# EVALUATION METHOD FOR INFORMATION CONTENT OF RASTER DATA USING FRACTAL DIMENSION

Toshihiro Osaragi[1]*

[1] School of Environment and Society, Tokyo Institute of Technology – osaragi.t.aa@m.titech.ac.jp

**Commission IV, WG IV/3**

**KEY WORDS:** raster data, geographic feature, information theory, fractal dimension, cell size, noisy channel.

## ABSTRACT:

Raster data are obtained by dividing an entire study area into a regular grid of cells. Hence, raster data information depends on cell size as well as the shapes of geographical features to be rasterized. In this paper, we report on the construction of a method for evaluating the information content obtained from raster data using information theory. We also provide some numerical examples showing that the resulting information content can be expressed as a logistic function of cell size and that its parameters can be written as simple linear models with the fractal dimensions of the geographic features to be rasterized.

## 1. INTRODUCTION

Real-world geospatial variations are infinitely complex. The closer we look, the more detail we can see, almost without limit, which means that we would need to use an infinitely large database to precisely capture the real world. Since this would be impossible, data must somehow be reduced to finite and manageable quantities by generalization or abstraction processes. As a result, geospatial variations must be represented in terms of discrete elements or objects, and the rules used to convert real geographical variations into discrete objects are called "data models." In their study, Tsichritzis and Lochovsky (1977) defined a data model as "a set of guidelines for the representation of the logical organization of the data in a database consisting of named logical units of data and the relationships between them."

Raster data constitute one of the data models used to present real geospatial variations by dividing the study space into isotropic and isochoric cells, in which each of the attribute information types (land-use, etc.) and numerical values (population density, etc.) are represented by one value. However, if the cell size (resolution or the side length of a cell) is large, detailed information related to geospatial variations will be lost, thus carrying ambiguity into the dataset. In contrast, if the cell size is too small, the data volume becomes too large and the labor and economic cost required to process the data grows exponentially, thus making the dataset practically useless. Given these considerations, it is important to analyze the relationships between cell size and the information content obtained from raster data (Jones, 1979). There are case studies analyzing it by counting the number of cells in which an error occurred, however, limited to a simple analysis on the relationship between the accuracy of raster data and the cell size (Liao and Bai, 2010; Bau et al., 2011). Osaragi (2015) examined the amount of information loss when visualizing spatial data, and proposed a method to minimize the amount of information loss based on information theory.

In this paper, we report on the development of a method to quantitatively evaluate how much geospatial information can be obtained from raster data. We also discuss how the quantity of obtained information relates to cell size and the shape complexity of geographic features.

## 2. INFORMATION CONTENT OBTAINED FROM RASTER DATA

### 2.1 Creation of raster data and the identification errors

To begin, let's suppose that all locations in a real geographical space (referred to hereafter as real space) are categorized by following a land-use type classification standard. The categories are denoted as $A=\{a_1, \ldots, a_n\}$ in the following paragraphs.

Next, imagine that real space is represented by raster data, and $B=\{b_1, \ldots, b_n\}$ are categories in the raster dataset that have been created to firmly distinguish them from the real space categories. If a location categorized as $a_1$ is rasterized as $b_1$, and another location categorized as $a_2$ is rasterized as $b_2$ and this process continues through the entire dataset, the real spaced information is presented consistently in the raster data. However, in practice, the correspondence of category sets $A$ and $B$ is not always consistent throughout the entire study area.

The probability that such inconsistent classifications will occur depends on the raster data cell size. Thus, an inconsistent classification (identification error) between the real space and the created raster data is presented in the form of a cross table, as shown in Figure 1. Denoting the cell size as $c$, $s_{ij}(c)$ in the cross table is the area whose category is $a_i$ in real space but is represented as $b_j$ in the raster data. If the cell size $c$ is infinitely small, the values of non-diagonal elements will be zero, the diagonal elements will be non-zero in the cross table, and the values of non-diagonal elements will increase as the cell size gets larger.

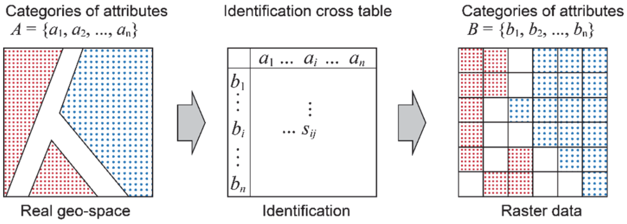


**Figure 1**. Identification error occurrence in the raster data creation process.

### 2.2 Formularization of information content

The basic concept of a "noisy channel" is well known in information theory (Figure 2) (Shannon, 1948; Brill et al., 2000). As an example, consider a case in which a string of symbols, $\{x_1, \ldots, x_n\}$, from the information source is transmitted to the receiver as $\{y_1, \ldots, y_n\}$ through a noisy channel. If there is no noise, the two strings will be the same. However, differences will arise if noises exist.

When considering real space as the information source, the raster dataset can be regarded as the received information transmitted through the channel. In such an analogy, the raster data identification process can be considered a noisy channel. Taken further, if cell size $c$ is small enough (high raster data resolution), the identification process is comparable to a channel with just a little noise. In contrast, if $c$ is large (low resolution), multiple categories will be packed into one cell and the process is similar to a channel with heavy noise.

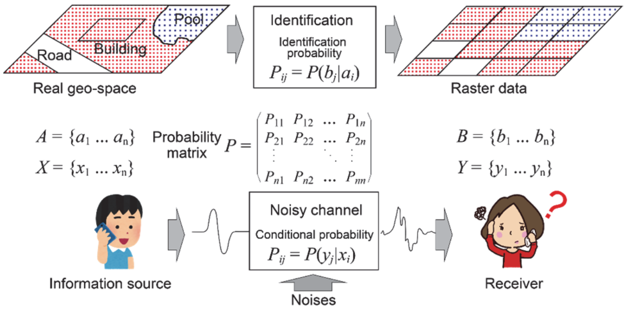


**Figure 2**. Noisy channel model in the raster data creation process.

In information theory, when an information source sends a symbol $x_i$, the conditional probability that $y_j$ is received, $P(y_j|x_i)$, is called the forward probability. Since a matrix with forward probabilities is called a channel matrix, we call it the rasterizing matrix in the case of raster data. Elements in a rasterizing matrix with cell size $c$, $P_c(b_j|a_i)$, can be calculated from the cross table shown in Figure 1.

In terms of information theory knowledge, when the conditional probability, $P_c(b_j|a_i)$, is given for any combination of $(i, j)$, the statistical properties of the noisy channel are fixed and known. As a result, the information content from a raster data with cell size $c$ can be formulated as shown below. Given a real space with status $A=\{a_1, \ldots, a_n\}$, the expected information content from the raster data ($B=\{b_1, \ldots, b_n\}$) of the space can be expressed by the following formula:

$$I_c(A;B) = H(A) - H_c(A|B) \tag{1}$$

where

$$H(A) = -\sum_i P(a_i)\log_2 P(a_i)$$

$$H_c(A|B) = -\sum_i \sum_j P_c(a_i, b_j)\log_2 P_c(a_i|b_j)$$

$$P_c(a_i|b_j) = \frac{P(a_i)P_c(b_j|a_i)}{\sum_i P(a_i)P_c(b_j|a_i)}$$

Here, $P(a_i)$ is the probability that a location in real space is $a_i$, $P_c(a_i,b_j)$ is the joint probability that a location is $a_i$ and identified as $b_j$, while $P_c(a_i|b_j)$ is the conditional probability that a location is actually $a_i$ when identified as $b_j$.

The first term, $H(A)$, in equation (1) is a well-known term called entropy in Shannon information theory for uncertainty assessment. It is also called the average self-information and indicates the uncertainty (ambiguity) related to the real space before the raster data values are known. More specifically, it is the uncertainty that a person will guess the categories of all locations correctly just by knowing the proportion of each category $A=\{a_1, \ldots, a_n\}$ in the space.

In contrast, $H_c(A|B)$ is the uncertainty about the real space after the raster data is known. The difference between the two values, $I_c(A;B)$, is the amount of information obtained from raster data with cell size $c$. In information theory, this is referred to as the average mutual information. Since it is hard to evaluate the relationship between the information-gain and the cell size $c$ using equation (1), we normalize this value with the Shannon's entropy, $H(A)$, which is equivalent to the average self-information.

$$R_c(A;B) = \frac{I_c(A;B)}{H(A)} = \frac{H(A)-H_c(A|B)}{H(A)} \tag{2}$$

$R_c(A;B)$ is then the proportion of information of the real space captured by raster data with cell size $c$. When cell size $c$ is infinitely small, the average mutual information is close to zero.

$$\lim_{c\to 0} R_c(A;B) = 1 \tag{3}$$

Hereafter, $R_c(A;B)$ is referred to as information-gain. In the following sections, we investigate how it changes with cell size $c$.

## 3. RELATIONSHIPS BETWEEN CELL SIZE AND THE INFORMATION-GAIN RATE

### 3.1 Cell size definition and information-gain rate measurement

For simplicity, we assume that the real space under consideration is a square with side length $L$, that is equally divided into $2^n$ cells horizontally and vertically, as shown in Figure 3. The cell size (side length) is given by the following equation with the number of splitting times, $n$ (≥1),

$$c(n) = \frac{L}{2^n} \tag{4}$$

In the following, we discuss the relationships between the information-gain rate and the number of splitting times, $n$.

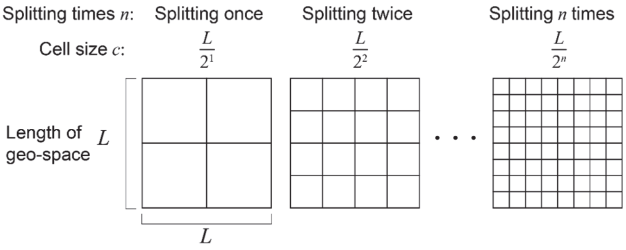


**Figure 3**. Cell size notation.

The study area (shown in Figure 4) is a square with a side length of about 1.6 km centered on the JR Ikebukuro Station, which is one of the largest terminal railway stations in Tokyo. We begin by obtaining the land-use vector data in this area. Since we are focusing on raster data, the real space vector data are regarded as the ground truth, even though they differ from the real space.

All six land-use categories listed in Table 1 are included in the study area, and Table 2 shows the correspondence between the number of splitting times and the cell size.

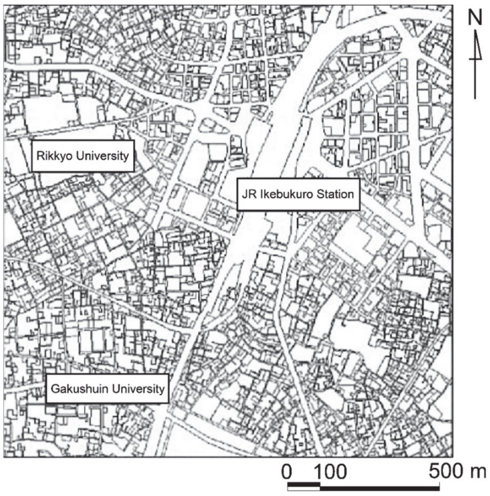


**Figure 4**. Study area.

| Land-use type | Area (km$^2$) | Percentage (%) |
|---|---|---|
| 1. Public | 0.31920 | 11.89 |
| 2. Commercial | 0.49398 | 18.40 |
| 3. Detached house | 0.46719 | 17.40 |
| 4. Apartment house | 0.34923 | 13.01 |
| 5. Empty or Others | 0.30633 | 11.41 |
| 6. Transportation | 0.74930 | 27.90 |
| Sum | 2.68522 | 100.00 |

**Table 1**. Land-use proportions.

| Splitting times *n* | Cell size *c* (m) |
|---|---|
| 1 | 819.2 |
| 2 | 409.6 |
| 3 | 204.8 |
| 4 | 102.4 |
| 5 | 51.2 |
| 6 | 25.6 |
| 7 | 12.8 |
| 8 | 6.4 |
| 9 | 3.2 |
| 10 | 1.6 |
| 11 | 0.8 |
| 12 | 0.4 |

**Table 2**. Cell size.

The relationship between the splitting time, $n$, and the information-gain rate, $R_c(A;B)$, is shown in Figure 5, which also shows the information-gain rate change. Here, it can be seen that the information-gain rate is small when there are only a few splitting times and that it increases together with the number of splitting times. However, after splitting more than seven times, the information-gain rate increment increase starts getting slower and slower. In particular, it is noteworthy that when the cell size is near 12 meters ($n$ = 7) and gets smaller, the information-gain rate increases rapidly. This indicates that additional cell splits would be possible before the peak is reached.

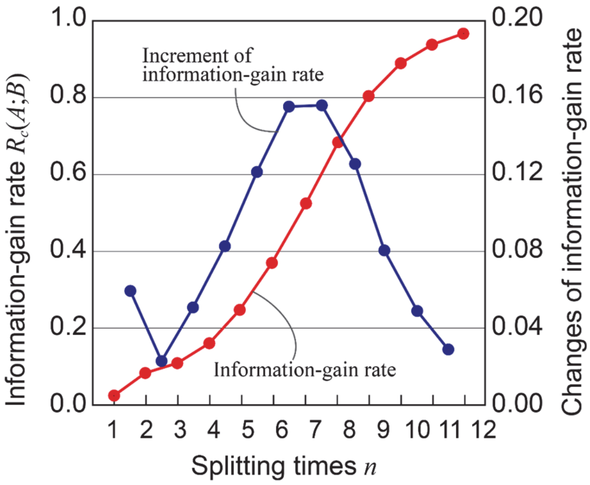


**Figure 5**. Information-gain rate.

### 3.2 Information-gain rate per code length

The information content per code length value is a basic metric for evaluating the balance between the quantity of received information and both transmission time and cost. As such, it is used to assess the quality of coding in the field of information theory. The data structure employed in the last section (see Figure 3) is generally called a quadtree data structure, and given the number of splitting times, $n$, the coordinate of a cell is represented with $n$ digits. In addition, one more digit is used to indicate the data category. As a result, $n$+1 digits are needed.
Figure 6 shows the information-gain rate per digit. In the case of the study area, peak efficiency hits when the number of splitting times, $n$, is 9 or 10, where the side length of the cell is about 2 meters. However, since this method is an approach to evaluate the efficiency of rasterizing the land-use of a randomly selected location, a different method needs to be employed to evaluate the rasterization efficiency of the entire study area.

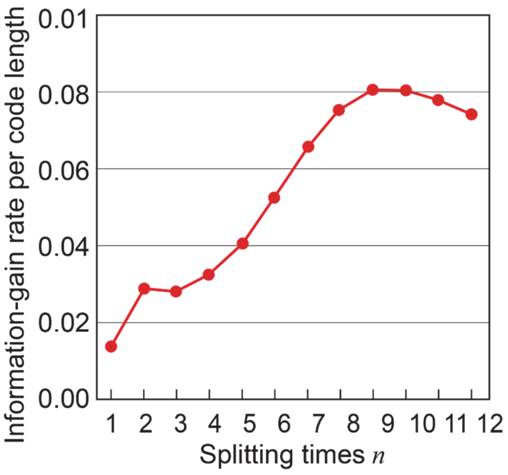


**Figure 6**. Information-gain rate per code length.

### 3.3 Logistic function for information-gain rate

Logistic functions can be adopted widely in the natural world. For example, the curve of the information-gain rate shown in Figure 5 might be a logistic curve. To confirm this hypothesis, we investigate whether the information-gain rate fits the following function using the number of splitting times, $n$, as the parameter:

$$f(n) = \frac{\exp[an+b]}{1+\exp[an+b]} \tag{5}$$

More specifically, if we could find a logistic formula like equation (5), the information-gain rate, $R_c(A;B)$, could be transformed into a simple linear model.

$$g(n) = \ln \frac{R_c(A;B)}{1-R_c(A;B)} = an + b \tag{6}$$

Figure 7 shows the results. Although some points deviate from the line when the number of splitting times, *n*, is small, the plot of the points almost follows a line. This indicates that the relationship between the information-gain rate and the splitting time can be approximately modeled by a curve.

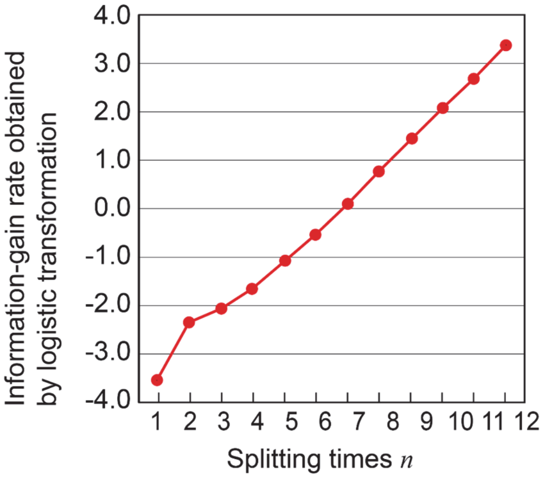


**Figure 7**. Logistic transformation.

## 4. RELATIONSHIPS BETWEEN SHAPE COMPLEXITY AND INFORMATION-GAIN RATE

### 4.1 Definition and properties of fractal dimension

In a logistic curve, parameter *a* is used to modify the logistic growth rate of the curve, while parameter *b* is used to set the sigmoid midpoint. We believe that these parameters are closely related to the shape characteristics of the target objects to be rasterized.

First of all, it should be noted that identification errors shown in Figure 1 occur at the boundaries of different categories, which means that the identification error value is related to the number of cells that lay across the boundaries of those categories. Here, we employ the idea of a fractal dimension, which is a ratio providing a statistical index of complexity comparing how the detail in a pattern changes with the scale at which it is measured in fractal geometry (Mandelbrot, 1967; Falconer, 2003).

The fractal dimension calculated using the box-counting method (Liu et al., 2003; Smith et al., 1996) is defined in the following paragraphs, in which the number of cells with cell size *c* is denoted by *N*(*c*) to cover an object. If the object has fractal properties, the following relation holds when *c* is sufficiently small.

$$N(c) \propto c^{-d} \tag{7}$$

If an object is rasterized by splitting *n* times, the relationship between the cell size *c* and *n* can be described as follows:

$$c \propto 2^{-n} \tag{8}$$

Thus, the number of cells needed to cover the object can be regarded as a function of *n*, *N*(*n*), and the relationship between *N*(*n*) and the fractal dimension, *d*, should be as follows:

$$\ln N(n) \propto nd \tag{9}$$

Here, if the object is composed of lines shaped like the coastline of an island, the perimeter, *L*, and the area surrounded by those lines, *A*, have the following relationship. The fractal dimension can easily be determined from the results of the following:

$$L \propto A^{\frac{d}{2}} \tag{10}$$

If the study area is covered by complex shapes with a large fractal dimension value, the information-gain rate cannot be high unless the number splitting times, *n*, is sufficiently large. In such cases, parameter *b* should be small. This implies the following relationship:

$$\frac{\partial b}{\partial d} < 0 \tag{11}$$

On the other hand, from equation (9), we can see that a larger value of *d* results in a larger increment in the number of boxes, *N*(*n*). In other words, a larger fractal dimension results in a larger area portion that can be recognized by splitting the time increment. This implies the following in the relationship between parameter, *a*, and *d*:

$$\frac{\partial a}{\partial d} > 0 \tag{12}$$

### 4.2 Logistic curve shape and fractal dimension

In this section, equations (11) and (12) are discussed in greater detail. If the target is a fractal shape, the fractal dimension can be calculated from the perimeter and area using equation (10). To investigate objects with diverse fractal dimensions, virtual maps are constructed by the following steps (a) to (c) below:

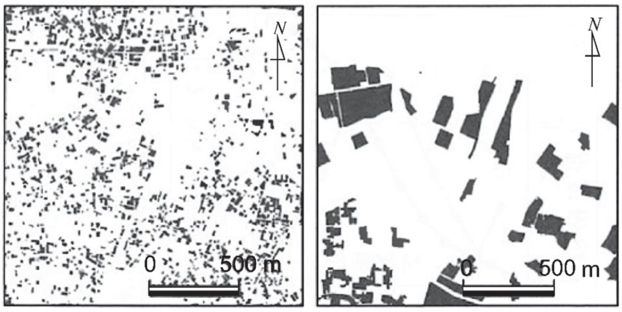
(a) Five classifications of fixed area lots.

Using the previously described land-use vector data (while excluding transportation-use land lots) we classify lots into five classes by summing their areas in ascending order so that each class consists of one-fifth of the total area. As a result, the areas of the five classes are almost the same but the lot perimeters are different. More specifically, the first class contains shapes with the longest perimeters since it is composed of a large number of small area shapes. Hence, the fractal dimension is the largest amongst the five classes. In contrast, the fifth class, which contains shapes with the shortest perimeters, has the smallest fractal dimension. Transportation-use land lots are excluded, since they are too large to be categorized in one-fifth of the total area.

(b) Five classes of fixed perimeter lots.

Similarly, using the same land lots (while still excluding transportation-use lots), we obtain five classes by summing the length of lot perimeters in ascending order so that each class consists of one-fifth of the total length of all perimeters. The perimeter lengths of the five classes are almost the same but their areas are different. Here, the first class, which contains lots with the smallest areas, has the largest fractal dimension value, while the fifth class, which contains shapes with the shortest areas, has the smallest fractal dimension.

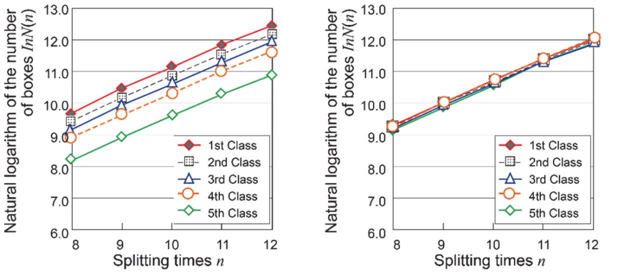
(c) Based on the above preparations, we then constructed maps using the above-mentioned shapes. More specifically, we created 10 virtual map types, examples of which are shown in Figure 8.



(a) Virtual map of the first class (b) Virtual map of the fifth class

**Figure 8**. Shape differences of classes having the same area.

We first investigate whether the prepared maps mentioned above (virtual maps) have fractal properties. The results are shown in Figure 9. If $n$ is sufficiently large, we can confirm that relationship expressed by equation (9) holds. Although there is no strict self-similarity (Mandelbrot, 1982), Figure 9 shows virtual maps could be viewed statistically as fractals.



(a) Classifications by same area (b) Classifications by same perimeter length

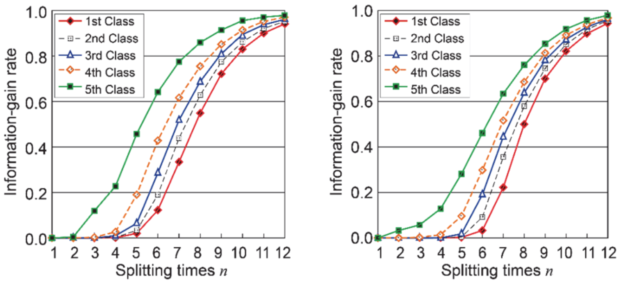
**Figure 9**. Conformation of fractal properties in virtual maps.

Next, Figure 10 shows the information-gain rates of raster data for the 10 virtual maps by number of splitting times. The transformed values produced by equation (6) are shown in Figure 11. Here, we can see that the values do not keep pace with the area proportion changes covered by each classification, especially when the number of splitting times, $n$, is small (Figure 11). Therefore, we calculate the values of $a$ and $b$ to fit equation (6) for $n \geq 9$.

We further investigated whether the relationships between parameters $a$, $b$, and fractal dimension, $d$, as expressed by equations (11) and (12), could be found. The results are shown in Figure 12, where a strong linear correlation is observed. This means that parameters $a$ and $b$ can be formulated functions of fractal dimension $d$:

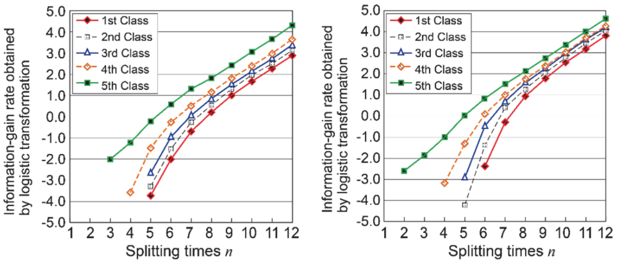
$$a(d) = \alpha_1 d + \beta_1 \tag{13}$$

$$b(d) = \alpha_2 d + \beta_2 \tag{14}$$



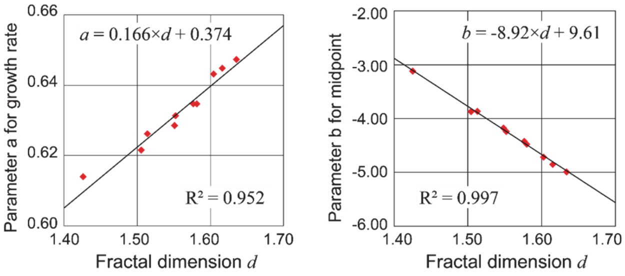
(a) Classifications by same area (b) Classifications by same perimeter length

**Figure 10**. Information-gain rate for virtual map.



(a) Classifications by same area (b) Classifications by same perimeter length

**Figure 11**. Information-gain rate for logistic transformation.

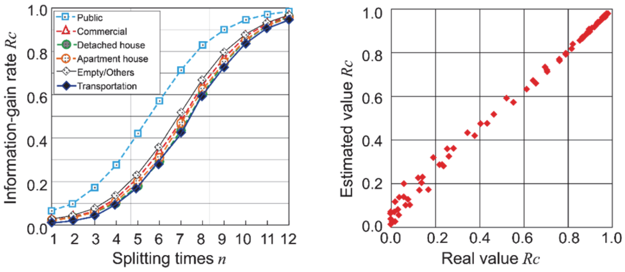


(a) Parameter to control steepness: $a$ (b) Parameter to set midpoint: $b$

**Figure 12**. Relationship between the fractal dimension and parameters.

### 4.3 Calculation of information-gain rate from fractal dimension

From the above analysis, given the fractal dimension, the information-gain rates of the raster data as a function of the number of splitting times, $n$, can be calculated. To confirm this, the previously used map showing six land-use types was employed. First, the fractal dimensions of these six types, which have various areas and perimeters, were estimated using parameters $a$ and $b$ in equations (13) and (14). Next, the information-gain rate was estimated using the estimated parameters $a$ and $b$ calculated in equation (5), as shown in Figure 13(a). Real values were then calculated from the change in the number of splitting times and the estimated values produced by equation (5). The results are shown in Figure 13(b). These two sets of values coincide closely, which numerically confirms our theory.



(a) Calculated information-gain rates (b) Estimated values and real values

**Figure 13**. Information-gain rate calculation.

## 5. PROPERTIES AND INFORMATION CONTENT OF GEOGRAPHIC FEATURES

### 5.1 Information content from adjacent cells

The above-discussed information content is the information obtained from one cell. However, when one cell is surrounded by other cells that belong to the same category, we may safely conclude that this cell is of the same type. However, if it is adjacent to various other cell types, this cell could belong to any one of those types. Therefore, it is important to consider the data together with information on the surrounding cells.

In this section, we discuss the information content of surrounding cells. We begin by noting that the confirmed theories presented in the previous sections can be easily applied to calculating the information content of surrounding cells. For example, the conditional probability from the mutual information shown by equation (1), $P_c(a_i|b_j)$, can be extended to the probability, $P_c(a_i|b_j, c_k)$, with two conditions. Here, $c_k$ are the symbols obtained from the surrounding cells. The information gain changes based on the category of the symbols (i.e., what we know from the surrounding cells). In the next subsection, numerical examples are introduced.

## 5.2 Numerical examples

Figs. 14 to 16 show the changes to the information-gain rate of "public", "detached house", and "apartment house" cells. More specifically, the information content obtained from the number of cells (in eight neighborhoods) for each category is shown. In other words, the values of the y-axis are estimated by the following equation:

$$D(A; b_j, c_k) = R_c(A; b_j, c_k) - R_c(A; b_j) \quad (15)$$

where

$b_j$: Public, detached house, apartment house

$c_k$: Number of surrounding cells of category $k$

Regarding Figure 14, which concerns "public" cells, the information content obtained from the number of cells of the same category is huge because "public" land-uses cover relatively huge areas. From Figure 15, we can see that for the "detached house" category, the most information is obtained from adjacent cells of the same type, but is still less than in the case of "public" cells. This is because the sizes of "detached house" lots are smaller than "public" lots.

In Figure 16, although "apartment house" gains the most information from the surrounding cells of the same category when the number of splitting times, $n$, is big, it gains the most information from "detached house" lots when the splits are fewer than eight, namely the cell size is greater than 6 meters. We think this is because "apartment house" lots are often distributed as scattered among detached houses, so "apartment house" gains much information from surrounding cells of "detached house" lots. Hence, as is stated above, the loss of information content due to the cell size can be countered, to some extent, if the symbols of neighboring cells are considered carefully.

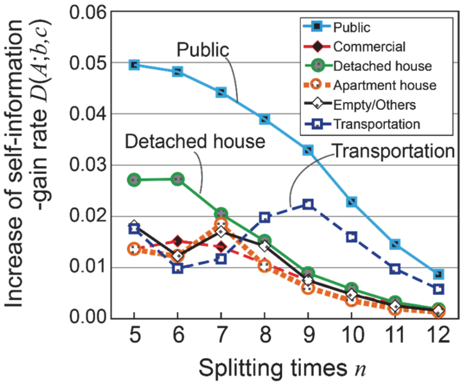


**Figure 14**. Self-information-gain rate increase for "public" lots.

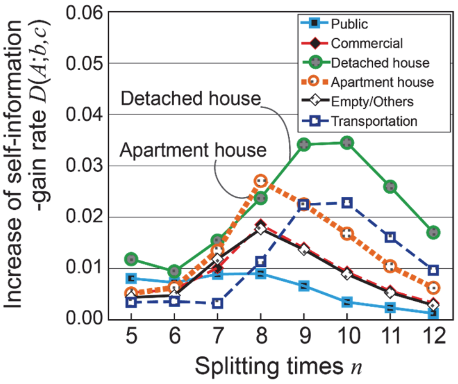


**Figure 15**. Self-information-gain rate increase for "detached house" lots.

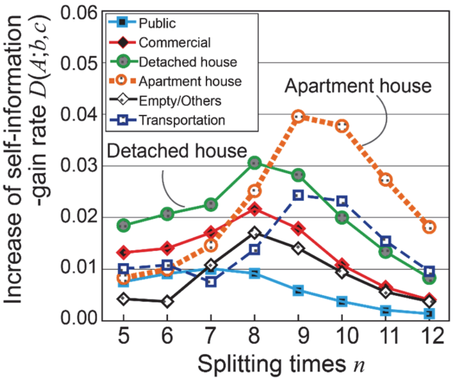


**Figure 16**. Self-information-gain rate increase for "apartment house" lots.

## 6. SUMMARY AND CONCLUSIONS

In this paper, we explored how much information content could be obtained from raster data. To accomplish this, we discussed how much information content could be lost in the process of creating raster data, and obtained the following results.

The information-gain rate was defined using the self-information and mutual information averages, which are the basic metrics used in information theory. Next, we showed that the information-gain rate can be expressed by a logistic curve with a single parameter $n$, which is the number of splitting times needed to determine the raster data resolution. We then demonstrated that the two parameters determining the shape of the logistic curve can be expressed as a linear function of the fractal dimension, $d$.

More specifically, given the raster data fractal dimension, the information content obtained by cell size, $c$, can be theoretically calculated. Conversely, the loss of information content in the process of creating raster data can be evaluated quantitatively. We also demonstrated that the loss of information content due to large cell size can be countered, to some extent, if the symbols of neighboring cells are considered carefully.

It is important to discuss the generality of the results obtained in this study. The authors believe that the results obtained in this study are valid for objects if they have fractal properties. However, it is not always clear what kind of object has the fractal properties. In this paper, we examined the spatial distribution of land-use as an example. It is obviously necessary to continue research on the spatial distribution of the various objects constructing cities.

We consider the advantages and disadvantages of the proposed solution as follows. Until now, data noise generated in the process of data creation has been often ignored or not enough considered in analysis. The advantage of the solution proposed in this paper is that it provides an idea of how much data noise is included or a method for reducing the noise. On the other hand, this solution cannot be applied when the objects to be analyzed do not have fractal properties. Namely, it is one of disadvantages that objects to be analyzed is limited.

## ACKNOWLEDGEMENTS

The author would like to extend special thanks to Ms. Naoko Kurisaki for her computer-based numerical calculations.

## REFERENCES

Tsichritzis, D.S., Lochovsky, F.H., 1977. *Database Management Systems*. Academic Press, New York.

Jones, D.S., 1979. *Elementary Information Theory*, Clarendon Press, Oxford, pp. 11-15.

Liao, S., Bai, Y., 2010. A new grid-cell-based method for error evaluation of vector-to-raster conversion, *Computational Geosciences*, 14(4), pp. 539–549.

Bai, Y., Liao, S., Sun, J., 2011. Scale effect and methods for accuracy evaluation of attribute information loss in rasterization, *J. Geogr. Sci.* 21(6), pp. 1089-1100.

Osaragi, T., 2019. Classification and Space Cluster for Visualizing GeoInformation, *International Journal of Data Warehousing and Mining*, ISPRS, 15, pp. 19–38.

Shannon, C.E., 1948. A Mathematical Theory of Communication, *Bell System Technical Journal*, 27, pp. 379–423 & 623–656.

Brill, E., Moore, R.C., 2000. An Improved Error Model for Noisy Channel Spelling Correction, *Proceedings of ACL* 2000.

Falconer, K., 2003. *Fractal Geometry: Mathematical Foundations and Applications*, Wiley, p. 308.

Mandelbrot, B., 1967. How Long is the Coast of Britain? Statistical Self-Similarity and Fractional Dimension, *Science*, 156(3775), pp. 636–638.

Liu, J.Z., Zhang, L.D., Yue, G.H., 2003. Fractal Dimension in Human Cerebellum Measured by Magnetic Resonance Imaging, *Biophysical Journal*, 85 (6), pp. 4041–4046.

Smith, T.G., Lange, G.D., Marks, W.B., 1996. Fractal methods and results in cellular morphology — dimensions, lacunarity and multifractals, *Journal of Neuroscience Methods*, 69 (2), pp. 123–136.

Mandelbrot, B.B., 1982. *The Fractal Geometry of Nature*, Spektrum Akademischer Verlag, p. 44.