Computers, Environment and Urban Systems 41 (2013) 65–74

Contents lists available at SciVerse ScienceDirect

Computers, Environment and Urban Systems

journal homepage: www.elsevier.com/locate/compenvurbsys

A method for comparing numerical variables defined in a region

Yukio Sadahiro *

Center for Spatial Information Science, University of Tokyo, 5-1-5, Kashiwanoha, Kashiwa-shi, Chiba 277-8568, Japan

ARTICLE INFO

Article history:
Received 13 September 2012
Received in revised form 13 March 2013
Accepted 13 March 2013
Available online 4 June 2013

Keywords:
Comparison
Transformation
Numerical variables

ABSTRACT

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

1. Introduction

This paper proposes a method for 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 by two-dimensional space, though not as smoothly as temperature distribution because they are often calculated by aggregating point data in spatial units. Traffic flow is a numerical variable defined by network space. The height of trees, the age of buildings, and the annual sales of supermarkets are numerical variables defined by discrete space.

There are several ways to compare 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 to each other. The Kullback–Leibler divergence (Kullback, 1959; Kullback & Leibler, 1951) is also useful to compare positive variables. If variables are defined by 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 a given spatial dimension (Haining, 1991; Hubert, Golledge, Constanzo, & Gale, 1985; Lee, 2001). There are two prominent groups of methods in the literature that incorporate the spatial aspect explicitly.

One group extends 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 (Lee, 2001; Stephane, Sonia, & Francois, 2008; Wartenberg, 1985), while others develop new measures of spatial autocorrelation (Haining, 1991; Hubert et al., 1985; Tjøstheim, 1978).

Another class of methods uses the earth mover's distance (Peleg, Werman, & Rom, 1989; Rubner, Tomasi, & Guibas, 2000; Zhao, Yang, & Tao, 2010). These methods consider the turning of a pile of dirt into another form with the least cost. This process 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 employed to compare numerical variables defined by discrete space. Consequently, they are not directly applicable to the analysis of numerical variables defined by 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 aspatial difference that exists in the original variables. Standardization prevents us from separating the differences in spatial and aspatial dimensions.

There are several papers that discuss the separation of spatial and aspatial factors, but their focus is not on the comparison of numerical variables. Pontius (2000, 2002) and Pontius and 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 aspatial factors permits us to deepen our understanding of the structure of spatial phenomena. Following

* Tel.: +81 471364310; fax: +81 471364292.
E-mail address: sada@csis.u-tokyo.ac.jp

0198-9715/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.compenvurbsys.2013.03.003

the line of these papers, this paper aims to evaluate the difference between numerical variables separately in spatial and aspatial 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 of the method used 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 $\mathbf{R} = \{R_1, R_2, \ldots, R_n\}$ ($\mathbf{N} = \{1, 2, \ldots, n\}$) in each of which two sets of numerical variables $U = \{u_1, u_2, \ldots, u_n\}$ and $V = \{v_1, v_2, \ldots, v_n\}$ 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 the *overall difference* given by

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

Though this measure is easy to calculate and understand, it does not recognize the spatial difference between variables as general statistical measures. In Fig. 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 does not recognize this difference, as it neglects the spatial dimension.

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.

2.1.1. Rearrangement transformation

Rearrangement transformation changes the location of U values so that its spatial distribution is similar to that of V as closely as possible. To this end, it relocates U values in the way that the rank of U coincides with 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:

$$\rho_{ij}(U, V) = \begin{cases} 1 & \text{if } r(u_i) = r(v_j) \\ 0 & \text{otherwise} \end{cases}. \tag{2}$$

Rearrangement transformation reduces the difference between U and V to

$$D_R(U, V) = \sum_{i \in \mathbf{N}} \sum_{j \in \mathbf{N}} \rho_{ij}(U, V) |u_i - v_j| \tag{3}$$

This indicates the summation of the absolute difference of elements between U and V of the same rank. We call the difference between $D_O(U,V)$ and $D_R(U,V)$ the *location difference*:

$$\begin{aligned} d_L(U, V) &= D_O(U, V) - D_R(U, V) \\ &= \sum_{i \in \mathbf{N}} \left(|u_i - v_i| - \sum_{j \in \mathbf{N}} \rho_{ij}(U, V) |u_i - v_j| \right) \end{aligned} \tag{4}$$

The location difference estimates the difference in the location of values of two variables, ranging from 0 to $D_O(U,V)$. 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 Fig. 1, for instance, we can transform U into V_{11} or V_{12} by only changing the location of values. In such a case, the location difference reaches its maximum.

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

$$r(u_i) = r(v_i) \quad \forall i \in \mathbf{N} \tag{5}$$

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

$$u_i \leqslant v_i \quad \forall i \in \mathbf{N}, \tag{6}$$

or

$$u_i \geqslant v_i \quad \forall i \in \mathbf{N} \tag{7}$$

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

$$D_O(U, V) = D_R(U, V). \tag{8}$$

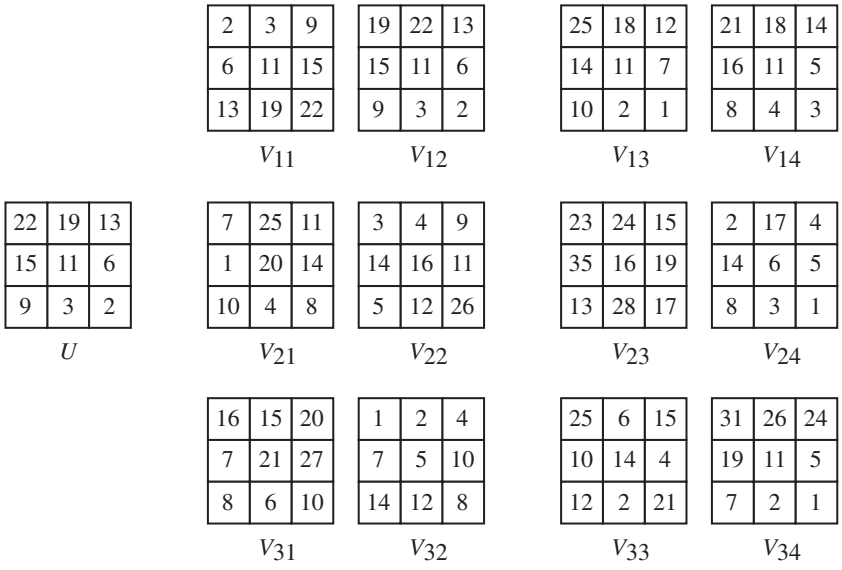

Fig. 1. The distributions of U and V.