Front. Earth Sci. 2013, 7(2): 206–216
DOI 10.1007/s11707-012-0350-y

RESEARCH ARTICLE

Intersection delay estimation from floating car data via principal curves: a case study on Beijing's road network

Xiliang LIU, Feng LU, Hengcai ZHANG (✉), Peiyuan QIU

State Key Laboratory of Resources and Environmental Information system, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China

Abstract It is a pressing task to estimate the real-time travel time on road networks reliably in big cities, even though floating car data has been widely used to reflect the real traffic. Currently floating car data are mainly used to estimate the real-time traffic conditions on road segments, and has done little for turn delay estimation. However, turn delays on road intersections contribute significantly to the overall travel time on road networks in modern cities. In this paper, we present a technical framework to calculate the turn delays on road networks with float car data. First, the original floating car data collected with GPS equipped taxies was cleaned and matched to a street map with a distributed system based on Hadoop and MongoDB. Secondly, the refined trajectory data set was distributed among 96 time intervals (from 0: 00 to 23: 59). All of the intersections where the trajectories passed were connected with the trajectory segments, and constituted an experiment sample, while the intersections on arterial streets were specially selected to form another experiment sample. Thirdly, a principal curve-based algorithm was presented to estimate the turn delays at the given intersections. The algorithm argued is not only statistically fitted the real traffic conditions, but also is insensitive to data sparseness and missing data problems, which currently are almost inevitable with the widely used floating car data collecting technology. We adopted the floating car data collected from March to June in Beijing city in 2011, which contains more than 2.6 million trajectories generated from about 20000 GPS-equipped taxicabs and accounts for about 600 GB in data volume. The result shows the principal curve based algorithm we presented takes precedence over traditional methods, such as mean and median based approaches, and holds a higher estimation accuracy (about 10%–15% higher in RMSE), as well as reflecting the changing trend of traffic congestion. With the estimation result for the travel delay at intersections, we analyzed the spatio-temporal distribution of turn delays in three time scenarios (0: 00–0: 15, 8: 15–8: 30 and 12: 00–12: 15). It indicates that during one's single trip in Beijing, average 60% of the travel time on the road networks is wasted on the intersections, and this situation is even worse in daytime. Although the 400 main intersections take only 2.7% of all the intersections, they occupy about 18% travel time.

Keywords intersection delay, float car data, trajectory, principal curves

Received July 9, 2012; accepted November 23, 2012

E-mail: zhanghc@lreis.ac.cn

1 Introduction

Nowadays traffic congestion has become a more and more serious problem in the cities around the world, especially in the metropolises. The skeleton of a city is composed by the roads with a high hierarchical level, such as expressways, arterial streets, collector streets, etc. The intersections in between play an essential role to demonstrate the spatial-temporal characters of city dynamics and contribute much to the travel time cost (Nielsen et al., 1998). The turn delays on the intersections reflect not only the drivers' discomfort, but also the logic of signal control and design (Heidemann, 1994). Hence estimating the time-dependent turn delays at intersections on city road networks at a fine spatial-temporal granularity is of striking importance, both for real-time route planning and traffic management.

In the field of traffic congestion research, much attention has been paid to whole travel time estimation, link travel time estimation (Hellinga and Fu, 2002; Skabardonis and Geroliminis, 2005), queue length (Heidemann, 1994; Liu et al., 2012; Ghaffarian et al., 2012), link travel speed (Cheu et al., 2002), and so on, to access the total time elapsed on the road links. However, these link-based delay

functions (the traditional cumulative input-output curves) may not properly capture the impact of an intersection in which intersection delays account for a large amount of total travel time (Fabritiis et al., 2008). Obtaining the intersection delay data are not a trivial task. The traditional fixed-point data collecting equipment, such as inductive loop detectors, ultra-sonic vehicle detectors, and closed circuit television (CCTV) cameras, covers only a small part of the traffic network and involves extra installation and maintenance cost, performing a sub-economic way which cannot adapt to the development of modern traffic dynamic demand. Moreover, the low density of the equipment and the high distribution in urban areas make these kinds of stationary detectors makes it impossible to produce reliable information about travel time within the network (Gühnemann et al., 2004). In parallel, some researchers rely on sophisticated micro-simulation and successfully apply it in some specific fields (Lu, 2010). However, it is easy to draw the wrong conclusions and falls short of the ability to generally imitate the real world if one is not fully familiar with the model (Long et al., 2011).

In recent years, real-time floating car data (FCD) of city road networks collected by operating vehicles (taxicabs, probe cars, buses, private cars, etc.) equipped with GPS-enabled computers has become the mainstream in the research of intersection delay estimation because of its cost-effectiveness and flexibility compared with other traffic data sources (Herring et al., 2010). We take taxicabs as our FCD source because:

1) Big cities usually have a large number of taxicabs traversing in urban areas which gains a huge volume of GPS trajectories that guarantee a wide coverage of the whole road network every day (Hunter et al., 2009);

2) These GPS trajectories also contain a prior knowledge about the spatial-temporal traffic volume distribution, that is, the knowledge of the experienced drivers from the physical world (Zheng and Zhou, 2011).

Boyce et al. (1991) argue that floating cars can provide a much more accurate travel time estimation than inductive loop detectors, with 99.4% reliability in 50000 reports, however, it should be noticed that the equipped GPS sensors, the signal transmission, and the data processing provide a noisy magnetic signature of the taxicabs mixed with various random factors at an exact timestamp which will influence the FCD quality (Kothuri et al., 2008). Moreover, the uneven density of the road network makes the taxicab distribution heterogeneous across the network, leading to an unavoidable data sparseness problem (Li and Rose, 2011). To reduce the raw errors, some data cleaning techniques such as mean\median filter, Kalman filter, particle filter, and so on are employed (Zheng and Zhou, 2011). The traditional means to deal with data sparseness problem is conducted by using current and near-past records from a historical perspective (Bejan et al., 2010; Herrera et al., 2010). Ban et al. (2009) have developed a least squares-based algorithm to estimate the delay patterns from sampled travel time records by recognizing the underlying characteristics of signalized intersection delays. Ban et al. also show that delay patterns can be represented as piecewise linear (PWL) curves. These curves are developed by using well-developed traffic flow theory on queue forming and discharging at signalized intersections. However, the missing data problem still remains incompletely solved (Ban et al., 2009).

Principal curves (Hastie and Stuetzle, 1989), which were first defined by Trevor Hastie and Werner Stuetzle as "self-consistent" smooth curves that pass through the "*middle*" of a d-dimensional probability distribution or data cloud, provide a novel perspective to solve the problems mentioned above. Principal curves can be used to create nonlinear one-dimensional description of data. They are well suited as a starting point for a general data presentation method as they can be defined for all types of multivariate data distributions. This method is nonparametric, and the shape of the curves is suggested by the original data. This method has shown its power in dealing with the raw data noise and non-uniform distribution, which are common phenomena in traffic data (Zhang and Wang, 2003). For example, many FCD sample points can be obtained during the morning\evening peaks, while fewer are collected in other time periods of a day. However, the potential of this method has not been fully mined in the field of intersection delay estimation.

In this paper, we try to overcome data sparseness and missing data problems with the help of a principal curves method, and build a turn time table of the intersections in Beijing city. With the calculation result, we are eager to find some patterns of the city's spatial-temporal turn delay distributions. Furthermore, we also focus on the average turn delay ratio during one trip in Beijing. The contributions permit the following statements:

1) We introduce the principal curves method into the floating cars' trajectories processing and solve the data sparseness and missing data problems regardless of intersection signals and traffic volume, and draw a detailed turn delay table for Beijing's intersections on the main roads (the expressways, arterial streets, collector streets), which provides a new perspective to accurately forecast urban travel time and understand the city's mobility.

2) We build a FCD real-time processing framework based on MongoDB (Chodorow and Dirolf, 2010) and Hadoop (White, 2010), which can deal with big data challenge and satisfy the demand of real-time estimation. The data-driven framework is also fit for further work as the data volume increases in the future.

3) The turn delay ratios both on the whole road network and on the 400 main intersections in Beijing are calculated by a data-driven procedure from the real world data, providing a further understanding of the real-time traffic status and is of great help for daily route planning.

This paper is organized as follows: Section 1 introduces the research background, problem description, and the

solution way. Section 2 introduces some definitions and conceptions in our approach from the literatures. Section 3 describes the technical framework based on the principal curves approach in detail. Section 4 demonstrates the data pre-processing procedure (including data cleaning and map matching) with a distributed computing architecture based on Hadoop and MongoDB. Section 5 implements all our methods with experiments. In section 6, we also discuss some problems faced in our work. Section 7 gives the conclusion and further outlook of our research.

2 Related works

2.1 What's the intersection delay?

The traditional intersection delay, i.e., the turn delay, is defined as the vehicle running time loss when turning or crossing the intersection from upstream traffic flow to downstream due to traffic interference, traffic management, and control facilities (Homburger et al., 2007). Xie et al. treat the turn delay as a special kind of link travel time, that is, the sampled travel times between two consecutive locations on arterial streets, one upstream and the other downstream, of a signalized intersection (Xie et al., 2001). Stephan Winter names the turn delay as turn cost, which is related to the continuation of travel in a node (Winter, 2002). Wang gives a clear definition: '*The floating car's travel time in excess of that of user-specified free flow travel is defined as the floating car's measured travel time delay*' (Wang, 2004). To calculate the exact elapsed time at the intersections, some researches also pay attention to the intersection range definition. Sun investigates Beijing's real road network and refines the range to 160 m (Sun, 2007). Zhang et al. considers the dynamic spatial distribution character of the floating cars, and sets the range to 200 m, which starts at the center of an intersection and extends 100 m both upstream and downstream (Zhang et al., 2011).

2.2 How to estimate the turn delay using FCD?

Floating car data (Ban et al., 2009) is used to infer turn delays from intersection delay patterns using the virtual trip line (VTL) technique. As a vehicle equipped with a GPS panel passes by a VTL location, the location and speed of the vehicle are sent to a secure server from which all vehicles' information is aggregated and transferred to traffic models. A least squares–based algorithm is developed to estimate the delay patterns from sampled travel time records by recognizing the underlying characteristics of signalized intersection delays. Ban et al. also show that delay patterns can be represented as PWL curves. These curves are developed by using well-developed traffic flow theory on queue forming and discharging at signalized intersections. Kwong et al. propose a new scheme in which wireless traffic sensors are deployed downstream (at a fixed distance of 12 m) of signalized intersections (Kwong et al., 2009). A unique feature of such a vehicle reidentification method is that traffic signal information is not required (Zhang et al., 2011). Wang argues that whether the FCD estimation is an accurate method, and suggests the floating car interval be 15 min to get a more accurate result (Wang, 2004); this result is also proven in the HCM manuals. Zhang et al. estimate turn delay at different times and turn types in the city road network based on personal GPS collected trajectories (the Geolife data set) and propose a prediction model based on Neural Networks to handle these records (Zhang et al., 2011), which paves a new way to denken on turn delay estimation.

2.3 Some limitations using FCD as the data source

The minimum coverage plays an important role in estimation using FCD. The reliability of travel time estimation based on FCD highly depends on the percentage of floating cars participating in the traffic flow (Hong et al., 2007). As a rule, a lower percentage of floating cars is required in more congested traffic conditions while a higher percentage is needed in low flow conditions (Fabritiis et al., 2008). Xiong et al. investigate the relationship between the floating car records and the data from the remote traffic microwave sensor (RTMS) system, and conclude that 7 to 9 FCD records are needed in the Beijing road network's 2nd Ring Road and 3rd Ring Road every kilometer in 5 min, and the current sample size FCD can meet the minimum sample size requirements by 73% and 87% in time and space separately (Xiong et al., 2010). Although the statistical average method can get a promising result from the floating car record in a time period if the minimum coverage of the road network is satisfied, the variance of the observed trajectories in the turn delay estimation still cannot go to zero as the records increase (Hellinga and Fu, 2002), hence the authors in this paper suggest that a probabilistic model should be carefully constructed in order to make good use of FCD.

Another limitation of FCD that cannot be overlooked in the intersection delay estimation is the randomness in the data sampling due to the dynamic nature of the traffic flow. For a specific intersection, the turn delay can be accessed if the number of traveling monitored cars achieves a significant penetration level and the single car transmission rate is adequate (Fabritiis et al., 2008). Therefore, the key in the FCD applications needs to study the relationship between the characteristics of the floating car operation and their traffic and road links and to reduce the influence of random errors (Cowan and Gates, 2002).

3 Principal curves approach

3.1 Definition of principal curves

Principal curves are the nonlinear generalization of principal components. They give a summarization of the data in terms of a 1- d space nonlinearly embedded in the data space. Intuitively, a principal curve '*passes through the middle of the* (*curved*) *data cloud*' (Hastie and Stuetzle, 1989). This method is nonparametric, and the shape of the curve is suggested by the original data. This method has shown its power in dealing with the raw data noise and non-uniform distribution, which is a common phenomenon in the traffic data. For example, many FCD sample points can be obtained during the morning\evening peaks, while fewer in other time periods of a day.

The definition of a principal curve typically depends on the principal component property one wants to generalize. Most of the time, this definition is first stated for an R^d-valued random variable $X=(X_1,X_2,...,X_d)$ with known distribution, and then adapted to the practical situation where one observes independent variables $X_1,X_2,...,X_n$ distributed as X.

The original definition of a principal curve relies on the self-consistency property of principal components. In other words, a smooth (infinitely differentiable) parameterized curve $f(t)=(f_1(t),f_2(t),...,f_d(t))$ is a principal curve for X if f satisfies the following requirements:

1) f does not intersect itself;

2) f has finite length inside any bounded subset of R^d;

3) f is self-consistent.

Here the last requirement means:

$$f(t)=E[X|t_f(X)=t], \tag{1}$$

where the so-called projection index $t_f(x)$ is the largest real number t minimizing the squared Euclidean distance between x and $f(t)$. More formally, the $t_f(x)$ is expressed as follows:

$$t_f(x)=\sup\{t:\|x-f(t)\|=\inf_{t'}\|x-f(t')\|\}. \tag{2}$$

The self-consistency property can also be interpreted by saying that each point of the curve f is the mean of the observations projecting on f around this point.

Figure 1 shows the ability of principal curves as a means to cross the '*middle*' of the data distribution.

3.2 Implementation of principal curves

The iterative learning process of the principal curves can be expressed as follows:

Step 1: Initialization. The initial principal curve $f^{(j)}(t)$ is defined as the first line principal component, here $j=1$.

Step 2: Projection. $\forall x\in R^d$, calculate the $t_f(x)$ in Eq.(2), here, we choose Euclidean distance as the metric.

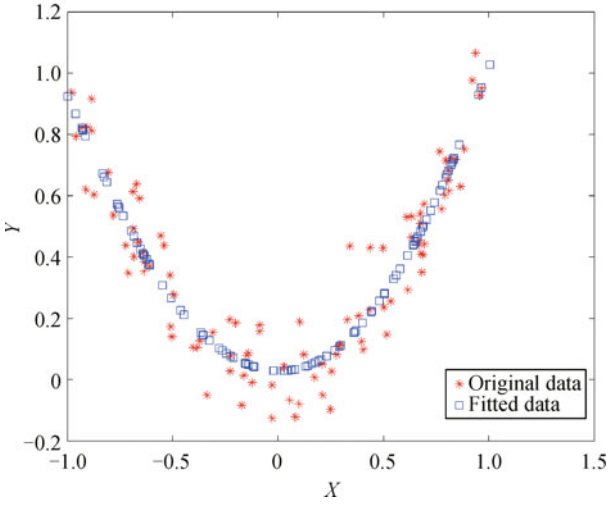

Fig. 1 An example of principal curves. The artificial red points represent the original data set with noise, and the blue curve stands for the fitting result of the principal curves, which represents the '*middle*' of the data cloud

Step 3: Expectation. Based on the self-consistent character, the first principal curve is re-calculated by

$$f^{(j)}(t)=E[X|t_{f^{(j)}}(X)=t]. \tag{3}$$

Step 4: Adjustion. If $1-\dfrac{\Delta(f^{(j+1)})}{\Delta(f^{(j)})}<\varepsilon$, the iteration is stopped. Else, $j=j+1$ and go to step 2. Here ε is a predefined threshold, we take $\varepsilon=0.01$ in our principal curves implementation.

4 Data set & preprocessing

4.1 Definitions

In the following section, first we introduce some terms and concepts used in our research, including GPS log file, FCD structure, and Road. Then the data set we used is described in detail. And last, we give our data preprocessing method based on the distributed architecture Hadoop and MongoDB.

GPS log file: A piece of GPS log file is a collection of GPS points which are collected from the GPS-enabled devices installed in the taxicabs. The data structure is as follows: $gpsfile=\{pt_1,pt_2,...,pt_n\}$ where n is the point record number, pt_i is a single log data which can be regarded as a triple $pt_i=\{lat_i,lon_i,timestamp_i\}$, indicating the exact latitude and longitude at the timestamp i, The GPS log file format in this paper we use contains the ordinary forms such as *gpx*, *kml*, *plt,* and *log*.

FCD structure: A FCD record contains a sequence of GPS log files, and represents the real travel log during a time period. The FCD file is defined as $traj=$

$\{gpsfile_1,gpsfile_2,...,gpsfile_n\}$. To ensure the estimation accuracy of the intersection delay, the condition $\Delta time_i <$ 10 s is required; here $\Delta time_i$ denotes the time interval between two consecutive GPS log files where $\Delta time_i = time_{i+1} - time_i$, $i = 1,2,...,n$.

Road: We define a road as a two-triple: $Road = \{Segment,Node\}$ where $Segment = \{rs_1,rs_2,...,rs_m\}$ stands for the road links and $Node = \{(node_1,node_2,...,node_n) | node_i = (lat_i,lon_i), lat_i \in R, lon \in R\}$ for the intersections in the road network. A piece of road segments rs_i is defined as $rs_i = \{(node_i,node_{i+1}) | node_i,node_{i+1} \in Node\}$. In our research, we use road centerline data and the different lanes are combined.

4.2 Data set & study area

The FCD data set contains 2636149 trajectories from about 20000 GPS-equipped taxicabs in Beijing collected by a business company from March to June in 2011, accounting for about 598 GB in data volume which leads a favorable coverage across the whole city of Beijing. We adopt the metropolitan road network of Beijing which is composed of 18857 nodes and 26621 road segments, holding an amount of 14614 intersections. To get a general idea of Beijing's congestion status, we choose 400 main intersections from all the 14614 ones, which mainly locate on the main roads, namely the expressways, the arterial streets and collector streets. The selected intersections in total represent the skeleton of Beijing's road network. The details are shown in the following Fig. 2:

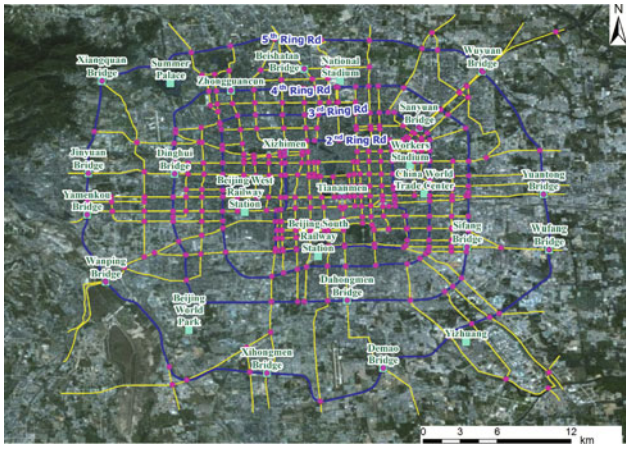

Fig. 2 Study area and main roads with 400 intersections. The blue lines represent the ring roads, the yellow lines stand for the expressways, arterial streets and collector streets, and the purple points are the selected intersections. Some significant landmarks are also labeled with their names, using cyan filled blocks

According to previous discussion in the second section, we adopt 200 m as the intersection's boundary which starts from the center of an intersection and extends 100 m in both directions on the road, as shown in Fig. 3.

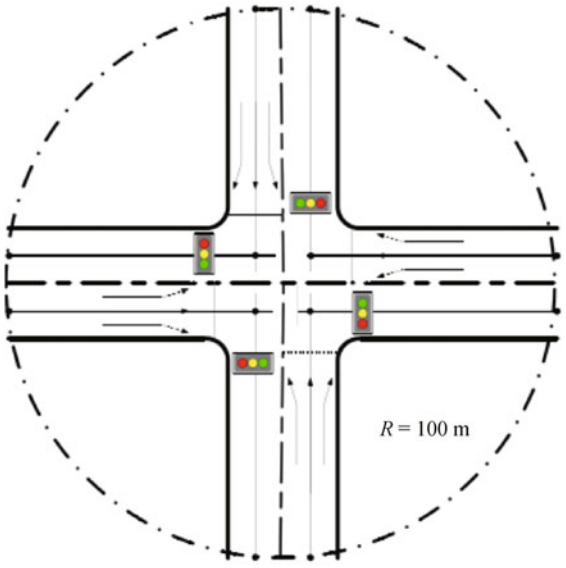

Fig. 3 Demonstration of the boundary of an intersection. In this figure, the circle's center is the junction of the two central lines of the crossing roads, with a radius of 100 m

Other types of intersections, such as T type, F type and so on, can also be treated like the intersection shown above.

4.3 Preprocessing via Hadoop & MongoDB

As shown in Sects. 4.1 and 4.2, the indefinite FCD record length and FCD volume become a big challenge in our study. The lengths of the FCD records are not the same because different taxicabs behave diversely during a day. So the traditional database based on SQL is not suitable for the storage and processing of the data set.

In our research, a novel kind of database, MongoDB, is employed to store the trajectories. MongoDB is part of the NoSQL family of database systems. Instead of storing data in tables as is done in a 'classical' relational database, MongoDB stores structured data as JSON-like documents with dynamic schemas (MongoDB calls the format BSON), making the integration of data in certain types of applications easier and faster. We also set up a distributed data extraction architecture based on the Hadoop framework that supports data-intensive distributed applications. This kind of processing architecture enables applications to work with thousands of computational independent computers and petabytes of data, in our research, we utilize five computers. The adhibition of MongoDB and Hadoop architecture not only solve the processing of our data set, but also provide a new tool to tackle with the data-intensive problem in the following research as the FCD volume increases in the future.

To extract the turn delay records for a given intersection from the database constructed above, we utilize two data

pre-processing procedures in this paper, namely the raw error filtering and the map-matching process. First, all the GPS log files are converted to WGS84 coordinate system according to their respective longitudes and latitudes. This transformation leads to a standard deviation of around 20 m based on empirical evaluation. To filtrate the stationary trajectory records considering that vacant taxis always park on roadside, we then remove the trajectories with large time interval (here we set 30 min as the threshold). Next, with the refined but low-sample-rate trajectories (the sampling rate is about 1 min), we carry out the map-matching procedure based on the theory of mode identification. We not only consider the spatial geometric and topological structure of the road network, but also take the speed constraints into account (we set 150 km/h as the threshold because the maximum speed limit is 120 km/h on the airport freeway. In the inner city, however, most of the arterial roads' speed is confined to 80 km/h). We take discrete GPS points or consecutive track curves of a vehicle for correct location points or road segments as the templates. By comparing each template with the sample, the most similar one is selected as the match result. For more details, please refer to (Lou, et al., 2009). Finally the turn records of all the floating cars on every intersection are calculated using the following algorithm and are stored in the ***RIRresult***, as shown in the following Table 1.

With the ***RIRresult***, we then calculate the elapsed time within a given intersection's boundary at a specific time interval using linear interpolation method, and store the result for the principal curve fitting in the next Section.

Table 1 Turn delay records extraction algorithm

Algorithm ExtractTurnRecord()

Input: road, road intersection, GPS trajectory

Output: Turn record data set:*RIResult*

1. *Traj* = ∅;
2. *RIResult* = ∅;
3. *RNIndex* = *buildSpatialIndex*(*Road*);
4. *RIIndex* = *buildSpatialIndex*(*RoadIntersection*);
5. *Trajectory* = *DataClean*(*GPSTrajectory*);
6. For *trajectory* in *Trajectory*
7. For $point_i$ in *trajectory*
8. $traj_i$ = *EndowGeo*($point_i$,*RNIndex, RIIndex*);
9. *Traj* = *Traj* ∪ $traj_i$;
10. End
11. End
12. For $traj_k$ *in Traj*
13. $result_k$ = *ExactIntersectionRecord*($traj_k$);
14. *RIResult* = *RIResult* ∪ $result_k$;
15. End
16. Return *RIResult;*

5 Experiment

5.1 Principal curves implementation

In this paper, we solve the data sparseness and missing data problems with principal curves based on the description in Sect. 3. First, for a given time period, the turn record number N is compared with the threshold 10. If $N \geqslant 10$, then the average turn delay is estimated as the average result of all the turn records. Otherwise, the missing data at every minute in this time interval is complemented giving consideration to the neighbor values with a result not only abuts the real average value but also reflects the trend changing along the timeline. Based on the fitted time values, the average time for this time interval is re-calculated. A demonstration of the principal curves' fitting ability in dealing with the data sparseness problem is shown in the following Fig. 4. Here the data come from the records of Beishatanqiao intersection with fewer taxicabs from road 704 to road 705 in a single day (2011.3.3, Thursday).

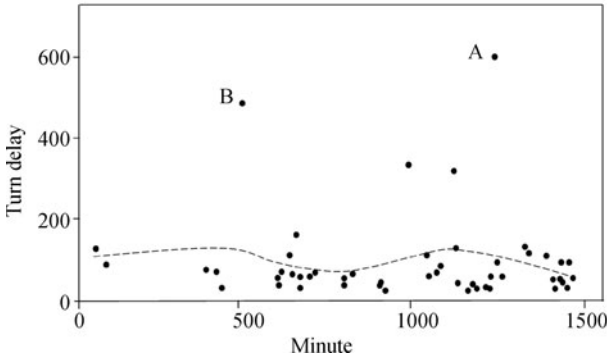

Fig. 4 Demonstration of the fitting ability of principal curves

The horizontal axis stands for the 1440 min during a day and the vertical the record number. From the figure we can see that there are only 51 turn records in this selected day with no taxicabs during the 2: 00–6: 00 time period and only fewer records for every time slots, causing a serious data sparseness and missing data problem. The fitting result via principal curves is shown by the black dashed curve, which passes through the '*middle*' of the data set while ignoring the outlier points shown in the figure as A and B which may be caused by accident, and generates a reliable result.

During our experiment with all the 2636149 trajectories, the new approach based on the principal curves generally surpasses the traditional mean method by 10%–15% in root mean square error (RMSE) for a given FCD file, more higher during the time period 2:00–6:00 and for the marginal intersections which have fewer taxicab records. The mathematical explanation is that the turn delays in a specific time interval may not obey the normal distribution

which is the premise of the mean method, while the principal curve approach is not affected by this precondition because the theory foundation is nonparametric estimation. Part of our turn delay table is also provided in Table 2.

Table 2 Turn delay table in Beijing (part)

ID	FID	TID	TTP	TIID	TD/sec
174	704	705	2	95	43.59
174	704	705	2	96	50.65
174	704	706	3	1	66.41
174	704	706	3	2	54.28
174	704	706	3	3	58.83
174	704	706	3	4	62.10
174	704	706	3	5	72.60
174	704	706	3	6	56.20
174	704	706	3	7	59.45
174	704	706	3	8	65.54
…	…	…	…	…	…

In the table above, the first column **ID** represents the intersection ID labeled from 1 to 400, and the second column **FID** stands for the road segment ID where the taxi comes from the upstream, while the **TID** indicates the downstream road segment ID. The fourth column **TTP** stands for the turn types of this FCD record, with "1" for "Turning left," "2" for "Turn right," and "3" for "Going straight." The fifth column **TIID** means the time interval ID starting from 1 to 96, which divides one day into 96 time slots with 15 min as the interval step. The last column **TD** is our turn delay value based on the principal curve algorithm (in seconds).

5.2 Spatio-temporal analysis of the turn delay table

Based on the turn delay estimation result, a turn delay table for the whole city is built. To diagnose the spatial-temporal characters of the turn delay table of Beijing's road network, we project all the intersection delays onto our study area with three kinds of driving types: turning left, going straight, and turning right. Three specific representative time slices 0: 00–0: 15, 8:15–8: 30 and 12: 00–12: 15 are selected to express the night, the peak, and the normal conditions of a day to get a temporal impression of all the turn delays' distribution. To get a clearer understanding of the turn delays under three turn types at selected time intervals, we utilize the kernel density estimate (KDE) in ArcGIS, with the scale factor setting as 0.009015. In our turn delay table derived from the principal curves method, the maximum turn delay value during these three time periods is no more than 140 s (138 s as the maximum), which is shown in the horizontal color bar below.

From Figs. 5–7, we can see that:

1) The turning left process always consumes the most time at a given time window comparing to the turning right and going straight processes. From the three time intervals shown in Fig. 5, the intersections which consume more time when turning left generally display a similar spatial distribution, but their intensities (the turn delays) behave differently. Obviously in the morning rush hour (8: 15–8: 30), a wider range of regions contain higher turn delays.

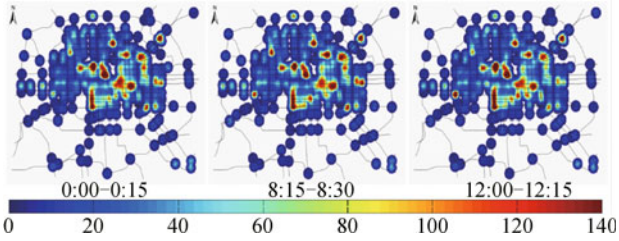

Fig. 5 Spatial distribution of turning left at three given time intervals

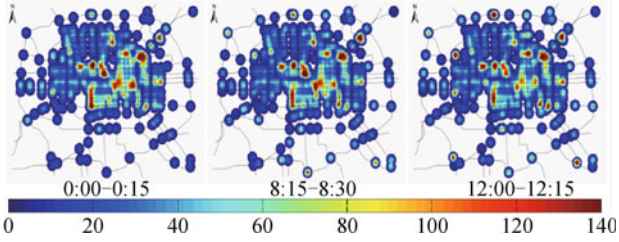

Fig. 6 Spatial distribution of turning right at three given time intervals

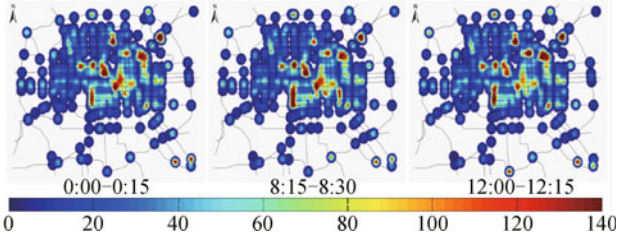

Fig. 7 Spatial distribution of going straight at three given time intervals

2) The turning right process always consumes the least time at a given time period. Unlike the spatial distribution of turning left process, in Fig. 6, the intersections which possess a higher turn delays seem more scattered, especially on the overpasses at the boundary of the city.

3) For the going straight pattern, there is no remarkable difference between the three time periods except for a slight variation in turn delays during the time period 8: 15–8: 30 and the time period 12: 00–12: 15. Although the average number of intersections which have a higher level

of going straight delays has declined during the time period 0: 00–0: 15, some intersections on the boundary still hold a higher level of turning delays in the deep night. These three pictures also prove that the novel fitting algorithm based on the principal curves is reliable and effective in dealing with the sparse data set, because in the time period 0: 00–0: 15 there are not so many taxicabs driving on the roads as during the other two time periods.

5.3 Turn delay ratio

The turn delay ratio represents the portion of the whole travel time wasted at intersections and gives a rough idea of the traffic condition of the road network regardless of the road level hierarchies and time slots. Based on the huge floating car data set, we first compute the turn delay ratio of the whole network. Every piece of the 2636149 trajectory records is extracted and matched to the detailed road network. Because there are so many alleyways across the whole road network with an average distance of 202 m between two intersections, the range of an intersection described in Sect. 4.2 cannot be used. Here we first calculate the elapsed time IT^j between two consecutive GPS log points with linear interpolation, and secondly the real elapsed time RT^j is calculated as the time difference between the two GPS log points. If there are intersections between these two successive points, then the turn delays between the two points is marked as T^j, where $T^j = RT^j - IT^j$. All the T^j along a single trajectory are then accumulated as the total turn delays T^i_{delay}, where $i = 1,2,...,2636149$. The turn delay ratio $Ratio_i$ is calculated as $Ratio_i = T^i_{delay}/T^i_{total}$, where T^i_{total} means the total travel time of this selected trajectory. Two approaches to calculate the final turn delay ratio for the whole network are employed:

$$RATIO1 = \sum_{i=1}^{n} T^i_{delay} / \sum_{i=1}^{n} T^i_{total}, \tag{4}$$

$$RATIO2 = \frac{1}{n} \sum_{i=1}^{n} T^i_{delay} / T^i_{total}. \tag{5}$$

The results make no significant differences: $RATIO1 = 0.5909$ and $RATIO2 = 0.6038$. Therefore we take the turn delay ratio for the whole network as $RATIO_{ALL} = 60\%$, which means almost 60% of time is generally wasted at the intersections during one's travel in Beijing. To our knowledge, this conclusion is much higher compared to the situation in the medium congested Municipality of Copenhagen (the turn delay ratio is about 17%–35%) (Nielsen et al., 1998) and previous probabilistic assumption in Beijing (the turn delay ratio is about 38%) (Zheng et al., 2010), indicating that the traffic congestion has become more serious. And the result also demonstrates the value of our turn delay estimation work.

Next, we also calculate the turn delay ratio on the selected 400 intersections on the main roads (the expressways, the arterial streets, and the collector streets), which compose the skeleton of the road network in our research. We utilize 200 m as the range of an intersection which is defined in Sect. 4.2, and the rest of the processing is the same to the whole road network's calculation. To our surprise, although the 400 selected intersections take only 2.7% of the total intersections, the final result $RATIO_{MAINROADS}$ contributes about 18% of total travel time during one trip in Beijing. This conclusion offers a deeper understanding of Beijing's real traffic conditions.

6 Discussion

The turn delay table provides a basic idea of Beijing's road network and can be regarded as an effective and reliable baseline to ameliorate the online travel time prediction and daily route planning. However, there are still some remaining limitations to discuss in the processing and implementation of our methods:

1) In our research, we take a piece of road as a whole while neglecting the different composition of roads, that is, different taxicab trajectories on different lanes with the same direction are combined together to calculate the turn delays at the same intersection. Another data confusion problem may occur on the double-deck arterial roads, different trajectories on different decks while heading direction during the same time period are also considered homogenous. To get a more accurate estimation, the FCD refining problem cannot be neglected. The data separation from diverse sources can be exacted and matched with a more elaborate road network map.

2) We adopt some predetermined parameters described in a list of previous literatures, however, the feasibility of these parameters should be carefully considered. First, Wang suggests the length between two timestamps be 15 min for the freeway facility study (Wang, 2004). Is this kind of time partition method fully suitable for our research? Considering the heterogeneity of taxicabs' spatial distribution and different record densities during a day, we suspect that the length of a time interval should not be limited by 15 min, that is, the time granularity should go along with the record numbers to get a more appropriate estimation. Secondly, Xiong et al. concluded that 7 to 9 FCD records are needed in Beijing road network's 2nd Ring Road and 3rd Ring Road every kilometer in 5 min (Xiong et al., 2010), however, in our research we take 10 records as our threshold. Our research area is also not confined within the 2nd Ring Road and 3rd Ring Road but with a much wider region (all the urban area and suburbs within the 5th Ring Road are included). Is this threshold setting can meet the need of further study? A more

meticulous test over the whole research region should be carried out in future experiments. Thirdly, following Zhang et al., the region of an intersection defined in this paper is 200 m which spreads from the center of an intersection and extends 100 m both upstream and downstream (Zhang et al., 2011). This definition of an intersection overlooks the differences of intersections' structures and the different spatial-temporal taxicabs' distribution, which should be re-examined in our future work.

3) Although Fabritiis et al. and Hong et al. have argued the importance of the percentage of floating cars participating in the traffic flow (Fabritiis, et al., 2008; Hong, et al., 2007), we insist that without extra devices such as real-time cameras and other online traffic flow information this is not a trivial task to examine the timely floating cars' participating percentage. Nevertheless, the taxicab drivers are a group of people with experienced skills and advance physical knowledge about the real world, so the trajectory records can be regarded as the ground-truth information which can well reflect the real traffic condition and thus the participating percentage becomes less significant in our study.

4) The roughly 60% turn delay ratio across the whole city seems higher than our expectation. We think there are three reasons: i) the taxicabs' spatial distribution is diverse: more taxicabs are congested in the downtown area, and fewer in the urban fringe areas; ii) the taxicabs' temporal distribution is also unbalanced during different time periods of a day, with more taxicabs operating in the daytime while fewer operate during the night. iii) the intersections without traffic lights, which are mainly located on the low level roads (e.g., the alleyways) are also included in the calculation, which may partly overestimate the real condition. Given consideration to these limitations, a more precise and time-dependent turn delay ratio which gives consideration to different confidence intervals as the FCD records variance in volume should be included in our future work.

5) We afford a static statistical intersection delay estimation based on the rich FCD which can be regarded as the base line of the road network. However, a dynamic turn delay table incorporating weather conditions, traffic accidents, big events(concerts, holidays etc.) and individual driving strategies (both of the fleet drivers and of the end users for whom the route is computed) is more promising. This research calls for extra auxiliary data. The micro blog may become a sound source for this further task. There is still a long way to go.

Despite these limitations, we believe that this study is useful because it focuses on the intersection delays on the main roads which reflect the road network's real-time status. In addition, we establish a distributed computing architecture based on Hadoop and MongoDB which can also meet the future need as the data volume increases, and derive the turn delay table from thousands of records. Furthermore, the turn delay ratios across the whole network and on the main 400 intersections provide an initial impression to understand the city's real traffic conditions and reflect the city's mobility, encouraging further study.

7 Conclusions

In this paper, we focus on the intersection delays on the road network in Beijing. We propose a data-intensive calculation framework based on the Hadoop and MongoDB, which is not only suitable for current research but also for future research. A novel fitting algorithm based on principal curves is adapted in our research to deal with the data sparseness and missing data problem which in return reduces the influence of outlier points and generally gains a 10%–15% higher accuracy in RMSE compared to traditional methods such as mean and median. We weave a turn delay table for Beijing's main intersections on the arterial roads which can provide supplementary turn delay information at any time for travel time planning and control. Furthermore, based on the turn delay table, the spatial-temporal characteristics of Beijing's road network at three representative time periods: 0: 00–0: 15, 8: 15–8: 30 and 12: 00–12: 15 are diagnosed with the result showing that the traffic condition has become a serious problem. In addition, we compute the turn delay ratios both on the whole network and on the selected arterial roads. The result shows that the average turn delay ratio during one trip in Beijing is about 60%, which is much higher compared to the situation in the medium congested Municipality of Copenhagen (the turn delay ratio is about 17%–35%) and previous probabilistic assumption in Beijing (the turn delay ratio is about 38%). Although the 400 selected intersections on the main roads (the expressways, the arterial streets, and the collector streets) take only about 2.7% of all the intersections distributed among the whole road network, cost about 18% of travel time, compared to the average of 60% on the whole road network. This conclusion also implies that the traffic congestion in Beijing has become a serious problem which needs our further research. Our future work will mainly focus on the traffic light controlled intersections, considering the differences of various spatio-temporal conditions and ameliorating our turn delay table with other kinds of non-parametric methods.

Acknowledgements This research was supported by the National Natural Science Foundation of China (Grant No. 41271408), the National Hi-tech Research and Development Program of China (No. 2012AA12A211) and State Key Laboratory of Resources and Environmental Information System Open Foundation (No. 088RA500KA). And we also thank the anonymous referees for their helpful comments and suggestions.

References

Ban X J, Herring R, Hao P, Bayen A M (2009). Delay pattern estimation for signalized intersections using sampled travel times. Transport Res Rec, 2130(1): 109–119

Bejan A, Gibbens R, Evans D, Beresford A, Bacon J, Friday A (2010). Statistical modelling and analysis of sparse bus probe data in urban areas. In:13th International IEEE Conference on Intelligent Transportation Systems (ITSC), September 19–22, Madeira Island, Portugal, 1256–1263

Boyce D E, Kirson A, Schofer J L (1991). Design and implementation of ADVANCE: the Illinois dynamic navigation and route guidance demonstration program. In: IEEE Vehicle Navigation and Information Systems Conference(VNIS'91), October 20–23, Dearborn, Michigan, United States, 2: 415–426

Cheu R L, Xie C, Lee D (2002). Probe vehicles population and sample size for arterial speed estimation. Comput-Aided Civ Inf, 17(1): 53–60

Chodorow K, Dirolf M (2010). MongoDB: The Definitive Guide. Sebastopol, O'Reilly Media, Inc

Cowan K, Gates G (2002). Floating Vehicle Data system-realisation of a commercial system. In: 11th International Conference on Road Transport Information and Control, March 19–21, London, the United Kingdom, 87–189

Fabritiis C D, Ragona R, Valenti G (2008). Traffic estimation and prediction based on real time floating car data. In: 11th International IEEE Conference on Intelligent Transportation Systems, October 12–15, Beijing, China, 197–203

Ghaffarian H, Fathy M, Soryani M (2012). Vehicular ad hoc networks enabled traffic controller for removing traffic lights in isolated intersections based on integer linear programming. Iet Intell Transp Sy, 6(2): 115–123

Gühnemann A, Schäfer R P, Thiessenhusen K U, Wagner P. (2004). Monitoring traffic and emissions by floating car data. Institute of Transport Studies Working Paper, Issue ITS-WP-04–07, Sydney, Australia

Hastie T, Stuetzle W (1989). Principal curves. J Am Stat Assoc, 84(406): 502–516

Heidemann D (1994). Queue length and delay distributions at traffic signals. Transport Res B—Meth, 28(5): 377–389

Hellinga B R, Fu L (2002). Reducing bias in probe-based arterial link travel time estimates. Transport Res C—Emer, 10(4): 257–273

Herrera J, Work D, Herring R, Ban X J, Jacobson Q, Bayen A M (2010). Evaluation of traffic data obtained via GPS-enabled mobile phones: the Mobile Century field experiment. Transport Res C—Emer, 18(4): 568–583

Herring R, Hofleitner A, Abbeel P, Bayen A (2010). Estimating arterial traffic conditions using sparse probe data. In: Proceedings of the 13th International IEEE Conference on Intelligent Transportation Systems, September 19–22, Madeira Island, Portugal, 929–936

Homburger W S, Hall J W, William R R, Edward C S, Michelle D, Loretta H, John J L, Matthew R, Vernon H W (2007). Fundamentals of traffic engineering. University of California, Berkeley, Institute of Transportation Studies

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

Hunter T, Herring R, Abbeel P, Bayen A (2009). Path and travel time inference from GPS probe vehicle data. In: NIPS Analyzing Networks and Learning with Graphs, December 7–10, British Columbia, Canada

Kothuri S M, Tufte K A, Fayed E, Robert L B (2008). Toward understanding and reducing errors in real-time estimation of travel times. Transport Res Rec, (2049): 21–28

Kwong K, Kavaler R, Rajagopal R, Varaiya P (2009). Arterial travel time estimation based on vehicle re-identification using wireless magnetic sensors. Transport Res C—Emer, 17(6): 586–606

Li R, Rose G (2011). Incorporating uncertainty into short-term travel time predictions. Transport Res C—Emer, 19(6): 1006–1018

Liu H X, Ma W, Wu X, Hu H (2012). Real-time estimation of arterial travel time under congested conditions. Transportmetrica, 8(2): 87–104

Long J, Gao Z, Zhao X, Lian A, Orenstein P (2011). Urban traffic jam simulation based on the cell transmission model. Netw Spat Econ, 11 (1): 43–64

Lou Y, Zhang C Y, Zheng Y, Xie X, Wang W, Huang Y (2009). Map-matching for low-sampling-rate GPS trajectories. In: Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. Seattle, Washington, 352–361

Lu C (2010). A travel time estimation for planning models considering signalized intersections.ITE J, 80(10): 34–39

Nielsen O A, Frederiksen R D, Simonsen N (1998). Using expert system rules to establish data for intersections and turns in road networks. Int Trans Oper Res, 5(6): 569–581

Skabardonis A, Geroliminis N (2005). Real-time estimation of travel times on signalized arterials. In: Hani S M, ed. Transportation and Traffic Theory. Flow, Dynamics and Human Interaction. 16th International Symposium on Transportation and Traffic Theory, July 19-21, College Park Maryland, United States, 387–406

Sun L (2007). An approach for intersection delay estimate based on floating vehicles. Dissertation for Ph.D. Degree. Beijing: Beijing University of Technology (in Chinese)

Wang Z (2004). Using floating cars to measure travel time delay: How accurate is the method? Transport Res Rec, 1870(1): 84–93

White T (2010). Hadoop: The Definitive Guide. Sebastopol: Yahoo Press, O'Reilly Media, Inc

Winter S (2002). Modeling costs of turns in route planning. GeoInformatica, 6(4): 345–361

Xie X, Cheu R L, Lee D H (2001). Calibration-free arterial link speed estimation model using loop data. J Transp Eng, 127(6): 507–514

Xiong J, Liu J, Guan J, Sun J, Liu X, Wen H (2010). The minimum sample size determination of floating car in Beijing expressway. J Transp Syst Eng Inf Technol, 4: 38–43 (in Chinese)

Zhang H C, Lu F, Zhou L, Duan Y Y (2011). Computing turn delay in city road network with GPS collected trajectories. In: Proceedings of the 2011 International Workshop on Trajectory Data Mining and Analysis. September 17–21, Beijing, China, ACM,45–52

Zhang J P, Wang J (2003). An overview of principal curves. Chinese Journal of Computers, 26(2): 129–146 (in Chinese)

Zheng N B, Lu F, Duan Y Y (2010). Dynamic dual graph model for turn

delays on road networks. Journal of Image and Graphics, 15(6): 915–920 (in Chinese)

Zheng Y, Zhou X F (2011). Computing with Spatial Trajectories. New York: Springer-Verlag

AUTHOR BIOGRAPHIES

Xiliang Liu obtained his B.S. degree in land resources management in Tongji University, China, 2007 and M.S. degree in geographic information engineering in China University of Mining & Technology (Beijing), China, 2011. Now he is a Ph.D. candidate in geographic information system in the Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences. His current research interests mainly focus on trajectory data mining and intersection delay estimation. E-mail: liuxl@lreis.ac.cn.

Dr. Feng Lu is a Professor of Institute of Geographical Sciences and Natural Resources Research, Chinese Academy of Sciences. He received his B.S. degree from Wuhan University and Ph.D. degree from Institute of Remote Sensing Applications, Chinese Academy of Sciences. He is the Deputy Director of State Key Laboratory of Resources and Environmental Information System, a guest Ph.D. Advisor for the Fuzhou University, a member of the Information Technology Committee of Chinese Transportation Association, a member of the Theory and Methodology Committee of the Chinese Association of GIS and a member of the New Technology Applications Committee of Chinese City Planning Association. Dr. Lu's research interests cover spatial data modeling, spatial DBMS, trajectory data mining, GIS for transportation and urban GIS application. During the past years, he has published over 120 referred journal and conference papers. E-mail: luf@lreis.ac.cn.

Hengcai Zhang received his B.S. degree in geographic science from Shandong Normal University and M.S. degree in cartography and geographic information system from Capital Normal University. Now he is a Ph.D. student from Institute of Geographical Sciences and Natural Resources Research, Chinese Academy of Sciences. His interests focus on moving objects database and spatial-temporal data mining. E-mail: zhanghc@lreis.ac.cn.

Peiyuan Qiu received the B.S. degree in geographic science from Qingdao University, China in 2009 and the M. E. degree in cartography and geographic information engineering from Beijing University of Civil Engineering and Architecture, China in 2012. He is currently a Ph.D. candidate in Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences. His current research interests focus on geographic semantics and trajectory mining. E-mail: qiupy@lreis.ac.cn.