Front. Earth Sci. 2013, 7(1): 34–42
DOI 10.1007/s11707-012-0343-x

RESEARCH ARTICLE

How many probe vehicles are enough for identifying traffic congestion?—a study from a streaming data perspective

Handong WANG[1,3], Yang YUE (✉)[1,2], Qingquan LI[1,2]

1 State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China
2 Engineering Research Center for Smart Acquisition and Applications of Spatiotemporal Data, Ministry of Education, Wuhan 430079, China
3 Changjiang Institute of Survey, Planning, Design and Research, Changjiang Water Resources Commission, Wuhan 430010, China

Abstract Many studies have been carried out using vehicle trajectory to analyze traffic conditions, for instance, identifying traffic congestion. However, there is a lack of a systematic study on the appropriate number of probe vehicles and their sampling interval in order to identify traffic congestion accurately. Moreover, most of related studies ignore the streaming feature of trajectory data. This paper first represents a novel method of identifying traffic congestion considering the stream feature of vehicle trajectories. Instead of processing the whole data stream, a series of snapshots are extracted. Congested road segments can be identified by analyzing the clusters' evolution among a series of adjacent snapshots. We then calculated a series of parameters and their corresponding congestion identification accuracy. The results have implications for related probe vehicle deployment and traffic analysis; for example, when 5% of probe vehicles are available, 85% identification accuracy can be reached if the sampling time interval is 10 s.

Keywords vehicle trajectory data, floating car data, streaming data, traffic congestion

1 Introduction

Recently, vehicle trajectory data has become an effective complement to traditional traffic detectors. Compared with these fixed traffic detectors, such as inductive loops and video cameras, probe vehicles theoretically can be used to identify traffic condition of most road links. Therefore, using their trajectory data (instantaneous speed and position) has become a trend on identifying and predicting traffic condition. Some approaches have been proposed to identify traffic congestion or traffic incident by retrieving the average travel time or travel speed along road links (Cheu et al., 2002; Kerner et al., 2005; Pattara-atikom et al., 2006; Yoon et al., 2007; Fabritiis et al., 2008; Sananmongkhonchai et al., 2008); however, the identified accuracy highly relies on the percentage of probe vehicles participating in traffic stream and their sampling time interval (Cheu et al., 2002; Hong et al., 2007; Fabritiis et al., 2008). On one hand, there is a lack of a systematic study on the appropriate number probe vehicles and their trajectory sampling interval in order to identify traffic congestion accurately and in a timely manner. On the other hand, although various statistics models were applied to improve the accuracy, such as artificial neural networks (Fabritiis et al., 2008), surface fitting (Shi and Kong, 2008), and a Kalman filter (Nanthawichit et al., 2003; Kong et al., 2009), the computations are usually complex and time-consuming.

To answer the questions of "how many probe vehicles are enough for an accurate identification of traffic congestions?", and "what is the proper trajectory data sampling interval?", we first need to put the discussion under a traffic congestion identification framework.

Vehicle trajectory data is a typical streaming data that, ordered in sequence of trajectory points p_1, p_2, …, p_n, arrive continuously and must be accessed in order. Compared with a traditional static data set, the trajectory data streams have unique characteristics such as continuously arriving data, unbounded data size, evolution over time, and inherent temporal component (Guha et al., 2003). Existing studies (Srinivasan and Jovanis, 1996; Cheu et al., 2002; Zhang et al., 2007a, 2007b) consider 4%–7% probe vehicles are the proper rate to ensure satisfactory identification accuracy, under various settings. To the best of our knowledge, most of the existing methods have not considered the streaming feature of vehicle trajectories.

Received June 23, 2012; accepted September 21, 2012

E-mail: yueyang@whu.edu.cn

So we discuss the problem from a streaming data perspective, and propose a snapshot approach to deal with the streaming feature of vehicle trajectories. The contribution of this paper is two fold: i) we propose a method that can deal with data streams, and ii) we investigate the influences of various parameter settings to identification accuracy, such as snapshot time interval, cluster overlap degree, percentage of probe vehicles, and trajectory sampling interval. These parameters are valuable for practical studies.

The following of the paper is organized as follows. Section 2 first introduces the challenges and requirements of streaming data processing and analysis. Section 3 presents the proposed methodology. Section 4 answers the questions based on a group of experiments. Section 5 concludes the paper.

2 Streaming data

In reality, vehicle trajectory data is a typical streaming data because they are continuously collected and with unbounded data size. Since it is impossible to load all the data into memory, traditional data processing and analysis approaches that use multiple passes are not suitable for the traffic data streams. A fast incremental processing of the continuously arriving data is of necessity to improve the processing and analysis efficiency.

Due to the inherent temporal nature of streaming data, synopsis data structures and a snapshot data scenario are two common methods used to process the newly arriving data. Most synopsis methods use statistical methods to capture the inherent features of data streams, such as sampling, wavelets, sketches, and histograms. A snapshot data scenario attempts to capture the evolution trends of data streams by selecting a series of snapshot data sets in data streams, which avoids processing the whole data stream and therefore can improve data processing efficiency (Aggarwal, 2007).

Here, we adopt a snapshot approach to deal with the trajectory data, focusing on the evolution of the underlying data, which processes the newly arriving data according to the timestamps, and does not compare it with all the data that have been processed. The following section will introduce the methodology in detail.

3 Methodology

The approach we propose is based on the fact that the speed of vehicles in congested road segments is lower and the density is higher. Moreover, the spread of congestion can be either spatial or temporal, or both. Therefore, we use clustering as the basic approach to find the road segments with relatively lower travel speed and higher vehicle density. The aim is to provide a representation of the clusters that are not only compact, but do not grow appreciably with the number of points processed (Barbará, 2002).

3.1 Definition

3.1.1 Definition 1: Trajectory point

Vehicle trajectories are constrained within a road network; therefore, we define trajectory points within the road network context using a four-tuple:

$$P = \{PtID,\ LinkID,\ SDist,\ t\}, \tag{1}$$

where P is the vehicle trajectory point, *PtID* is the identifier of P, *LinkID* is the identifier of the road link on which the vehicle trajectory point is located, *SDist* is the distance between the vehicle point and the start node of the road link, and t is the timestamp when the vehicle trajectory point was collected.

Here, the relative distance *SDist* is used, instead of the point coordinates, because this representation facilitates calculating the network distance between points, which has the advantage over Euclidean distance in traffic analysis (Zou et al., 2012).

3.1.2 Definition 2: Cluster

Because of the road network constraint, the conventional two-dimensional cluster convex-hull is degenerated into an one-dimensional road segment, which can be described as a three-tuple:

$$C_t = \{PtNum,\ Segs,\ t\}, \tag{2}$$

where C_t is the cluster generated at timestamp t, *PtNum* is point number in the cluster, *Segs* is the road segment set where the trajectory points were located, and t is the snapshot timestamp when the points in the cluster were collected. Figure 1 shows the representation of cluster at timestamp t, where $PtNum = 40$, $Segs = \{L_1, L_2, L_3, L_4\}$,

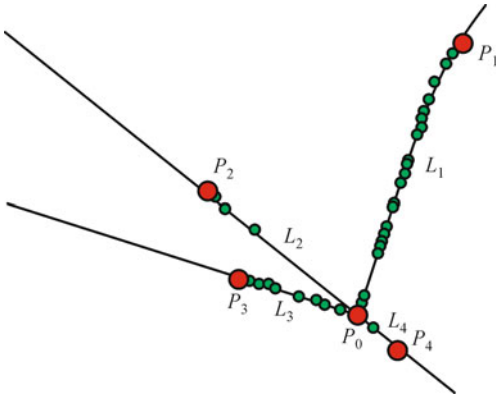

Fig. 1 An example of a cluster snapshot

the cluster C_t can be represented as $C_t = \{40, \{L_1, L_2, L_3, L_4\}, t\}$.

3.1.3 Definition 3: Cluster overlap

Let C_t, C_{t+1} be two clusters at two continuous timestamps: t and $t+1$. The overlap degree of C_t and C_{t+1} is the normalized proportion between the overlap length of the two clusters with their total length:

$$Overlap(C_t, C_{t+1}) = \frac{\sum_{l \in C_t \cap C_{t+1}} Len(l)}{\sum_{l \in C_t \cup C_{t+1}} Len(l)}, \quad (3)$$

where $\sum_{l \in C_t \cap C_{t+1}} Len(l)$ is the overlapped, or intersected length of the clusters at time t and $t+1$; while $\sum_{l \in C_t \cup C_{t+1}} Len(l)$ is the total length of the clusters.

3.1.4 Definition 4: Congestion

Let $C = \{c_1, c_2, \ldots, c_k\}$ be a sequence of cluster snapshots such that for each i $(1 \leqslant i < k)$, and the cluster overlap $Overlap(c_i, c_{i+1}) \geq \theta_S$, where θ_S $(0 < \theta_S \leqslant 1)$ is a given threshold. Then if the length of C is longer than sl, C is defined as congestion:

$$Len(C) = \sum_{i=1}^{k} Len(c_i) \geqslant sl, \quad (4)$$

where $Len(c_i)$ is the total length of all identified road segments in the *Segs* of c_i. sl is a given threshold, such a queue length.

Or, if the duration of the cluster exceeds a certain time extent:

$$Dur(C) = MaxTstamp(C) - MinTstamp(C) \geqslant td, \quad (5)$$

where $MaxTstamp(C)$ and $MinTstamp(C)$ are the latest and the first timestamp of C, respectively. td is a given time threshold to define the duration of congestion.

3.2 Methodology

Fig. 2 illustrates the proposed methodology. First, we generate a snapshot data set from vehicle trajectory streams based on certain predefined time interval. Then, clustering is applied to the snapshots to retrieve the clusters that consist of trajectory points with high density and low speed. Results are stored in a snapshot cluster list. The clusters in the snapshot cluster list are treated as candidate congestion segments. Evolution analysis is conducted between clusters on the continuous snapshots to construct and maintain a list of moving congestion segments. We use 'overlap degree' to determine whether the cluster survives in the next continuous snapshot. Finally, each moving congestion segment is checked; only those moving congestion segments whose total queue length or temporal duration exceeds the thresholds can be identified as traffic congestion.

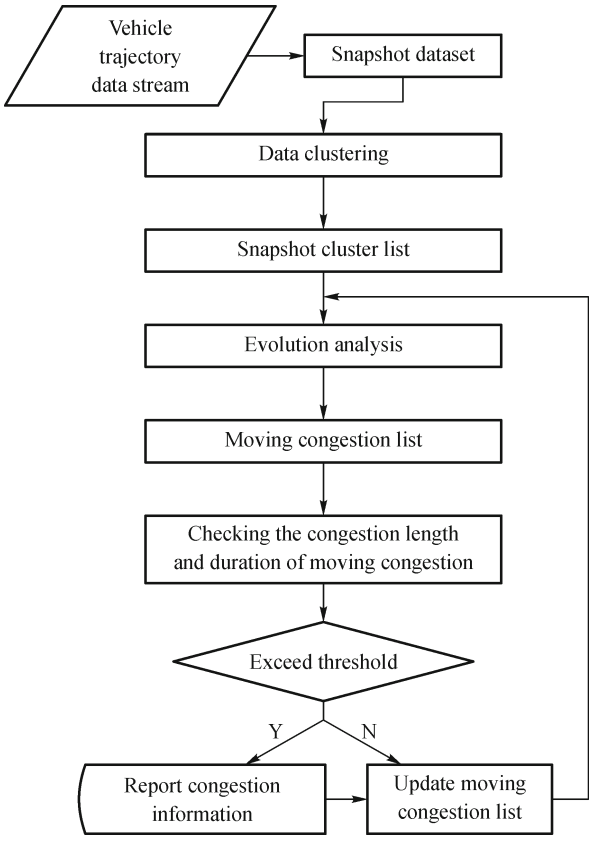

Fig. 2 Flowchart of our proposed method

There are many clustering algorithms such as the partition based algorithm K-Means (Hartigan and Wong, 1979), density based algorithm DBSCAN (Ester et al., 1996), OPTICS (Ankerst et al., 1999), and grid based algorithm (Wang et al., 1997). We chose the DBSCAN algorithm because we are more focused on density.

4 Experiment

4.1 Data

Although probe vehicle trajectory data, such as those collected by taxi (floating car data, FCD), is increasing available, the data itself is not enough for conducting this research. This is because most of the data are used in an offline manner and most of the studies do not have real traffic condition data to validate the study results. So in this study, we made a compromise: we adopted simulated data to obtain the parameters first, to facilitate further studies on real-world data. Meanwhile, it is feasible to test various data sampling rates using simulation data, which is not possible for most FCD data.

This study generated simulated vehicle trajectory data

using VISSIM, which can simulate traffic flow at a microscopic level. Each trajectory point has the following properties: position coordinates, speed, vehicle ID, link ID, and associated timestamp. The sketch of the network is shown in Fig. 3 which has 110 road links, and the total length of the lines is about 90 km. The simulation period was 70 min, in which the first 10 min was a network warm-up period. The simulation data started from 10 min and thus the total analysis time period is 60 min (1.0 h). During this period, 13783 vehicles were generated, with the average speed of around 37 km/h (10.26 m/s).

Fig. 3 The sketch of simulated urban road network

4.2 The setting of parameters

First, it is necessary to determine the threshold of "lower speed" as aforementioned, to define traffic congestion. In China, urban traffic state is divided into five levels: very smooth, smooth, light congestion, congestion and serious congestion (Chi, 2007). The corresponding speed ranges are shown in Table 1. According to this criterion, we use 25 km/h as the threshold, i.e., only those vehicle trajectories points whose speed is under 25 km/h are considered for clustering.

Second, we need to clarify "congestion," i.e., what kinds of states can be defined as " traffic congestion?" The Ministry of Public Security of the People's Republic of China gives the definitions of traffic congested intersection and traffic congested road segment as follows (Lu and Zhu, 2001): the queue length of vehicles exceeds 250 min any intersection without traffic control; or vehicles in any traffic-signal-controlled intersection cannot pass the intersection within three signal cycles; or a roadway with over 1000 m queue length is a congested road segment. Therefore, this paper sets the spatial length and temporal duration thresholds of traffic congestion as 250 m and 120 s. It is also necessary to determine other parameters, such as snapshot time interval and the cluster overlap threshold θ_S between clusters. A series of experiments were conducted to examine the identification accuracy under different parameter settings.

Table 1 Vehicle speed and traffic state (Chi, 2007)

Average speed/(km·h^{-1})	Traffic state
[0, 15]	Serious congestion
(15, 25]	Congestion
(25, 35]	Light congestion
(35, 45]	Smooth
>45	Very smooth

4.3 Results and discussion

Due to space constraints, only a group of experimental results are given to illustrate the relationship between parameter selection and identified traffic congestion. We first set the cluster overlap threshold θ_S from 0.5 to 0.9, with the snapshot time intervals at 5, 10, 15, and 20 s, respectively.

Because the output of DBSCAN is sensitive to the two parameters: the minimum number of points (MinPt) and the size of neighborhood (Eps). Following the method proposed by (Ester et al., 1996), we determine the two parameters using training data under different percentages of probe vehicles and data sampling time intervals. Some of the results are shown in Table 2.

Table 2 DBSCAN parameters

Percentage/%	Sampling interval/s	MinPt	Eps/m
5	10	3	3645
10	5	3	938
10	10	5	3181
15	10	3	1077
20	10	3	948
20	20	3	2014

Fig. 4 shows identified congestion under different values of snapshot interval time and cluster overlap, which includes the number of identified congested areas, queue lengths, and duration of congestion. It can be observed from Fig. 4(a) that when the cluster overlap value is 0.5, the number of identified congested areas sharply declines as snapshot time interval increases; however, the impact of snapshot time interval decreases as the cluster overlap value increases. For example, when the cluster overlap value increases to 0.9, the number of identified congested areas changes slightly with the increasing of snapshot interval time. It is similar for the identified queue length and duration as shown in Figs. 4(b) and 4(c).

Different snapshot time intervals and cluster overlaps also result in different spatial distributions of traffic congestion. Fig. 5(a)–5(e) depict the spatial distributions of identified traffic congestions with the overlap of 0.5, 0.6,

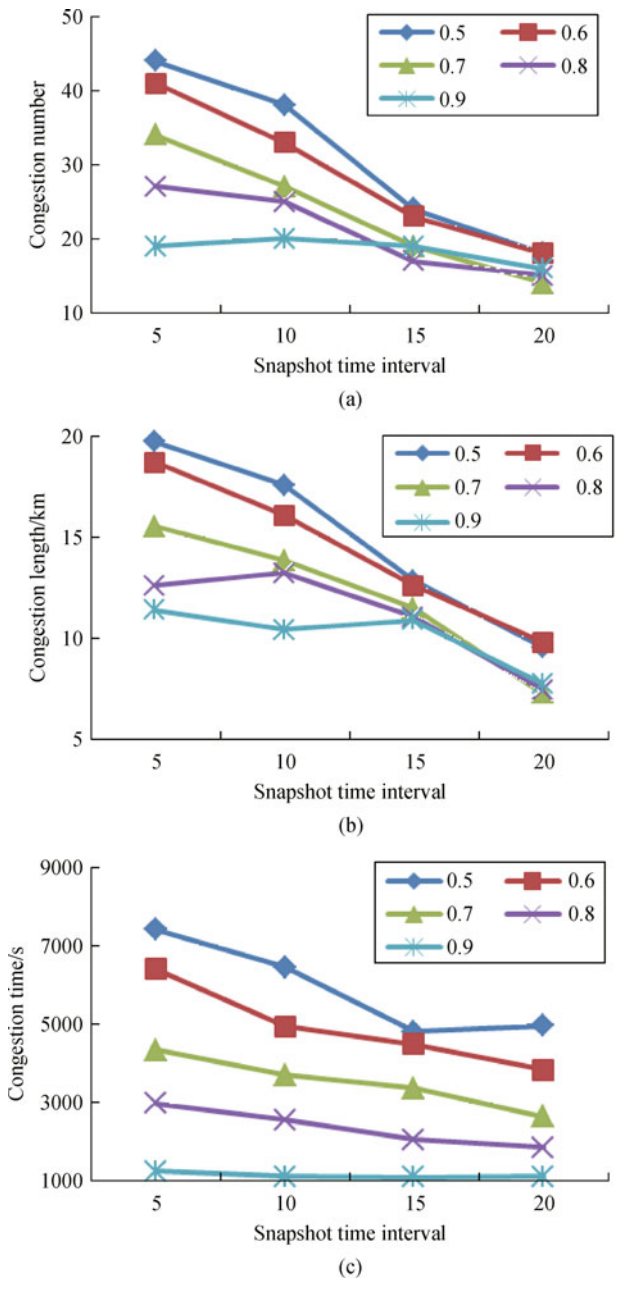

Fig. 4 Identified congestions under different values of cluster overlap and snapshot time interval. (a) Number of identified traffic congested areas; (b) queue length; (c) traffic congestion duration

0.7, 0.8, and 0.9, respectively, under the same snapshot time interval of 10 s. As shown in Fig. 5, the smaller the value of cluster overlap, the more traffic congestion can be identified near road intersections. With the increase of overlap value, the number of congestions drops; while the congestions that can always be identified under all clusters overlap values, are those spreading over 250 m or lasting over 120 s.

Figs. 6(a)–6(d) represent the identified traffic congestions with the snapshot time interval at 5, 10, 15 and 20 s, respectively; with the same overlap value of 0.7.

As expected, it can be observed from Fig. 6 that the number of identified congested areas drops as snapshot time interval increases. In general, a smaller time interval is helpful for identifying intersection congestion and retrieving local traffic state. A larger time interval is suitable for capturing relatively severe congestion. So the selection of snapshot time interval is a trade-off between details of traffic state and data process efficiency. Fig. 7 summarizes congestion identification accuracy under different parameter settings.

Table 3 depicts the relationship between the percentages of probe vehicles with the accuracy of identification rate under different sampling time intervals. It can be observed that around 85% identification accuracy can be reached when 5% of probe vehicles are used, and sampling time interval is 10 s. There are some fluctuations in terms of the identification accuracy; it is mainly because of the uneven distribution of vehicles, which leads to variation of vehicle density at road segments.

To examine the impact of the selection of DBSCAN parameters to the identification accuracy, we use the minimum point, Minpt, as an example. The result is shown in Fig. 8. It illustrates the congestion identification accuracy under the combination of different percentage of probe vehicles (5%, 10%, 12%) and sampling time interval (5 s, 10 s, 15 s), when MinPt ranges from 3 to 6. The selection of the parameters should be determined in view of both application requirement (to what extent, congestion is defined), and system capability.

5 Conclusions

To answer the question of "how many probe vehicles are enough for an accurate identification of traffic congestion?", and "what is the proper data sampling interval?", in this paper, we proposed a novel method to identify urban traffic congestion using simulated vehicle trajectories in the context of streaming and real-time manners. We first retrieve the candidate traffic congestion by clustering the low speed and high density trajectory points. Since trajectory data streams generally have huge volume, we select a series of snapshot data sets which hold the evolution characteristics of the whole stream. Evolution analysis is conducted on the clusters among continuous snapshot data sets to capture the spatial and temporal spread of congestion. Experiment results show that the proposed method can identify traffic congestion efficiently for the whole road network.

We investigated the relationship of identified congested areas considering snapshot time interval and cluster overlap degree. For example, if overlap degree = 0.8, or snapshot time interval = 15 s, those non-intersection congested areas can be detected. Results also show that, when 5% of probe vehicles are available, 85% identifica-

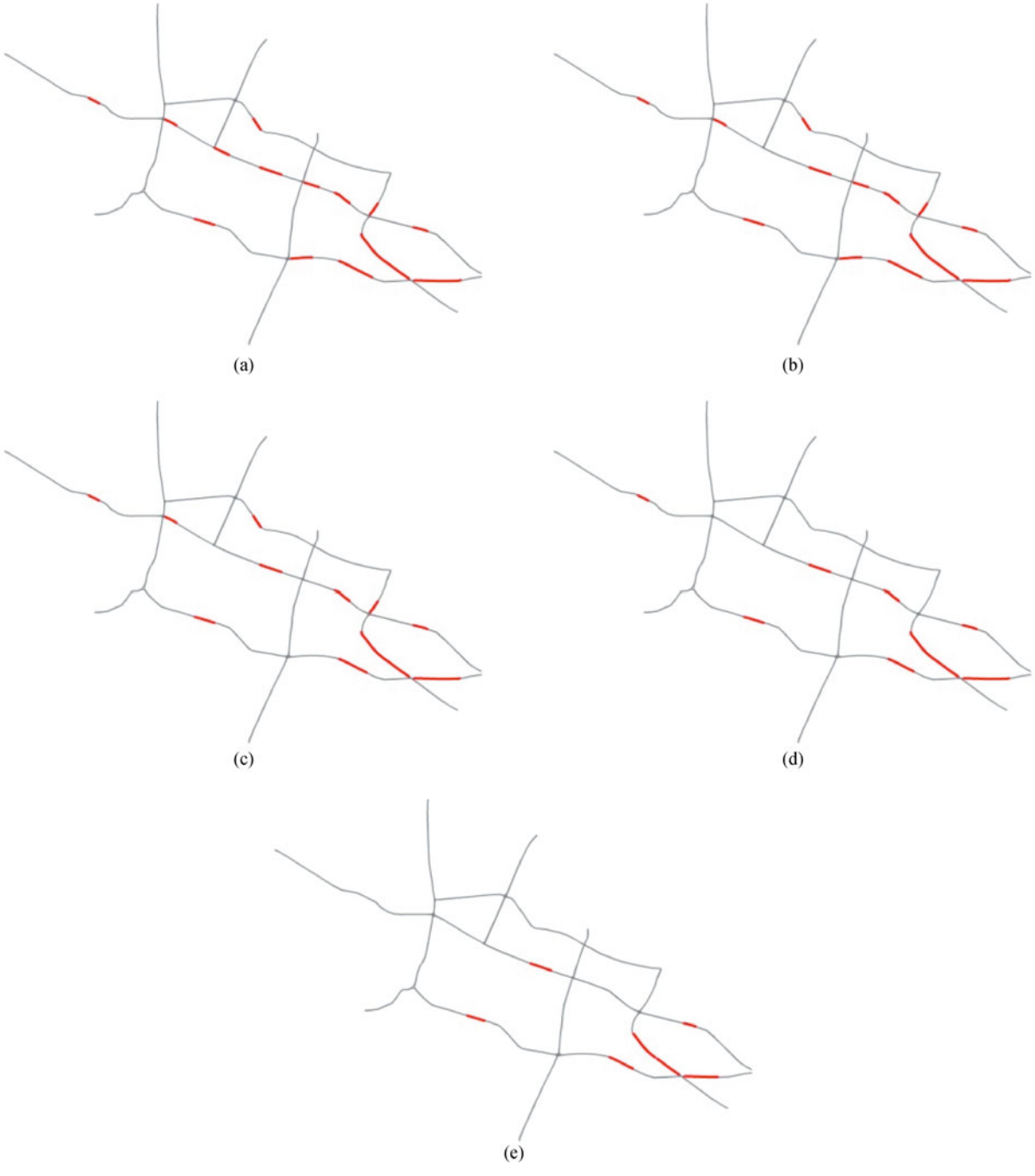

Fig. 5 The impact of different cluster overlap. Snapshot time interval = 10 s. (a) Overlap value = 0.5; (b) overlap value = 0.6; (c) overlap value = 0.7; (d) overlap value = 0.8; (e) overlap value = 0.9

Table 3 Relationship between percentage of probe vehicle and sampling interval with identification rate

Sampling time interval/s	percentage of probes/%	Identification rate/%
1	⩾3	⩾98.72
5	⩾3	⩾89.74
10	⩾5	⩾84.62
15	⩾10	⩾80.77
20	⩾15	⩾78.21

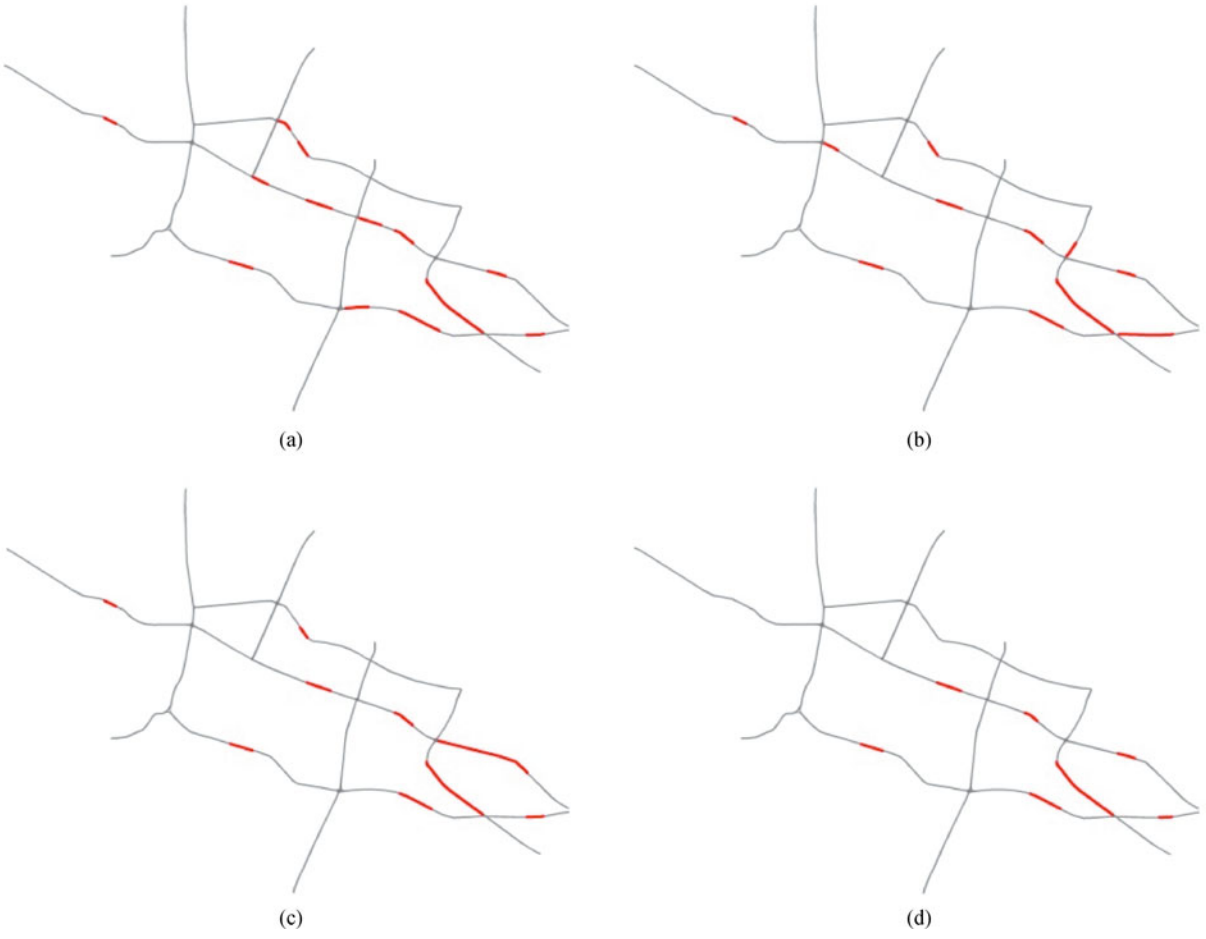

Fig. 6 The impact of different snapshot time interval. Cluster overlap value = 0.7. (a) Time interval = 5 s; (b) time interval = 10 s; (c) time interval = 15 s; (d) time interval = 20 s

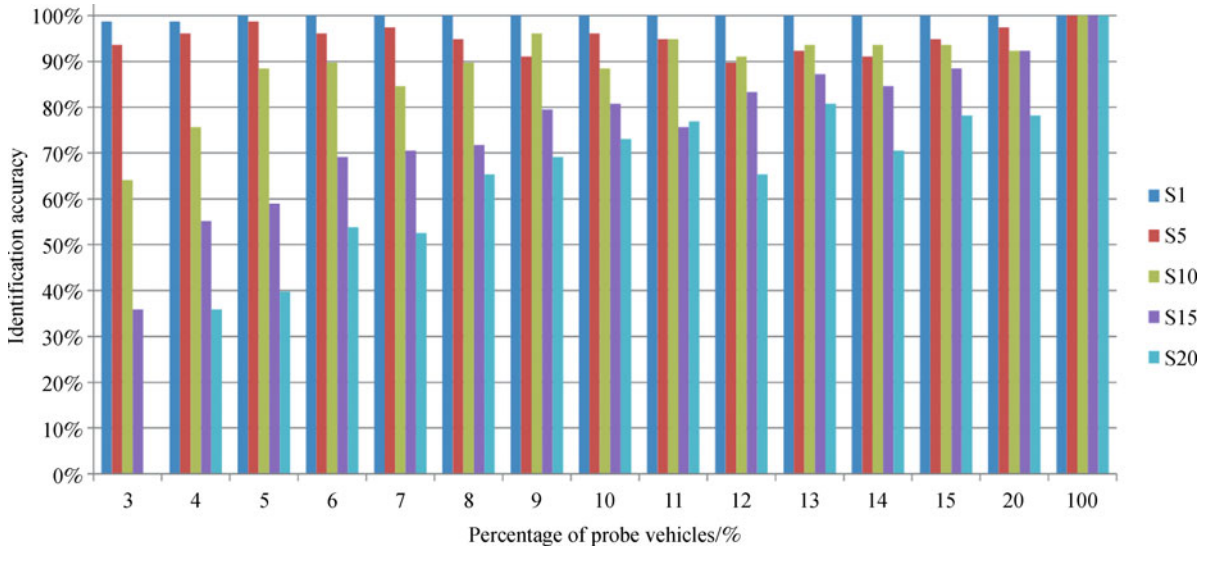

Fig. 7 Identification accuracy under different parameter settings

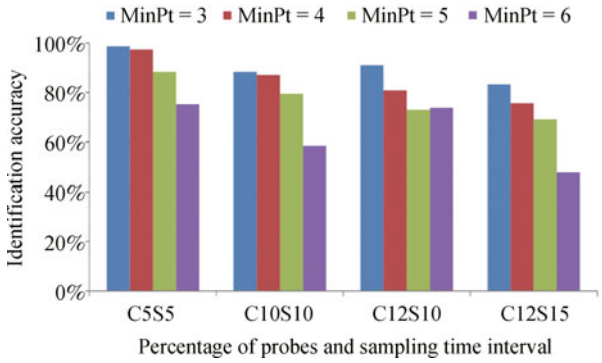

Fig. 8 The impact of DBSCAN parameters (Minpt) to identification accuracy

tion accuracy can be reached if the sampling time interval is 10 s. We trained the DBSCAN parameters accordingly. The results are valuable for studies using real-world data.

Acknowledgements This work was supported by the National Natural Science Foundation of China (Grant Nos. 41171348, 40830530 and 60872132).

References

Aggarwal C C (2007). An introduction to data streams. Data Streams, 31: 1–8

Ankerst M, Breunig M M, Kriegel H P, Sander J (1999). OPTICS: ordering points to identify the clustering structure. In: Proceedings of ACM SIGMOD'99 International Conference on Management of Data, New York: ACM

Barbará D (2002). Requirements for clustering data streams. ACM SIGKDD Explorations Newsletter, 3(2): 23–27

Cheu R L, Xie C, Lee D (2002). Probe vehicles population and sample size for arterial speed estimation. Computer Aided Civil and Infrastruct Engineering, 17(1): 53–60

Chi C (2007). The research about the traffic congestion evaluation system of sections in Beijing. Dissertation for Master Degree, Beijing: Beijing Jiaotong University

Ester M, Kriegel H P, Sander J, Xu X (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of ACM KDD'96, Menlo Park, CA: AAAI Press

Fabritiis C D, Ragona R, Valenti G (2008). Traffic estimation and prediction based on real time floating car data. In: Proceedings of International IEEE Conference on Intelligent Transportation Systems, 12–15 Oct. 2008, Beijing, China

Guha S, Meyerson A, Mishra N, Motwani R, O'Callaghan L (2003). Clustering data streams: theory and practice. IEEE Transactions on Knowledge and Data Engineering, 15(3): 515–528

Hartigan J, Wong M (1979). Algorithm as136: a *k*-means clustering algorithm. Journal of the Royal Statistical Society, Series C (Applied Statistics), 28(1): 100–108

Hong J, Zhang X, Wei Z, Li L, Ren Y (2007). Spatial and temporal analysis of probe vehicle-based sampling for real-time traffic information system. In: Proceedings of International IEEE Conference on Intelligent Vehicles Symposium, 13–15 Jun. 2007, Istanbul, Turkey

Kerner B S, Demir C, Herrtwich R G, Klenov S L, Rehborn H, Aleksic M, Haug A (2005). Traffic state detection with floating car data in road networks. In: Proceedings of International IEEE Conference on Intelligent Transportation Systems, 13–16 Sept. 2005, Vienna, Austria

Kong Q J, Li Z P, Chen Y K, Liu Y C (2009). An approach to urban traffic state estimation by fusing multisource information. IEEE Transactions on Intelligent Transportation Systems, 10(3): 499–511

Lu H P, Zhu J (2001). Analysis of the Urban Traffic. Beijing: China Water Power Press (in Chinese)

Nanthawichit C, Nakatsuji T, Suzuki H (2003). Application of probe vehicle data for real-time traffic state estimation and short-term travel time prediction on a freeway. In: TRB 2003 Annual Meeting CD-ROM, 11–15 Jan. 2003, Washington, DC, USA

Pattara-atikom W, Pongpaibool P, Thajchayapong S (2006). Estimating road traffic congestion using vehicle velocity. In: Proceedings of International Conference ITS Telecommunications, 21–23 Jun. 2006, Chengdu, China

Sananmongkhonchai S, Tangamchit P, Pongpaibool P (2008). Road traffic estimation from multiple GPS data using incremental weighted update. In: Proceedings of the 8th International Conference ITS Telecommunications, 24 Oct. 2008, Phuket, Thailand

Shi W, Kong Q (2008). A GPS/GIS integrated system for urban traffic flow analysis. In: Proceedings of International IEEE Conference on Intelligent Transportation Systems, 12–15 Oct. 2008, Beijing, China

Srinivasan K, Jovanis P (1996). Determination of the number of probe vehicles required for reliable travel time measurement in an urban network. Transportation Research Record, 1537: 15–22

Wang W, Yang J, Muntz R (1997). STING: a statistical information grid approach to spatial data mining. In: Proceedings of the 23rd VLDB Conference, 25–29 Aug., 1997, Athens, Greece

Yoon J, Noble B, Liu M (2007). Surface street traffic estimation. In: Proceedings of International Conference Mobile Systems, Applications and Services, 11–14 Jun. 2007, San Juan, Puerto Rico

Zhang C B, Yang X, Yan X (2007a). Probe vehicles sample size for mobile traffic detection system. China Journal of Highway and Transport, 20(1): 96–101 (in Chinese)

Zhang C B, Yang X, Yan X (2007b). Method for floating cars sampling cycle optimization. Journal of Transportation Systems Engineering

and Information, 7(3): 100–104

Zou H X, Yue Y, Li Q Q, Yeh A G O (2012). An improved distance metric for the interpolation of link-based traffic data using Kriging: a case study of a large-scale urban road network. International Journal of Geographical Information Science, 26(4): 667–689

AUTHOR BIOGRAPHY

Handong Wang obtained his Ph.D. in photogrammetry and remote sensing from Wuhan University, China. He is an Engineer at Changjiang Institute of Survey, Planning, Design and Research, Changjiang Water Resources Commission. Currently he is focusing on the research of water resources informatization.