CSIS Discussion Paper No. 117

A method of comparing numerical variables defined in a region

Yukio Sadahiro

February 2013

Center for Spatial Information Science, University of Tokyo
5-1-5, Kashiwanoha, Kashiwa-shi, Chiba 277-8568, Japan
sada@csis.u-tokyo.ac.jp

Keywords: comparison, transformation, numerical variables

Abstract

This paper develops a new method of comparing numerical variables defined in a region. The method covers numerical variables defined over a two-dimensional space such as temperature and humidity distributions and those defined on a discrete space such as the height of trees and the age of buildings. To evaluate the difference between two variables, the method considers three types of transformations. The transformations convert one variable so that it fits the other as well as possible. The result gives a basis for the separate evaluation of spatial and non-spatial differences between the variables. The transformations also permit us to describe the spatial difference in more detail. To test the validity of the method, the paper applies it to an analysis of three spatial datasets of different sizes. The result showed that the proposed method is effective for evaluating and visualizing the difference between numerical variables.

1. Introduction

This paper proposes a method of comparing numerical variables defined in a region. Temperature, humidity, and the density of carbon dioxide are represented as numerical variables defined continuously over a two-dimensional space. Population counts and population density are also variables defined on a two-dimensional space, though not as smooth as temperature distribution because they are often calculated by aggregating point data in spatial units. Traffic flow is a numerical variable defined on a network space. The height of trees, the age of buildings, and the annual sales of supermarkets are numerical variables defined on a discrete space.

There are several ways of comparing these numerical variables. One method is to employ general statistical measures. Correlation coefficients, both rank and product-moment, tell us whether two variables are correlated with each other. The Kullback-Leibler divergence (Kullback & Leibler, 1951; Kullback, 1959) is also useful to compare positive variables. If variables are defined on a discrete space, we can use the χ^2 test to evaluate the difference between the variables from a statistical perspective.

Unfortunately, however, the above statistical measures do not recognize differences in the spatial dimension (Hubert *et al.*, 1985; Haining, 1991; Lee, 2001). There are mainly two groups of methods in the literature that incorporate the spatial aspect explicitly.

One group extends the Pearson’s correlation coefficient to consider the correlation between variables and their spatial autocorrelation simultaneously. Some of the methods employ Moran’s I to evaluate the spatial autocorrelation (Wartenberg, 1985; Lee, 2001; Stephane *et al.*, 2008), while the others develop new measures of spatial autocorrelation (Tjøstheim, 1978; Hubert *et al.*, 1985; Haining, 1991).

Another class of methods uses the earth mover’s distance (Peleg *et al.*, 1989; Rubner *et al.*, 2000; Zhao *et al.*, 2010). The methods consider the turning of a pile of dirt into another form with the least cost. It is formulated as a transportation problem, and the solution is used as a measure of the difference between two variables. Although the earth mover’s distance is primarily used in image processing, it is also useful in spatial analysis.

The above existing methods are motivated to compare numerical variables defined on a discrete space. Consequently, they are not directly applicable to the analysis of numerical variables defined over a continuous space. In addition, the above methods implicitly assume variables with the same total volume. When the volume is different, they divide each variable by its total volume. Though such a standardization permits us to focus on the spatial difference between variables, it conceals the non-spatial difference that existed in the original variables. The standardization prevents us from separating the differences in the spatial dimension and non-spatial dimensions.

There are several papers that discuss the separation of spatial and non-spatial factors, though their focus is not on the comparison of numerical variables. Pontius (2000, 2002) and Pontius & Millones (2011) propose statistical measures for comparing categorical variables. Assuming a stochastic process, these measures evaluate the degree to which the observed number and location of each category differ from the expected ones. Wong (2011) proposes a new framework that considers the spatial and attribute dimensions separately when measuring the spatial autocorrelation. The separation of spatial and non-spatial factors permits us to deepen our understanding of the structure of spatial phenomena. Following the line of these papers, this paper aims to evaluate the difference between numerical variables separately in spatial and non-spatial dimensions.

Section 2 proposes several measures for evaluating the difference between numerical variables. It also discusses an extension of the measures to treat the difference between categorical variables. Section 3 applies the proposed approaches to an analysis of three datasets of different sizes in order to demonstrate the effectiveness the method in exploratory spatial analysis. Section 4 summarizes the conclusions with discussion.

2. Method

This paper discusses the comparison of numerical variables defined over a two-dimensional continuous space and those defined on a discrete space. We first discuss the latter and then proceed to the former.

Suppose n regions $\mathbf{R}=\{R_1, R_2, ..., R_n\}$ ($\mathbf{N}=\{1, 2, ..., n\}$) in each of which two sets of numerical variables $U=\{u_1, u_2, ..., u_n\}$ and $V=\{v_1, v_2, ..., v_n\}$ are defined. The location of R_k is indicated by that of its representative point denoted by $\mathbf{z}_k$.

2.1 Separation of the differences between variables

A simple method of comparing two variables is to sum up the difference between the variables in every region. We call this *overall difference* given by

$$D_O\left(U,V\right)=\sum_{i\in\mathbf{N}}\left|u_i-v_i\right| . \tag{1}$$

Though this measure is easy to calculate and understand, it does not recognize the spatial difference between variables as general statistical measures. In Figure 1, for instance, variables U, V_{11} and V_{12} have the same configuration of values, which results in $D_O(U, V_{11})=D_O(U, V_{12})$. Their spatial distribution, however, is different in that both U

and V_{12} have a peak in the top row whereas the peak of V_{11} is at the lower-right corner. The measure D_O do not recognize this difference since it neglects the spatial dimension.

2	3	9
6	11	15
13	19	22

V_{11}

19	22	13
15	11	6
9	3	2

V_{12}

25	18	12
14	11	7
10	2	1

V_{13}

21	18	14
16	11	5
8	4	3

V_{14}

22	19	13
15	11	6
9	3	2

U

7	25	11
1	20	14
10	4	8

V_{21}

3	4	9
14	16	11
5	12	26

V_{22}

23	24	15
35	16	19
13	28	17

V_{23}

2	17	4
14	6	5
8	3	1

V_{24}

16	15	20
7	21	27
8	6	10

V_{31}

1	2	4
7	5	10
14	12	8

V_{32}

25	6	15
10	14	4
12	2	21

V_{33}

31	26	24
19	11	5
7	2	1

V_{34}

Figure 1. The distributions of *U* and *V*.

To resolve the problem, we consider three types of transformations: 1) rearrangement, 2) moving, and 3) addition/deletion. We apply a transformation to *U* so that it fits *V* as well as possible. This permits us to evaluate the difference between *U* and *V* in the spatial dimension.

1) Rearrangement transformation

Rearrangement transformation changes the location of *U* values so that its spatial distribution is similar to that of *V* as well as possible. To this end, it relocates *U* values in the way that the rank of *U* coincides that of *V* in every cell. Let $r(u_i)$ be the function indicating the rank of u_i in *U*. Rearrangement is represented by a binary function defined by:

$$\rho_{ij}(U) = \begin{cases} 1 & \text{if } r(u_i) = r(v_j) \\ 0 & \text{otherwise} \end{cases} .$$

(2)

Rearrangement transformation reduces the difference between *U* and *V* to

$$D_R\left(U,V\right)=\sum_{i\in\mathbf{N}}\sum_{j\in\mathbf{N}}\rho_{ij}\left(U\right)\left|u_i-v_j\right| .$$

(3)

We call the difference between $D_O(U, V)$ and $D_R(U, V)$ the *location difference*:

$$\begin{aligned}d_L\left(U,V\right)&=D_O\left(U,V\right)-D_R\left(U,V\right)\\&=\sum_{i\in\mathbf{N}}\left(\left|u_i-v_i\right|-\sum_{j\in\mathbf{N}}\rho_{ij}\left|u_i-v_j\right|\right).\end{aligned}$$

(4)

Location difference estimates the difference in the location of values of two variables. If all the elements of *U* are identical to those of *V*, the rearrangement transformation can completely resolve the difference between *U* and *V*. In Figure 1, for instance, we can transform *U* into V_{11} or V_{12} by only changing the location of values. In such a case, we have

$$d_L\left(U,V\right)=D_O\left(U,V\right) .$$

(5)

The rearrangement transformation, on the other hand, completely fails in two circumstances. One is the case when

$$r\left(u_i\right)=r\left(v_i\right)\quad\forall i\in\mathbf{N}$$

(6)

holds as shown in V_{13} and V_{14} in Figure 1. The other is when

$$u_i\le v_i\quad\forall i\in\mathbf{N} ,$$

(7)

or

$$u_i\ge v_i\quad\forall i\in\mathbf{N}$$

(8)

holds as shown in V_{23} and V_{24} in Figure 1. In both cases we have

$$D_O\left(U,V\right)=D_R\left(U,V\right) .$$

(9)

and

$$d_L\left(U,V\right)=0 \quad . \tag{10}$$

2) Moving transformation

The rearrangement transformation changes the location of variables with keeping their values. This prevents us from transforming U into V_{13}, V_{14}, V_{21}, and V_{22}. We thus introduce the *moving transformation*, a special case of earth moving mentioned earlier. The moving transformation permits us to partially transfer U values between regions so that the difference between U and V is minimized.

Earth moving, in its original form, assumes two variables with the same total volume such as the probability density distribution. It converts the distribution of one variable into that of another variable by transferring values between regions at the least cost. The transfer can occur between every pair of regions, where the cost is given by the total weighted volume of transfer. Earth moving makes two distributions completely identical.

The moving transformation, on the other hand, permits variables with different total volumes. In addition, it only considers the transfer of a variable between adjacent regions. The moving transformation converts the distribution of one variable into that of another variable so that the summation of the difference at each region is minimized at the least cost. As a result, the moving transformation leaves the difference in the total volume of variables.

To derive the moving transformation, we have to calculate the volume of transfer between every adjacent regions. At present, however, we only need the result of the transformation, which can be obtained without calculating the actual volume of transfer.

The difference that remains after the moving transformation is given by

$$D_M\left(U,V\right)=\left|\sum_{i\in\mathbf{N}}u_i-\sum_{i\in\mathbf{N}}v_i\right|. \tag{11}$$

The moving transformation can completely convert U into $V_{13,}$ V_{14}, V_{21} and V_{22} in Figure 1, each of which has the same total volume as U.

We call the difference between $D_R(U, V)$ and $D_M(U, V)$ the *configuration difference*:

$$d_C\left(U,V\right)=D_R\left(U,V\right)-D_M\left(U,V\right)$$
$$=\sum_{i\in\mathbf{N}}\sum_{j\in\mathbf{N}}\rho_{ij}\left(U\right)\left|u_i-v_j\right|-\left|\sum_{i\in\mathbf{N}}u_i-\sum_{i\in\mathbf{N}}v_i\right|.$$

(12)

It is equal to $D_M(U, V)$ when the total volume of U is equal to that of V.

The moving transformation does not work when inequality (7) or (8) holds as shown in V_{23} and V_{24} in Figure 1. In addition, after the use of the rearrangement transformation, the moving transformation cannot further reduce the difference between variables if either

$$u_i\le v_i\rho_{ij}\left(U\right)\quad\forall i\in\mathbf{N},$$

(13)

or

$$u_i\ge v_i\rho_{ij}\left(U\right)\quad\forall i\in\mathbf{N}$$

(14)

holds. Examples include V_{31} and V_{32} in Figure 1. In the above cases, we have

$$D_R\left(U,V\right)=D_M\left(U,V\right).$$

(15)

and

$$d_C\left(U,V\right)=0.$$

(16)

3) The addition/deletion transformation

To remove the difference in the total volume between variables, we finally employ the *addition/deletion* transformation. It changes the value of U into that of V at individual locations. As a result, U becomes completely identical to V.

The difference reduced by the addition/deletion transformation is called the *volume difference*:

$$d_V\left(U,V\right)=D_M\left(U,V\right)$$
$$=\left|\sum_{i\in\mathbf{N}}u_i-\sum_{i\in\mathbf{N}}v_i\right|.$$

(17)

The relationship among the difference measures is shown in Figure 2. Relative measures are also useful to compare the location, configuration, and volume differences between different pairs of variables:

$$\delta_L(U,V) = \frac{d_L(U,V)}{D_O(U,V)},$$

(18)

$$\delta_C(U,V) = \frac{d_C(U,V)}{D_O(U,V)},$$

(19)

and

$$\delta_V(U,V) = \frac{d_V(U,V)}{D_O(U,V)}.$$

(20)

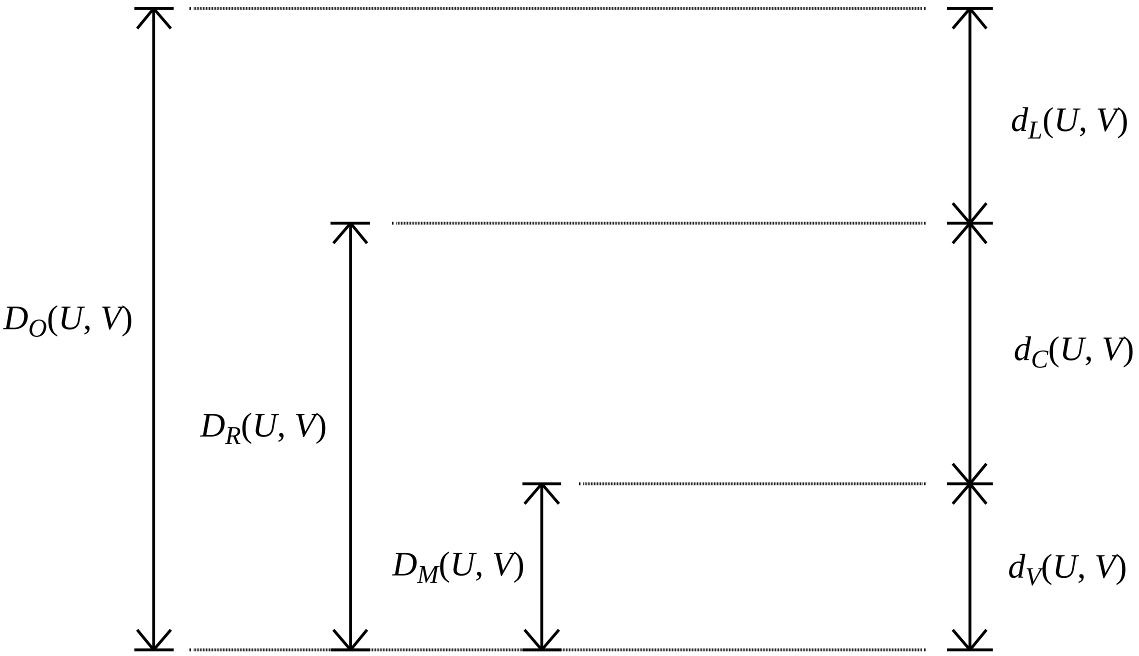


Figure 2. The relationship between the differences.

The three transformations described above permit us to evaluate the difference between variables in the spatial and non-spatial dimensions separately. The location and configuration differences indicate the spatial difference, while the volume difference represents the non-spatial one. The relative measures allow us to evaluate the proportion of each component in the overall difference.

The proposed measures can be calculated with or without the standardization of variables, depending on the objective of the analysis. If the separation of the spatial and non-spatial difference is critical, variables should be compared without standardization. If an emphasis is on the spatial difference, variables should be compared after standardization. In the latter case, we have

$$D_M\left(U,V\right)=d_V\left(U,V\right)=0 \quad . \tag{21}$$

After standardization, inequalities (7) or (8) can hold only when U and V are identical. This implies that while the moving transformation is almost always effective, the rearrangement transformation may not work when V=V_{13} or V=V_{14} in Figure 1.

2.2 Evaluation of the spatial difference between variables

The difference measures proposed in the previous subsection permit us to consider the difference between variables in the spatial and non-spatial dimensions separately. Evaluation is performed basically based on the difference of variables in individual regions. Consequently, though the transformations on which the difference measures are defined consider the spatial dimension explicitly, their summary measures do not reflect the spatial distribution of variables. The difference measures, for instance, fail to distinguish the difference between U and V_{11} and that between U and V_{12} in Figure 1.

To complement the difference measures, this subsection introduces two distance measures using the two spatial transformations, that is, the rearrangement and moving transformations. The distance measures are defined as the weighted average distance of a particular spatial transformation.

The rearrangement transformation defines a distance measure by

$$\lambda_R(U,V)=\frac{\sum_{i\in\mathbf{N}}\sum_{j\in\mathbf{N}}\rho_{ij}(U)\left|\mathbf{z}_i-\mathbf{z}_j\right|}{n\frac{\sum_{i\in\mathbf{N}}\sum_{j\in\mathbf{N}}\left|\mathbf{z}_i-\mathbf{z}_j\right|}{n^2}}=\frac{n\sum_{i\in\mathbf{N}}\sum_{j\in\mathbf{N}}\rho_{ij}(U)\left|\mathbf{z}_i-\mathbf{z}_j\right|}{\sum_{i\in\mathbf{N}}\sum_{j\in\mathbf{N}}\left|\mathbf{z}_i-\mathbf{z}_j\right|}. \tag{22}$$

In the moving transformation, we cannot calculate the weighted average distance as we have not yet derived the actual volume of transfer between regions. We thus formulate the moving transformation as an optimization problem where the moving cost is minimized.

Let us assume that U value is transferred from a region only to its adjacent regions. Let $\mathbf{NA}_i$ be the set of regions adjacent to R_i. The volume of U value transferred from R_i to R_j is denoted by x_{ij} ($x_{ij}\geq 0$). We formulate the moving transformation as the following optimization problem where the total volume of transfer is minimized.

Problem MT$_0$ (Moving Transformation):

$$\begin{aligned}&\underset{x_{ij},i,j\in\mathbf{N}}{\text{minimize}}\quad \sum_{i\in\mathbf{N}}\sum_{j\in\mathbf{NA}_i}x_{ij}\\&\text{subject to}\\&\quad\sum_{i\in\mathbf{N}}\left|u_i-\sum_{j\in\mathbf{N}}x_{ij}+\sum_{j\in\mathbf{N}}x_{ji}-v_i\right|=\left|\sum_{i\in\mathbf{N}}u_i-\sum_{i\in\mathbf{N}}v_i\right|\\&\quad x_{ij}\geq 0\quad i\in\mathbf{N},\, j\in\mathbf{NA}_i\\&\quad x_{ij}=0\quad i\in\mathbf{N},\, j\notin\mathbf{NA}_i\end{aligned}$$

Problem MT$_0$ is not solvable by using a general optimization technique since it contains the absolute functions in the constraints. To remove the absolute functions in the constraints, we define variable w_i as

$$w_i=u_i-v_i-\sum_{j\in\mathbf{N}}\left(x_{ij}-x_{ji}\right). \tag{23}$$

Substituting Equation (23) into the first constraint of Problem MT$_0$, we obtain

$$\begin{aligned}\sum_{i\in\mathbf{N}}|w_i| &= \left|\sum_{i\in\mathbf{N}}(u_i - v_i)\right| \\ &= \left|\sum_{i\in\mathbf{N}}\left\{w_i + \sum_{j\in\mathbf{N}}(x_{ij} - x_{ji})\right\}\right| \\ &= \left|\sum_{i\in\mathbf{N}} w_i + \sum_{i\in\mathbf{N}}\sum_{j\in\mathbf{N}}(x_{ij} - x_{ji})\right|. \\ &= \left|\sum_{i\in\mathbf{N}} w_i\right|\end{aligned} \tag{24}$$

Problem MT$_0$ then becomes

Problem MT:

$$\begin{aligned}&\underset{y_{ij},i,j\in\mathbf{N}}{\text{minimize}} \quad \sum_{i\in\mathbf{N}}\sum_{j\in\mathbf{NA}_i} x_{ij} \\ &\text{subject to} \\ &\quad \sum_{i\in\mathbf{N}}(u_i - v_i) \le 0 \Rightarrow w_i \le 0, \forall i \in \mathbf{N} \\ &\quad \sum_{i\in\mathbf{N}}(u_i - v_i) \ge 0 \Rightarrow w_i \ge 0, \forall i \in \mathbf{N} \\ &\quad x_{ij} \ge 0 \quad i \in \mathbf{N}, j \in \mathbf{NA}_i \\ &\quad x_{ij} = 0 \quad i \in \mathbf{N}, j \notin \mathbf{NA}_i\end{aligned}$$

Problem MT is equivalent to Problem MT$_0$ in that both yield the same solution. Problem MT is a linear optimization problem whose computational complexity is O(n). It is thus solvable in a linear time.

The solution of Problem MT permits us to calculate the weighted average distance of moving transformation. Let $\mu_A(\mathbf{R})$ be the average distance between all the pairs of adjacent regions of **R**. It is given by

$$\mu_A(\mathbf{R}) = \frac{\sum_{i\in\mathbf{N}}\sum_{j\in\mathbf{NA}_i}\left|\mathbf{z}_i - \mathbf{z}_j\right|}{2n_A}, \tag{25}$$

where n_A is the number of pairs of adjacent regions. The distance measure is then defined as

$$\lambda_M\left(U,V\right)=\frac{2\sum_{i\in\mathbf{N}}\sum_{j\in\mathbf{NA}_i}\left|\mathbf{z}_i-\mathbf{z}_j\right|x_{ij}}{\mu_A\left(\mathbf{R}\right)\sum_{i\in\mathbf{N}}\left|u_i-v_i\right|}$$

$$=\frac{4n_A\sum_{i\in\mathbf{N}}\sum_{j\in\mathbf{NA}_i}\left|\mathbf{z}_i-\mathbf{z}_j\right|x_{ij}}{\sum_{i\in\mathbf{N}}\sum_{j\in\mathbf{NA}_i}\left|\mathbf{z}_i-\mathbf{z}_j\right|\sum_{i\in\mathbf{N}}\left|u_i-v_i\right|}\quad .$$

(26)

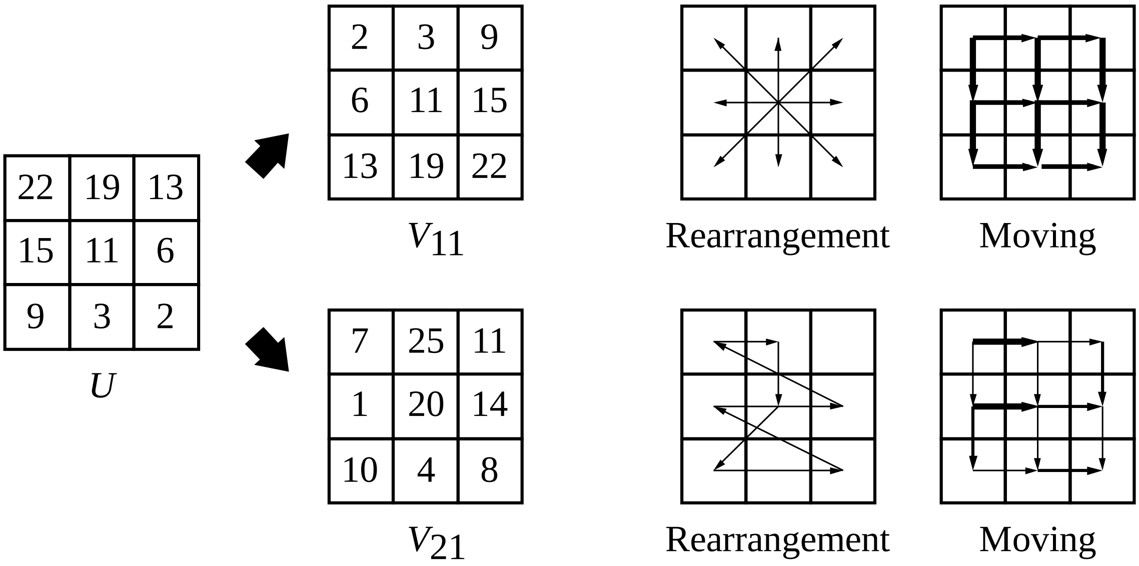
In the moving transformation, it is also useful to calculate the total volume of transfer. Its standardized form is given by

$$\gamma_M\left(U,V\right)=\frac{\sum_{i\in\mathbf{N}}\sum_{j\in\mathbf{NA}_i}x_{ij}}{\sum_{i\in\mathbf{N}}\left|u_i\right|+\sum_{i\in\mathbf{N}}\left|v_i\right|}\,.$$

(27)

2.3 Visualization of the difference between variables

The spatial difference between variables can be visualized effectively by a map based representation. The rearrangement and moving transformations are visualized by arrows that indicate the degree and direction of transformations as shown in Figure 3.



22	19	13
15	11	6
9	3	2

U

2	3	9
6	11	15
13	19	22

V_{11}

7	25	11
1	20	14
10	4	8

V_{21}

Figure 3. Rearrangement and moving transformations from *U* to *V*. The width of arrows indicates the volume of transfer in the moving transformation.

Visualization of rearrangement transformation, however, does not work when variables are defined in a large number of regions. In such a case, the map often contains numerous crossing arrows so that it becomes too complicated to interpret. Visualization of moving transformation, on the other hand, is relatively free from this problem because the arrows are drawn only between adjacent regions. In addition, the moving transformation has a wider variation of map representations such as streamlines, divergence, and critical points because the moving transformation can be regarded as a vector field. This paper recommends the moving transformation as a basis for visualizing the spatial difference because of its wide applicability and flexibility.

Visualization by the vector field and streamlines may remind of a method for displaying the migration pattern proposed by Tobler (1987, 1995). These papers, however, treat the actual movement of people while this paper considers a hypothetical movement of a variable as a means of evaluating the difference between variables. We should note that the vector field and streamlines that visualize the moving transformation do not indicate the actual movement of a variable in the real world.

2.4 Comparison of continuous variables

This subsection extends the method proposed above to the comparison of variables defined over a continuous two-dimensional space. Instead of discrete variables *U* and *V*, we consider two continuous variables $f(\mathbf{z})$ and $g(\mathbf{z})$ defined as functions of location $\mathbf{z}$ in region *R*.

The difference measures proposed in Subsection 2.1 are defined based on the summation of values calculated in individual regions. Consequently, their extension to continuous dimension is accomplished by replacing the summations with integrals. The difference measures $D_O(U, V)$ and $D_M(U, V)$ become

$$D_O\left(f(\mathbf{z}), g(\mathbf{z})\right) = \int_{\mathbf{z}\in R} \left| f(\mathbf{z}) - g(\mathbf{z}) \right| \mathrm{d}\mathbf{z} \tag{28}$$

and

$$D_M\left(f(\mathbf{z}), g(\mathbf{z})\right) = \left| \int_{\mathbf{z}\in R} f(\mathbf{z})\mathrm{d}\mathbf{z} - \int_{\mathbf{z}\in R} g(\mathbf{z})\mathrm{d}\mathbf{z} \right|, \tag{29}$$

respectively.

Extension of the measure $D_R(U, V)$ is as follows. Let $\rho(s, f(\mathbf{z}))$ be a binary function defined by

$$\rho\left(s, f\left(\mathbf{z}\right)\right) = \begin{cases} 1 & \text{if } f\left(\mathbf{z}\right) = s \\ 0 & \text{otherwise} \end{cases} .$$

(30)

Using the function, we evaluate the area where $f(\mathbf{z})$ and $g(\mathbf{z})$ and are smaller than t, that is,

$$a\left(f\left(\mathbf{z}\right)\right) = \int_{\mathbf{y}\in R}\int_{s\in\left[0, f\left(\mathbf{z}\right)\right]} \rho\left(s, f\left(\mathbf{y}\right)\right) \mathrm{d}s\mathrm{d}\mathbf{y}$$

(31)

and

$$a\left(g\left(\mathbf{z}\right)\right) = \int_{\mathbf{y}\in R}\int_{s\in\left[0, g\left(\mathbf{z}\right)\right]} \rho\left(s, g\left(\mathbf{y}\right)\right) \mathrm{d}s\mathrm{d}\mathbf{y} ,$$

(32)

respectively. We then define another binary function that indicates the rearrangement of U at $\mathbf{z}_1$ to $\mathbf{z}_2$:

$$\eta\left(\mathbf{z}_1, \mathbf{z}_2\right) = \begin{cases} 1 & \text{if } a\left(f\left(\mathbf{z}_1\right)\right) = a\left(g\left(\mathbf{z}_2\right)\right) \\ 0 & \text{otherwise} \end{cases} .$$

(33)

Equation (33) corresponds to equation (2) in Subsection 2.1. Using this function, we define $D_R(f(\mathbf{z}), g(\mathbf{z}))$ as

$$D_R\left(f\left(\mathbf{z}\right), g\left(\mathbf{z}\right)\right) = \int_{\mathbf{z}_1\in R}\int_{\mathbf{z}_2\in R} \eta\left(\mathbf{z}_1, \mathbf{z}_2\right) \left| f\left(\mathbf{z}\right) - g\left(\mathbf{z}\right) \right| \mathrm{d}\mathbf{z}_2\mathrm{d}\mathbf{z}_1 .$$

(34)

We can similarly define the other measures of difference. The distance measure based on the rearranging transformation is given by

$$\lambda_R\left(f\left(\mathbf{z}\right), g\left(\mathbf{z}\right)\right) = \frac{\int_{\mathbf{z}_1\in R}\int_{\mathbf{z}_2\in R} \eta\left(\mathbf{z}_1, \mathbf{z}_2\right) \left|\mathbf{z}_1 - \mathbf{z}_2\right| \mathrm{d}\mathbf{z}_2\mathrm{d}\mathbf{z}_1}{A \frac{\int_{\mathbf{z}_1\in R}\int_{\mathbf{z}_2\in R} \left|\mathbf{z}_1 - \mathbf{z}_2\right| \mathrm{d}\mathbf{z}_2\mathrm{d}\mathbf{z}_1}{A^2}}$$
$$= \frac{A\int_{\mathbf{z}_1\in R}\int_{\mathbf{z}_2\in R} \eta\left(\mathbf{z}_1, \mathbf{z}_2\right) \left|\mathbf{z}_1 - \mathbf{z}_2\right| \mathrm{d}\mathbf{z}_2\mathrm{d}\mathbf{z}_1}{\int_{\mathbf{z}_1\in R}\int_{\mathbf{z}_2\in R} \left|\mathbf{z}_1 - \mathbf{z}_2\right| \mathrm{d}\mathbf{z}_2\mathrm{d}\mathbf{z}_1} ,$$

(35)

where A is the area of **R**. The distance measure based on the moving transformation cannot be represented in an analytical form since it is calculated from the solution of an optimization problem. It is thus represented as a discrete approximation:

$$\lambda_M\left(f(\mathbf{z}),g(\mathbf{z})\right)=\frac{4n_A\sum_{i\in\mathbf{N}}\sum_{j\in\mathbf{NA}_i}\left|\mathbf{z}_i-\mathbf{z}_j\right|x_{ij}}{\sum_{i\in\mathbf{N}}\sum_{j\in\mathbf{NA}_i}\left|\mathbf{z}_i-\mathbf{z}_j\right|\left|\int_{\mathbf{z}\in R}\right|f(\mathbf{z})-g(\mathbf{z})\left|\mathrm{d}\mathbf{z}\right.}\ .$$

(36)

2.5 Comparison of categorical variables

The above method is also applicable to the comparison of categorical variables. Suppose two categorical variables $U=\{u_1, u_2, ..., u_n\}$ and $V=\{v_1, v_2, ..., v_n\}$ defined in n regions $\mathbf{R}=\{R_1, R_2, ..., R_n\}$. Each variable takes one of W categories represented by W integers $\mathbf{W}=\{1, 2, ..., W\}$.

Let φ(i, j) be a measure of distance between different categories i and j. The simplest definition of φ(i, j) is

$$\varphi\left(u_i,v_j\right)=\begin{cases}1 & \text{if } u_i\neq v_j\\ 0 & \text{otherwise}\end{cases}.$$

(37)

A distance matrix **Γ** is defined by

$$\mathbf{\Gamma}=\begin{bmatrix}0 & \varphi(1,2) & \cdots & \varphi(1,W)\\ \varphi(2,1) & 0 & & \varphi(2,W)\\ \vdots & & \ddots & \vdots\\ \varphi(W,1) & \varphi(W,2) & \cdots & 0\end{bmatrix}.$$

(38)

Using the matrix, we define the overall difference between U and V as

$$D_O\left(U,V\right)=\sum_{i\in\mathbf{N}}\varphi\left(u_i,v_i\right)\ .$$

(39)

The rearrangement transformation is defined as the transformation that relocates U values so that they U and V coincide as well as possible. Let x_{ij} be a binary function indicating the rearrangement of U:

$$x_{ij}=\begin{cases}1 & \text{if } u_i \text{ is relocated to } R_j\\ 0 & \text{otherwise}\end{cases}.$$

(40)

The rearrangement transformation is then obtained as a solution of an assignment problem:

Problem RT (Rearrangement Transformation):

$$\begin{aligned} &\underset{x_{ij},i,j\in\mathbf{N}}{\text{minimize}} && \sum_{i\in\mathbf{N}}\sum_{j\in\mathbf{N}} x_{ij}\varphi\left(u_i,v_j\right) \\ &\text{subject to} && x_{ij}\geq 0, i, j\in\mathbf{N} \\ & && \sum_{i\in\mathbf{N}} x_{ij}=\sum_{j\in\mathbf{N}} x_{ij}=1 \end{aligned}$$

Problem RT is a linear assignment problem, where every value of *U* is assigned to one of *V* without overlap (Burkard & Dragoti-Cela, 1999; Burkard *et al*., 2009). Since its computational complexity is O(*n*), it is solvable in a linear time. The rearrangement transformation reduces the difference between *U* and *V* to

$$D_R\left(U,V\right)=\sum_{i\in\mathbf{N}} x_{ij}\varphi\left(u_i,v_i\right) \quad . \tag{41}$$

The moving transformation is not defined for categorical variables because a partial transfer of a variable is not meaningful. The addition/deletion transformation defined for numerical variables corresponds to the *replacement transformation* that changes u_i into v_i in every region. The replacement transformation completely removes the difference between *U* and *V*.

Using the above difference measures, we can evaluate the spatial and non-spatial difference separately in categorical variables. The *location* and *attribute* differences are defined as

$$d_L\left(U,V\right)=D_O\left(U,V\right)-D_R\left(U,V\right) \tag{42}$$

and

$$d_A\left(U,V\right)=D_R\left(U,V\right), \tag{43}$$

respectively. The former indicates the spatial difference while the latter is the non-spatial difference.

The distance measure is calculated based on the rearrangement transformation. It is defined as

$$\lambda_R\left(U,V\right)=\frac{n\sum_{i\in\mathbf{N}}\sum_{j\in\mathbf{N}} x_{ij}\left|\mathbf{z}_i-\mathbf{z}_j\right|}{\sum_{i\in\mathbf{N}}\sum_{j\in\mathbf{N}}\left|\mathbf{z}_i-\mathbf{z}_j\right|} .$$

3. Applications

To test the validity of the methods proposed in the previous section, this section applies it to the analysis of three datasets of different sizes. The computation was done on a machine with an Intel Core i7-2620M 2.70GHz processor and 8 GB memory running under Windows 7 Professional. We used a mathematical programming software NUOPT ver.14 (Mathematical Systems Inc.) for solving Problem MT.

3.1 A small dataset

We first apply the method to the comparison of variables shown in Figure 1. A primary objective is to understand the properties and behavior of the difference and distance measures.

We start with examining difference measures. We can completely transform U into V_{11} or V_{12} by only the rearrangement transformation. Consequently, $\delta_L(U, V)$=1 and $\delta_C(U, V)$=$\delta_V(U, V)$=0 as seen in Table 1. The rearrangement transformation, on the other hand, does not work at all for transforming U into V_{13}, V_{14}, V_{23}, or V_{24} as indicated by $\delta_L(U, V)$=0 in Table 1. The moving transformation is more broadly effective than the rearrangement transformation. The contrast is very clear in the cases of V_{13} and V_{14}, where the rearrangement transformation completely fails while the moving transformation totally removes the difference between the variables. We can confirm this by $\delta_L(U, V)$=0 and $\delta_C(U, V)$=1 in these cases. The moving transformation is not at all effective only in the transformation of U into V_{23} or V_{24}, where $\delta_C(U, V)$=0. Though $\delta_C(U, V)$=0 holds also for V_{31} and V_{32}, this does not imply that the moving transformation is ineffective in these cases. It is because the moving transformation cannot further reduce the difference between variables after the rearrangement transformation is applied. In other words, the rearrangement and moving transformations are equally effective for transforming U into V_{31} and V_{32}. The addition/deletion transformation completely removes the difference between variables in any case. It works even if both the rearrangement and moving transformation are ineffective such as the cases of V_{23} and V_{24}.

Table 1 Difference and distance measures calculated for variables shown in Figure 1.

	V_{11}	V_{12}	V_{13}	V_{14}	V_{21}	V_{22}	V_{23}	V_{24}	V_{31}	V_{32}	V_{33}	V_{34}
$D_O(U, V)$	98	6	10	8	62	86	90	40	68	85	51	36
$D_R(U, V)$	0	0	10	8	12	12	90	40	30	37	9	36
$D_M(U, V)$	0	0	0	0	0	0	90	40	30	37	9	26
$D_{AD}(U, V)$	0	0	0	0	0	0	0	0	0	0	0	0
$\delta_L(U, V)$	1.000	1.000	0.000	0.000	0.806	0.860	0.000	0.000	0.559	0.565	0.824	0.000
$\delta_C(U, V)$	0.000	0.000	1.000	1.000	0.194	0.140	0.000	0.000	0.000	0.000	0.000	0.278
$\delta_V(U, V)$	0.000	0.000	0.000	0.000	0.000	0.000	1.000	1.000	0.441	0.435	0.176	0.722
$\lambda_R(U, V)$	1.477	0.153	0.000	0.000	1.125	1.296	0.617	0.774	0.909	1.175	0.927	0.356
$\lambda_M(U, V)$	2.653	1.000	1.200	1.000	1.903	2.698	0.000	0.000	1.529	1.600	1.373	0.389
$\gamma_M(U, V)$	1.300	0.030	0.060	0.040	0.590	1.160	0.000	0.000	0.442	0.834	0.335	0.062

We then turn to the distance measures. Though both V_{11} and V_{12} can be obtained from U by only the rearrangement transformation, they are different in the spatial arrangement of values. The peak of U is closer to that of V_{12} than that of V_{11}. Transformation from U into V_{11} requires the relocation of longer distance, and consequently, $\lambda_R(U, V_{11})>\lambda_R(U, V_{12})$ as seen in Table 1. It also applies to the moving transformation that is reflected by $\lambda_M(U, V_{11}) >\lambda_M(U, V_{12})$.

When the moving transformation is completely ineffective, Problem MT does not yield a valid solution. The distance measure $\lambda_M(U, V)$ inevitably becomes zero as seen in $\lambda_M(U, V_{23})$ and $\lambda_M(U, V_{24})$. On the other hand, $\lambda_R(U, V)$ does not necessarily become zero when the rearrangement transformation fails. The rearrangement transformation relocates U values so that their ranks coincide those of V in every cell, even if it does not reduce the differences between variables at all. This results in $\delta_L(U, V)=0$ and $\lambda_R(U, V)>0$ as seen in the cases of V_{23} and V_{24}.

3.2 A larger dataset

We then apply the method to the comparison of a larger dataset shown in Figure 4. Our objective is to examine the validity of the visualization method of moving transformation. To focus on the spatial difference, we set the variables to have the same total volume. Variables U, V_1, V_3, V_5, and V_6 are unimodal, that is, they all has only a single peak. Variable V_4 has two peaks, which is often called bimodal. Variable V_2 has a doughnut-shaped distribution.

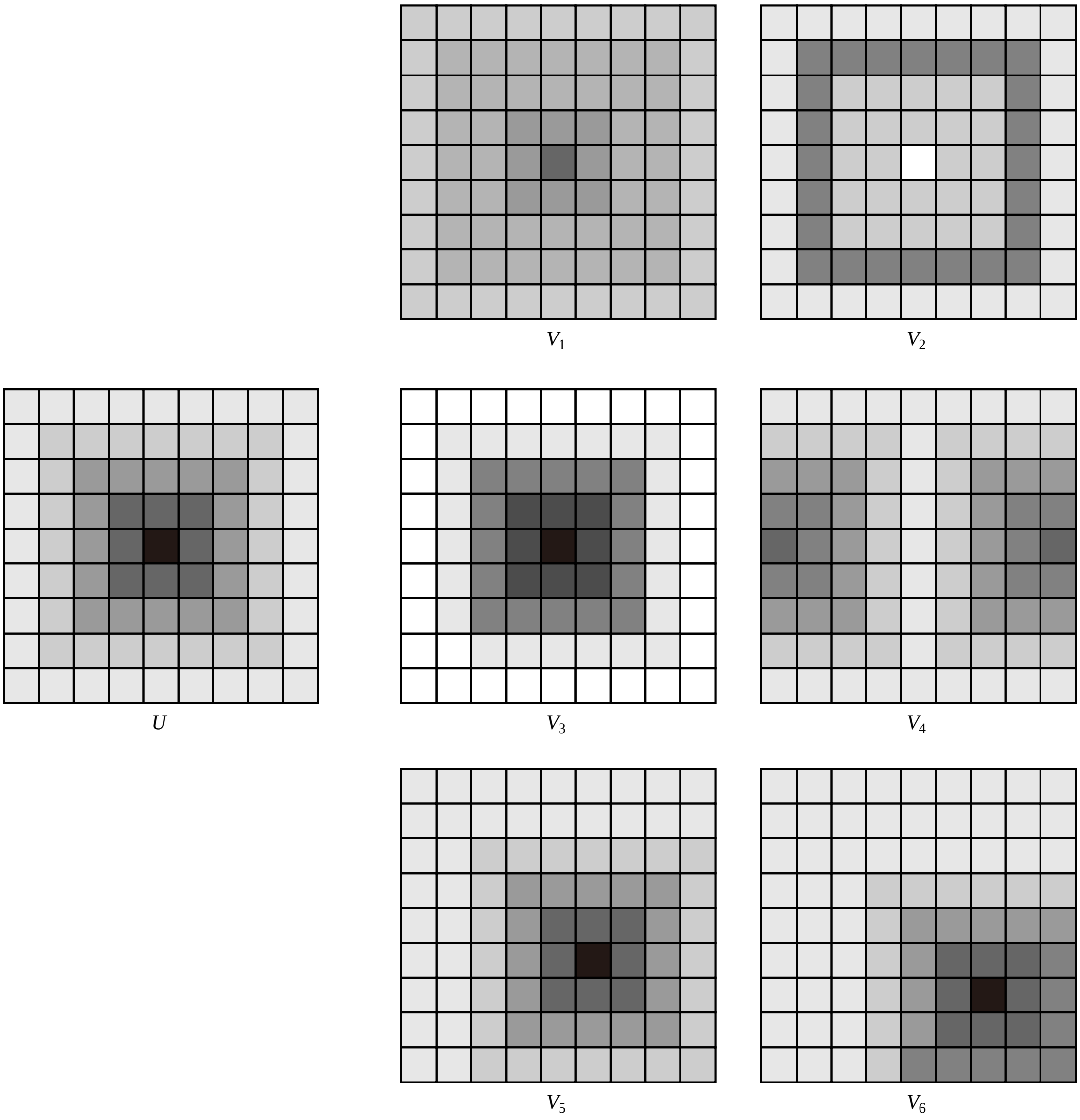


Figure 4. A larger dataset. The values of variables are indicated by gray shades.

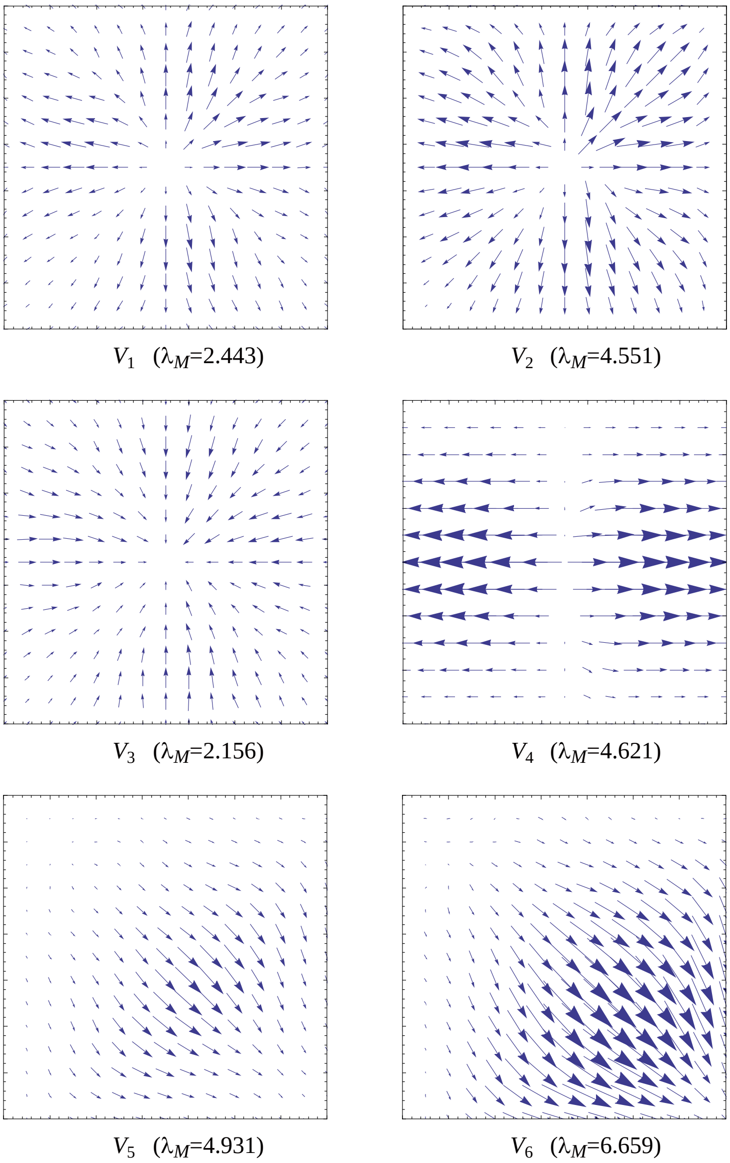
Figure 5 visualizes the moving transformation by the vector plot. The direction and length of arrows indicate the direction and volume of moving transformation, respectively.

Both U and V_1 are unimodal whose peak is located at the center of the region. Since the slope is gentler in the latter, transformation of U into V_1 can be regarded as a slight collapse of a unimodal distribution. This is clearly visualized in Figure 5, where all the arrows are pointing outward from the center.

The moving transformation of U into V_2 looks quite similar in Figure 5, though the distribution of V_2 is quite different from that of V_1. This implies that the transformation into V_2 lies on an extension of that into V_1. It is reflected in longer arrows and the larger value of the distance measure λ_M.

Transformation into V_3 is the centralization of U from the surrounding to the center. Transformation into V_4 is a morphological change from a unimodal to a bimodal distribution. Arrows clearly visualize these transformations. The distance measure is larger in the latter case because the change is more substantial. Transformations into V_5 and V_6 are both the shift of a peak. Since the peak moves further in V_6, arrows are longer and the distance measure is larger.

The moving transformation helps us to grasp intuitively the spatial difference between variables. Besides the vector plot, streamlines are also effective as seen later in the following subsection.



V_1 (λ_M=2.443)

V_2 (λ_M=4.551)

V_3 (λ_M=2.156)

V_4 (λ_M=4.621)

V_5 (λ_M=4.931)

V_6 (λ_M=6.659)

Figure 5. Visualization of moving transformation that changes *U* into *V* in Figure 4. The direction and length of arrows indicate the direction and volume of moving transformation, respectively.

3.3 A real dataset

We finally apply the proposed method to the analysis of a real dataset to test its practical feasibility. We analyzed the change of population distribution from 1970 to 2005 in Chiba Prefecture, Japan. Figure 6 shows the population distribution in 2005 and the railway network in Chiba. The data is obtained as a 1km resolution raster dataset that consists of 80*150 cells. Since Chiba is adjacent in the east of Tokyo Metropolis, its population density is higher in the northwestern area that is densely inhabited by people working in Tokyo.

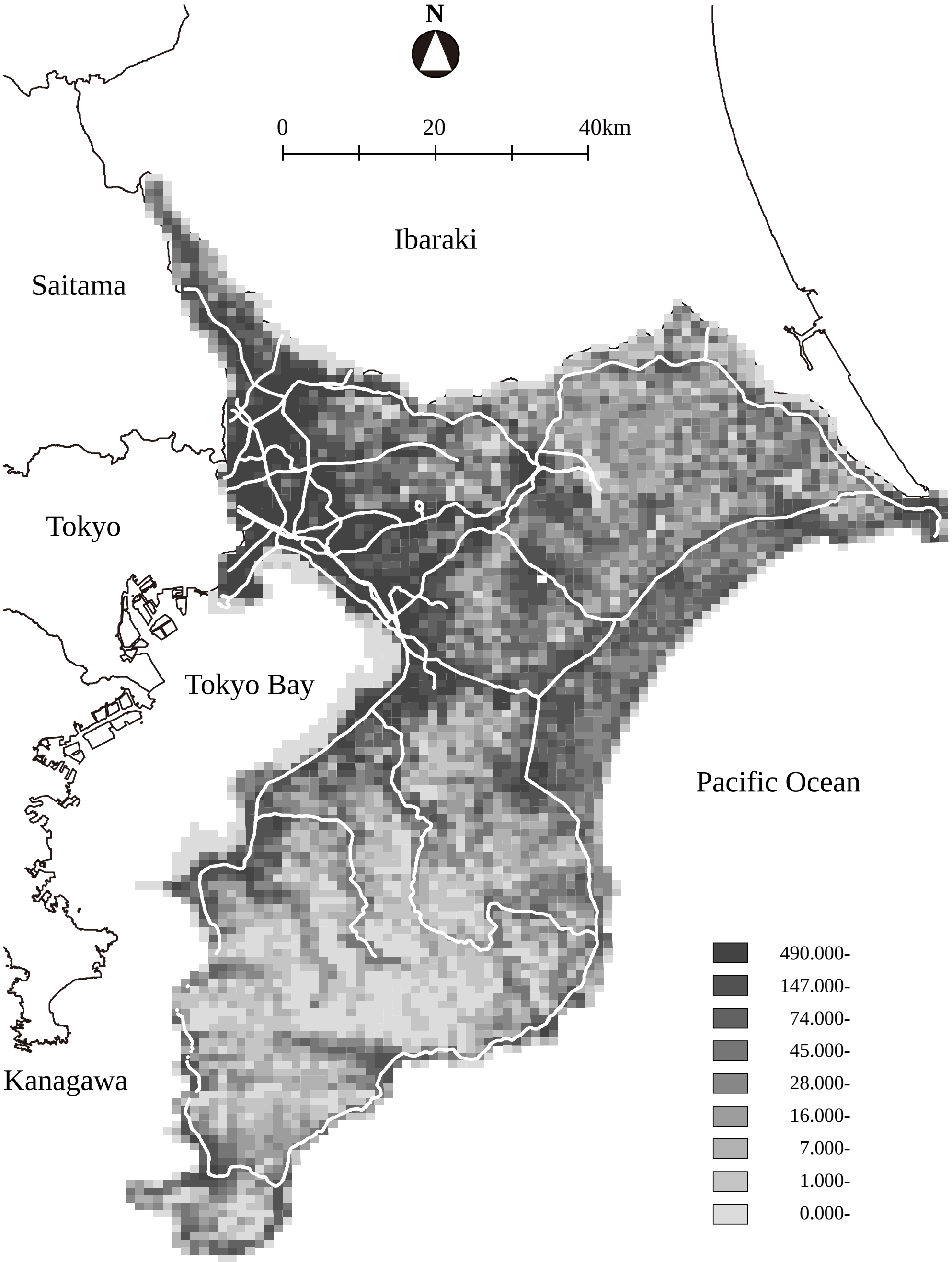


Figure 6. Population distribution in Chiba prefecture, Japan in 2005. Chiba is adjacent to Tokyo, Saitama, and Ibaraki prefectures. White lines indicate railway lines.

Figure 7 shows the change of population distribution from 1970 to 2005. To compare the spatial structure of the distributions between different years, we standardized them so that their summation is all equal. Population density is higher in the northwestern area all through the period. Population constantly decreases in the south area, while its change is not so clear in other areas.

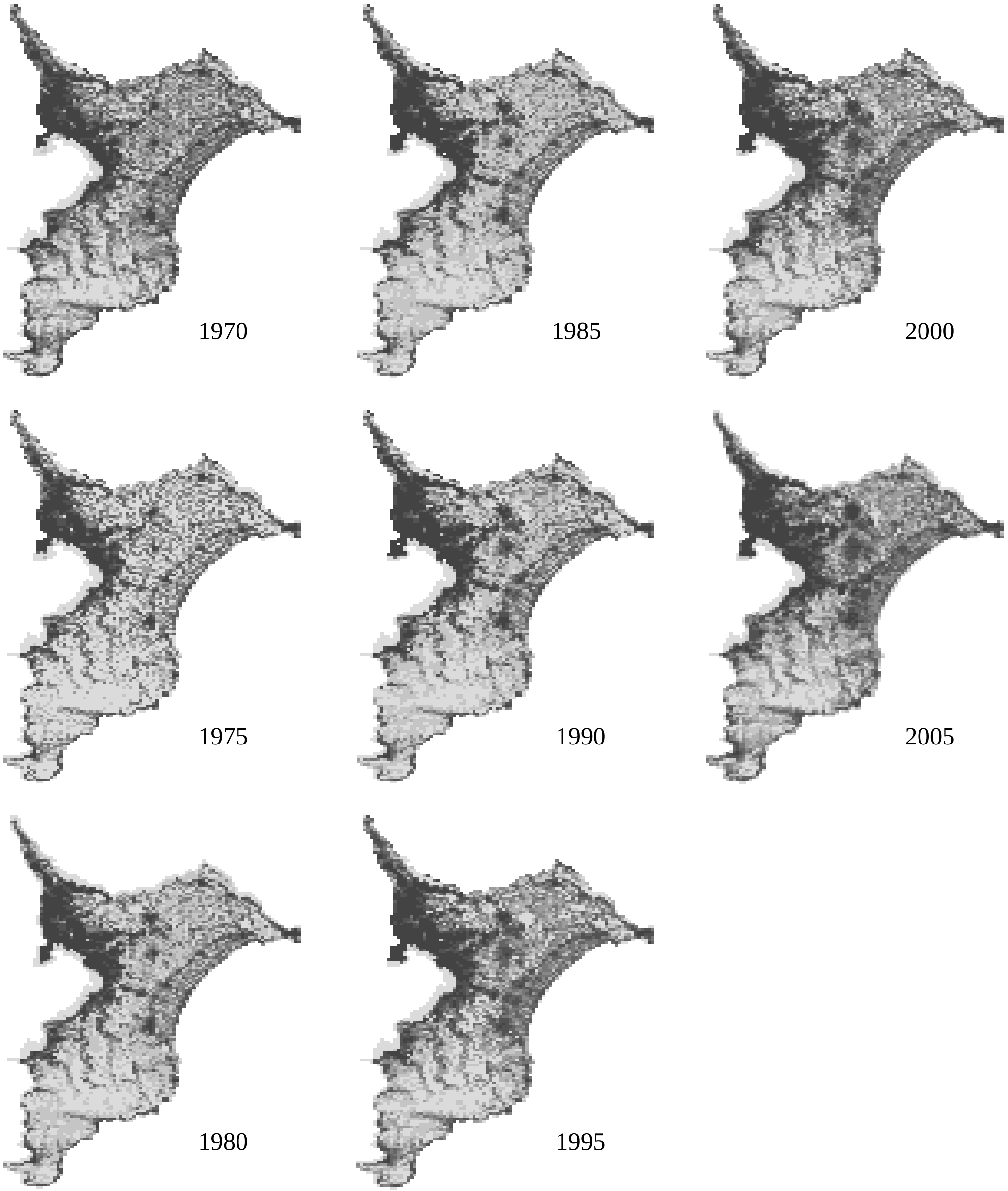
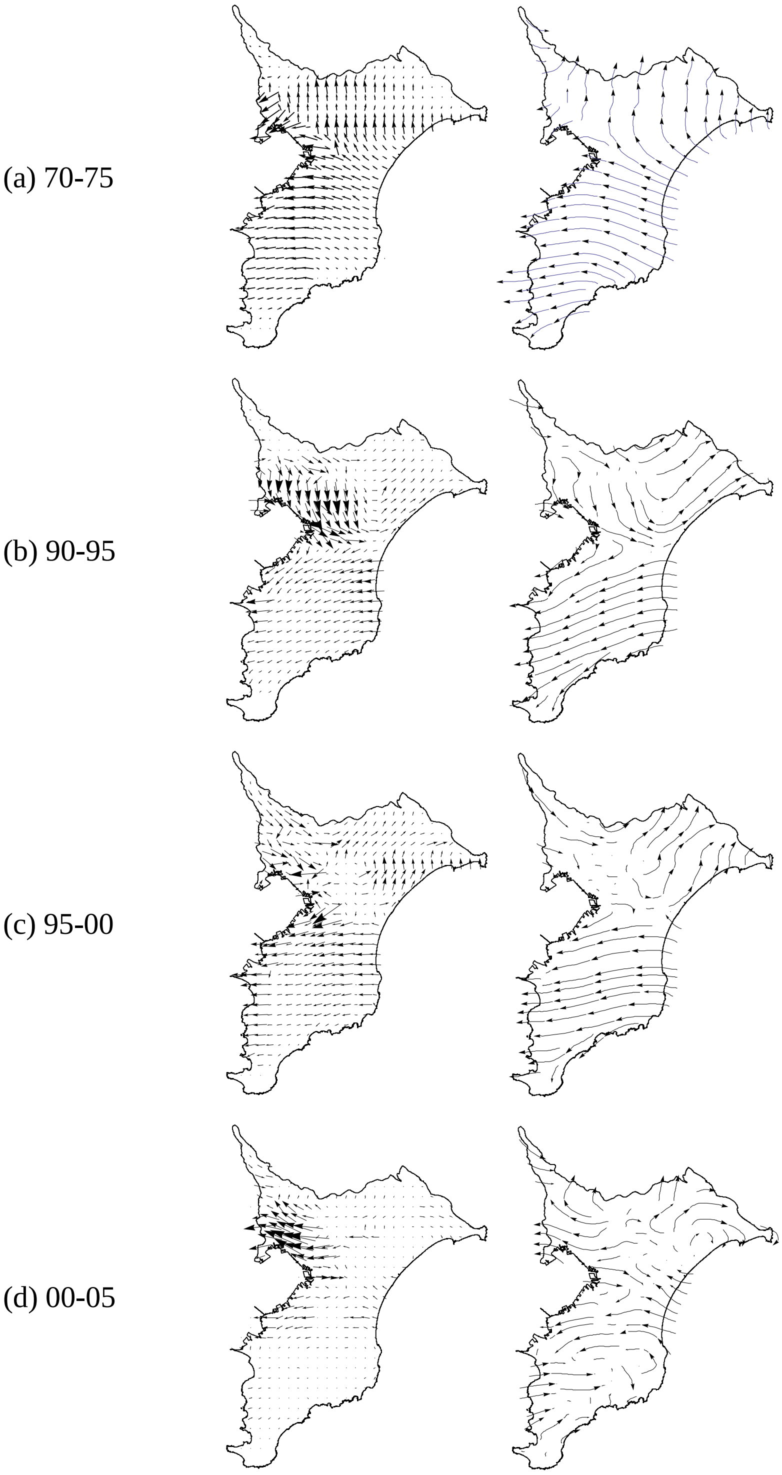


Figure 7. The change of population distribution in Chiba from 1970 to 2005. The figures are comparable with each other since they show the relative distribution of population in each year.

We analyzed the change of population distribution by comparing the distributions in every five-year period. The result is shown in Figure 8 and Table 2. The former visualizes the moving transformation while the latter summarizes the difference and distance measures. Since the moving transformation is very similar in the 70-75, 75-80, 80-85, and 85-90 periods, we present only that of ‘70-75’ (Figure 8).


(a) 70-75

(b) 90-95

(c) 95-00

(d) 00-05

Figure 8. Vector plots and streamlines representing the change in population distribution in five-year periods.

Table 2 Difference and distance measures of five-year periods.

	70-75	75-80	80-85	85-90	90-95	95-00	00-05
$D_O(U, V)$	0.298	0.297	0.186	0.129	0.097	0.070	0.089
$D_R(U, V)$	0.082	0.098	0.055	0.044	0.031	0.014	0.018
$D_M(U, V)$	0.000	0.000	0.000	0.000	0.000	0.000	0.000
$D_{AD}(U, V)$	0.000	0.000	0.000	0.000	0.000	0.000	0.000
$\delta_L(U, V)$	0.723	0.672	0.705	0.659	0.679	0.797	0.799
$\delta_C(U, V)$	0.277	0.328	0.295	0.341	0.321	0.203	0.201
$\delta_V(U, V)$	0.000	0.000	0.000	0.000	0.000	0.000	0.000
$\lambda_R(U, V)$	0.441	0.329	0.293	0.301	0.268	0.247	0.300
$\lambda_M(U, V)$	0.438	0.394	0.352	0.247	0.247	0.230	0.243
$\gamma_M(U, V)$	1.258	0.722	0.648	0.581	0.599	0.402	0.516

The Tokyo suburban area rapidly expanded from the 1960s to 1980s. During this period, population increased primarily in two areas in Chiba. One is the northwestern area that is adjacent to Tokyo Metropolis. Arrows are facing west in this area as seen in Figure 8a. The other is the northern end where a railway line is running directly connected to Tokyo. In this area arrows are facing to the railway line running along the northern boundary of Chiba.

The expansion of Tokyo began to slow down from 1991 when an economic bubble collapsed. In parallel with this, new towns have been built in the north-central area of Chiba, especially from 1990 to 1995. Population increased in this area more rapidly than in the other areas, and consequently, arrows in this area were turned around to the south as seen in Figure 8b.

The new towns, however, could not attract people as expected. Moreover, the collapse of economic bubble caused a drastic drop in land prices in Tokyo. As a result, people returned to the central area of Tokyo and its close suburbs. Figure 8c visualizes the transitional phase where the vector field looks complicated in the center of Chiba. Figure 8d, on the other hand, clearly indicates that people returned to Tokyo from 2000 to 2005.

Let us then look at Table 2 in comparison with Figure 8. Population distribution drastically changed in the 70-75 period as indicated by long arrows in Figure 8a. This is reflected as large values of the difference and distance measures during this period in Table 2. Conversely, the measures show the smallest values in the 95-00 period when the change in population distribution is not significant as seen in Figure 8c.

5. Conclusion

This paper proposed a new method of comparing numerical variables defined in a region. The method introduced three types of transformations called the rearrangement, moving, and addition/deletion transformations. The transformations convert one variable so that it fits the other as well as possible. The result provides a basis for evaluating the differences between variables in terms of spatial and non-spatial dimensions separately. The distance measures permit us to investigate the spatial difference between variables in more detail. To test the validity of the method, the paper applied it to an analysis of three spatial datasets of different sizes. The result indicated that the method is effective for evaluating and visualizing the difference between variables.

We finally discuss some limitations and extensions of the paper for future research.

First, we should extend the method to compare variables defined on different spaces. For comparing continuous variables, numerous methods have been proposed in image analysis and mathematical morphology (Serra, 1984, 1988; Goutsias & Heijmans, 2000; Shih, 2009l; Soille, 2010). Concerning discrete variables, statistical tests are available such as the *t*-test and the Wilcoxon-Mann-Whitney test. However, these methods do not explicitly consider the spatial and non-spatial differences separately. Extension to this direction should be further examined.

Second, this paper considers three aspects of the difference between variables: location, configuration, and volume. The difference, however, can also be evaluated from other perspectives. Spatial analysis, for instance, often focuses on the topological structure of the distribution of numerical variables (Warntz, 1966; Warntz & Waters, 1975; Okabe & Masuda, 1984). Difference in arrangement of local maxima and minima has also drawn the attention of geographers. The difference between variables should be discussed from a wider variety of perspectives.

Third, comparison of more than two variables is also an important topic. Though we can analyze multiple variables simultaneously by applying our method to every pair of variables, it is time consuming and the integration of the obtained results is not straightforward. Wartenberg (1985) develops a measure of multivariate spatial correlation by borrowing the scheme of principal factor components analysis. We should extend our method along with this line.

References

Burkard, R. & Dragoti-Cela, E. (1999). Linear assignment problems and extensions. In P. M. Pardalos, D.-Z. Du, & R. L. Graham (Eds.), *Handbook of Combinatorial Optimization, Supplement Volume A* (pp. 75-149). Berlin: Springer.

Burkard, R., Dell'Amico, M., & Martello, S. (2009). *Assignment Problems*. Philadelphia: Society for Industrial and Applied Mathematics.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37-46.

Goutsias, J. & Heijmans, H. J. A. M. (2000). *Mathematical morphology*. Fairfax, VA: IOS Press.

Haining, R. (1991). Bivariate correlation with spatial data. *Geographical Analysis*, 23, 210-227.

Hubert, L. J., Golledge, R. G., Constanzo, C. M., & Gale, N. (1985). Measuring association between spatially defined variables: an alternative procedure. *Geographical Analysis*, 17, 36-46.

Kullback, S. & Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, 22, 79-86.

Kullback, S. (1959). *Information theory and statistics*. New York: Wiley.

Lee, S-I. (2001). Developing a bivariate spatial association measure: an integration of Pearson's *r* and Moran's *I*. *Journal of Geographical Systems*, 3, 369-385.

Okabe, A. & Masuda, S. (1984). Qualitative analysis of two-dimensional urban population distributions in Japan. *Geographical Analysis*, 16, 301-312.

Peleg, S., Werman, M., & Rom, H. (1989). A unified approach to the change of resolution: space and gray-level. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11, 739-742.

Pontius, R. G. Jr. (2000). Quantification error versus location error in comparison of categorical maps. *Photogrammetric Engineering & Remote Sensing*, 66, 1011-1016.

Pontius, R. G. Jr. (2002). Statistical methods to partition effects of quantity and location during comparison of categorical maps at multiple resolutions. *Photogrammetric Engineering & Remote Sensing*, 68, 1041-1049.

Pontius, R. G. Jr. & Millones, M. (2011). Death to Kappa: birth of quantity disagreement and allocation disagreement for accuracy assessment. *International Journal of Remote Sensing*, 32, 4407-4429.

Rubner, Y., Tomasi, C., & Guibas, L. J. (2000). The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40, 99-121.

Serra, J. (1984). *Image processing and mathematical morphology, Volume 1*. New York: Academic Press.

Serra, J. (1988). *Image processing and mathematical morphology, Volume 2: Theoretical Advances*. New York: Academic Press.

Shih, F. Y. (2009). *Image processing and mathematical morphology: Fundamentals and applications*. Boca Raton: CRC Press.

Soille, P. (2010). *Morphological image analysis: principles and applications*. Berlin: Springer.

Stephane, D., Sonia, S., & Francois, D. (2008). Spatial ordination of vegetation data using a generalization of Wartenberg's multivariate spatial correlation. *Journal of Vegetation Science*, 19, 45-56.

Tjøstheim, D. (1978). A measure of association for spatial variables. *Biometrika*, 65, 109-114.

Tobler, W. (1987). An experiment in migration mapping by computers. *The American Cartographer*, 14, 155-163.

Tobler, W. (1995). Migration: Ravenstein, Thorntwaite, and beyond. *Urban Geography*, 16, 327-343.

Warntz, W. (1966). The topology of a socioeconomic terrain and spatial flows. *Papers in Regional Science*, 17, 47-61.

Warntz, W. & Waters, N. (1975). Network representations of critical elements of pressure surface. *Geographical Review*, 65, 476-492.

Wartenberg, D. (1985). Multivariate spatial correlation: a method for exploratory geographical analysis. *Geographical Analysis*, 17, 263-283.

Wong, D. W. (2011). Exploring spatial patterns using an expanded spatial autocorrelation framework. *Geographical Analysis*, 43, 327-338.

Zhao, Q., Yang, Z., & Tao, H. (2010). Differential earth mover's distance with its applications to visual tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32, 274-287.