Urban link travel time estimation using large-scale taxi data with partial information

ARTICLE *in* TRANSPORTATION RESEARCH PART C EMERGING TECHNOLOGIES · AUGUST 2013

Impact Factor: 2.82 · DOI: 10.1016/j.trc.2013.04.001

4 AUTHORS:

Xianyuan Zhan
Purdue University
8 PUBLICATIONS **21** CITATIONS
SEE PROFILE

Samiul Hasan
The Commonwealth Scientific and Ind…
20 PUBLICATIONS **123** CITATIONS
SEE PROFILE

Satish V Ukkusuri
Purdue University
117 PUBLICATIONS **949** CITATIONS
SEE PROFILE

Camille N. Kamga
City College of New York
33 PUBLICATIONS **60** CITATIONS
SEE PROFILE

Transportation Research Part C 33 (2013) 37–49

Contents lists available at SciVerse ScienceDirect

Transportation Research Part C

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

Urban link travel time estimation using large-scale taxi data with partial information

Xianyuan Zhan [a], Samiul Hasan [b], Satish V. Ukkusuri [a,*], Camille Kamga [c]

[a] School of Civil Engineering, Purdue University, 550 Stadium Mall Drive, West Lafayette, IN 47907, USA
[b] School of Civil Engineering, Purdue University, 550 Stadium Mall Drive, West Lafayette, IN 47907-2051, USA
[c] Civil Engineering, Marshak Hall, Suite 910, 160 Convent Avenue, The City College of New York, New York, NY 10031, USA

ARTICLE INFO

Article history:
Received 16 August 2012
Received in revised form 1 April 2013
Accepted 1 April 2013

Keywords:
Traffic state estimation
Path inference
Large scale data analysis
GPS-enabled taxicab
Probe vehicle
Urban networks

ABSTRACT

Taxicabs equipped with Global Positioning System (GPS) devices can serve as useful probes for monitoring the traffic state in an urban area. This paper presents a new descriptive model for estimating hourly average of urban link travel times using taxicab origin–destination (OD) trip data. The focus of this study is to develop a methodology to estimate link travel times from OD trip data and demonstrate the feasibility of estimating network condition using large-scale geo-location data with partial information. The data, collected from the taxicabs in New York City, provides the locations of origins and destinations, travel times, fares and other information of taxi trips. The new model infers the possible paths for each trip and then estimates the link travel times by minimizing the error between the expected path travel times and the observed path travel times. The model is evaluated using a test network from Midtown Manhattan. Results indicate that the proposed method can efficiently estimate hourly average link travel times. This research provides new possibilities for fully utilizing the partial information obtained from urban taxicab data for estimating network condition, which is not only very useful but also is inexpensive and has much better coverage than traditional sensor data.

1. Introduction

Accurate estimation and prediction of urban link travel times are important for improving urban traffic operations and identifying key bottlenecks in the traffic network. They can also benefit users by providing accurate travel time information, thereby allowing better route choice in the network and minimizing overall trip travel time. However, to accurately assess link travel times, it is important to have good real-time information from either in-road sensors such as loop detectors, microwave sensors, or roadside cameras, or mobile sensors (e.g. floating cars) or Global Positioning System (GPS) devices (e.g. cell phones). However, in most of these cases, only limited information is available related to speed or location, hence, one has to develop appropriate methodologies to accurately estimate the performance metric of interest at the link, path or network level.

In the last few years, there has been a growing trend of implementing GPS installed taxicabs in urban areas. While GPS-equipped taxicabs have many advantages, including the ability to locate taxis and track lost packages, they also serve as useful real-time probes in the traffic network. Taxis equipped with GPS units provide a significant amount of data over days and months thereby providing a rich source of data for estimating network wide performance metrics. However, currently there

* Corresponding author.
E-mail addresses: zhanxianyuan@gmail.com (X. Zhan), samiul.hasan@gmail.com (S. Hasan), sukkusur@purdue.edu (S.V. Ukkusuri), ckamga@utrc2.org (C. Kamga).

0968-090X/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.trc.2013.04.001

are limited methodologies making use of this new source of data to estimate link or path travel times in the urban network. Within this context, this paper proposes a new method for estimating hourly urban link travel times using large-scale taxicab data with partial information. The taxicab data used in this research provides limited trip information, which only contains the origin and destination location coordinates, travel time and distance of a trip. However, the extensive amount of data records compensates for the incompleteness of the data and makes the link travel time estimation possible. A novel algorithm for estimating the link travel times will be presented and tested in this paper using a test network in New York City.

1.1. Related work

Previous research on urban link travel time estimation and prediction has largely relied on various data sources, including: loop detectors (Coifman, 2002; Zhang and Rice, 2003; Oh et al., 2003; Wu et al., 2004), automated vehicle identification (AVI) (Park and Rilett, 1998; Li and Rose, 2011; Sherali et al., 2006), video camera, Remote Traffic Microwave Sensors (RTMS) (Yeon et al., 2008), and automated number plate recognition (Hasan et al., 2011). All of these data collection methods require installing corresponding sensors to retrieve data. Therefore a large number of sensors are required to achieve a reasonable accuracy level based on these data sources. The cost of installing and maintaining such a large number of sensors is prohibitive. Hence predicting link travel times with reasonable accuracy and network coverage based on sensor data could be expensive.

On the other hand, there is a significant potential to use emerging large-scale data sources to estimate dynamic demand and dynamic network conditions in urban areas. For instance, GPS devices in dedicated fleets of vehicles or in users' mobile phones can be viable sources of data for monitoring traffic in large cities (Herrera et al., 2010). Industry models, such as Inrix,[1] have also gained popularity in recent years where private entities install, collect, utilize and sell "large-scale" historical traffic data from GPS-equipped vehicles or mobile phones. With an increasing amount of GPS data available from taxi, transit, and mobile phones, a new option of using such large-scale decentralized data for link travel time estimation becomes realistic. Herring et al. (2010) used GPS traces data from a fleet of 500 taxis in San Francisco, CA. to estimate and predict traffic conditions. However, in this work, instead of link travel times, discrete traffic states were predicted. Zheng and Van Zuylen (in press) also proposed an ANN model to estimate urban link travel times based on sparse probe vehicle data (e.g., GPS traces from GPS-equipped vehicles or smartphones). Hunter et al. (2009) proposed a statistical approach for path and travel time inference using GPS probe vehicle trajectory data. The GPS data used in their study has been recorded each minute, where the inferred path consists of at most five link segments. This method is not applicable if the GPS data has a longer recording interval or only has the starting and ending coordinates. Estimating link travel times from GPS data provides a much cheaper and a larger coverage area in the urban network compared with approaches using fixed sensor data. However, all of the above mentioned approaches are only applicable for GPS trace data, in which the trajectories of vehicles are available. To the best of our knowledge, there is no study found in literature that used OD level GPS data for urban link travel time estimation, even though extensive amount of such less detailed data (e.g. taxicab data) is generated and recorded every day.

1.2. Background and objectives

In New York City, GPS devices are installed in each taxicab. The taxicab data is collected and archived by the New York City Taxi and Limousine Commission (NYTLC), an agency that is responsible for all taxi related issues in New York City. The New York City has the largest market for taxis in North America with 12,779 (in 2006) yellow medallion taxicabs serving about 240 million passengers a year. The taxi service transports 25% of all fare-paying bus, subway, taxi and for-hire vehicle passengers that are traveling within Manhattan (Schaller Consulting, 2006; King et al., 2012).

In this paper, data collected from New York City taxicabs is used to estimate the link travel times. The dataset provides an extensive amount of taxi trip data, which records the trip starting and ending geo-location, along with information about trip distance, time and fare. Unlike the detailed GPS trajectory data used in previous studies, the dataset only provides the trip origin and destination information (i.e. starting, ending location and time) without the exact trajectory of the taxicab; only path travel time and distance are known. However, the advantage of the massive amount of data (the number of observations recorded within a day range between 450,000 and 550,000) makes it possible to infer the possible routes that the taxicab is taking and further, to estimate the link travel times in the New York City network. There is potential bias associated with measuring network link travel times from taxis, as taxi drivers are just one particular group of all drivers in the network. However, given the high penetration rate of taxicabs, it is reasonable to assume that taxis are good probe vehicles and therefore taxi travel times are a good representation of the actual network condition.

In this research we propose a methodology to estimate urban link travel times based on taxi GPS data that includes only the information about the origin and destination of the trip and total travel time to reach the destination. The goal of this study is to show the potential of using taxicab data as a complimentary data source in urban transportation operation and management. The link travel times estimated from taxicabs provide an hourly aggregate measure of the urban network condition, which can be fused with the information from other existing data sources such as fixed sensors in the future.

[1] Inrix, Inc. http://www.inrix.com.

The paper is organized as follows: the next section describes the methodological approach developed in the paper to estimate link travel times; the subsequent sections present the test data and network, and the model results respectively. The final section presents the concluding remarks.

2. Methodology

This section presents the proposed link travel time estimation model. We treat the path taken by a taxi as *latent* and derive the expected path travel time as a summation of each of the probable path travel time multiplied by the probability of taking that particular path. Link travel time estimation problem then becomes estimating the link travel times that minimize the least square error between the observed and expected path travel times. An MNL model is embedded to compute the probability that a taxi driver chooses a given path in the constructed reasonable path set, and the expected path travel time is computed for each trip record. The data are first processed to run the model, which include two steps: data mapping and constructing reasonable path set. The taxi trip origin and destination points are first mapped to the nearest links in the network. Instead of using all possible paths between each origin and destination points, we use *k*-shortest path algorithm to construct 20 shortest paths for each OD nodal pair of a trip, referred as the reasonable path set. The generated reasonable path sets serve as the basis for the link travel time estimation process.

2.1. Link travel time estimation

Link travel times in the network are estimated by minimizing the least squared difference between expected path travel times and the observed path travel times. We consider the actual path choice of the taxi as a latent variable and the link travel times as the model parameters to be estimated, the expected path travel time for observation i, $E(Y_i|R_i)$ can be written as:

$$E(Y_i|R_i) = \sum_{m\in R_i} g_m(\vec{t}) P_m(\vec{t}, d, \theta) \tag{1}$$

where Y_i is the variable of path travel time for observation i; R_i is the set of possible paths of a OD trip observation i; $\vec{t}$ is the vector of link travel times; d is the path distance set for R_i; $g_m(\vec{t})$ is the path travel time for path m; $P_m(\cdot)$ is the probability of selecting path m; and θ is a positive scale parameter.[2]

For a given path, the path distance is fixed, the variables to be estimated are the vector of link travel times $\vec{t}$ and the scale parameter θ. Then, $E(Y_i|R_i)$ can be represented by a function of $R_i, \vec{t}$ and θ,

$$E(Y_i|R_i) = f(R_i, \vec{t}, \theta) \tag{2}$$

The error between observed path travel time y_i and expected path travel time $E(Y_i|R_i)$ is defined as the residual for observation i, which is:

$$r_i = y_i - f(R_i, \vec{t}, \theta) \tag{3}$$

Link travel times are estimated by minimizing the square difference between the expected path travel times and the actual path travel times observed in the data set D, defined as $S(\vec{t}, \theta)$,

$$S(\vec{t}, \theta) = \sum_{i\in D} r_i^2 = \sum_{i\in D} (y_i - f(R_i, \vec{t}, \theta))^2 \tag{4}$$

$$\vec{t} = \arg\min_{\vec{t}} S(\vec{t}, \theta) \tag{5}$$

2.2. Route choice model

Due to the absence of any information on the path taken by the taxicab drivers, the actual path needs to be inferred. Thus a route choice model is developed to find the path choice of the taxicab drivers. Due to the lack of social or behavioral characteristics of taxi drivers in the dataset, traditional econometric models cannot be estimated. Hence, we build the route choice model using the limited cost variables from the dataset. We implement an MNL model to serve as the route choice model and consider the trip cost C_m in terms of both trip time and distance. The route choice model is defined as

$$P_m(\vec{t}, d, \theta) = \frac{e^{-\theta C_m(\vec{t}, d_m)}}{\sum_{j\in R_i} e^{-\theta C_j(\vec{t}, d_j)}} \tag{6}$$

The parameter θ scales the perceived path cost. A large θ indicates a small perception error, and drivers will tend to select the path with minimum cost; while a small θ suggests a large perception variance, larger cost path gets more probability of being

[2] This will be further discussed in the route choice model.

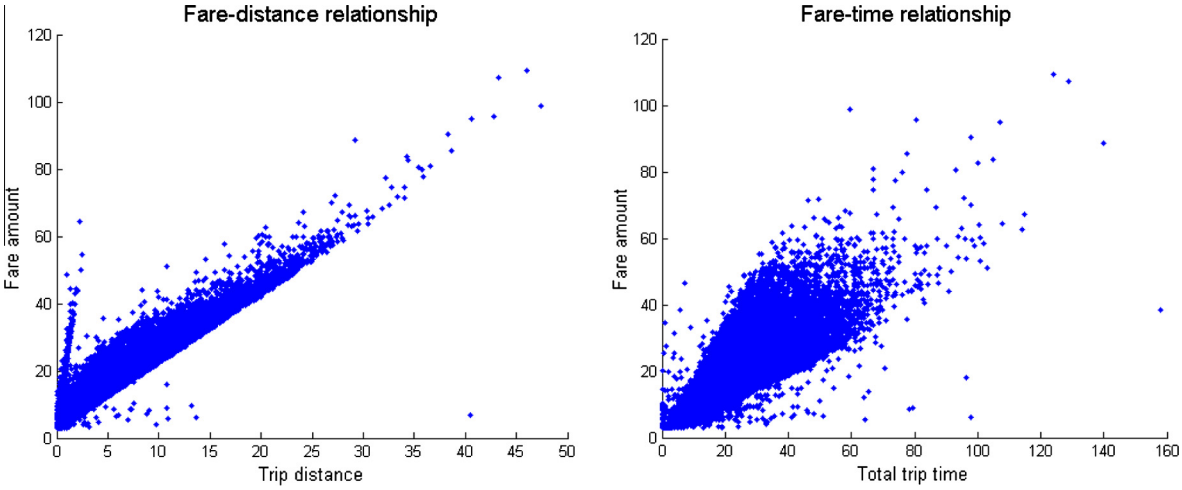

Fig. 1. Fare-time-distance relationship.

selected. In this model, θ is estimated together with the link travel times, which captures the variation in drivers' perceived path cost in different time period and network conditions.

The path cost C_m can be assumed as a function of trip fare. This is based on the assumption that each driver minimizes both trip time and distance, so that the driver can make more trips and thus make more revenue. We introduce a threshold ratio when constructing the reasonable path sets to exclude the trips that violate the aforementioned route choice behavior assumption. That is, if the taxi driver takes a much longer route to make more revenue in a single trip, then none of the paths in the reasonable path set will fall within the threshold given the observed path distance. These records are removed from the model estimation to ensure the input data matches with the route choice behavior assumption.

According to the taxicab fare rates provided by New York Taxi and Limousine Commission, the taxicab fare calculation involves both trip time and distance.[3] For standard city rate (taxi trips within Manhattan all follow this rate), fare (exclude surcharge and tax) include \$2.50 upon entry, and \$0.5 for each additional unit. The unit fare is:

- One-fifth of a mile, when the taxicab is traveling at 6 miles an hour or more; or
- 60 s when not in motion or traveling at less than 6 miles per hour.
- The taximeter shall combine fractional measures of distance and time in accruing a unit of fare.

The taxi rate of fare suggests a linear relationship with trip time and distance. The actual fare-time-distance relationship from the data is illustrated in Fig. 1. Considering the complicated traffic condition and fare calculating method in actual situations, a linear model for the trip fare-time-distance relationship estimated from the data is used rather than the rate of fare provided by NYTLC:

$$fare = \beta_0 + \beta_1 \cdot time + \beta_2 \cdot distance \tag{7}$$

The estimated coefficients of β_0, β_1 and β_2 are listed in Table 1. The units for time and distance are minute and mile respectively; the fare used in the calculation does not include surcharge and tax. The estimation result shows that time and distance are highly significant in determining the trip fare. The model has a R^2 value of 0.99, suggesting that the data is well fitted using this simple linear model. The path cost used in the route choice model is therefore modeled as:

$$C_m(\vec{t}, d_m) = \beta_1 \cdot g_m(\vec{t}) + \beta_2 \cdot d_m \tag{8}$$

where d_m is the distance for path m, and the path travel time of path $m, g_m(\vec{t})$ is defined as

$$g_m(\vec{t}) = \alpha_1 t_0 + \alpha_2 t_D + \sum_{l \in L} \delta_{ml} t_l \tag{9}$$

where t_0 is the travel time of the link where the trip starting point lies; t_D is the travel time of the link where the trip ending point lies; L is the set of the links; t_l is the travel time of the link l; δ_{ml} is the link-path incident relationship, 1 if link l is in path m, 0 otherwise; and α_1, α_2 are the distance proportions.

The simple linear form of the path cost function is used for two reasons: (1) the linear fare-time-distance relationship is supported by data, and distance and time are identified as significant factors that impact the trip fare; and (2) a simple form of path cost function ensures the model is computationally tractable for large-scale input data and the short term link travel

[3] Taxicab rates from New York Taxi & Limousine Commission: http://www.nyc.gov/html/tlc/html/passenger/taxicab_rate.shtml.

Table 1
Linear model for fare-time-distance relationship.

	Coefficient	Standard deviation	*P* value
β_0 (intercept)	2.143	0.00161	0.000
β_1 (coefficient for time)	0.275	0.00021	0.000
β_2 (coefficient for distance)	1.563	0.00058	0.000
Number of observations		415,561	
R-squared		0.99	
Adjusted *R*-squared		0.99	

time estimation purpose. The constant term is not included since this common component cancels out in the MNL model. Further, as the starting and ending points lie within the starting and ending links, a taxi only experiences a part of the total link travel times to traverse those links. In this study, the proportion of this part of link travel time to the total link travel time is assumed to be the distance proportions α_1 and α_2 defined in the data mapping section.

2.3. Data mapping

It is common in urban environments such as New York City that taxicabs often travel in the GPS shadow of tall buildings causing errors in the GPS data. Thus a data mapping process is introduced to pre-process the raw GPS data. There are two purposes in this step: first, to map the data to nearest links in the road network to reduce GPS errors; second, to match the starting and ending points to the actual road network and transform the raw data into usable data for network level analysis.

Fig. 2 illustrates the data mapping procedure. The raw origin and destination points (black points in Fig. 2) are mapped to the perpendicular foot of the nearest link (blue points in Fig. 2), and the new points are then used in the later analysis. The locations (represented by distance ratio between two endpoints of a link) of the new points on the link are also computed to calculate accurate *k*-shortest path distance in later step.

The new origin and destination points correspond to four endpoints of two links. In big cities like New York, a great proportion of the links in the urban grid network are one-way streets. For origin and destination points that lie on one-way streets, the actual two intermediate nodes in abovementioned four endpoints are easily identified given the directional information of the link. For any point lies on the two-way street, both the two endpoints of this link are used as intermediate origin/destination nodes for this record. All the combination of the intermediate origin and destination nodes and the corresponding shortest path sets are then used to generate the reasonable path sets for this record. These identified intermediate points serve as intermediate origin and destination nodes. The distance proportion to the total length of the link from the new origin point to the intermediate origin node is defined as α_1, and the distance proportion from the new destination point to the intermediate destination node is defined as α_2. For points lying on the two-way streets, different combination of α_1 and α_2 are allowed for the same record, depending on the combination of intermediate origin/destination node.

2.4. Constructing reasonable path sets

Given the origin and destination of a taxi trip, the number of paths in urban network between the origin and destination are potentially large, especially for downtown grid networks of big cities. Since the actual path taken by a taxi driver is

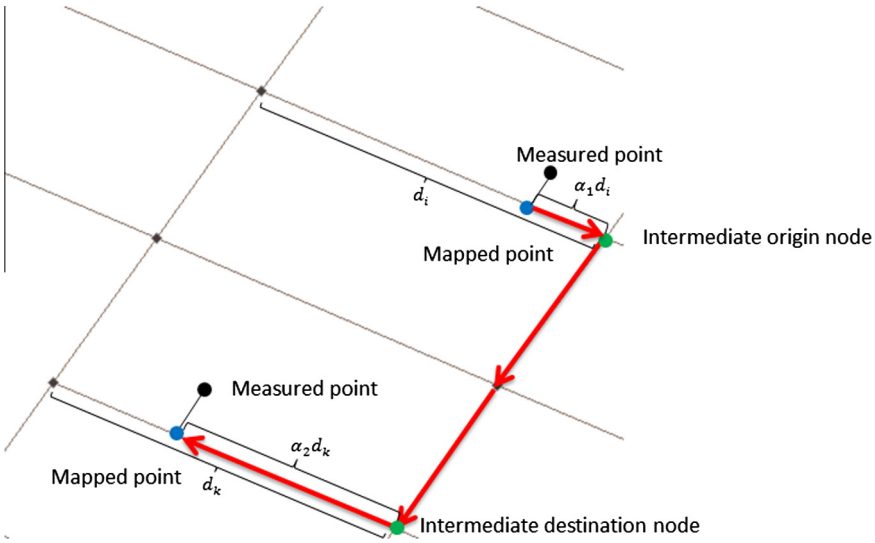

Fig. 2. Illustration of data mapping.

unknown, an important sub-question of the analysis is to infer the possible path set of a given taxi trip. Considering the large number of observations available in a large network, the overall search space for the possible path sets are huge. It is necessary to reduce the size of the possible path sets. In this study, Yen's k-Shortest Path algorithm (Yen, 1971) ($k = 20$) is used to generate the initial path sets, and the trip distance recorded in the data is then used to eliminate unreasonable paths. Only the paths that do not have excessively high or low lengths compared to the observed taxi trip distance will be used.

Because the trip distance recorded in the data is not very accurate (only accurate to 0.1 mile), a threshold ratio of 15–25% for weekday, and 20–25% for weekend (both upper and lower) is used, depending on the amount of data available during one hour. The threshold ratio is used to filter out the unreasonable paths whose measured lengths deviate significantly from the recorded trip distance.

2.5. Solution approach

To solve this non-linear least square problem, the Levenberg–Marquardt (LM) method (Nocedal and Wright, 2006; Fletcher, 1971) is used. The Levenberg–Marquardt method is a widely used optimization algorithm in solving least square curve fitting and nonlinear programming problems. It outperforms the simple gradient decent method and the well-known Gauss–Newton (GN) methods in a wide variety of problems. The traditional Gauss–Newton method uses a line-search method, which is computationally expensive for solving this problem, since the objective function is huge. The updating method in Gauss–Newton method is similar to Newton's method, which has numerical issues when the approximated Hessian is near singular and easily fails to converge to the optima if improper initial value is used. Levenberg–Marquardt method on the other hand, uses a trust-region strategy instead of the line search method, which determines the step size before the updating step. The different Hessian approximation method used in LM also helps to ensure the positive definiteness of the approximated Hessian in each iteration. This results in a more robust performance, which means that in many cases Levenberg–Marquardt method finds a solution even if it starts very far off the final minimum. It is showed in Nocedal and Wright (2006) that Levenberg–Marquardt enjoys rapid local convergence near optima, and under ideal cases, the convergence is actually quadratic.

For simplicity, define

$$p_m(\vec{t},\theta) = e^{\theta(-\beta_1 \cdot g_m(\vec{t}) - \beta_2 \cdot d_m)} \tag{10}$$

Thus we can write the denominator of Eq. (6) as:

$$S_{R_i}(\vec{t},\theta) = \sum_{j \in R_i} e^{\theta(\beta_1 \cdot g_j(\vec{t}) - \beta_2 \cdot d_j)} = \sum_{j \in R_i} p_j(\vec{t},\theta) \tag{11}$$

Then, the expected path travel time can be written as,

$$E(Y_i|R_i) = f(R_i,\vec{t},\theta) = \sum_{m \in R_i} g_m(\vec{t}) \frac{p_m(\vec{t},\theta)}{S_{R_i}(\vec{t},\theta)} \tag{12}$$

Define

$$\begin{aligned} J_{ik} &= \frac{\partial f(R_i,\vec{t},\theta)}{\partial t_k}, \quad k = 1,2,\ldots,N \\ J_{iN+1} &= \frac{\partial f(R_i,\vec{t},\theta)}{\partial \theta} \end{aligned} \tag{13}$$

Thus $\boldsymbol{J}$ forms a $N_D \times (N+1)$ matrix, where N_D is the number of observations in data set D, N is the number of links in the network. The vector of link travel times and the scale parameter θ are updated iteratively using

$$\begin{aligned} \vec{t} &\approx \vec{t}^{k+1} = \vec{t}^{k} + \vec{p}_t^{(k)} \\ \theta &\approx \theta^{k+1} = \theta^{k} + p_\theta^{(k)} \end{aligned} \tag{14}$$

$\vec{p}^{(k)} = (\vec{p}_t^{(k)^T}, p_\theta^{(k)})^T$ is the update direction in kth iteration, which is obtained by solving the following linear system,

$$(\boldsymbol{J}^{(k)T}\boldsymbol{J}^{(k)} + \lambda \boldsymbol{I})\vec{p}^{(k)} = \boldsymbol{J}^{(k)T} r_i \tag{15}$$

where $\boldsymbol{J}^{(k)T}\boldsymbol{J}^{(k)}$ is the first order approximation of the Hessian matrix of the problem, and λ is referred to as damping factor, which adjusted at each iteration under a trust-region strategy. A modified Levenberg–Marquardt method replaces the identity matrix $\boldsymbol{I}$ with the diagonal matrix with the diagonal element of $\boldsymbol{J}^{(k)T}\boldsymbol{J}^{(k)}$, which shows as follows

$$(\boldsymbol{J}^{(k)T}\boldsymbol{J}^{(k)} + \lambda diag(\boldsymbol{J}^{(k)T}\boldsymbol{J}^{(k)}))\vec{p}^{(k)} = \boldsymbol{J}^{(k)T} r_i \tag{16}$$

This study uses the modified version of Levenberg–Marquardt method, as it avoids slow convergence in the direction of small gradient. Detailed description of the updating scheme of damping factor λ and implementation is discussed by Fletcher (1971).

In the above equations, J_{ik} is computed as

$$J_{ik} = \sum_{m \in R_r} \left\{ \frac{p_m(\vec{t},\theta) S_{R_r}(\vec{t},\theta) \frac{\partial g_m(\vec{t})}{\partial t_k} [1 - \beta_1 \theta g_m(\vec{t})] - g_m(\vec{t}) p_m(\vec{t},\theta) \frac{\partial S_{R_r}(\vec{t},\theta)}{\partial t_k}}{[S_{R_r}(\vec{t},\theta)]^2} \right\} \tag{17}$$

in which $k = 1, 2, \ldots, N$, and for $k = N + 1$, the J_{iN+1} is defined as

$$J_{iN+1} = -\sum_{m \in R_r} \frac{p_m(\vec{t},\theta) S_{R_r}(\vec{t},\theta)}{S_{R_r}(\vec{t},\theta)} \left[\beta_1 g_m(\vec{t}) + \beta_2 d_m + \frac{\frac{\partial S_{R_r}(\vec{t},\theta)}{\partial \theta}}{S_{R_r}(\vec{t},\theta)} \right] \tag{18}$$

$\frac{\partial g_m(\vec{t})}{\partial t_k}$, and $\frac{\partial S_{R_r}(\vec{t},\theta)}{\partial t_k}$, are defined as follows:

$$\frac{\partial g_m(\vec{t})}{\partial t_k} = \begin{cases} \alpha_1 & \text{if } t_k \in L_O \\ \alpha_2 & \text{if } t_k \in L_D \\ \delta_{mk} = \begin{cases} 1 & \text{if link } k \text{ on path } m \\ 0 & \text{if link } k \text{ not on path } m \end{cases} \end{cases} \tag{19}$$

$$\frac{\partial S_{R_r}(\vec{t},\theta)}{\partial t_k} = -\beta_1 \sum_{j \in R_r} \left[p_j(\vec{t},\theta) \frac{\partial g_j(\vec{t})}{\partial t_k} \right] \tag{20}$$

$$\frac{\partial S_{R_r}(\vec{t},\theta)}{\partial \theta} = -\sum_{j \in R_r} [\beta_1 g_j(\vec{t}) + \beta_2 d_j] p_j(\vec{t},\theta) \tag{21}$$

One can observe that the problem is not convex and hence may have multiple local optima. A proper initial point is needed to ensure the convergence to the most probable solution. A preprocessing step is used to search for the network wide optimal mean speed. In this step, all the links in the network are assumed to have the same mean speed v_m, thus a 1-dimensional search algorithm can be implemented to find the v_m that minimizes the objective function. The obtained mean speed is then used to calculate the initial values of the link travel times. In Table 2, data from 3/15/2010 (Monday) 21:00–22:00 are used to test the choice of different initial link speeds for link travel time estimation. The result shows that using the network wide optimal mean speed as an initial point yields the lowest objective value and RMSE, which suggests that the preprocessing step is an effective approach of finding desirable link travel time estimates.

3. Testing data and network

The data used in this research was collected by New York City Taxi and Limousine Commission on a trip by trip basis. The data records each trip origin and destination GPS coordinate, trip distance and duration, fare, payment method, and other related information. The data set contains data from February 2008 to November 2010. In this study, a week's data (from 3/15/2010 to 3/21/2010) is selected to test the proposed method.

A small region in the southeast of Central Park of Midtown Manhattan is selected to serve as the study region, which is a 1370 m × 1600 m rectangle area. The corresponding network is also extracted (Fig. 3), which contains 193 nodes and 381 directed links. The network has 331 road segments and only 50 of them are two-way streets. From the original data set, all the records that fall within the region are extracted. Fig. 4 presents the number of observations inside the study region in a typical weekday (3/15/2010, Monday) and weekend (3/20/2010, Saturday) respectively. We obtain as many as 1000 observations in one hour on a typical weekday (Monday) and about 500 observations in one hour in a weekend (Saturday) inside the study region.

Table 2
Test results for different choices of model initial values.

Initial speed (mph)	Objective function value	Iteration used	RMSE	MAPE (%)
10[a]	779.830	20	1.372	21.87
8	1215.410	17	1.713	29.30
12	783.143	16	1.375	21.52
8–12 Uniformly distributed	801.487	20	1.391	22.49
8–12 Uniformly distributed	805.075	16	1.394	22.68
6–14 Uniformly distributed	805.146	27	1.395	22.55
6–14 Uniformly distributed	807.044	23	1.396	22.35

[a] Network wide optimal mean speed.

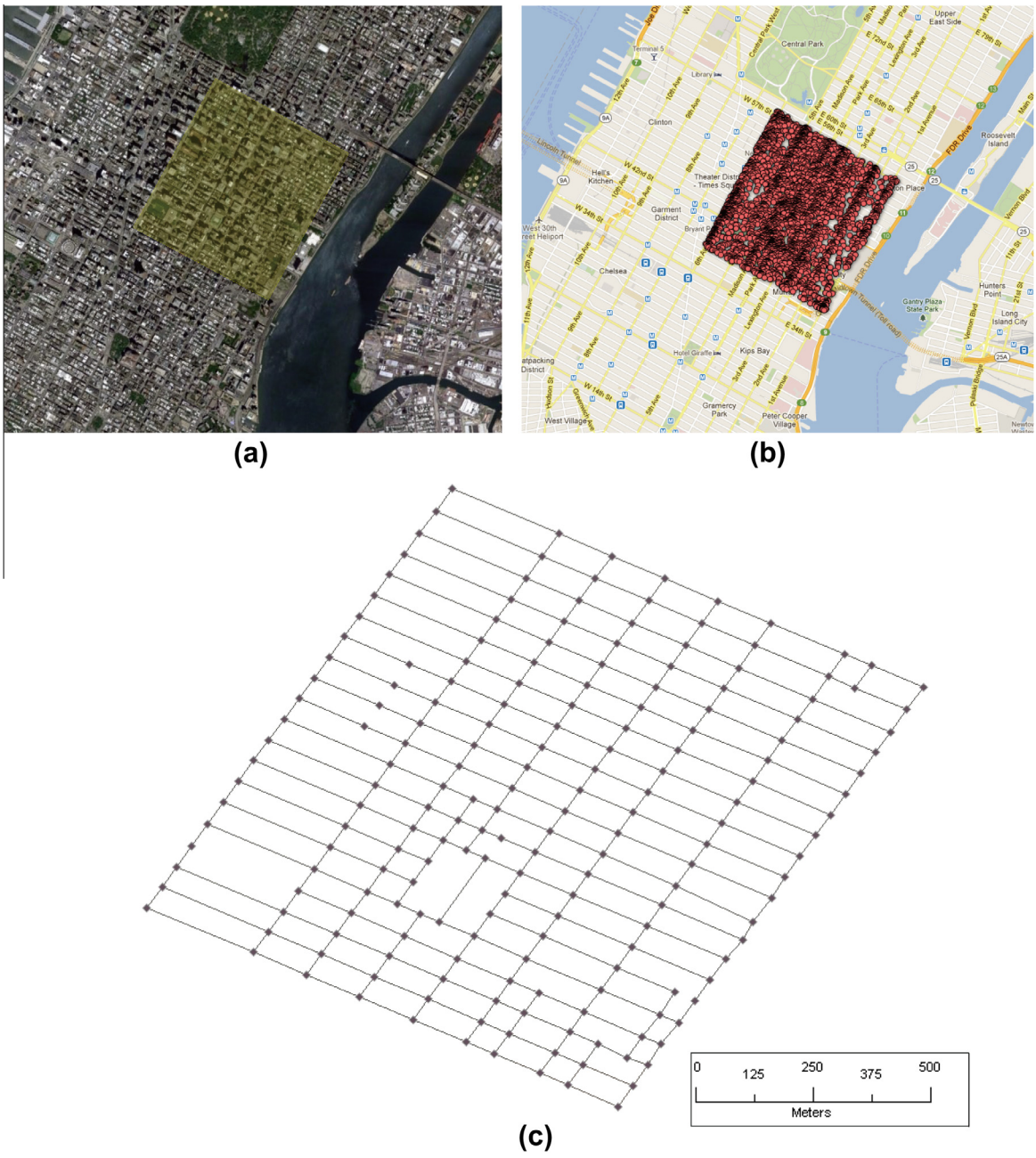

Fig. 3. Test network of the study region: Midtown Manhattan.

In this study, the data is split into hourly intervals, and link travel times are estimated using the data from the corresponding hour. Although traffic conditions can change rapidly during one hour, a shorter time period will not guarantee a good statistical significance due to the insufficient amount of observations given the limited information in the data.

4. Model results

To implement the model discussed in the previous section, a Matlab code is written using Parallel Computing Toolbox. A *k*-shortest path set is required to be computed for each nodal pair in the network and this process takes a considerable amount of time. But once the process is complete, the path sets are stored and needs no further computation. The steps of data mapping and constructing reasonable path sets take little time to complete, as they make use of the information from

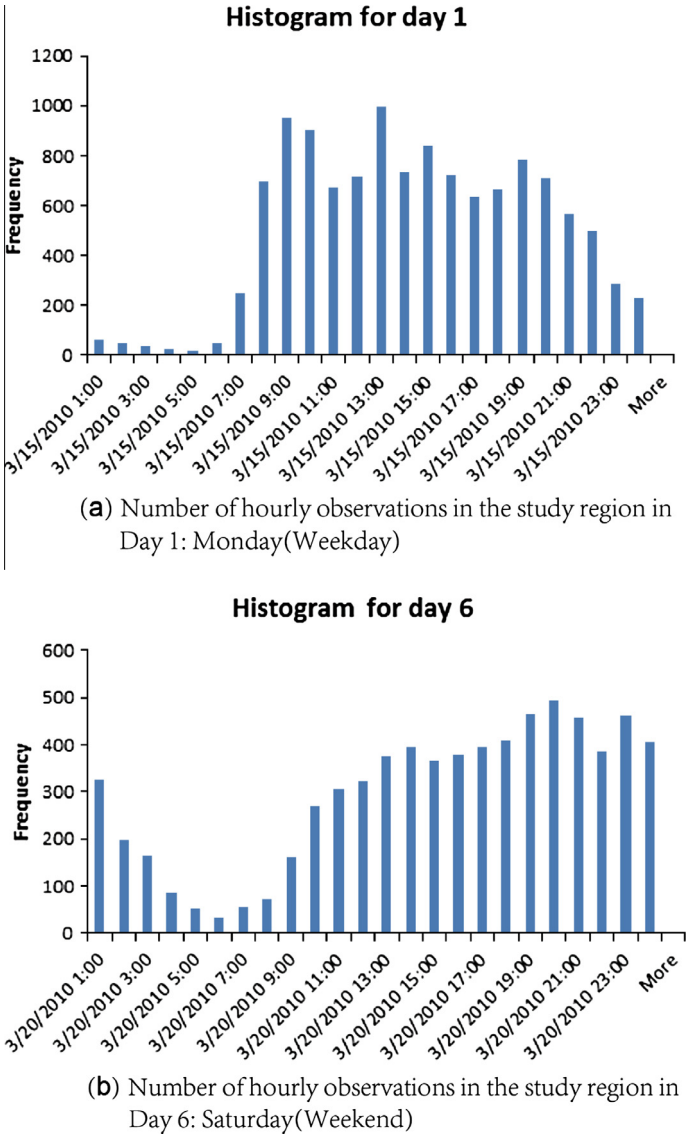

Fig. 4. Histogram for number of hourly observations in the study region.

already computed *k*-shortest path sets of the network. The Levenberg–Marquardt method provides good convergence properties, and the entire optimization process can be efficiently solved within 15 min using an Intel i7 CPU laptop. The computation time can be further reduced by using Matlab C/MEX code or a more powerful computer.

Link travel times for four time periods (9:00–10:00, 13:00–14:00, 19:00–20:00, and 21:00–22:00) in a day are estimated based on a week's Taxi GPS data (from 3/15/2010 to 3/21/2010). The time period from 9:00 to 10:00 represents the morning peak period, as the highest number of taxi trips are observed in this period on weekdays; while 21:00–22:00 is tested for the off-peak hour situation. A lower bound of speed (one mile per hour) is used to ensure that we do not obtain unreasonably large travel times; an upper bound of speed (30 miles per hour) is used as the free flow speed to set a lower bound for the estimated link travel times. We use the link speeds instead of the link travel times to give a more intuitive representation of the link travel time estimation results. Fig. 5 presents the estimated link speeds and correlation plots of observed and estimated path travel times for Monday, Tuesday, Wednesday and Saturday, and the results for Thursday, Friday and Sunday are attached in Fig. 6 in the Appendix A.

Based on model estimation results, for weekdays, it is found that most of the links have speeds between 4 and 8 miles/h in the 9:00–10:00 morning peak hour. During the 13:00–14:00 period, the distribution of speed is slightly improved and peaks around 7 miles/h. In the 19:00–20:00 period, the mean speed is observed between 6 and 8 miles/h. However, in the 21:00–22:00 off-peak period, a great number of links are observed to have speeds around 10 miles/h. In contrast, during weekends,

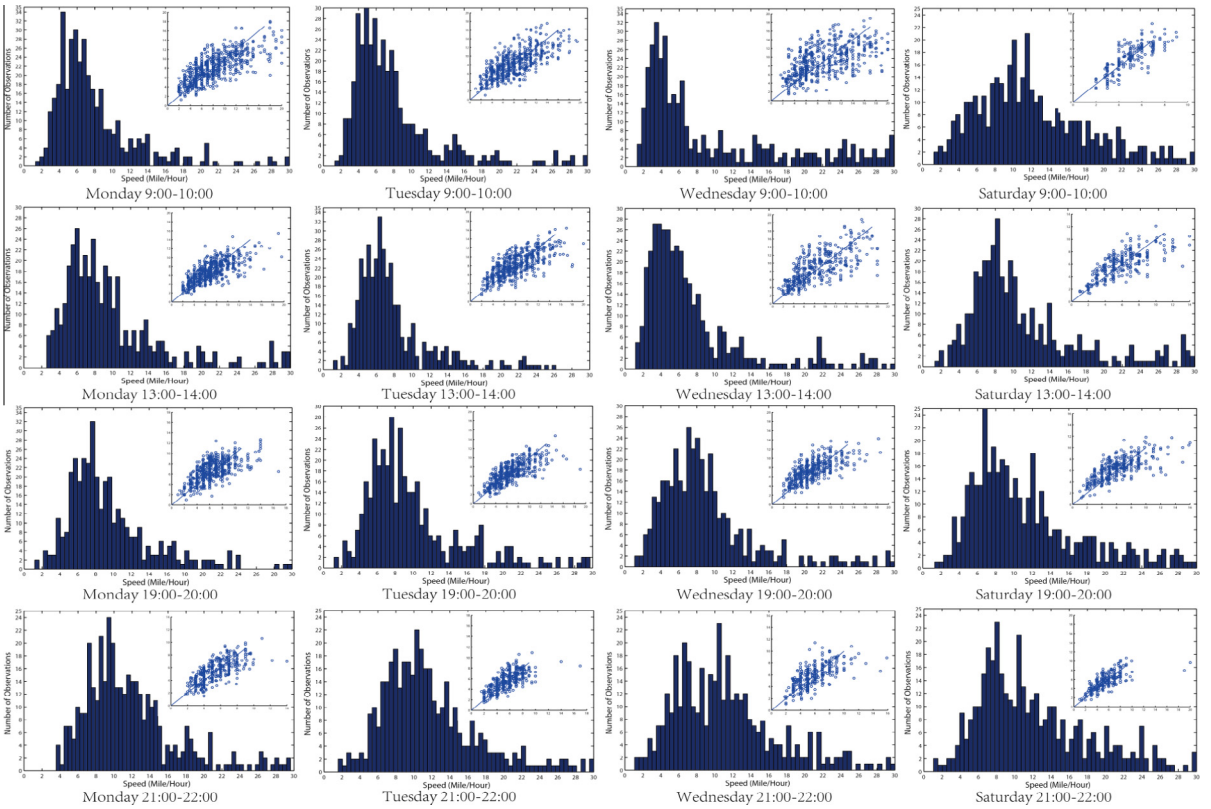

Fig. 5. Histogram of estimated link speed and correlation plot of observed and estimated path travel time for Monday, Tuesday, Wednesday and Saturday (inside plot, *X*-axis: observed path travel time (min), *Y*-axis: estimated path travel time (min)).

a relatively higher average speed (8–10 miles/h) is observed during 9:00–10:00 in the morning, and relatively lower average speed (about 8 miles/h) is observed during 19:00–20:00 pm period. These values are consistent with a previous study on New York City traffic speeds where it is reported that on weekdays in the daytime, in east Midtown, average traffic speed is 6.3 mph whereas on Saturdays, the average speed is about 8.5 mph (Grynbaum, 2010).

The root mean square Error (RMSE) and mean absolute percentage error (MAPE) are used to evaluate the estimation results:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(T_i^{Pr} - T_i^{Ob})^2} \tag{20}$$

$$\mathrm{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{T_i^{Pr} - T_i^{Ob}}{T_i^{Ob}}\right| \times 100\% \tag{21}$$

where T_i^{pr} is the model estimated trip path travel time; T_i^{ob} is the observed trip path travel time; and n is the number of observations.

The model estimation errors and the estimated values of the scale parameter θ are presented in Table 3. As showed in the result, except for 2 time intervals (Wednesday 9:00–10:00 and 13:00–14:00), all the link travel time estimation results have MAPE below 30%, and for some off-peak hours, e.g. 21:00–22:00 pm, the MAPE is only around 22%. It is noticeable that Wednesday 9:00–10:00 and 13:00–14:00 have much larger errors and lower link travel speeds compared with other days. It is found that this Wednesday (3/17/2010) happened to have St. Patrick's Day Parade. The parade was from 8:00 to 15:00 and marched down the 5th street (contained in the test network). Few roads were temporarily closed and huge crowds were drawn along the parade routes. This caused huge disruption in traffic network and explained the high estimation errors on Wednesday. It is also observed that after the parade ended, the estimation results for 19:00–20:00 and 21:00–22:00 restore to normal condition, which have RMSE under 2.5 min and MAPE under 30%. It is found that in congested time period (e.g. 9:00–10:00 on weekdays), the results have relatively higher estimation errors. This could be the effect of the rapid changes in the network condition, since the model estimates hourly average link travel times. The link speed estimates also confirm that in high estimation error time period, a greater proportion of links have relatively lower traveling speed.

The estimation results for the scale parameter θ show relatively large variance in drivers' route choice behaviors. All the estimated values for θ are smaller than 1.2, and very small θ (0.003) is observed. The change in θ reflects the relatively large

Table 3
Model estimation error and estimated value for the scale parameter θ.

Day	Error	Time period			
		9:00–10:00	13:00–14:00	19:00–20:00	21:00–22:00
Monday	RMSE (min)	2.614	1.981	1.937	1.372
	MAPE (%)	29.51	24.22	26.27	21.87
	θ	0.165	0.063	0.435	0.068
Tuesday	RMSE (min)	2.461	2.302	1.827	1.437
	MAPE (%)	29.63	25.59	23.33	22.20
	θ	1.082	0.049	0.329	0.003
Wednesday	RMSE (min)	3.827[a]	3.216[a]	2.180	1.691
	MAPE (%)	41.32[a]	34.97[a]	28.73	24.40
	θ	1.030	0.867	1.153	0.539
Thursday	RMSE (min)	2.468	2.699	2.490	1.382
	MAPE (%)	27.28	27.92	28.54	21.05
	θ	0.469	0.037	0.264	0.499
Friday	RMSE (min)	2.260	2.179	1.692	1.334
	MAPE (%)	27.76	27.04	25.17	22.26
	θ	0.075	0.010	0.717	0.245
Saturday	RMSE (min)	1.034	1.690	1.839	1.584
	MAPE (%)	16.84	24.58	27.14	21.61
	θ	0.469	0.287	0.081	0.087
Sunday	RMSE (min)	2.041	1.518	1.395	1.160
	MAPE (%)	25.44	23.70	22.72	19.87
	θ	0.166	0.239	0.190	0.615

[a] Traffic disturbance caused by Patrick's Day Parade.

variance in taxi drivers' route choice behaviors in different time periods, days and network conditions. The wide range for the estimated values of θ (0.003–1.153) could be the result of several reasons. One plausible explanation could be that as traffic conditions change in different time periods of a day, taxi drivers may have different levels of perception error, which are reflected in their route choice behaviors. However θ only considers the overall variance in taxi drivers' perceived path costs and treats all taxi drivers as homogeneous individuals, which does not capture the behavioral heterogeneity among the taxi drivers.

Three consecutive Mondays in 2009 (2009/9/14, 2009/9/21, 2009/9/28) are also investigated to see if repeatability exist across weeks (due to space limitation, these results are not included). However, no significant pattern is found in terms of link speed profile and travel time variation during a day. The findings of the three Mondays agrees with the general pattern found on weekdays discussed above, but variation in terms of the distribution of link travel speeds is also observed, and no conclusive inference can be made across weeks.

In this model, intersection delay is not modeled due to the lack of detailed vehicle trajectory information in the data. In the testing network, most of the links have lengths ranging from 80 to 300 m; assuming the vehicle traveling at a speed of 8 miles per hour, a great number of links will have a travel time less than 1 min. However, the intersection delay at a traffic signal sometimes can be greater than the link travel time itself. In a 10 min trip, it is very likely to have at least 2 min of intersection delay on average, which partly explains the RMSE of around 2 min in the model. This is a potential source of errors of the model. The intersection delay causes inconsistency in the link travel time estimation and leads to overestimation of actual link travel times. However, given only origin and destination information provided in the data, modeling intersection delay separately will introduce excessive complexity in travel time estimation, which makes the short term estimation intractable. Also, there is no guarantee on the quality of the estimated intersection delay, since too little information is available to separate the intersection delay from the total link travel time. Thus given the incompleteness of the data, we combine the intersection delay into the link travel times and focus on estimation the hourly average link travel times.

Furthermore, because the link travel times are estimated as hourly average values, variations in link travel times within one hour can introduce errors in the model estimation (Fosgerau and Fukuda, 2012). The heterogeneity among the drivers' behaviors (e.g. some drivers prefer to drive fast and choose the shortest path, some drivers prefer to drive at a moderate speed and take a relatively long path, etc.) may also contribute to the estimation errors. Certain trips are observed to take as much as 20 min in the testing network, which involve a lot of uncertainty in path choices, leading to some errors in estimation results as well.

5. Discussion and conclusion

In this study, a new model is proposed to use the limited information provided in the taxi GPS data to estimate urban link travel times. The taxicab data used in this study lacks the information of actual paths taken by the taxi drivers. The proposed model treats the path taken as latent, constructs a reasonable path set, formulates an MNL model to compute the probability

of a path being taken by the driver, and estimates the link travel times by optimizing a nonlinear least square problem. Model estimation results indicate that the proposed method can efficiently estimate hourly average link travel times.

It is recommended to split the whole urban region into smaller zones (e.g. 1.5 km $\times$ 1.5 km) to implement this model, because of the following reasons: (1) Larger zones contain longer trips, which involve more uncertainties in path choices, thus long trips are less reliable in the link travel time estimation given this type of data. (2) Preparing the k-shortest path set for all the nodal pairs in a large network is computationally expensive. The number of nodal pairs grows as n^2 as the number of nodes in the network, and a greater k value is also needed to ensure a good representation of reasonable paths. By reducing the zone size, we can ensure the computational tractability for short term link travel time estimation. (3) The data provides a large number of records in an hour even in a 1.37 km $\times$ 1.6 km size zone, thus the amount of data is enough for the model.

This model can be further verified using the actual trajectory information of the taxi trips. Although this information has been collected by NYLTC, it is currently unavailable to the researchers. The model is also applicable to use trajectory data (treating two intermediate trajectory points as origin and destination point). The accuracy of the model can be improved with more detailed data and greater number of observations.

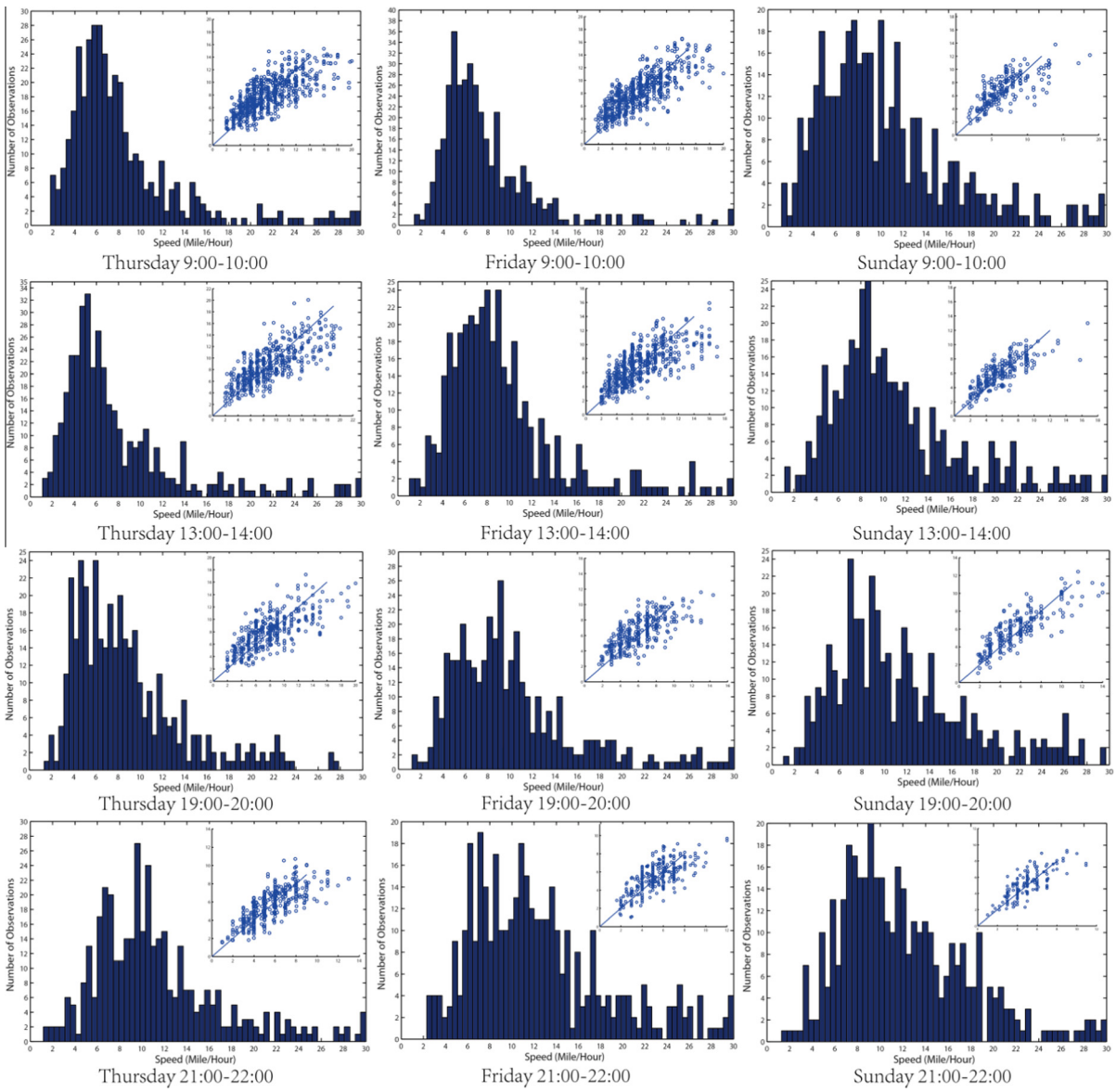

Fig. 6. Histogram of estimated link speed and correlation plot of observed and estimated path travel time for Thursday, Friday and Sunday (inside plot, *X*-axis: observed path travel time (min), *Y*-axis: estimated path travel time (min)).

In this model, only the data in the current time period are used in the estimation, and historical data are not used. Further research can be done to investigate a hybrid approach of using historical data as well as optimizing current estimation error. Another research direction in the future is to improve the route choice model to account for more realistic route choice behaviors of the taxi drivers. The current route choice model only considers drivers who minimize trip time and distance in each trip, and records that do not comply with this assumption are filtered out. A more comprehensive route choice model would utilize more data records and provide less estimation bias. Furthermore, intersection delays are important causes of irregularity of link travel times, which may lead to bias in the estimated travel times. Future research can be done to incorporate the effects of intersection delays in the link travel time estimation, and thus improve the estimation accuracy. All these efforts would provide a more accurate and reliable way to estimate urban network conditions using the partial information provided by the taxicab data.

Acknowledgments

This research presented in this paper was supported by RITA/USDOT project "The Use of Large Scale Datasets for Understanding Network State" for which the authors are grateful. The authors are solely responsible for the findings of the research work.

Appendix A.

See Fig. 6.

References

Coifman, B., 2002. Vehicle reidentification and travel time measurement on congested freeways. Transportation Research Part A: Policy and Practice 36 (10), 899–917.

Fletcher, R., 1971. A Modified Marquardt Subroutine for Nonlinear Least Squares. Rpt. AERE-R 6799, Harwell. Matlab Parallel Computing Toolbox. Mathworks, Inc.

Fosgerau, M., Fukuda, D., 2012. Valuing travel time variability: characteristics of the travel time distribution on an urban road. Transportation Research Part C: Emerging Technologies 24, 83–101.

Grynbaum, M.M., 2010. Gridlock May Not Be Constant, but Slow Going Is Here to Stay. New York Times. Retrieved July 31, 2012. <http://www.nytimes.com/2010/03/24/nyregion/24traffic.html?ref=nyregion>.

Hasan, S., Choudhury, C.F., Ben-Akiva, M.E., Emmonds, A., 2011. Modeling of travel time variations on urban links in London. Transportation Research Record: Journal of the Transportation Research Board 2260 (-1), 1–7.

Herrera, J.C., Work, D., Ban, X., Herring, R., Jacobson, Q., Bayen, A.M., 2010. Evaluation of traffic data obtained via GPS-enabled mobile phones: the mobile century field experiment. Transportation Research Part C: Emerging Technologies 18, 568–583.

Herring, R., Hofleitner, A., Abbeel, P., 2010. Estimating arterial traffic conditions using sparse probe data. Proceedings of the ITS, 19–22.

Hunter, T., Herring, R., Abbeel, P., 2009. Path and travel time inference from GPS probe vehicle data. Neural Information Processing Systems Foundation (NIPS), Vancouver, Canada, (December).

Inrix, Inc. http://www.inrix.com.

King, David, Peters, J., 2012. Taxicabs for Improved Urban Mobility: Are We Missing an Opportunity? Transportation Research Board 91st Annual Meeting, 19p.

Li, R., Rose, G., 2011. Incorporating uncertainty into short-term travel time predictions. Transportation Research Part C: Emerging Technologies 19 (6), 1006–1018.

Nocedal, J., Wright, S.J., 2006. Numerical Optimization. second ed. Springer, pp. 258–264.

Oh, J.S., Jayakrishnan, R., Recker, W., 2003. Section travel time estimation from point detection data. In: 82nd Annual Meeting of Transportation Research Board, Washington, DC, USA.

Park, D., RILETT, L.R., 1998. Forecasting multiple-period freeway link travel times using modular neural networks. Journal of the Transportation Research Board 98, 163–170.

Schaller Consulting, 2006. The New York City Taxicab Fact Book, (March). <www.schallerconsult.com>.

Sherali, H.D., Desai, J., Rakha, H., 2006. A discrete optimization approach for locating automatic vehicle identification readers for the provision of roadway travel times. Transportation Research Part B: Methodological 40 (10), 857–871.

Taxicab Rates from New York Taxi & Limousine Commission: <http://www.nyc.gov/html/tlc/html/passenger/taxicab_rate.shtml>.

Wu, C.-H., Ho, J.-M., Lee, D.T., 2004. Travel-time prediction with support vector regression. IEEE Transactions on Intelligent Transportation Systems 5 (4), 276–281.

Yen, J.Y., 1971. Finding the K shortest loopless paths in a network. Management Science 17, 712–716.

Yeon, J., Elefteriadou, L., Lawphongpanich, S., 2008. Travel time estimation on a freeway using discrete time Markov chains. Transportation Research Part B: Methodological 42 (4), 325–338.

Zhang, X., Rice, J., 2003. Short-term travel time prediction. Transportation Research Part C: Emerging Technologies 11 (3–4), 187–210.

Zheng, F., Van Zuylen, H., in press. Urban link travel time estimation based on sparse probe vehicle data. Transportation Research Part C: Emerging Technologies 13. Elsevier Ltd. http://dx.doi.org/10.1016/j.trc.2012.04.00.