Contradiction Hidden in Values of Urban Raster Data

Toshihiro Osaragi[1], Yoshitsugu Aoki[1]

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

Keywords: Urban raster data, Land use zone, Floor-area ratio zone, Identification.

Abstract

This paper discusses that raster data identified by the independent largest fraction method may include both manifest and hidden contradictions, and describes the underlying mechanisms of contradictions. Using the examples of zones of land use and floor-area ratio, we demonstrate that the cells including more than three different zone category combinations might be identified with a non-existing combination of the zone categories. For the zones in which the ratio of contradictory cells is comparatively large, this problem can be quite significant. We model, therefore, the probability of contradictory identification using the adjacency relationship of different categories, and demonstrate the good fitness of the model using actual urban raster data. The results reveal that there can be a significant proportion of hidden contradictions.

1. Introduction

A common method of creating raster data is to identify the cell as the one with the largest faction within the cell. Contradictions may arise, for instance, when two different regulations of land use categories, such as land use zone and floor-area ratio zone, are identified separately for the cell as illustrated in Figure 1. Suppose that there are three zoning categories within a cell: “First residential/100%”, “Commercial/400%”, and “Neighborhood commercial/400%”. In this case, “First residential” covers the largest fraction, and thus the land use zone of the cell is identified as “First residential”. On the other hand, if the floor-area ratio zone is identified independently, “400%” covers the largest area and thus the floor-area ratio zone is identified as “400%”. Namely, this cell is identified as “First residential/400%”, which is an illegal combination of land use zone and floor-area ratio zone and does not exist in Japan's Building Standards Act.

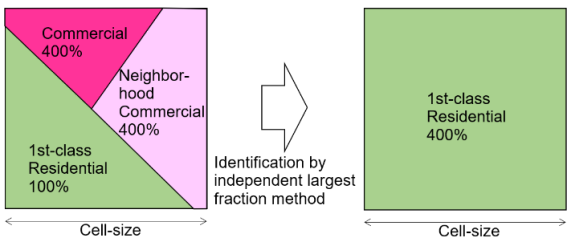

Figure 1. An example of contradictory identification (manifest contradiction).

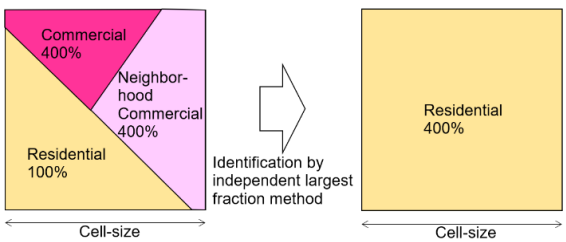

Figure 2. An example of contradictory identification (hidden contradiction).

Similarly, the example depicted in Figure 2 presents the same problem. Despite the absence of “Residential/400%” in the cell, it is identified as “Residential/400%”. While the categorization constitutes a legal combination, the cell is represented improperly. Such inaccuracies may significantly impact subsequent analyses. For instance, when the correlation between land use zone and floor-area ratio zone is investigated in residential areas, the resulting findings may be distorted, because such combination is never existing in reality. No problem occurs if the data on the land-use zone and floor-area ratio zone are used separately. However, if both categories of data are combined and used simultaneously, a logical contradiction will occur, making analysis difficult.

The issue mentioned above arises from the independent identification of two variables. This research considers the method where “a cell is identified as the pairwise combination of ‘land use zone/floor-area ratio zone’ whose area is largest in it”. We call this identification as the “pairwise largest fraction method”, and differentiate this method from the conventional approach, which we refer to as the “independent largest fraction method.” In this research, we define “contradictory identification” as “differences between identified results by pairwise largest fraction method and independent largest fraction method (i.e., for a cell, the combination of independently identified categories is not the same as that of the largest combination by pairwise identification).” We investigate and discuss mathematical notations and contradictory patterns that occur in the identification process of urban raster data.

2. Related Work

Various types of errors occur in the data handled by Geographic Information Systems (GIS) during the collection, conversion, and operation processes. Therefore, a great deal of research has been conducted to date on how errors in geographic information arise and affect spatial analysis results, and how we should treat and reduce errors. We start by reviewing the previous work from the viewpoint of two main data models (vector data and raster data). For vector data, Maras et al. provided a detailed study of geometric and topological errors and a visualization method to detect the sources of errors to increase the accuracy and reliability of GIS analysis (Maras et al., 2010). Bartonek et al. analyzed the error rate and accuracy improvement of identification results in automatic image identification, and demonstrated the error rate of identification results falls within the range of 2% to 3% by their proposed method (Bartonek et al., 2014). In addition, numerous studies have been conducted on the potential impact of

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-4-2024-247-2024 | © Author(s) 2024. CC BY 4.0 License.

data uncertainty and errors in specific analytical fields (Rae et al., 2007).

In order to understand the results taken from raster data correctly, much research has been done on errors in raster data. In particular, research has been attempted from various perspectives on the Digital Elevation Model (DEM). For example, Canters et al. evaluated the effects of raster data uncertainty in landscape identification using simulation technology to evaluate errors (Canters et al., 2002). Specifically, they used Monte Carlo simulation to evaluate the effects of DEM errors, uncertainties in land use identification, and the combined effects of both. Nackaerts et al. similarly used Monte Carlo simulation to analyze the effects of DEM errors on Boolean viewshed maps implemented in GIS software (Nackaerts et al., 1999). Dolan and Lucieer showed how to visualize uncertainty in terrain data using Monte Carlo simulation and analyzed how uncertainty affects calculations such as slope angle (Dolan and Lucieer, 2014). Lee et al. used simulation to analyze the effect of DEM errors on the accuracy of topographic feature extraction (Lee et al., 1992). Specifically, they evaluated the influence of the magnitude and spatial pattern of DEM errors on the extraction results of floodplain cells. Furthermore, Wu and Huang noted that the quality of DEM varies according to data sources in terms of horizontal resolution and vertical accuracy, and analyzed the problem of DEM uncertainty in hydrological simulations supported by GIS (Wu and Huang, 2008). Based on these empirical studies, Wechsler reviewed many excellent research related to DEM uncertainty (Wechsler, 2007).

Vector-raster conversion is one of the classic research topics in the field of GIS. Zhou et al. analyzed the increase and loss of polygon area that occurs during vector-raster conversion and proposed an equal area conversion model based on the area compensation optimization principle, which minimizes the distortion of the area of the entire dataset (Zhou et al., 2007). Bettinger et al. used vegetation distribution data to demonstrate the influence of grid-cell size on the vector-raster-vector conversion process (Bettinger et al., 1996). In addition, Auradkar et al. evaluated the accuracy loss associated with format conversion algorithms available in open-source GIS using land use and land cover (LULC) map data as an example, and showed that the number of vertices and shape complexity of vector data is correlated with conversion error (Auradkar et al., 2021).

When we use GIS to perform data operations on a map with errors, the errors will propagate. Errors existing in the map are propagated one after another through repeated operations, which may increase the uncertainty in the validity of the conclusions drawn. Therefore, much research has been conducted on the effects of error propagation. Arbia et al. analyzed how source map errors that occur as a result of overlay operations propagate (Arbia et al., 2010). Lunetta et al. used remote sensing data to identify potential sources of error at each step of the data integration process and assessed the impact of error propagation on decision-making and implementation processes (Lunetta et al., 1991). Furthermore, Choudhry and Morad analyzed the features and impacts of spatial errors in GIS and discussed ways to reduce the risk of error propagation in digital hydrological models (Choudhry and Morad, 1998). Biljecki et al. focused on the level of detail (LOD) and position errors and performed a multiple error propagation analysis that combines both types of errors (Biljecki et al., 2018). Specifically, they used a 3D city model to isolate errors in three spatial analyses (computing gross volume, envelope area, and solar irradiation of buildings) and showed how they propagate. As a more generalized theoretical study, Mitchell and Daley formulated a generalized Kalman filter consisting of model error and observation error to investigate the influence of discretization error on data assimilation (Mitchell and Daley, 2002). Based on many of these studies, Ouédraogo et al. attempted to compare topographic data collection techniques for high-resolution DEM generation and reviewed methods for DEM error propagation and its removal (Ouédraogo et al., 2014).

As another comprehensive theoretical study, Liu et al. focused on sampling strategy, sampling error estimation, and error evaluation model, and showed how to estimate the appropriate sample size using the Boltzmann curve (Liu et al., 2017). Thapa and Bossler also provided an overview of the various standards and specifications used in data collection methods and the various errors that occur during the data collection process by reviewing previous work (Thapa and Bossler, 1992).

As mentioned above, a large body of studies have been conducted on errors in GIS data from various perspectives, there are however no study that has discussed inconsistent errors (combinations of categories that do not logically exist) latent in the attribute information of raster data. Attribute values (categorical values) of raster data are generally identified with the largest area in each cell. However, when different types of information are identified separately, the resulting data set may contain some inconsistencies. Given this background, this paper investigates the mechanisms by which logical inconsistencies occur during the raster data creation process and discusses the extent to which inconsistencies exist in existing raster data.

3. Modeling the Contradictory Identification

3.1 Effects of Cell Size and Target Domain of This Research

In a two-dimensional space of area S, there are multiple closed sub-areas $C_1, C_2, \dots, C_n$. The probability of randomly selecting a point within C_1, denoted as $p(c_1)$, equals the ratio of the area of C_i, denoted as Sc_i, to the overall area, S (i.e., $\frac{Sc_i}{S}$). However, when the ratio (Sc_i) is calculated from data with multiple sub-area, $p(c_i)$ may involve errors. The accuracy of $p(c_i)$ estimation depends generally on the cell size, as a small cell size results in a larger number of boundary cells and thus a more accurate estimation, while a large cell size leads to a less accurate estimation. This fundamental features were investigated by Goodchild and Moy (1976) and Crapper (1980, 1984), and it was shown that, with fixed cell size, the variance of estimated $p(c_i)$ value can be calculated using characteristics of the space such as the area, perimeter, shape parameters, etc. Also, Osaragi discussed the relationships between the information loss and cell size from the viewpoint of information theory (Osaragi, 2022). In this study, we focus on contradictory identifications rather than errors raised by cell size. Specifically, we investigate the extent to which identification contradictions impact the results when the cell size is fixed.

3.2 Notations of Contradictory Identifications

The floor-area ratio zone (item A) and land use zone (item B) in each cell is notated as “A_0: unspecified, A_1: 50%, A_2: 60%, …” and “B_0: unspecified, B_1: 1st-class residential, B_2: 2nd-clas residential, …”. For simplicity, we represent floor-area ratio zone with i (or m) and land use zone with j (or k) using the subscript only. A pairwise category is represented by (i, j), and the set of legal combinations is denoted by U. Hence, (i, j) should be a member of a set U existing in urban regulation. Next, the area of the pairwise category (i, j) in a cell is denoted by $S_{i,j}$. The areas of categories i and j in the cell are therefore:

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-4-2024-247-2024 | © Author(s) 2024. CC BY 4.0 License.

$$a_i = \sum_j S_{i,j} \quad (1)$$
$$b_j = \sum_i S_{i,j} \quad (2)$$

According to the “independent largest fraction method”, each cell is identified in the following way:

$$\delta_i = \begin{cases} 1 & if\ a_i \geq a_m\ for\ all\ m \neq i \\ 0 & otherwise \end{cases} \quad (3)$$
$$\delta_j = \begin{cases} 1 & if\ b_j \geq b_k\ for\ all\ k \neq j \\ 0 & otherwise \end{cases} \quad (4)$$

In other words, if $\delta_i = 1$, the cell is identified as A_i, and if $\delta_j = 1$, the cell is identified as B_j. On the other hand, the “pairwise largest fraction method” indicates the following:

$$\delta_{i,j} = \begin{cases} 1 & \text{if } S_{i,j} \geq S_{m,k} \text{ for all } m \neq i, k \neq j \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

If $\delta_{i,j} = 1$, the cell is identified as (A_i, B_j) combination. Aggregated values in this research based on Equations (3) to (5) are:

$N_{i,j,*}$: The number of cells where $\delta_{i,j} = 1$.
$N_{i,j}$: The number of cells where $\delta_i = \delta_j = 1$.
$\underline{N_{i,j}}$: The number of cells where $\delta_{i,j} = 1 \wedge \delta_i = \delta_j = 1$.
$E_{i,j,*}$: The number of cells where $\delta_{i,j} = 1 \wedge \delta_i = \delta_j = 0$.
$E_{i,j}$: The number of cells where $\delta_{i,j} = 0 \wedge \delta_i = \delta_j = 1$. (6)

$N_{i,j,*}$ and $N_{i,j}$ are the number of cells using the pairwise largest fraction method and independent largest fraction method; $\underline{N_{j,j}}$ is the number of cells whose identifications are not impacted by the methods used to identify. $E_{i,j,*}$ is the number of cells whose category is (i,j) by the pairwise largest fraction method but was misidentified as other categories by the independent largest fraction method. $E_{i,j}$ is the number of cells whose category is not be (i,j) by pairwise largest fraction method but is identified as (i,j) by independent identifications.

Here, the contradictions accounted by $E_{i,j}$ is referred to as “manifest contradictions” if $(i,j) \notin U$. On the other hand, if $(i,j) \in U$ and the contradictions cannot be identified based on the identification results, it is referred to as “hidden contradictions”. Figures 1 and 2 show examples of “manifest contradictions” and “hidden contradictions” respectively. We have the following relationships for the above values:

$$N_{i,j,*} = \underline{N_{i,j}} + E_{i,j,*} \quad (7)$$
$$N_{i,j} = \underline{N_{i,j}} + E_{i,j} \quad (8)$$

Dividing the above values by the total number of cells (N), we get the expected values for one cell:

$$p(i,j)_* = \underline{p}(i,j) + e(i,j)_* \quad (9)$$
$$p(i,j) = \underline{p}(i,j) + e(i,j) \quad (10)$$

In the following, we will analyze and construct a model for the quantity of identification error $e(i,j)$ included in $p(i,j)$ from the independent largest fraction method. Additionally, we will assess the values of $p(i,j)_*$, $p(i,j)$, $e(i,j)_*$, and $e(i,j)$ to evaluate the contradictory identifications in existing data.

3.3 Cases Where Contradictory Identifications Arise

Contradictory identifications may arise when a cell contains more than three pairwise categories. In the following, to simplify the discussion, we assume that each cell contains at most three pairwise combinations. Under this assumption, there exist four patterns of contradictory identifications that are accounted by $E_{i,j}$ as shown in Figure 3.

(a) There exist three zones in a cell whose pairwise categories are (m,j), (i,k_1), (i,k_2). Their areas (i.e., $S_{m,j}$, S_{i,k_1}, S_{i,k_2}) satisfies the following:

$$\left.\begin{array}{c} S_{i,k_1} < S_{m,j} \\ S_{i,k_2} < S_{m,j} \\ S_{m,j} < S_{i,k_1} + S_{i,k_2} \end{array}\right\} \quad (11)$$

That is, in the shaded area in Figure 3(a), the cell is identified as $\delta_i = \delta_j = 1$ while it should be identified as $\delta_{m,j} = 1$. In this case, the identification (the independent largest fraction method) is different from the pairwise category (the pairwise largest fraction method).

(b) As a special case of (a), there exist three zones, (m,j), (i,j), (i,k_2) in a cell, and their areas (i.e., $S_{m,j}$, S_{i,k_1}, S_{i,k_2}) satisfies the following:

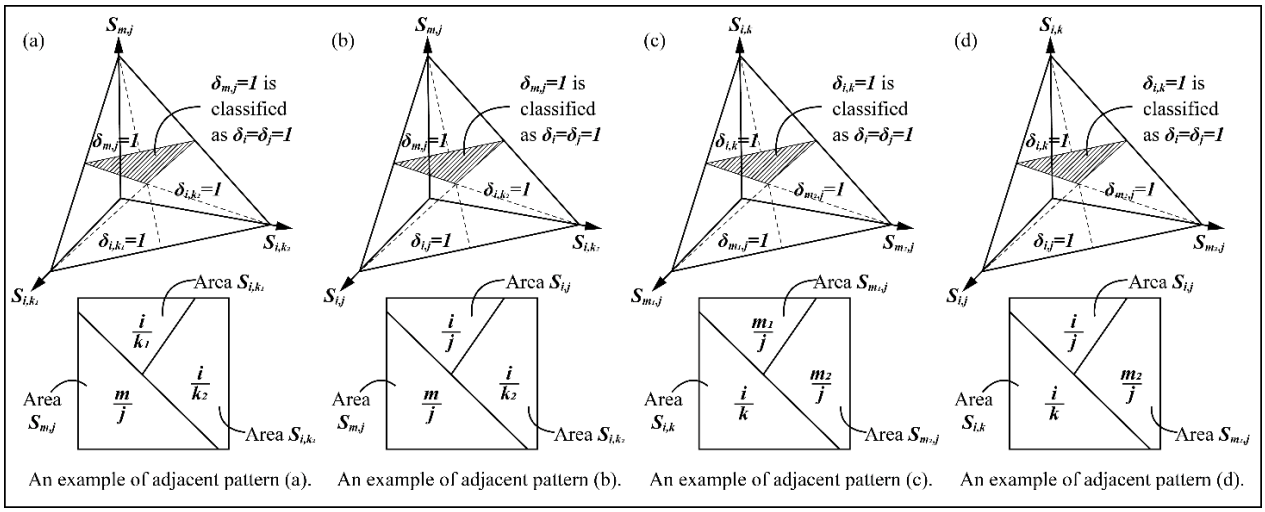

Figure 3. Patterns of contradictory identifications.

$$\left.\begin{array}{c} S_{i,j} < S_{m,j} \\ S_{i,k_2} < S_{m,j} \\ S_{m,j} < S_{i,j} + S_{i,k_2} \end{array}\right\} \quad (12)$$

That is, in the shaded area in Figure 3(b), the pairwise category (i,j) is existing in the cell but is not the category with the largest area.

(c) There exist three zones, (i,k), (m_1,j), (m_2,j) in a cell. Their areas (i.e., $S_{i,k}$, $S_{m_1,j}$, $S_{m_2,j}$) satisfies the following:

$$\left.\begin{array}{c} S_{m_1,j} < S_{i,k} \\ S_{m_2,j} < S_{i,k} \\ S_{i,k} < S_{m_1,j} + S_{m_2,j} \end{array}\right\} \quad (13)$$

In the shaded area in Figure 3(c), the cell should be identified as $\delta_{i,k} = 1$, but is identified as $\delta_i = \delta_j = 1$ which does not exist in the cell.

(d) As a special case of (c), there exist three zones, (i,k), (i,j), (m_2,j) in a cell, and their areas (i.e., $S_{i,k}$, $S_{i,j}$, $S_{m_2,j}$) satisfies the following:

$$\left.\begin{array}{c} S_{i,j} < S_{i,k} \\ S_{m_2,j} < S_{i,k} \\ S_{i,k} < S_{i,j} + S_{m_2,j} \end{array}\right\} \quad (14)$$

In the shaded area in Figure 3(d), the cell is identified as (i,j) which is not the largest zone in the cell. We can see the contradictory identifications arise.

Assuming that there are no notable differences in the adjacent zone of each category pair, and the origin of the cell line is randomly set, one–twelfth of the cell (the proportion of the area of each shaded part in the triangle) indicates the adjacency relationship of the three category pairs that will result in contradictory identifications in Figure 3.

As stated above, the probability of contradictory identifications can be determined by knowing the probability of each adjacency pattern. In the following, we will present our approach to modeling this probability.

3.4 Modeling the Probability of Contradictory Identifications

Figure 3 shows contradictory identifications where the cell is $\delta_{i,j} = 0$ but was wrongly identified as $\delta_i = \delta_j = 1$, which is accounted by $E_{i,j}$. By referring to the adjacency relationship of category pairs, we first consider the probability of contradictory identification $e(i,j)$.

If we denote the probability of occurrence of contradictory identifications in Figures 3(a) and (b) with $e_1(i,j)$ and that in Figures 3(c) and (d) as $e_2(i,j)$, we have:

$$\begin{aligned} e_1(i,j) &= \sum_{m=1}\sum_{k_1}\sum_{k_2\neq k_1} p(m,j)_* s(i:m) t(k_1,k_2:j) \\ &= (r(i,j) - s(i:i)p(i,j)_*)\beta_j \end{aligned} \quad (15)$$

$$\begin{aligned} e_2(i,j) &= \sum_{k=1}\sum_{m_1}\sum_{m_2\neq m_1} p(i,k)_* t(j:k) s(m_1,m_2:i) \\ &= (r(i,j) - t(j:j)p(i,j)_*)\alpha_i \end{aligned} \quad (16)$$

where

$p(i,j)_*$: Probability that a location is of category A_i and B_j.

$s(i:m)$: The probability that a location of A_m is adjacent to A_i.

$t(j:k)$: The probability that a location of B_k is adjacent to B_j.

$s(m_1,m_2:i)$: The probability that a location of A_i is adjacent to A_{m_1} and A_{m_2} at the same time.

$t(k_1,k_2:j)$: The probability that a location of B_j is adjacent to B_{k_1} and B_{k_2} at the same time.

$$\alpha_i = \sum_{m_1}\sum_{m_1\neq m_2} s(m_1,m_2:i)$$

$$\beta_j = \sum_{k_1}\sum_{k_1\neq k_2} t(k_1,k_2:j)$$

The probability that a location is of category A_i and B_j is assumed to be close to the value taken from the pairwise largest fraction method. The accuracy of the value depends on the cell size, but we do not examine the effect of cell size here. Given Equations (15) and (16), $e(i,j)$ can be expressed by:

$$\begin{aligned} e(i,j) &= \big(e_1(i,j) + e_2(i,j)\big)/12 \\ &= \big(\alpha_i + \beta_j\big)r(i,j)/12 - \Big(\alpha_i t(j:j) + \beta_j s(i:i)\Big)p(i,j)_*/12 \end{aligned} \quad (17)$$

Similarly, we have the following for $e(i,j)_*$.

$$e(i,j)_* = \Big(\alpha_i\big(t(j) - t(j:j)\big) + \beta_j\big(s(i) - s(i:i)\big)\Big)p(i,j)_*/12 \quad (18)$$

where $s(i) = \sum_m s(i:m)$ and $t(j) = \sum_k t(j:k)$. Subtracting Equation (10) from Equation (9) and using Equations (17) and (18), we get:

$$p(i,j)_* = \frac{\big(p(i,j) - \big(\alpha_i + \beta_j\big)r(i,j)/12\big)}{1 - \big(t(j)\alpha_i + s(i)\beta_j\big)/12} \quad (19)$$

Once we get $p(i,j)_*$, we get other variables using Equations (9), (10), (17), and (18).

4. Discussion Using Existing Raster Data

4.1 Raster Data and the Distribution of Contradictory Identifications

Using the actual raster data on urban regulations, we examine the proposed model. The data on land use zone and floor-area ratio zone are obtained from the Digital Detailed Information (Map Center of Japan: 1988) and identified using the independent largest fraction method with cell size of 100m × 100m. The target area is the Tokyo Metropolitan Area, excluding river, lake, and sea areas, where any zone is designated, and we exclude cells that are adjacent to the excluded areas. The adjacency of zones is computed for the remaining cells.

Table A1 (in Appendix) shows the distribution of $N_{i,j}$,where legal combinations (i.e., U) are colored in yellow. It is observed that there are many cells including "manifest contradictions" (i.e., combinations that are not included in the set U).

4.2 Fitness of the Model

First, we assess the fitness of the model. It is necessary to know the observation values of contradictory identifications. If $(i,j) \in U$, the distribution of $e(i,j)$ is not directly observed. However, if $(i,j) \notin U$, which means that the errors are raised by "manifest contradiction" and $p(i,j) = e(i,j)$. Thus, we will examine the model using data in this case with 101 samples, which are members of $(i,j) \notin U$.

To estimate the value of $e(i,j)$ using Equation (17), α_i, β_j, $r(i,j)$, $t(j:k)$, and $s(i:m)$ should be calculated. For calculating adjacency probabilities between categories, the data created by the independent largest fraction method is preferred than that by the pairwise largest area, since it is more accurate in terms of the area of each category within the cell. Thus, these values are calculated first using the given dataset.

The estimated values of α_i and β_j are presented in Table 1, where β_j is particularly large in residential, neighborhood commercial, and commercial areas. In other words, these areas are likely to be adjacent to other categories.

Estimated value of α_i ($\times 10^{-3}$)					
i=0: unspecified	10.5	i=5: 150%	315.7	i=10: 600%	390.8
i=1: 50%	201.1	i=6: 200%	200.3	i=11: 700%	338.5
i=2: 60%	257.3	i=7: 300%	428.5	i=12: 800%	243.0
i=3: 80%	274.0	i=8: 400%	45.8	i=13: 900%	236.2
i=4: 100%	283.9	i=9: 500%	484.4	i=14: 1000%	175.7
Estimated value of β_j ($\times 10^{-3}$)					
j=0: unspecified	15.1	j=3: residential	349.4	j=6: semi–industry	242.0
j=1: 1st-class residential	294.6	j=4: neighbour-hood coml.	541.5	j=7: industry	186.3
j=2: 2nd-class residential	297.3	j=5: commercial	372.2	j=8: industry use only	102.7

Table 1. Estimated values of α_i and β_j.

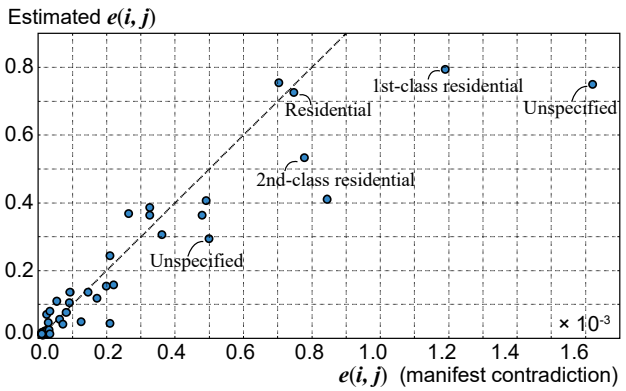

Figure 4. Fitness of the model.

When $(i,j) \notin U$, the relationship between $p(i,j)$ (i.e., manifest contradiction $e(i,j)$) and $e(i,j)$ calculated using Equation (17) is shown in Figure 4. When the observed value of $e(i,j)$ is small, the model performs well. Nevertheless, when $e(i,j)$ is large, there are more contradictions in the real data than in the model estimation. There are several potential reasons. (i) We assume that there are at most three categories of zones that are adjacent to each other in a cell, while in fact there can be more. (ii) The probability of the occurrence of contradictory identifications can be higher than $1/12$ due to the shape characteristics of zones. (iii) There might be contradictory identifications because of the low resolution of the map or other measurements. (iv) There can be effects from variables aggregated from existing datasets, such as α_i and β_j. Specifically, the deviations from the model prediction are particularly prominent in the “unspecified” and “1st-class residential” categories, but further examination indicates that contradictory identifications arise frequently in areas where only the two categories are adjacent. In other words, there can be other contradictions other than those shown in Figure 3 (e.g., errors caused by the image scanner).

4.3 Calculated Values of Contradictory Identifications

Next, the values of $p(i,j)_*$ and $\underline{p(i,j)}$ are calculated and presented in Tables A2 and A3 (in Appendix). To understand the characteristics of contradictions, the value of $e(i,j)$ is normalized with $p(j)$ ($= \sum_i p(i,j)$), and $e(i,j)/p(j)$ is reported in Table A4 (in Appendix). The value of $\sum_i e(i,j)/p(j)$ (i.e., the possibility of the occurrence of contradictory identifications for each land use) is larger in “Neighborhood commercial” and “Commercial” zones. This finding aligns with Table 1, where β_j is larger in these areas. “Neighborhood commercial” and “Commercial” zones are often designated along the road and extend in a linear form, and hence they are adjacent to other categories of zones more frequently (Figure 5). As a result, contradictory identifications happen often. On the contrary, industrial zones are situated in the suburban areas where the land use zone is less diverse, and they are often designated as a complete zone, similar to the city center. Cells in such areas are rarely adjacent to other zone categories, and thus contradictory identifications are fewer. Moreover, for “Neighborhood commercial/ 200%”, the $\sum_i e(i,j)/p(j)$ value is large and thus it is likely that “hidden contradictions” arise in such areas. The same applies to “1st-class residential/ 200%”.

As is discussed above, datasets that are identified using the independent largest fraction method may include many manifest contradictions, as well as hidden contradictions. When two different datasets are combined, we should analyze them with caution.

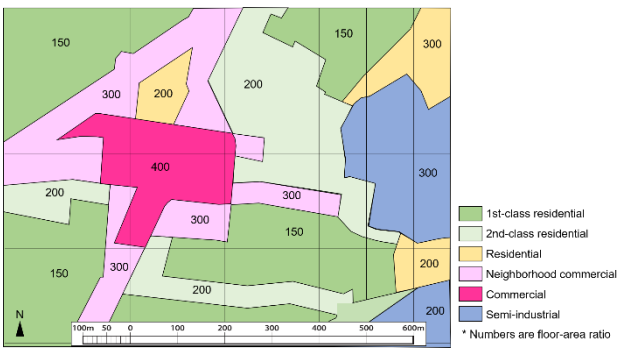

Figure 5. Example of land use zone and floor-area ratio zone.

5. Summary and Conclusions

In this study, we demonstrated that raster data identified by the independent largest fraction method may include both manifest and hidden contradictions and identified the underlying mechanisms of contradictions. The probability of contradictory identification is modeled using the adjacency relationship. Despite the good fitness of the model, the original datasets exhibited more contradictions than our theoretical predictions, highlighting the necessity to consider additional factors in the model. Moreover, our analysis of the distribution of contradictory identifications reveals that there can be a significant proportion of hidden contradictions. Accordingly, even if the combination of land use zone and the floor-area ratio zone is legal, the dataset may not reflect the real condition. Specifically, areas with linear

patterns, such as “Neighborhood commercial”, are more likely to have contradictions. Therefore, it can be risky to analyze the correlation between land use zone and floor-area ratio zone using raster data from such areas. In the future, one of the topics that should be addressed to use the current data efficiently is the detection and removal of contradictions one by one to recover the datasets.

Acknowledgements

A portion of this paper was presented at a domestic conference held by the Urban Planning Society of Japan (Aoki and Osaragi, 1993). The paper is written in Japanese and in a format of non-available automatic translation. The authors believe that the manuscript added the latest research reviews and much discussion would be of great value to publish in English.

References

Aoki, Y., Osaragi, T., 1993: Contradictory Classification of Urban Lattice Data, Papers on City Planning, *City Planning Institute of Japan*, 28, 379-384.

Arbia, G., Griffith, D., Haining, R., 2010: Error propagation modelling in raster GIS: overlay operations, *International Journal of Geographical Information Science,* 12(2), 145–167.

Auradkar, P.K., et al., 2021: Accuracy assessment and performance analysis of raster to vector conversions on LULC data – India, *Journal of Engineering, Design and Technology*, 20(6), 1787–1809.

Bartonek, D., et al., 2014: Method of error assessment in image identification, *14th International Multidisciplinary Scientific Geoconference* (*SGEM*), III, 745–752.

Bettinger, P., Bradshaw, G.A., Weaver, G,W., 1996: Effects of geographic information system vector–raster–vector data conversion on landscape indices, *Canadian Journal of Forest Research*, 26(8), 1416–1425.

Biljecki, F., et al., 2018: The effect of acquisition error and level of detail on the accuracy of spatial analyses, *Cartography and Geographic Information Science*, 45(2), 156–176.

Canters, F., Genst, W., Dufourmont, H., 2002: Assessing effects of input uncertainty in structural landscape identification, *International Journal of Geographical Information Science*, 16(2), 129–149.

Choudhry, S., Morad, M., 1998: GIS Errors and Surface Hydrologic Modeling: An Examination of Effects and Solutions, *Journal of Surveying Engineering*, 124(3), 134–143.

Crapper, P.F., 1980: Errors Incurred in Estimating an Area of Uniform Land Cover Using Landsat, *Photogrammetric Engineering and Remote Sensing,* 46(10), 1295–1301.

Crapper, P.F., 1984: An estimate of the Number of Boundary Cells in a Mapped Landscape Coded to Grid-cells. *Photogrammetric Engineering and Remote Sensing*, 50(10), 1497–1503.

Dolan, M.F.J., Lucieer, V.L., 2014: Variation and Uncertainty in Bathymetric Slope Calculations Using Geographic Information Systems, *Marine Geodesy*, 37(2), 187–219.

Goodchild, M.F., Moy, W.S., 1976: Estimation from grid data: the map as a stochastic process, *Proceedings of the Commission on Geographical Data Sensing and Processing*, Moscow, 67–81.

Lee, J., Snyder, P.K., Fisher, P., 1992: Modeling the effect of data errors on feature extraction from digital elevation models, *Photogrammetric Engineering and Remote Sensing*, 58(10), 1461–1467.

Liu F., et al., 2017: Sampling strategy and error estimation for evaluation of quadratic form error using Cartesian coordinate data, *IET Science, Measurement & Technology*, 11(7).

Lunetta, R.S., et al., 1991: Remote Sensing and Geographic Information System Data Integration: Error Sources and Research Issues, *Photogrammetric Engineering & Remote Sensing*, 57(6), 677–687.

Maras, S.S., et al., 2010: Topological error correction of GIS vector data, *International Journal of the Physical Sciences*, 5(5), 476–483.

Mitchell, H., Daley, R., 2002: Discretization error and signal/error correlation in atmospheric data assimilation, *Tellus A: Dynamic Meteorology and Oceanography*, 49(1).

Nackaerts, K., Govers, G., Van Orshoven, J., 1999: Accuracy assessment of probabilistic visibilities, *International Journal of Geographical Information Science*, 13(7), 709–721.

Ouédraogo, M.M., Degré, A., Debouche, C., 2014: The high resolution digital terrain model, its errors and their propagation. *A review, Biotechnologie, Agronomie, Société et Environnement*, 18(3), 407–421.

Osaragi, T., 2022: Evaluation method for information content of raster data using fractal dimension, *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences,* Volume V-4-2022 XXIV ISPRS Congress (2022 edition), 6–11 June 2022, Nice, France.

Rae, C., Rothley, K., Dragicevic, S., 2007: Implications of error and uncertainty for an environmental planning scenario: A sensitivity analysis of GIS–based variables in a reserve design exercise, *Landscape and Urban Planning*, 79(3–4), 210–217.

Thapa, K., Bossler, J., 1992: Accuracy of Spatial Data Used Information Systems, *Photogrammetric Engineering and Remote Sensing*, 58(6), 835–841.

Wu, S., Li, J., Huang, G.H., 2008: Characterization and evaluation of elevation data uncertainty in water resources modeling with GIS, *Water Resource Management*, 22(8), 959–972.

Wechsler, S.P., 2007: Uncertainties associated with digital elevation models for hydrologic applications: a review, *Journal of Earth System Science*, 11, 1481–1500.

Zhou, C.H., et al., 2007: An equal area conversion model for rasterization of vector polygons, *Science in China Series D: Earth Sciences*, 50, 169–175.

Appendix

(i,j)	Unspeci–fied	1st-class residential	2nd-class residential	Residen-tial	Neighbor-hood comercial	Commer-cial	Semi–Industrial	Industrial	Industrial use only	Sum
Unspecified	14329	0	0	0	0	0	0	0	0	14329
50%	86	6565	142	49	11	0	5	3	0	6861
60%	118	24110	571	142	43	1	43	3	0	25031
80%	244	36420	526	476	135	10	57	6	0	37874
100%	337	28212	2359	505	149	26	22	7	3	31620
150%	62	6879	6409	221	100	23	17	18	0	13729
200%	1099	2689	46834	64181	5422	325	19921	9816	13660	163947
300%	9	180	2005	3684	5344	220	2948	21	61	14472
400%	354971	807	331	1232	1089	5596	808	101	121	365056
500%	2	4	19	64	39	2669	26	2	0	2825
600%	0	1	8	30	17	1945	18	0	0	2019
700%	0	0	4	6	0	704	2	0	0	716
800%	1	0	2	5	0	492	0	0	0	500
900%	0	1	0	1	0	107	0	0	0	109
1000%	0	0	0	0	0	101	0	0	0	101
Sum	371258	105868	59210	70596	12349	12219	23867	9977	13845	679189

Table A1. The distribution of $N_{i,j}$ from existing data
(Legal combinations (i.e., U) are colored in yellow).

(i,j)	Unspecified	1st-class residential	2nd-class residential	Residential	Neighborhood commercial	Commercial	Semi–Industrial	Industrial	Industrial use only
Unspecified	0.02115								
50%		0.00989							
60%		0.03662							
80%		0.05524							
100%		0.04198	0.00303						
150%		0.01019	0.00976						
200%		0.00133	0.07148	0.09919	0.00842		0.03023	0.01478	0.02016
300%			0.00287	0.00558	0.00918		0.00449	0.00001	0.00008
400%	0.52180			0.00158	0.00164	0.00849	0.00114	0.00012	0.00014
500%						0.00441			
600%						0.00285			
700%						0.00108			
800%						0.00077			
900%						0.00017			
1000%						0.00016			
Sum	0.54295	0.15525	0.08714	0.10635	0.01924	0.01793	0.03586	0.01491	0.02038

Table A2. Estimated value of $p(i,j)_*$ (pairwise largest fraction method)
(Legal combinations (i.e., U) are colored in yellow).

(i,j)	Unspecified	1st-class residential	2nd-class residential	Residential	Neighborhood commercial	Commercial	Semi–Industrial	Industrial	Industrial use only
Unspecified	0.02115								
50%		0.00989							
60%		0.03662							
80%		0.05524							
100%		0.04198	0.00303						
150%		0.01019	0.00976						
200%		0.00133	0.07148	0.09919	0.00842		0.03023	0.01478	0.02016
300%			0.00287	0.00558	0.00918		0.00449	0.00001	0.00008
400%	0.52180			0.00158	0.00164	0.00849	0.00114	0.00012	0.00014
500%						0.00441			
600%						0.00285			
700%						0.00108			
800%						0.00077			
900%						0.00017			
1000%						0.00016			
Sum	0.54295	0.15525	0.08714	0.10635	0.01924	0.01793	0.03586	0.01491	0.02038

Table A3. Estimated value of $\underline{p}(i,j)$ (these elements are not impacted by identification methods)
(Legal combinations (i.e., U) are colored in yellow).

(i,j)	Unspecified	1st-class residential	2nd-class residential	Residential	Neighborhood commercial	Commercial	Semi–Industrial	Industrial	Industrial use only
Unspecified	0.00001	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
50%	0.00009	0.00078	0.00051	0.00042	0.00021	0.00003	0.00017	0.00023	0.00002
60%	0.00022	0.00254	0.00472	0.00231	0.00318	0.00026	0.00157	0.00063	0.00000
80%	0.00057	0.00307	0.00610	0.00708	0.00801	0.00097	0.00214	0.00057	0.00020
100%	0.00055	0.00240	0.00426	0.00680	0.00825	0.00105	0.00136	0.00055	0.00023
150%	0.00020	0.00213	0.00153	0.00363	0.00713	0.00172	0.00089	0.00051	0.00006
200%	0.00138	0.01929	0.00247	0.00018	0.02193	0.02034	0.00096	0.00039	0.00014
300%	0.00001	0.00239	0.00422	0.00442	0.00549	0.02042	0.00481	0.00122	0.00052
400%	0.00006	0.00510	0.00470	0.00707	0.00947	0.00568	0.00401	0.00352	0.00199
500%	0.00000	0.00013	0.00080	0.00127	0.00565	0.00831	0.00225	0.00013	0.00000
600%	0.00000	0.00003	0.00035	0.00058	0.00263	0.00997	0.00068	0.00016	0.00001
700%	0.00000	0.00000	0.00007	0.00019	0.00024	0.00566	0.00011	0.00000	0.00000
800%	0.00000	0.00000	0.00001	0.00003	0.00006	0.00254	0.00001	0.00001	0.00000
900%	0.00000	0.00000	0.00000	0.00001	0.00000	0.00057	0.00000	0.00000	0.00000
1000%	0.00000	0.00000	0.00000	0.00001	0.00000	0.00026	0.00000	0.00000	0.00000
Sum	0.00309	0.03786	0.02974	0.03400	0.07225	0.07778	0.01896	0.00792	0.00317

Table A4. Estimated value of $e(i,j)/p(j)$
(Legal combinations (i.e., U) are colored in yellow).