Computers, Environment and Urban Systems 45 (2014) 24–33

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

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

Exploratory analysis of time series data: Detection of partial similarities, clustering, and visualization

Yukio Sadahiro [a,*], Tetsuo Kobayashi [b]

[a] Center for Spatial Information Science, University of Tokyo, Japan
[b] Department of Geography, Florida State University, Japan

ARTICLE INFO

Article history:
Received 5 May 2013
Received in revised form 27 January 2014
Accepted 2 February 2014
Available online 6 March 2014

Keywords:
Time series data
Partial similarity
Clustering

ABSTRACT

A new exploratory method for analyzing time series data is proposed. A computational algorithm detects partial similarities between simultaneously occurring time series data and clusters the data into groups based on their similarities. A graphical representation that visualizes the data clustering process helps us understand similarity between time series data and classifies them into smaller subgroups. Numerical measures evaluate the effectiveness of clusters and provide a means for testing their statistical significance. The proposed method was applied to an analysis of the change of population distribution during a day in Salt Lake County, Utah, USA. This application supports the technical soundness of the method and provides empirical findings.

1. Introduction

Exploratory analysis of time series data is an important topic in urban and environmental analysis. Graphical visualization of time series data is helpful in identifying and interpreting the relationships between data, for example, the relationship between the economic environment, labor market, demographic situation, and migration patterns (Bronars & Jansen, 1987; Gauthier, Tanaka, & Smith, 1992; Lane, 2010; Lane, 2012; Mielke, Relethford, & Eriksson, 1994; Unwin A., 1996). Comparison of periodic patterns, such as air pollution over a day and vegetation type over a year, helps us investigate their autocorrelation, detect anomalies, and predict unexpected changes (Nowak, Crane, & Stevens, 2006; Viovy, 2000; Xia, Rui, Bing, & Qingxi, 2008).

Classification is an effective and powerful method of exploratory analysis (Antunes & Oliveira, 2001; Gaber, Zaslavsky, & Krishnaswamy, 2005; Liao, 2005; Xing, Pei, & Keogh, 2010). Classification of time series remote sensing data permits us to generate highly accurate land cover maps (Gray & Song, 2013; Okabe & Masuyama, 2001; Petitjean, Inglada, & Gancarski, 2012; Verhoef, Meneti, & Azzali, 1996). Climatologists employ cluster analysis to find climatologically homogeneous geographical regions (Bengtsson & Cavanaugh, 2008; Fovell & Fovell, 1993; Gong & Richman, 1995; Richman & Lamb, 1985). Classification of temporal water quality patterns allows us to delineate homogeneous regions in an ecosystem and optimize the location of water monitoring sites (Henderson, 2006; Ignaccolo, Ghigo, & Giovenali, 2008; Pastres, Pastore, & Tonellato, 2011). In addition, the classification and visualization of traffic accident patterns are useful to detect region-specific causes of accidents (Lavrač, Jesenovec, Trdin, & Kosta, 2008).

There are two types of classification methods that are applicable to time series data. Whole matching methods evaluate the overall similarity between the data during the same time period (Keogh, Chakrabarti, Mehrotra, & Pazzani, 2001; Keogh & Lin, 2005), and subsequence matching methods classify time series data based on their partial similarities (Chen J. R., 2007; Denton, Besemann, & Dorr, 2008; Fu, Chung, Luk, & Ng, 2003; Goldin, Mardales, & Nagy, 2006; Keogh & Lin, 2005). Fig. 1 illustrates the difference between these methods. Whole matching methods classify only T_1 and T_2 into the same group based on their overall similarity, and subsequence matching methods detect partial similarities between the data, which are indicated by bold lines, and cluster T_1, T_2, T_3, and T_4 as one group. If scaling along the vertical axis is permitted, subsequence matching methods can find the similarity between T_1 and T_5, which is indicated by solid bold lines and dotted lines in Fig. 1.

Whole matching methods run faster; however, subsequence matching methods are capable of detecting a wider variety of partial similarities in temporal patterns. One drawback of the latter is the increased computational complexity due to their flexibility in the temporal dimension. They permit time lags between similar patterns, as can be seen for T_1 and T_4 in Fig. 1. However, this flexibility is not always essential in geography, ecology, and climatology because similar patterns at different times often have

* Corresponding author. Tel.: +81 358416273; fax: +81 471364310.
E-mail addresses: sada@ua.t.u-tokyo.ac.jp, sada@csis.u-tokyo.ac.jp (Y. Sadahiro).

http://dx.doi.org/10.1016/j.compenvurbsys.2014.02.001
0198-9715/

different meanings. For example, climatology discriminates rainfall peaks for different seasons. The comparison of temporal patterns is performed within the same or close time period across different locations. In consideration of their computational cost, subsequence matching methods are too flexible to serve this purpose.

To resolve this problem, this paper proposes a new method for analyzing time series data. The method clusters time series data into similar groups based on their partial similarities within the same time period. This improves the computational efficiency at the expense of flexibility in the temporal dimension. The proposed method classifies T_1, T_2, and T_3 into the same group in Fig. 1 while it discriminates T_4.

In this paper, time series data is referred to as *trends* for simplicity. The remainder of this paper is organized as follows. Section 2 proposes an exploratory method for analyzing trends. Section 3 applies the method to the analysis of changing population distribution during a day in Salt Lake County, Utah, USA, and Section 4 summarizes the findings from the case study and provides a discussion.

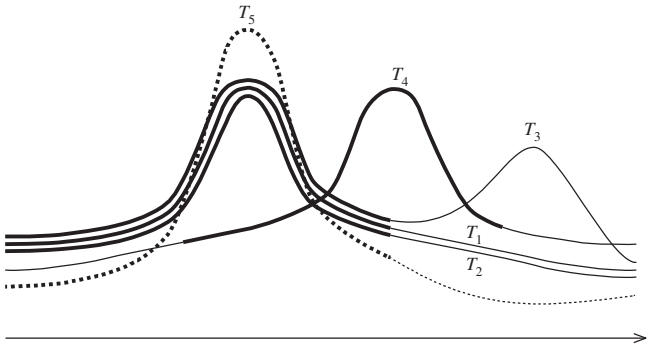

Fig. 1. Evaluation of similarity between time series data. Whole matching methods classify only T_1 and T_2 into the same group. Subsequence matching methods detect partial similarities between the data, which are indicated by bold lines, and cluster T_1, T_2, T_3, and T_4 as a group. If the scaling along the vertical axis is permitted, subsequence matching methods even find the similarity between T_1 and T_5, which is indicated by bold solid and dotted lines.

2. Method

Suppose a set of M trends during the time period $[\tau_S, \tau_E]$. The ith trend is denoted by T_i and is expressed as the numerical function $f_i(t)$ $(t \in [\tau_S, \tau_E])$.

2.1. Preprocessing

The *Neighborhood* of T_i, denoted by N_i, is the buffer area bound by the following four functions:

$$y_{iU}(t) = f_i(t) + b, \tag{1}$$

$$y_{iL}(t) = f_i(t) - b, \tag{2}$$

$$t = \tau_S, \tag{3}$$

and

$$t = \tau_E, \tag{4}$$

where b is the buffer distance. Eqs. (1)–(4) define the upper, lower, left, and right bounds of N_i, respectively. Fig. 2a shows the neighborhoods of trends T_1, T_2, T_3 and T_4. Note that the neighborhood is not identical to the buffer area usually generated in GIS packages. The boundaries of neighborhoods are given by vertically shifted trend functions and the vertical lines at τ_S and τ_E.

The intersection of all neighborhoods yields fragmented small polygons, as is shown in Fig. 2b. This paper calls them *polygons* and denotes as $\Lambda = \{P_1, P_2, \ldots, P_K\}$. The earliest and latest times of polygon Q are given by $t_S(Q)$ and $t_E(Q)$, respectively. The *length* of Q is defined as

$$l(Q) = t_E(Q) - t_S(Q). \tag{5}$$

Trends are regarded as similar if their neighborhoods overlap. The length of an overlap indicates the degree of similarity between trends. In Fig. 2a, for instance, T_1 is more similar to T_2 than T_3, which is represented by the length of their overlaps, i.e., the length of the neighborhood overlap for T_1 and T_2 is longer than that of T_1 and T_3. The number of neighborhoods overlapping on a polygon indicates the number of similar trends associated with the polygon. Polygons of dark shades in Fig. 2a suggest that many trends are partially similar around the polygons.

A *center* is a long set of adjacent polygons on which many neighborhoods overlap. Centers represent partial similarities between trends. The minimum length and minimum number of neighborhoods are given as $L_{\min}$ and α a priori.

Any neighborhood can be represented as a unique set of polygons. Let ϑ_i be a set of polygons that comprise neighborhood N_i. The set of neighborhoods is denoted by $\vartheta = \{\vartheta_1, \vartheta_2, \ldots, \vartheta_M\}$. The number of elements and the jth element in the set ϑ_i are denoted by $\#(\vartheta_i)$ and $e(\vartheta_i, j)$, respectively.

2.2. Detection of centers

Since centers represent partial similarities between trends, their detection allows us to cluster trends based on these partial similarities. To this end, this subsection proposes a computational algorithm. The detection of a center is described as an expansion process of a set of polygons. We first choose a polygon shared by the most neighborhoods as an initial set. This set gradually expands by incorporating its longest adjacent polygons for as long as possible. The length of an adjacent polygon is measured by its length outside the set of polygons. Expansion terminates when the set of polygons becomes longer than $L_{\min}$ or the number of its overlapping neighborhoods becomes smaller than α. Once we detect a center, we remove the polygons in the set and repeat the same process until no further center is detected.

Fig. 3 illustrates the detection of a center. We first choose P_1 as an initial set because it is contained in the neighborhoods of all the trends (Fig. 3a). We compare its adjacent polygons P_2 and P_3 by their length outside the set $\{P_1\}$, which are indicated by bold solid lines and dotted lines in Fig. 3b. Since the dotted lines are longer than the solid lines, the set expands to $\{P_1, P_2\}$ (Fig. 3c). We then evaluate its adjacent polygons $\{P_3, P_4, P_5, P_6\}$ by their length outside the set, and choose the longest polygon P_4. Since the set is longer than $L_{\min}$, detection terminates to yield the final results $\{P_1, P_2, P_4\}$ (Fig. 3d). The set of polygons grows from $\{P_1\}$ to $\{P_1, P_2\}$ and $\{P_1, P_2, P_4\}$ while its related trends reduce from $\{T_1, T_2, T_3, T_4\}$ to $\{T_1, T_2, T_3\}$ and $\{T_1, T_2\}$.

The set of centers is denoted by $\Omega = \{C_1, C_2, \ldots, C_N\}$, each of which consists of a set of polygons. A trend is said to be *assigned* to C_i if its neighborhood contains all the polygons composing C_i. Let Γ_i be the set of neighborhoods containing center C_i. The set of neighborhood sets is denoted by $\Gamma = \{\Gamma_1, \Gamma_2, \ldots, \Gamma_N\}$. Centers relate trends and polygons. This paper refers to this relationship as an *assignment*. A trend can be assigned to multiple centers since trends are clustered based on their partial similarities.

Fig. 4 shows a computational algorithm for extracting centers. Algorithm Center Extraction (CE) is an extension of Algorithm TE proposed by Sadahiro, Lay, and Kobayashi (2013), which was originally developed by Kharrat, Popa, Zeitouni, and Faiz (2008). Since