Computers, Environment and Urban Systems 67 (2018) 80–96

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

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

Detecting anomalies in spatio-temporal flow data by constructing dynamic neighbourhoods

Yan Shi [a,b], Min Deng [c,*], Xuexi Yang [c], Jianya Gong [a,b]

[a] State Key Laboratory of Information Engineering in Surveying, Mapping & Remote Sensing, Wuhan University, Wuhan, China
[b] Collaborative Innovation Center of Geospatial Technology, Wuhan University, Wuhan, China
[c] Department of Geo-informatics, Central South University, Changsha, Hunan, China

ARTICLE INFO

Article history:
Received 16 January 2017
Received in revised form 23 August 2017
Accepted 23 August 2017
Available online xxxx

Keywords:
Spatio-temporal flow data
Dynamic spatio-temporal flow
Dynamic neighbourhoods
Spatio-temporal flow anomalies

ABSTRACT

In massive spatio-temporal datasets, anomalies that deviate from the global or local distributions are not just useless noise but possibly imply significant changes, surprising patterns, and meaningful insights, and because of this, detection of spatio-temporal anomalies has become an important research hotspot in spatio-temporal data mining. For spatio-temporal flow data (e.g., traffic flow data), the existing anomaly detection methods cannot handle the embedded dynamic characteristic. Therefore, this paper proposes the approach of constructing dynamic neighbourhoods to detect the anomalies in spatio-temporal flow data (called spatio-temporal flow anomalies). In this approach, the dynamic spatio-temporal flow is first modelled based on the real-time attribute values of the flow data, e.g., the velocity of vehicles. The dynamic neighbourhoods are then constructed by considering attribute similarity in the spatio-temporal flow. On this basis, global and local anomalies are detected by employing the idea of the G^* statistic and the problem of multiple hypothesis testing is further addressed to control the false discovery rate. The effectiveness and practicality of our proposed approach are demonstrated through comparative experiments on traffic flow data from the central road network of central London for both weekdays and weekends.

1. Introduction

The wide usage of geo-location sensors and network connectivity makes it easier to capture enormous amounts of spatio-temporal data spanning certain spatial regions over a period of time. Spatio-temporal anomalies are a collection of records that significantly deviate from the global or local distributions with the consideration of non-spatial attributes (Shekhar, Lu, & Zhang, 2001). In most geographical fields, spatio-temporal anomalies are not just useless noises but imply significant changes, surprising patterns, and meaningful insights. Therefore, the detection of spatio-temporal anomalies has become a research hotspot in the field of spatio-temporal data mining (Miller & Han, 2009; Shekhar, Vatsavai, & Celik, 2009; Tan, Steinbach, & Kumar, 2006) and has received more attention in the detection of disease and crime hotspots (Kulldorff, Heffernan, Hartman, Assunção, & Mostashari, 2005; Delmelle, Dony, Casas, Jia, & Tang, 2014; Brunsdon, Corcoran, & Higgs, 2007; Nakaya & Yano, 2010; Shiode & Shiode, 2013; Cheng & Adepejue, 2013), the discovery of climate change (Sun, Xie, Ma, & Jin, 2005; Barua & Alhajj, 2007; Wu, Liu, & Chawla, 2010; Liu, Deng, Wang, Peng, and Mei, 2011; Telang, Deepak, Joshi, Deshpande, & Rajendran, 2014), the monitoring of environment change (Birant & Kut, 2006; Cheng & Li, 2006), the extraction of anomalous trajectories (Ge et al., 2010; Lee, Han, & Li, 2008; Zhang et al., 2011) and the detection of traffic congestion (Li, Han, Lee, & Gonzalez, 2007; Liu, Deng, et al., 2011; Pang, Chawla, Liu, & Zheng, 2011; Chawla, Zheng, & Hu, 2012; Pan, Zheng, Wilkie, & Shahabi, 2013).

The types of spatio-temporal data can vary with different applications. In general, the detection methods of existing spatio-temporal anomalies are mostly designed for spatio-temporal sequence data, e.g., climate spatio-temporal sequences. Spatial and non-spatial attributes are both embedded in the spatio-temporal sequences. Specifically, spatial attributes determine the geographical locations of entities by X-Y coordinates or latitude-longitude, while non-spatial attributes are described by time series. Taking Fig. 1(a) as an example, the entity P_1 can be denoted as $P_1 = (x_1, y_1, P_1 \cdot nsa_{t_1}, P_1 \cdot nsa_{t_2}, \ldots)$, where x_1, y_1 and $P_1 \cdot nsa_{t_1}, P_1 \cdot nsa_{t_2}, \ldots$ represent the spatial location and non-spatial attribute values of P_1, respectively. In the real world, there exists a kind of special spatio-temporal sequence data that has the characteristics of directionality and being dynamic, e.g., traffic flow data from a road network (also called 'spatio-temporal flow data' in this paper). In traffic flow data, the vehicles keep running on the roads, as shown in Fig. 1(b), and the traffic flow in the upstream direction can interact with

* Corresponding author.
E-mail addresses: csu_shiy@csu.edu.cn (Y. Shi), dengmin@csu.edu.cn (M. Deng), studyang@sina.cn (X. Yang), gongjy@whu.edu.cn (J. Gong).

https://doi.org/10.1016/j.compenvurbsys.2017.08.010
0198-9715/

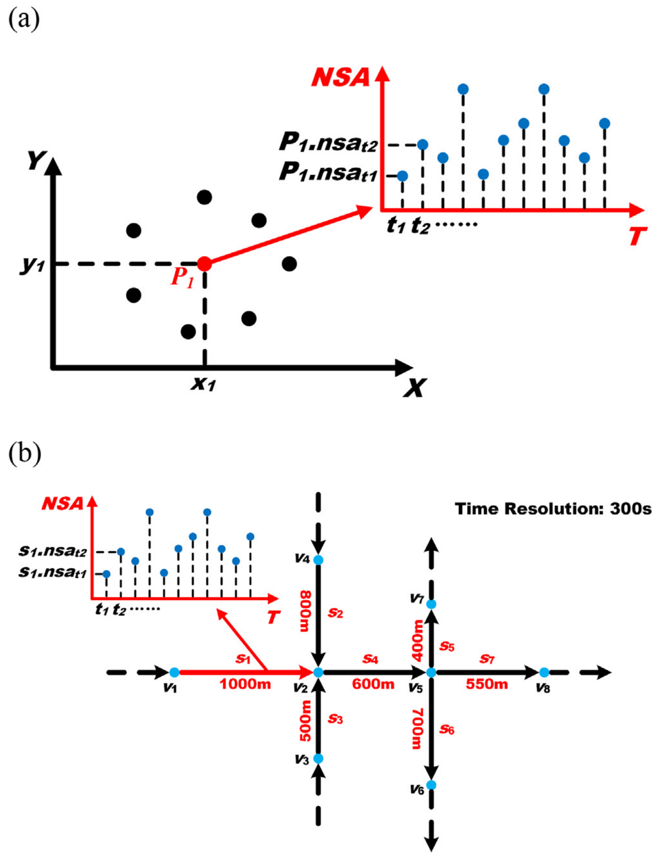

Fig. 1. Two types of spatio-temporal sequence data. (a) Traditional spatio-temporal sequence data; (b) dynamic spatio-temporal flow data, for example, traffic flow data on the road network.

that in the downstream direction. This paper specifically focuses on spatio-temporal flow data and presents the approach of constructing dynamic neighbourhoods to detect spatio-temporal flow anomalies. The major contributions are as follows:

- Constructing dynamic spatio-temporal neighbourhoods.
- Detecting various anomalies by integrating dynamic spatio-temporal neighbourhoods.

The rest of this paper is organized as follows. Section 2 reviews the related work about the detection of spatio-temporal anomalies and proposes our strategy for detecting spatio-temporal flow anomalies. In Section 3, the proposed method for detecting spatio-temporal flow anomalies is fully elaborated. In Section 4, extensive experiments on real-life data are analysed to demonstrate the effectiveness and practicability of the proposed method. The interesting findings are summarized and future research works are highlighted in Section 5.

2. Related work and our strategy for detecting spatio-temporal flow anomalies

2.1. Related work

Hawkins first defined an anomaly as 'an observation that deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism' (Hawkins, 1980). Because there are enormous collections of spatio-temporal datasets, Shekhar proposed a method of detecting anomalies in spatio-temporal datasets. He defined a spatio-temporal anomaly as 'an entity whose non-spatial attribute values are significantly different from those in its spatio-temporal neighbourhood' (Shekhar et al., 2001; Shekhar, Lu, & Zhang, 2003). Subsequently, various methods have been developed for spatio-temporal anomaly detection. These methods may be divided into two categories according to the type of spatio-temporal object, i.e., methods of detection of spatio-temporal anomaly points and detection of spatio-temporal anomaly regions.

2.1.1. Detection of spatio-temporal anomaly points

Spatio-temporal anomaly points are a handful of entities that have inconsistent non-spatial attribute values compared with their spatio-temporal neighbourhoods. Cheng and Li proposed a multi-scale approach for spatio-temporal anomaly detection (Cheng & Li, 2006). Two steps consisting of classification and aggregation are first performed to separate spatio-temporal anomalies. By comparison, those entities that exist in classification but disappear in aggregation are potential spatio-temporal anomalies. Furthermore, the true spatio-temporal anomalies are those potential ones that are significantly different from their temporal neighbourhoods. Sun et al. (2005) employed the distance-based strategy proposed by Shekhar to detect spatio-temporal anomalies in climate spatio-temporal sequences. Birant et al. developed a density-based spatio-temporal clustering method (called ST-DBSCAN) by integrating non-spatial attributes into spatio-temporal dimensions, by which those entities not belonging to any clusters were extracted as spatio-temporal anomalies (Birant & Kut, 2006). Barua and Alhajj employed wavelet analysis for scalable anomaly detection in large climate spatio-temporal sequences. Spatial and temporal anomalies were detected separately by focusing on climate change using latitude and time series, respectively (Barua & Alhajj, 2007). Liu et al. detected spatio-temporal anomalies within a space-time framework (Liu, Zheng, and Chawla, 2011). In this framework, spatial anomalies were extracted by a distance-based detection method. By verifying the spatio-temporal neighbourhood of each spatial anomaly, spatio-temporal anomalies were further determined as those entities with significantly different attribute values from others in their spatio-temporal neighbourhoods.

2.1.2. Detection of spatio-temporal anomaly regions

Spatio-temporal anomaly regions are constituted by a group of entities that together have significant differences from the behaviour of their surroundings. As a matter of fact, spatio-temporal anomaly regions belong to a kind of cluster that has significantly larger spatio-temporal scopes than spatio-temporal anomaly points. With respect to the detection of spatio-temporal anomaly regions, there are two main categories of strategies: 1) *Spatio-temporal cluster-based methods*. Spatio-temporal clusters are first detected, and then, anomaly regions are determined as those clusters that depart significantly from their spatio-temporal neighbours. For example, Telang et al. focused on detecting localized anomalous homogeneous regions over spatio-temporal sequences (Telang et al., 2014). Homogeneous regions were captured by spatial clustering, and then, their significant differences from the corresponding spatial neighbourhoods were evaluated to detect spatial anomaly regions. In addition, spatio-temporal anomaly regions were considered as those spatial anomaly regions spanning a certain time interval. 2) *Spatio-temporal statistics-based methods*. Spatio-temporal scan statistics is a classical method of detection of spatio-temporal anomaly regions, which is an extension of spatial scan statistics (Kulldorff, 1997; Kulldorff et al., 2005). The objective of spatio-temporal scan statistics is to identify one spatio-temporal cylinder containing those entities that have the most statistically significant difference from the ones outside the cylinder. Wu et al. modified the traditional spatio-temporal scan statistics method to detect k anomalous regions instead of only one region using precipitation datasets (Wu et al., 2010). Shiode and Shiode extended the spatio-temporal scan statistics method from the two-dimensional Euclidean plane to the network space by defining a network-based space-time search window to identify hotspots of street-level crime incidents (Shiode & Shiode, 2013). In addition,