and the number of cells is 100. Figure 4-a shows us that the value of AIC is minimum when the number of classes is seven. Hence, we can say that optimum classification has been achieved from a statistical viewpoint.

Next, this method is applied to the same kind of raster data as "*the number of companies, 1994 Establishment and Enterprise Census*". The cell-size is about 500 m by 500 m, and the number of cells is 3,220. The result is shown in figure 4-b. Since many samples are observed in this case, the likelihood of model (equation 3) becomes more dominant than the model's degree of freedom. Therefore, the optimum number of classes tends to become very large. Even if we classify the data into too many classes and represent it, the small degree of color-difference in a map cannot be distinguishable, and the legends may be complicated. It is also difficult to give significant meaning to the optimum number of classes, since we cannot find any big difference in

the values of AIC. Furthermore, much time is required to calculate an optimum solution when many classes are considered. Therefore, we cannot call it a realistic classification method.

Considering the above discussion, it may be more realistic to examine only step (ii) described in section 2.2, that is, to examine how we should set up the suitable boundary-rank by fixing the number of classes. As for step (i), GIS users should set up a-priori the number of classes to be employed according to their demands.

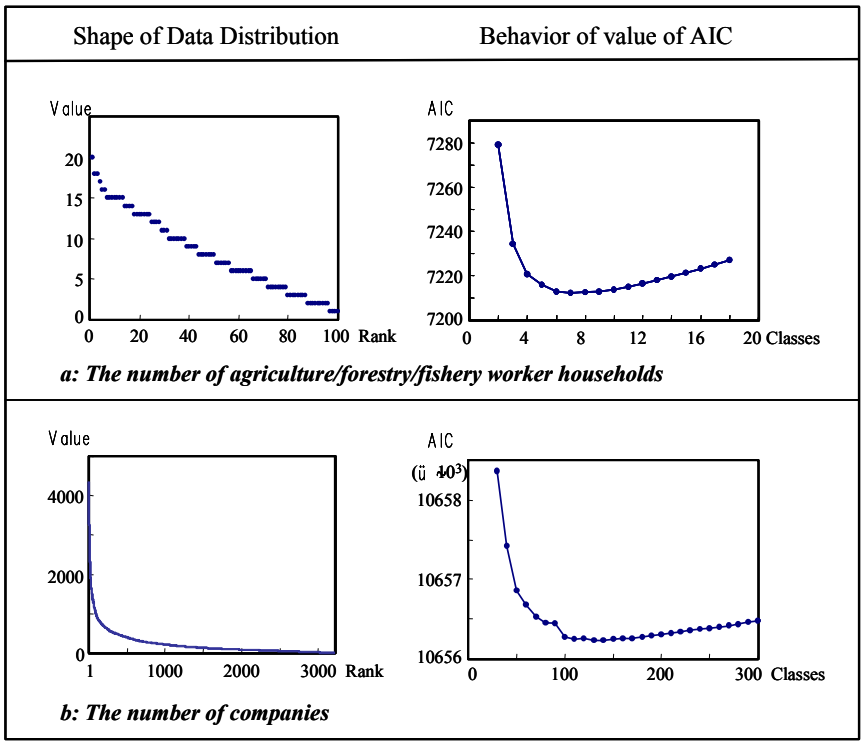

Figure 4: Shapes of data distribution and behavior of the vale of AIC in minimization process

3. Classification Method based on Minimization of Information Loss

When the different attribute values are classified into the same class, a part of the inherent information contained in original data must be lost. A classification method which keeps the loss of inherent information as low as possible can be considered an effective method in a sense of lessening the error of judgment. Roy et al. (1982) have presented a criterion to indicate a strategy for cumulatively combining corresponding discrete rows and columns of a square array to form an array of reduced size so that the loss of information thereby incurred is minimized. We employ

their strategy to our classification problem. In the following, we define the information loss caused by classification first.

We define the averaged information, denoted by I_0, for spatial data obtained by counting an attribute value within a certain spatial unit as follows:

$$I_0 = \sum_i p_i \log p_i \,, \tag{9}$$

where $p_i = \dfrac{x_i}{X}$. This equation shows the amount of information when any classification is not carried out. That is, equation (9) gives the total information contained in original data, when each value is classified into individual n classes. On the other hand, the averaged information, denoted by I, can be described as follows, when the data is classified into m classes.

$$I = \sum_{k=1}^{m} \sum_{i \in Gk} q_k \log q_k \qquad , \tag{10}$$

where $\hat{q}_k$ is given by $\sum_{i \in Gk} x_i \Big/ (X\, N_k)$.

The above-mentioned concept about the averaged information of discrete attribute variables can be naturally extended to the case of continuous variables. Denote x as the random variable of continuous data sources, and $p(x)$ as its density function. The averaged information I_0 of continuation sources of information can be defined as follows (Minami, 1995).

$$I_0 = -\int_{-\infty}^{\infty} p(x) \log_2 p(x) dx \tag{11}$$

However, actual spatial data are not necessarily obtained in a continuous form, but only the values aggregated in a certain space range are acquired in many cases. Then, in actual calculation, it is considered as follows.

First, the aggregated value in a certain space range Δx_i can be expressed by $p(x_i)\Delta x_i$ (i= 1, 2, ..., n). The averaged information I_0 can be defined by the following equation, since the integration range of equation (11) is equivalent to the whole space.

$$I_0 = -\sum_{i=1}^{n} p(x_i)\Delta x_i \log_2 p(x_i)\Delta x_i \tag{12}$$

Here, the value of $p(x_i)\Delta x_i$ is given by $x_i \Big/ \sum_{k=1}^{n} x_k$.

Consider the case that the whole space is classified into *m* class divisions. The averaged information *I* of this case can be defined by the following equation:

$$I = -\sum_{i=1}^{n} \sum_{i \in Gk} q(x_k)\Delta x_i \log_2 q(x_k)\Delta x_i \qquad (13)$$

The following statistical value calculated using equations (9), (10) or equations (12), (13), is defined as "the ratio of information loss", denoted by *L,* as follows:

$$L = \frac{I_0 - I}{I_0} \times 100\,(\%). \qquad (14)$$

If the value estimated by equation (14) is small enough, we can accept this classification from the viewpoint of information loss. Comparing equation (3) with equation (10) or (13), it turns out that the evaluation measurement of equation (14) is equivalent to the definition of AIC in the case where the number of model's parameters is not taken into consideration.

The value of AIC is, basically, the relative index used in order to judge the superiority or inferiority of models. Hence, there is no absolute meaning in the value of AIC itself. On the other hand, the ratio *L* of information loss in equation (14) is an index showing how much information of original data is lost. Therefore, this index *L* can help us as a reference, when we understand a map drawn using the classified data.

The classification method based on minimization of information loss can be achieved using the equation (14) as an evaluation index. Namely, the optimum classification should be performed according to the procedures described at "2.2 Method of boundary setting for classification ".

4. Comparison of Classification Methods

The above method will be verified through some comparisons with existing classification methods. Before the examinations, the features and capability of existing methods are reviewed briefly. The *Natural Breaks* classification method identifies breakpoints by looking for groups and patterns inherent in the data. One of the most popular GIS software programs, Arc View, employs Jenks'

optimization in this method, which minimizes the variation within each class (Jenks 1967). The features of the data are divided into classes whose boundaries are set where there are relatively big jumps in the values. As for the *Quantile* method, each class is assigned the same number of features. This may be misleading because low values are often included in the same class as high values. However, it is the best suited for the data that is linearly distributed, namely, data that does not have disproportionate numbers of features with similar values. In addition, it is suitable when we want to emphasize the relative position of a feature among other features. The *Equal Area* method classifies polygon features so that the total area of polygons in each class is approximately the same. The *Equal Interval* method divides the range of attribute values into equal sized sub-range. It is useful when we want to emphasize the amount of an attribute value relative to the other values, and ideal for data whose range is already familiar, such as percentage or temperature. Finally, the *Standard Deviation* method, it shows us the extent to which an attribute's values diverge from the mean of all the values (ESRI, 1996).

In order to investigate the characteristics of existing classification methods and our method, we have attempted to apply each method to the variety types of actual spatial data, i.e., seven different sets of data from *Digital Mesh Statistic* compiled by *Statistic Bureau & Statistic Center of Japan*. Each data is classified into nine classes using five existing classification methods mentioned above and our method based on the minimization of information loss. The ratio L of information loss by each method is shown in Table 1. Hatches are attached to the smallest or second smallest results among the existing classification methods. In order to grasp the characteristics of original data, the data is sorted from the largest to the smallest by the size of the attribute values, and the shape of its distribution is shown in figure 5. The ratio L of each classification method is also shown in figure 5 correspondingly. Furthermore, examples of spatial data representation are shown in figure 6 with the values of the ratio L.

Table 1: The ratio L of Information-loss: comparison of existing classification methods and a method based on minimization of information-loss

Classification Methods / Spatial Data	Quantile	Equal Area	Equal Interval	Standard Deviation	Natural Breaks	Minimized Information-loss
Industry	0.986	0.986	2.615	0.520	0.356	0.210
*	0.975	0.975	2.653	0.505	0.349	0.215
Company	3.080	3.080	6.323	1.899	0.583	0.332
*	2.945	2.945	6.322	1.866	0.584	0.335
Factory	1.194	1.194	2.353	0.708	0.465	0.209
*	1.315	1.315	2.393	0.515	0.497	0.242
Private Shop (ü 10^{-1})	0.759	0.759	3.100	0.513	0.373	0.219
*	0.792	0.792	3.105	0.507	0.390	0.224
Shops	1.376	1.376	5.461	1.407	0.553	0.280
*	1.323	1.323	5.427	0.915	0.589	0.283
Population (Tokyo ü 10^{-1}) **	1.052	1.052	3.282	5.851	1.134	0.837
	1.058	1.058	3.241	5.875	1.129	0.840
Population (Yoko-hama ü 10^{-1}) *	0.639	0.639	1.233	1.360	0.932	0.562
	0.645	0.645	1.275	1.415	0.922	0.563

the smallest Information-loss

the second smallest Information-loss

Upper / Lower Normal classification

* Round down the first figure / ** Round down the second figure

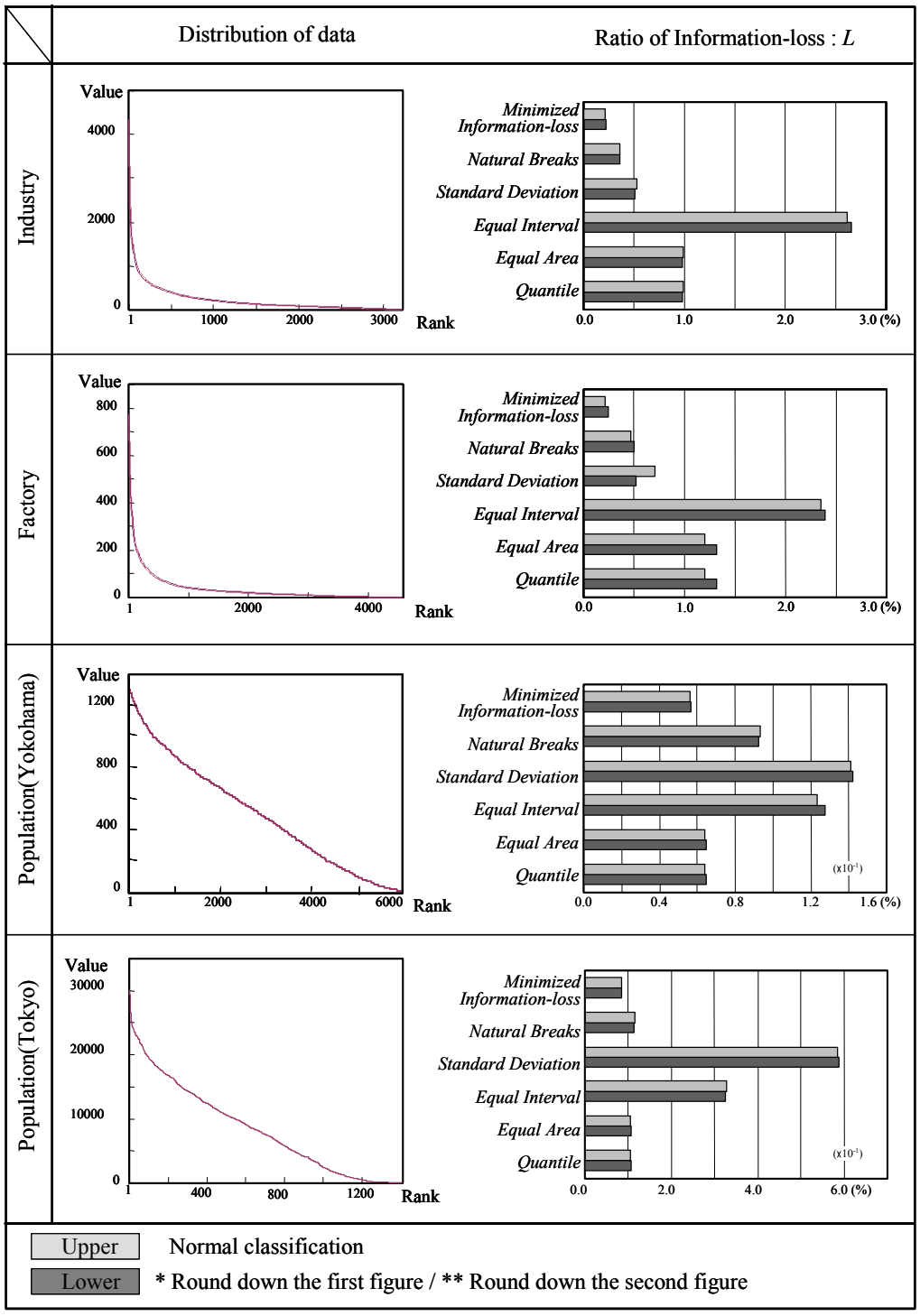

Figure 5: Shapes of data distribution and ration of information-loss

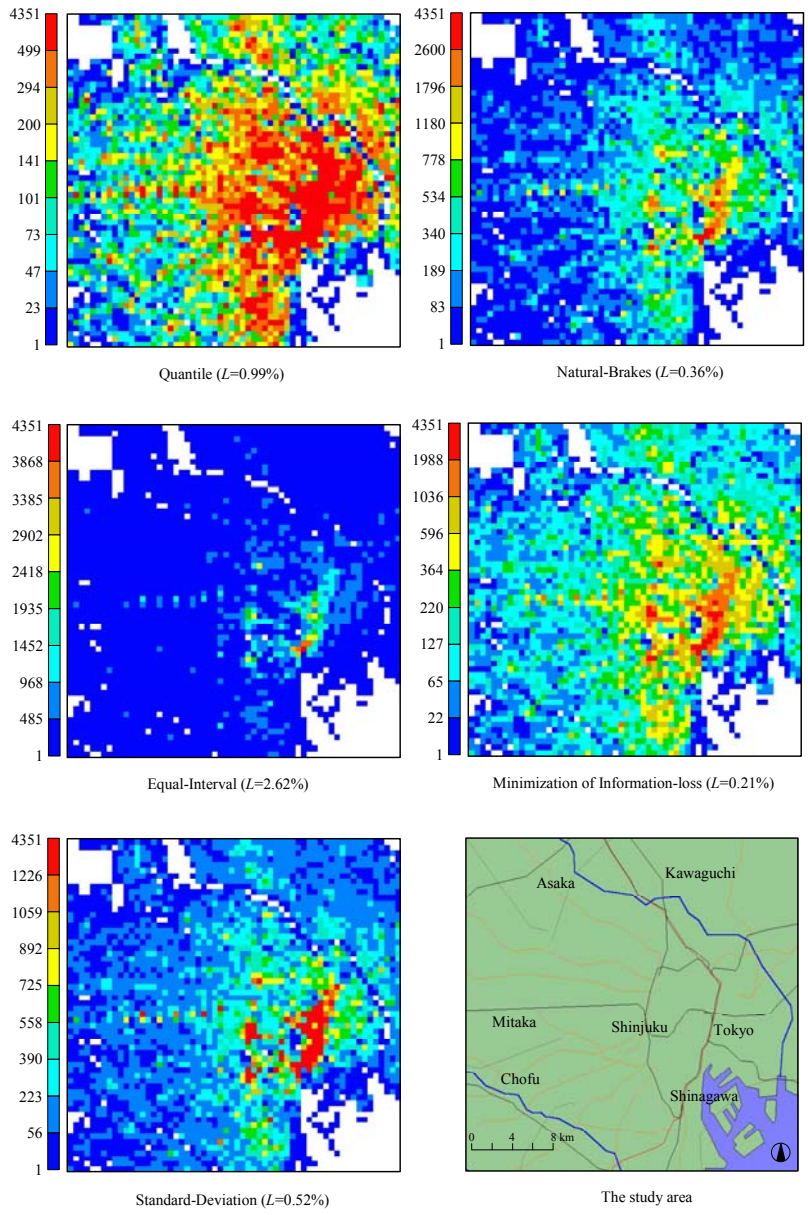

Figure 6: Visualization of spatial data by existing classification methods and minimization method of information-loss: the number of classes is 9, the number of cells is 3,220

Table 1 and figure 5 show us that the ratio *L* of information loss deeply depends on the feature of original data. For example, the *Natural-Breaks* classification method is effective for the data that has clear breakpoints such as the data "*the number of companies*". The *Natural-Breaks* can be acceptable to the variety types of data, since its information loss is comparatively smaller than that of the other methods. However, it should be noted that the *Natural-Breaks* is unsuitable for the data such as "*Yokohama population*" which has an unclear breakpoint. As for the *Standard-Deviation* classification method, it appears out that there can be a risk of being accompanied by a large amount of information loss. Furthermore, although it is easy to understand the legends in the *Regular-Intervals* method, due to the regular intervals in the range of boundary values, it must be recognized that there is a tendency to lose large amounts of information. Moreover, the *Quantile* method can be excellent for data that shows linear distribution, such as the data "*Tokyo population*", since its information loss is suppressed. However, in the case of data such as "*the number of factories*", the *Quantile* shows quite large information loss.

Considering the above discussion, it is necessary to examine the distribution characteristics of spatial data, in order to determine which classification method should be employed. If this process is neglected, we might have a risk of overlooking the nature of the original data. However, according to the classification method based on the minimization of information loss, we can correspond with flexibility in regards to any spatial data.

5.Summary and conclusions

The classification method based on AIC is proposed in order to grasp the nature of spatial data from the viewpoint of statistical meaning. However, this method is not effective for data with a large number of observations. Then, we proposed another classification method based on the minimization of information loss. This method is examined through the application to actual spatial data and comparing it with the existing classification methods. The results of numerical analysis show the flexibility and validity of the proposed method. However, further considerations regarding the efficient algorithm for minimizing information loss should be discussed.

6. Acknowledgements

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 and drawing figures.

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

ESRI, 1996, “ArcView GIS – The Geographic Information System for Everyone”, Environmental Systems Research Institute, USA

Goodchild M F, Guoqing S and Shiren Y, 1992, “Development and test of error model for categorical data”, Int. J. Geographical Information Systems, Vol.6, No.2 pp87-104.

Jenks G F, 1967, “The Data Model Concept in Statistical Mapping" International Yearbook of Cartography, Vol.7, pp.186-190

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)

Umesh R M, 1988, “A technique for cluster formation”, Pattern Recognition Vol.21, No.4, pp.393-400.