Hindawi Publishing Corporation
Discrete Dynamics in Nature and Society
Volume 2016, Article ID 7348705, 13 pages
http://dx.doi.org/10.1155/2016/7348705

Research Article

Urban Link Travel Time Estimation Based on Low Frequency Probe Vehicle Data

Xiyang Zhou,[1,2] Zhaosheng Yang,[1,2,3] Wei Zhang,[1,2,4] Xiujuan Tian,[1,2,3] and Qichun Bing[1,2,3]

[1] *College of Transportation, Jilin University, Changchun 130025, China*
[2] *State Key Laboratory of Automobile Simulation and Control, College of Transportation, Jilin University, Changchun 130025, China*
[3] *Jilin Province Key Laboratory of Road Traffic, College of Transportation, Jilin University, Changchun 130025, China*
[4] *Shandong High-Speed Group Co., Ltd., Jinan 250000, China*

Correspondence should be addressed to Wei Zhang; zhangwei_txj@126.com

Received 31 August 2015; Accepted 2 November 2015

Academic Editor: Filippo Cacace

To improve the accuracy and robustness of urban link travel time estimation with limited resources, this research developed a methodology to estimate the urban link travel time using low frequency GPS probe vehicle data. First, focusing on the case without reporting points for the GPS probe vehicle on the target link in the current estimation time window, a virtual report point creation model based on the K-Nearest Neighbour Rule was proposed. Then an improved back propagation neural network model was used to estimate the link travel time. The proposed method was applied to a case study based on an arterial road in Changchun, China: comparisons with the traditional artificial neural network method and the spatiotemporal moving average method revealed that the proposed method offered a higher estimation accuracy and better robustness.

1. Introduction

Accurate estimation of urban link travel times is essential for traffic operators and travellers, not only because link travel time is an important index for monitoring and evaluating the state of the traffic on an urban road network, but also because it is a critical input to dynamic route guidance systems which helps travellers make better route choices and avoid congestion. The estimation of urban link travel times relies on traffic data collection. In the past, traffic data were mainly collected by loop detectors [1–4]. However, due to the high cost of installation and maintenance, loop detectors are often only installed on a few links in the urban road network, which leads to unavailability of most of the network traffic data.

In recent years, most vehicles are equipped with GPS devices such as GPS navigators or smartphones, which provide a type of probe vehicle which can collect traffic data from the entire road network at low cost. These GPS probe vehicles can continuously collect traffic data by travelling on the road network and reporting their positions, instantaneous speeds, and movement directions at specific sampling frequencies. Chakroborty and Kikuchi [5] proposed a method of utilising buses equipped with GPS locators to estimate travel times along urban corridors. Liu et al. [6] discussed the feasibility of using a taxi dispatch system as a probe with which to collect traffic information. Zhan et al. [7] proposed an urban link travel time estimation model using large-scale taxi data with partial information. Zheng and McDonald [8] proposed two fuzzy clustering algorithms to estimate the travel time, which greatly reduced the influence of random error. Guessous et al. [9] proposed a probabilistic model to estimate the travel time under different traffic conditions, which took into account the levels-of-service.

However, most of these methods require the data collected by GPS probe vehicles with high sampling frequencies (e.g., 10 s intervals or less), which not only needs a large amount of data storage space but is also computationally expensive. Therefore, using low frequency probe vehicle data has become a new challenge for link travel time estimation

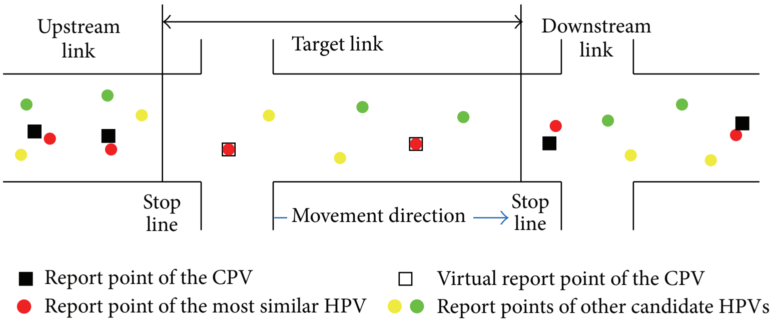

Upstream link
Target link
Downstream link
Stop line
Movement direction
Stop line
■ Report point of the CPV
□ Virtual report point of the CPV
● Report point of the most similar HPV
● ● Report points of other candidate HPVs

Figure 1: Schematic of the virtual report point creation model.

[10]. When the sampling interval is large (e.g., 60, 90, and 120 s), with the addition of data loss caused by signal dropouts or communication failure, the current probe vehicle (CPV) may have traversed one or more links between two consecutive available status reports, which means there could be no report point of the CPV on the target link. Jenelius and Koutsopoulos [11] developed a maximum likelihood model to allocate the observed travel time to each link based on link attributes and trip conditions. Hellinga et al. [12] developed an analytical model to estimate link travel time, which considers both stopping and congestion probabilities. Both methods can solve the problem to some extent; however, the models are extremely complex and the parameter calibration is laborious. Zheng and Van Zuylen [13] proposed a traditional artificial neural network (ANN) model and compared it with Hellinga et al.'s model. The results suggested that the ANN method performs better. However, the ANN method is limited when there is no report point on two consecutive links. Sanaullah et al. [14] developed a spatiotemporal moving average method by assuming that the CPV travels at the speed limit on the target link when it has no report point thereon. The assumption significantly simplifies the model. However, it is not consistent with fact, which leads to unsatisfactory accuracy in urban road networks.

The present study will therefore focus on the case in which there is no report point of the CPV on the target link. The methodology proposed in this paper extends previous work on travel time estimation using sparse GPS probe vehicle data by using historical GPS probe vehicle data to impute missing report points on the target link, which consists of two layers: first, the virtual report point creation model based on pattern recognition is proposed to recognise the historical probe vehicle which has the most similar travelling characteristic to the CPV and then use its report points to create the virtual report points for the CPV on the target link; second, an improved back propagation neural network model is used for estimating the travel time on the target link based on both virtual and real report points of the CPV.

The paper is organised as follows: Section 2 describes the proposed methodology: the virtual report point creation model is presented in Section 2.1 and the improved back propagation neural networks model is presented in Section 2.2. In Section 3, the performance of the proposed methodology is evaluated using data from a case study from Changchun, China. The results are compared with those from the artificial neural network method and the spatiotemporal moving average method. Section 4 outlines the conclusions and presents recommendations for future research.

2. Methodology

2.1. Virtual Report Point Creation Model. The main idea of the virtual report point creation model is to use the pattern recognition method to select the historical probe vehicle (HPV) that has the most similar travel characteristics to the current probe vehicle (CPV) from the candidate HPVs which have report points on the target link and then use the report points of the most similar HPV to create virtual report points for the CPV on the target link, as shown in Figure 1.

Many researchers have emphasised pattern recognition methods: existing methods mainly include the Bayesian [15], Principle Component Analysis [16], Linear Discriminant Analysis [17], Nonnegative Matrix Factorisation [18], Gaussian Mixture Model [19], Artificial Neural Networks [20], Support Vector Machines [21], and K-Nearest Neighbour Rule methods [22]. Among these methods, the K-Nearest Neighbour Rule method is not only nearly optimal in a large sample but also relatively easy to implement with no need for estimation of its parameters or a training process. Therefore, the K-Nearest Neighbour Rule (K-NNR) method is selected to search for the most similar HPV.

The main idea of K-NNR is to compare the distances between the test sample and the training samples that belong to different patterns and then select k training samples nearest to the test sample and determine to which patterns they belong; therefore, the pattern which contains the most selected training samples is recognised as the most similar pattern to the test sample. For the case of the most similar HPV as recognised here, the test samples are the report points of the CPV on the upstream and downstream links, the training samples are the report points of different candidate HPVs on the upstream and downstream links, and different candidate HPVs denote different patterns, assuming that all report points have been map-matched onto the links.

2.1.1. Training and Test Set Construction. The candidate HPVs (in the same estimation time window on the same day with the CPV) of the past few weeks are selected from the historical database following two rules: first, they must have the same movement direction as the CPV; second, they must have report points on the upstream link, the target link, and the downstream link. Since most similar history probe vehicle recognition needs to consider the report points on the upstream and downstream links, two training sets must be constructed.

The upstream training set is

$$T_{\text{training}}^{u} = \left\{M_{h,1}^{u,1}, \ldots, M_{h,1}^{u,l_1}, \ldots, M_{h,i}^{u,m}, \ldots, M_{h,i}^{u,l_i}, \ldots\right\} \tag{1}$$

and the downstream training set is

$$T_{\text{training}}^{d} = \left\{M_{h,1}^{d,1}, \ldots, M_{h,1}^{d,l_1'}, \ldots, M_{h,i}^{d,m}, \ldots, M_{h,i}^{d,l_i'}, \ldots\right\}, \tag{2}$$

where $M_{h,i}^{u,m}$ and $M_{h,i}^{d,m}$ denote the mth map-matched report points of the ith candidate HPV on the upstream link and the downstream link, respectively; l_i is the number of map-matched report points of the ith candidate HPV on the upstream link; l_i' is the number of map-matched report points of the ith candidate HPV on the downstream link. Correspondingly, two test sets must be constructed.

The upstream test set is

$$T_{\text{test}}^{u} = \left\{M_{c}^{u,1}, \ldots, M_{c}^{u,j}, \ldots, M_{c}^{u,l_j}\right\} \tag{3}$$

and the downstream test set is

$$T_{\text{test}}^{d} = \left\{M_{c}^{d,1}, \ldots, M_{c}^{d,j}, \ldots, M_{c}^{d,l_j'}\right\}, \tag{4}$$

where $M_c^{u,j}$ and $M_c^{d,j}$ denote the jth map-matched report points of the CPV on the upstream and downstream links, respectively; l_c is the number of map-matched report points of the CPV on the upstream link; l_c' is the number of map-matched report points of the CPV on the downstream link.

2.1.2. Characteristic Vector Construction. The traffic data collected by a map-matched report point include the position, the instantaneous speed, the time stamp, and the azimuth angle relative to the North. Since all candidate HPVs have the same movement direction as the CPV, the azimuth angle is unable to capture the travelling characteristic of the probe vehicles. On the contrary, the position, instantaneous speed, and time stamp can adequately capture the travelling characteristic of the probe vehicles; therefore, these parameters are selected as elements of the characteristic vectors. Due to the fact that these characteristic parameters have different dimensions, a zero-mean normalisation process is needed for every characteristic parameter before construction of the characteristic vectors.

The characteristic vector of the training sample:

$$\begin{gathered} \mathbf{M}_{h,i}^{u,m} = \begin{bmatrix} X_{h,i}^{u,m} \\ V_{h,i}^{u,m} \\ T_{h,i}^{u,m} \end{bmatrix}, \\ \mathbf{M}_{h,i}^{d,m} = \begin{bmatrix} X_{h,i}^{d,m} \\ V_{h,i}^{d,m} \\ T_{h,i}^{d,m} \end{bmatrix}, \end{gathered} \tag{5}$$

where $\mathbf{M}_{h,i}^{u,m}$ and $\mathbf{M}_{h,i}^{d,m}$ denote the characteristic vectors of $M_{h,i}^{u,m}$ and $M_{h,i}^{d,m}$, respectively; $X_{h,i}^{u,m}$ and $X_{h,i}^{d,m}$ denote the normalised values of the positions of $M_{h,i}^{u,m}$ and $M_{h,i}^{d,m}$ along the link, respectively; $V_{h,i}^{u,m}$, $T_{h,i}^{u,m}$, and $V_{h,i}^{d,m}$, $T_{h,i}^{d,m}$ denote the normalised values of the instantaneous speed and time stamp of $M_{h,i}^{u,m}$ and $M_{h,i}^{d,m}$, respectively.

Similarly, the characteristic vectors of the test samples are constructed as follows:

$$\begin{gathered} \mathbf{M}_{c}^{u,j} = \begin{bmatrix} X_{c}^{u,j} \\ V_{c}^{u,j} \\ T_{c}^{u,j} \end{bmatrix}, \\ \mathbf{M}_{c}^{d,j} = \begin{bmatrix} X_{c}^{d,j} \\ V_{c}^{d,j} \\ T_{c}^{d,j} \end{bmatrix}, \end{gathered} \tag{6}$$

where $\mathbf{M}_c^{u,j}$ and $\mathbf{M}_c^{d,j}$ denote the normalised values of the characteristic vectors of $M_c^{u,j}$ and $M_c^{d,j}$, respectively; $X_c^{u,j}$ and $X_c^{d,j}$ denote the normalised values of the positions of $M_c^{u,j}$ and $M_c^{d,j}$ along the link, respectively; $V_c^{u,j}$, $T_c^{u,j}$ and $V_c^{d,j}$, $T_c^{d,j}$ denote the instantaneous speed and time stamp of $M_c^{u,j}$ and $M_c^{d,j}$, respectively.

It is worth noting that the time stamp of a report point is recorded in year, month, day, hour, minute, and second terms. For the candidate HPVs and the CPV, the year, month, and day are not needed during most similar HPV recognition; therefore, the time stamp needs to be processed before characteristic vector construction to remove the year, month, and day: it is then converted into seconds for calculation purposes.

2.1.3. Characteristic Distance Function Construction. The characteristic distances between the test samples and the training samples are able to capture the degree of similarity between them. There are several distances to be measured such as the Euclidean distance, Manhattan distance, Bhattacharyya distance, and Mahalanobis distance. The characteristic distance functions for the most similar HPV recognition are based on the classic Euclidean distance.

The function of characteristic distance for the upstream link is

$$d\left(M_{h,i}^{u,m},\mathrm{CPV}\right) = \sum_{j=1}^{l_c}\sqrt{\left(X_{h,i}^{u,m}-X_c^{u,j}\right)^2+\left(V_{h,i}^{u,m}-V_c^{u,j}\right)^2+\left(T_{h,i}^{u,m}-T_c^{u,j}\right)^2} \tag{7}$$

and the function of characteristic distance for the downstream link is

$$d\left(M_{h,i}^{d,m},\mathrm{CPV}\right) = \sum_{j=1}^{l'_c}\sqrt{\left(X_{h,i}^{d,m}-X_c^{d,j}\right)^2+\left(V_{h,i}^{d,m}-V_c^{d,j}\right)^2+\left(T_{h,i}^{d,m}-T_c^{d,j}\right)^2}, \tag{8}$$

where $d(M_{h,i}^{u,m},\mathrm{CPV})$ denotes the characteristic distance between the CPV and the mth upstream map-matched report point of the ith candidate HPV; $d(M_{h,i}^{d,m},\mathrm{CPV})$ denotes the characteristic distance between the CPV and the mth downstream map-matched report point of the ith candidate HPV.

2.1.4. Most Similar HPV Recognition. Due to the low sampling frequency, the number of the map-matched report points of each candidate HPV is usually less than 10. Therefore, the value of k for K-NNR is set to 10.

For the upstream link, the 10 map-matched report points that have the shortest characteristic distances from the CPV are selected from all map-matched report points of all candidate HPVs on the upstream link; if the most of them belong to HPV i, then HPV i is recognised as the most similar HPV on the upstream link. Similarly, the most similar HPV on the downstream link is recognised in the same way.

2.1.5. Creation of Virtual Report Points. After the identification of the most similar HPVs on the upstream and downstream links, the virtual report points of the CPV on the target link are able to be created based on the map-matched report points of the most similar HPV.

The identification of the most similar HPVs would result in two scenarios: (1) the most similar HPV on the upstream link is also the most similar one on the downstream link; (2) the most similar HPV on the upstream link is different from the one on the downstream link. Therefore, the creation of the virtual report points in both scenarios is discussed as follows.

Scenario 1 (HPV i is the most similar HPV to the CPV on both the upstream and downstream links). In this case, only those report points of HPV i are used to create the virtual report points of the CPV on the target link, as shown in Figure 2(a).

In Figure 2(a), $M_c^{t,m}$ denotes the mth virtual report points of the CPV on the target link; $M_{h,i}^{t,m}$ denotes the mth map-matched report points of HPV i on the target link. The characteristic parameters of $M_c^{t,m}$ are determined as follows:

$$\begin{aligned} x_c^{t,m} &= x_{h,i}^{t,m}, \\ v_c^{t,m} &= v_{h,i}^{t,m}, \end{aligned} \tag{9}$$

where $x_c^{t,m}$ and $x_{h,i}^{t,m}$ denote the position along the link of $M_c^{t,m}$ and $M_{h,i}^{t,m}$, respectively; $v_c^{t,m}$ and $v_{h,i}^{t,m}$ denote the instantaneous speed of $M_c^{t,m}$ and $M_{h,i}^{t,m}$, respectively.

It is worth noting that the time stamps of the virtual report points of the CPV on the target link are supposed to be earlier than the time stamps of the real map-matched report points of the CPV on the downstream link but later than those on the upstream link. Therefore, the time stamp of $M_c^{t,m}$ is determined as follows:

$$t_c^{t,m} = \begin{cases} t_c^{u,l_j} + \dfrac{x_{h,i}^{t,1}-x_c^{u,l_j}}{x_{h,i}^{t,1}-x_{h,i}^{u,l_i}}\left(t_{h,i}^{t,1}-t_{h,i}^{u,l_i}\right) & m=1 \\ t_c^{t,m-1}+t_{h,i}^{t,m}-t_{h,i}^{t,m-1} & m>1, \end{cases} \tag{10}$$

where $t_c^{t,m}$ is the time stamp of $M_c^{t,m}$; $t_{h,i}^{t,m}$ is the time stamp of $M_{h,i}^{t,m}$; t_c^{u,l_j} and $t_{h,i}^{u,l_i}$ denote the time stamps of the last map-matched report point of the CPV and HPV i on the upstream link, respectively; x_c^{u,l_j} and $x_{h,i}^{u,l_i}$ denote the position along the link of the last map-matched report point of the CPV and HPV i on the upstream link, respectively.

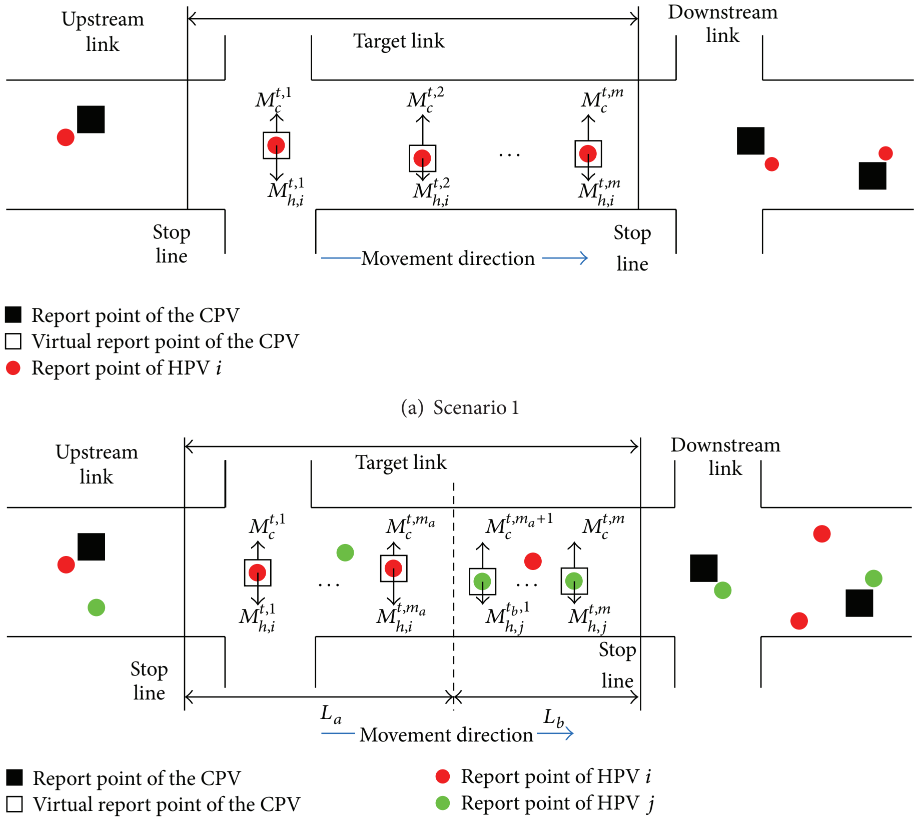
$t_c^{d,1}$ is the time stamp of the first map-matched report point of the CPV on the downstream link. If $t_c^{t,m} > t_c^{d,1}$, then discard $M_c^{t,m}$. Equation (10) shows that $t_c^{t,1}$ is always less than $t_c^{d,1}$; therefore, at least one virtual report point of the CPV would remain on the target link.

Scenario 2 (HPV i and HPV j are the most similar HPVs to the CPV on the upstream link and downstream links, respectively ($i \neq j$)). For this case, the target link L is divided into two segments, L_a and L_b, as shown in Figure 2(b). The map-matched report points of HPV i are used to create the virtual report points on L_a and the map-matched report points of HPV j are used to create the virtual report points on L_b:

$$\begin{aligned} L_a &= \frac{s_i}{s_i+s_j}L, \\ L_b &= \frac{s_j}{s_i+s_j}L, \end{aligned} \tag{11}$$

where s_i is the similarity between HPV i and the CPV; s_j is the similarity between HPV j and the CPV; s_i and s_j are determined on the basis of the reciprocal of the characteristic distance:

$$\begin{aligned} s_i &= \frac{1}{\sum_{m=1}^{l_i} d\left(M_{h,i}^{u,m},\mathrm{CPV}\right)}, \\ s_j &= \frac{1}{\sum_{m=1}^{l'_j} d\left(M_{h,j}^{d,m},\mathrm{CPV}\right)}. \end{aligned} \tag{12}$$



(a) Scenario 1

(b) Scenario 2

Figure 2: Creation of virtual report points in different scenarios.

The characteristic parameters $x_c^{t,m}$, $y_c^{t,m}$, and $v_c^{t,m}$ are determined in the same way as was proposed in Scenario 1. The time stamp $t_c^{t,m}$ is determined as follows:

$$t_c^{t,m} = \begin{cases} t_c^{u,l_j} + \dfrac{x_{h,i}^{t,1} - x_c^{u,l_j}}{x_{h,i}^{t,1} - x_{h,i}^{u,l_i}} \left(t_{h,i}^{t,1} - t_{h,i}^{u,l_i}\right) & m = 1 \\ t_c^{t,m-1} + t_{h,i}^{t,m} - t_{h,i}^{t,m-1} & 1 < m \leq m_a \\ t_c^{t,m_a} + \dfrac{x_{h,j}^{t_b,1} - x_c^{t,m_a}}{x_{h,j}^{t_b,1} - x_{h,j}^{t_a,l_a}} \left(t_{h,j}^{t_b,1} - t_{h,j}^{t_a,l_a}\right) & m = m_a + 1 \\ t_c^{t,m-1} + t_{h,j}^{t_b,m-m_a} - t_{h,j}^{t_b,(m-m_a)-1} & m > m_a + 1, \end{cases} \quad (13)$$

where m_a is the number of virtual report points from the CPV on L_a; $x_{h,j}^{t_b,1}$ and $t_{h,j}^{t_b,1}$ denote the position and time stamp of the first report point of HPV j on L_b, respectively; $x_{h,j}^{t_a,l_a}$ and $t_{h,j}^{t_a,l_a}$ denote the position and time stamp of the last report point of HPV j on L_a, respectively; $t_{h,j}^{t_b,m-m_a}$ is the time stamp of the $(m-m_a)$th report point of HPV j on L_b. As discussed in Scenario 1, if $t_c^{t,m} > t_c^{d,1}$, then discard $M_c^{t,m}$.

The proposed model is able to be extended to cases where there is no report point of the CPV on two or more consecutive links. In this case, the travelling trajectory of the CPV needs to be inferred before using the virtual report point creation model. Rahmani and Koutsopoulos [23] developed a path inference method from low frequency probe vehicle data for urban networks, which could be introduced into this case.

2.2. Improved Back Propagation Neural Networks Model. Artificial Neural Networks (ANN) have been widely used in parameter estimation. However, traditional ANN methods have many shortcomings, such as slow convergence, local optimum identification, and long training times. Aiming at overcoming the shortcomings of ANN models, Xiao et al. proposed an improved back propagation neural network (BPNN) that combined the momentum item and Levenberg-Marquardt algorithm to improve the generalisation ability [24]. The weight equation for their improved BPNN is as follows:

$$\omega(t+1) = \omega(t) + \Delta\omega(t) + \lambda(t) \times \alpha(t) \times \Delta\omega(t-1), \quad (14)$$

where $\omega(t)$ denotes the weight vector in the tth iteration time step; $\Delta\omega(t)$ denotes the variation of weights in the tth iteration time step; $\omega(t+1)$ denotes the weight vector in the next

Table 1: Parameter adjustment in the improved BPNN model.

Variation of error	Adjustment of momentum coefficient	Adjustment of learning rate
$E(t) < E(t-1)$	$\alpha(t) = 1.2\alpha(t-1)$	$\lambda(t) = 1.2\lambda(t-1)$
$E(t) = E(t-1)$	$\alpha(t) = \alpha(t-1)$	$\lambda(t) = \lambda(t-1)$
$E(t) > E(t-1)$	$\alpha(t) = \dfrac{1.2\alpha(t-1)}{1.2}$	$\lambda(t) = \dfrac{1.2\lambda(t-1)}{1.2}$

iteration time step; $\Delta\omega(t-1)$ denotes the variation of weights in the previous time step; $\alpha(t)$ is the momentum coefficient at the tth iteration, where $0 < \alpha(t) < 1$; and $\lambda(t)$ is the learning rate at the tth iteration. The adjustment of $\alpha(t)$ and $\lambda(t)$ is shown in Table 1.

Since the momentum coefficient is supposed to be within the interval $(0, 1)$, if $\alpha(t) > 1$, then set $\alpha(t)$ to 0.01.

As discussed in Zheng and Van Zuylen [13], the travel time along the target link is correlated with the travel times along both upstream and downstream links. Therefore, in the improved BPNN model, the report points of the probe vehicle on the upstream and downstream links are incorporated with the report points (real or virtual) on the target link. According to the discussion in Section 2.1, positions, instantaneous speeds, and their time stamps can form the input data set in the improved BPNN model. Figure 3 shows the structure of the improved BPNN model. The mathematical description of the improved BPNN model is as follows.

(1) Input Layer. Consider

$$\mathbf{X}(i) = \begin{bmatrix} X_1(i) \\ \vdots \\ X_M(i) \end{bmatrix} = \begin{bmatrix} \mathbf{x}(i) \\ \mathbf{v}(i) \\ \mathbf{t}(i) \\ N \\ \delta \end{bmatrix}, \tag{15}$$

$$\mathbf{x}(i) = \begin{bmatrix} x_1(i) \\ \vdots \\ x_n(i) \end{bmatrix}, \quad \mathbf{v}(i) = \begin{bmatrix} v_1(i) \\ \vdots \\ v_n(i) \end{bmatrix}, \quad \mathbf{t}(i) = \begin{bmatrix} t_1(i) \\ \vdots \\ t_n(i) \end{bmatrix},$$

where $\mathbf{X}(i)$ denotes the input data vector of CPV i; $\mathbf{x}(i)$ denotes the position vector of CPV i on the upstream link, target link, and downstream link; $\mathbf{v}(i)$ denotes the instantaneous speed vector of CPV i; $\mathbf{t}(i)$ denotes the time stamp vector of CPV i; n denotes the number of report points taken into consideration for CPV i; N is the link number of the target link; δ is equal to 1 if there are virtual report points of CPV i on the target link and 0 otherwise; and M is the number of input neurons which can be determined as follows:

$$M = 3n + 2. \tag{16}$$

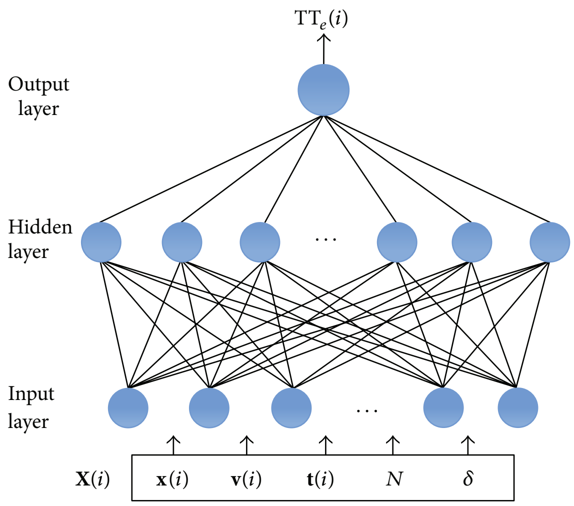


Figure 3: Structure of the improved BPNN model.

(2) Hidden Layer. Consider

$$H(i) = \begin{bmatrix} h_1(i) \\ \vdots \\ h_q(i) \end{bmatrix} = \begin{bmatrix} f\left(\sum_{j=1}^{M} \omega_{j,1} X_1(i) + b_1\right) \\ \vdots \\ f\left(\sum_{j=1}^{M} \omega_{j,q} X_1(i) + b_q\right) \end{bmatrix}, \tag{17}$$

where $h_q(i)$ is the value of the qth hidden neuron; $\omega_{j,q}$ denotes the weight connecting the jth input neuron and the qth hidden neuron; b_q denotes the bias with a fixed value for the qth hidden neuron; f denotes the activation function. Usually, the sigmoid function is selected as the activation function [13]:

$$f(y) = \frac{1}{1 + e^{-y}}. \tag{18}$$

The number of hidden neurons in the improved BPNN model Q can be determined from

$$Q = \sqrt{M + K} + \sigma, \tag{19}$$

where K is the number of output neurons. Here, there is only one output neuron: the estimated link travel time; thus, $K = 1$ and σ is a constant such that $0 \le \sigma \le 10$.

(3) Output Layer. Consider

$$Y(i) = \mathrm{TT}_e(i) = f\left(\sum_{q=1}^{Q} \omega_q h_q(i) + b\right), \tag{20}$$

where $Y(i)$ and $\mathrm{TT}_e(i)$ denote the estimated travel time of CPV i on the target link; ω_q denotes the weight connecting the qth hidden neuron and the output neuron; and b is the bias with a fixed value for the output neuron [13].

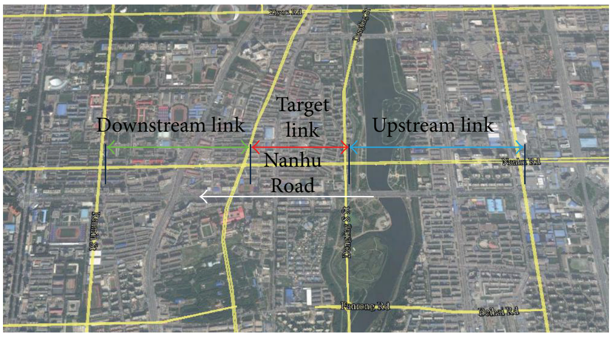


Figure 4: The case study arterial road in Changchun, China.

Finally, the estimated link travel time during the current time window is determined as follows:

$$\mathrm{TT}_e = \frac{1}{N_c}\sum_{i=1}^{N_c}\mathrm{TT}_e(i), \qquad (21)$$

where N_c is the number of CPVs in the current time window.

3. Model Application

The methodology proposed in Section 2 was applied to a route along Nanhu Road in Changchun, China, as shown in Figure 4. Changchun is the capital of the Chinese province Jilin, and the network in Changchun City contains approximately 5700 links and 3400 nodes. Nanhu Road is a typical urban arterial road, and the studied route is located in one of the busier areas of Changchun city.

The studied route is about 3.16 km long, divided into three links, and each link typically contains a signalised intersection; the red arrow indicated in Figure 4 is the target link for travel time estimation (about 0.74 km long); the blue arrow indicates the upstream link (about 1.33 km long); the green arrow indicates the downstream link (about 1.09 km long); the white arrow shows the driving direction taken into consideration in this case study.

3.1. Data Source and Description. The GPS probe vehicle data were obtained from the fleet dispatching system of a taxi company in Changchun city. A daily average of 2500 taxis were driving with GPS devices along the studied route. The default sampling frequency is one report *per* 30 s. To evaluate the effect of the proposed method at lower sampling frequencies, three different sampling frequency GPS probe data sets (i.e., 60, 90, and 120 s) were extracted from the original data set (i.e., 30 s). All GPS probe vehicle data were map-matched onto the road network using a method developed elsewhere [25]. GPS probe data for three consecutive Mondays (time interval: 6:00 a.m. to 6:00 p.m.) from 1 September, 2014, to 15 September, 2014, were used to construct the training sets for virtual report point construction, which takes the early and late peak periods into consideration. Correspondingly, GPS probe data for the time interval from 6:00 a.m. to 6:00 p.m. on 22 September, 2014 (a Monday), were used to construct the testing sets. The time window length for link travel time estimation is 5 minutes. Therefore, for each sampling frequency, there are 3 × 144 = 432 groups of data forming the training sample and 144 groups of data forming the testing sample: each group of data is divided into upstream and downstream sets.

The real link travel times of individual vehicles were recorded by high-resolution digital video cameras through license plate reidentification. The average value was chosen to be the reference link travel time. A total of 5528 vehicles were recorded during the experiment from 6:00 a.m. to 6:00 p.m. on 22 September, 2014. Therefore, the reference link travel time set with 144 values was constructed.

3.2. Selection of Experimental Parameters. For the improved BPNN model, the initial momentum item $\alpha = 0.01$; the initial learning rate $\lambda = 0.01$; the maximum number of iterations maxint $= 20{,}000$; the maximum permissible error $E_m = 1.0 \times 10^{-6}$; the value of σ used to determine the number of the hidden neurons was 8.

3.3. Performance Evaluation Indices. For the purpose of evaluating the performance of the link travel time estimation method proposed here, three widely used evaluation indices are introduced: Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE):

$$\begin{aligned}
\mathrm{RMSE} &= \sqrt{\frac{1}{N_T}\sum_{i=1}^{N_T}\left(\mathrm{TT}_{r,i} - \mathrm{TT}_{e,i}\right)^2}, \\
\mathrm{MAE} &= \frac{1}{N_T}\sum_{i=1}^{N_T}\left|\mathrm{TT}_{r,i} - \mathrm{TT}_{e,i}\right|, \qquad (22) \\
\mathrm{MAPE} &= 100 \times \frac{1}{N_T}\sum_{i=1}^{N_T}\left|\frac{\mathrm{TT}_{r,i} - \mathrm{TT}_{e,i}}{\mathrm{TT}_{r,i}}\right|,
\end{aligned}$$

where $\mathrm{TT}_{r,i}$ is the reference link travel time during the time window i; $\mathrm{TT}_{e,i}$ is the estimated travel time during time window i; N_T is the total number of time windows (here, $N_T = 144$).

3.4. Model Performance and Analysis. The link travel times estimated by the proposed method with different sampling frequency GPS probe data sets (i.e., 30, 60, 90, and 120 s) are compared with the reference link travel times shown in Figure 5. The trend in the estimated link travel times at different sampling frequencies is, on the whole, consistent with the reference link travel times. When the reference link travel time significantly increases during the early and late peak periods, the estimated link travel time curves at different sampling frequencies all show the same characteristics as the reference travel time curve.

Figures 6–8 show the correlation between the reference link travel times and the estimated link travel times under different sampling frequencies based on the proposed method,

Table 2: Performance measurements of the three methods.

Sampling interval (s)	Proposed method		ANN method		Spatiotemporal moving average method	
	MAE (s)	RMSE (s)	MAE (s)	RMSE (s)	MAE (s)	RMSE (s)
30	4.96	5.88	6.68	8.24	8.04	10.46
60	6.05	7.37	8.30	9.93	15.56	18.15
90	6.56	8.03	10.14	12.57	16.65	19.65
120	7.28	8.96	10.68	12.84	17.38	20.27

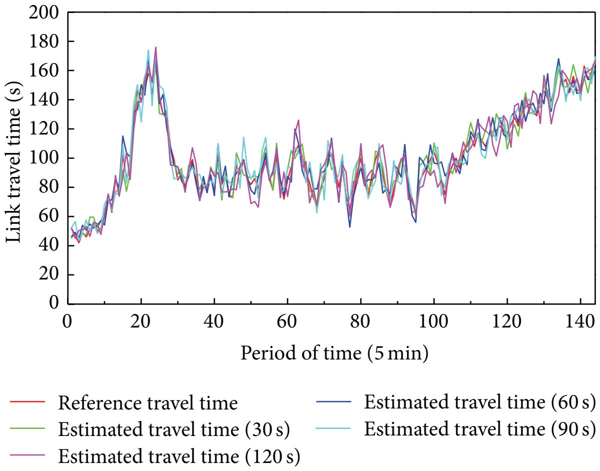


Figure 5: Estimation results based on the proposed method and different sampling frequencies.

the ANN model, and the spatiotemporal moving average method, respectively. It can be seen that the three methods all perform well at a sampling interval of 30 s; the link travel times estimated by the three methods all have high correlation with the reference link travel time (R^2 is more than 0.9 as shown in Figures 6(a), 7(a), and 8(a)) and, in particular, the proposed method (R^2 is more than 0.95). When the sampling frequency decreases from one report *per* 30 s to one report *per* 120 s, the link travel times estimated by the ANN model and the spatiotemporal moving average method both deviate from the reference link travel time significantly (R^2 for the ANN model decreases from 0.9263 to 0.83653 and R^2 for the spatiotemporal moving average method decreases from 0.91451 to 0.80118 as shown in Figures 7(b), 7(c), 7(d), 8(b), 8(c), and 8(d)), but the proposed method still performs well (R^2 for the proposed method is always more than 0.9 at sampling intervals of 60, 90, and 120 s as shown in Figures 6(b), 6(c), and 6(d)).

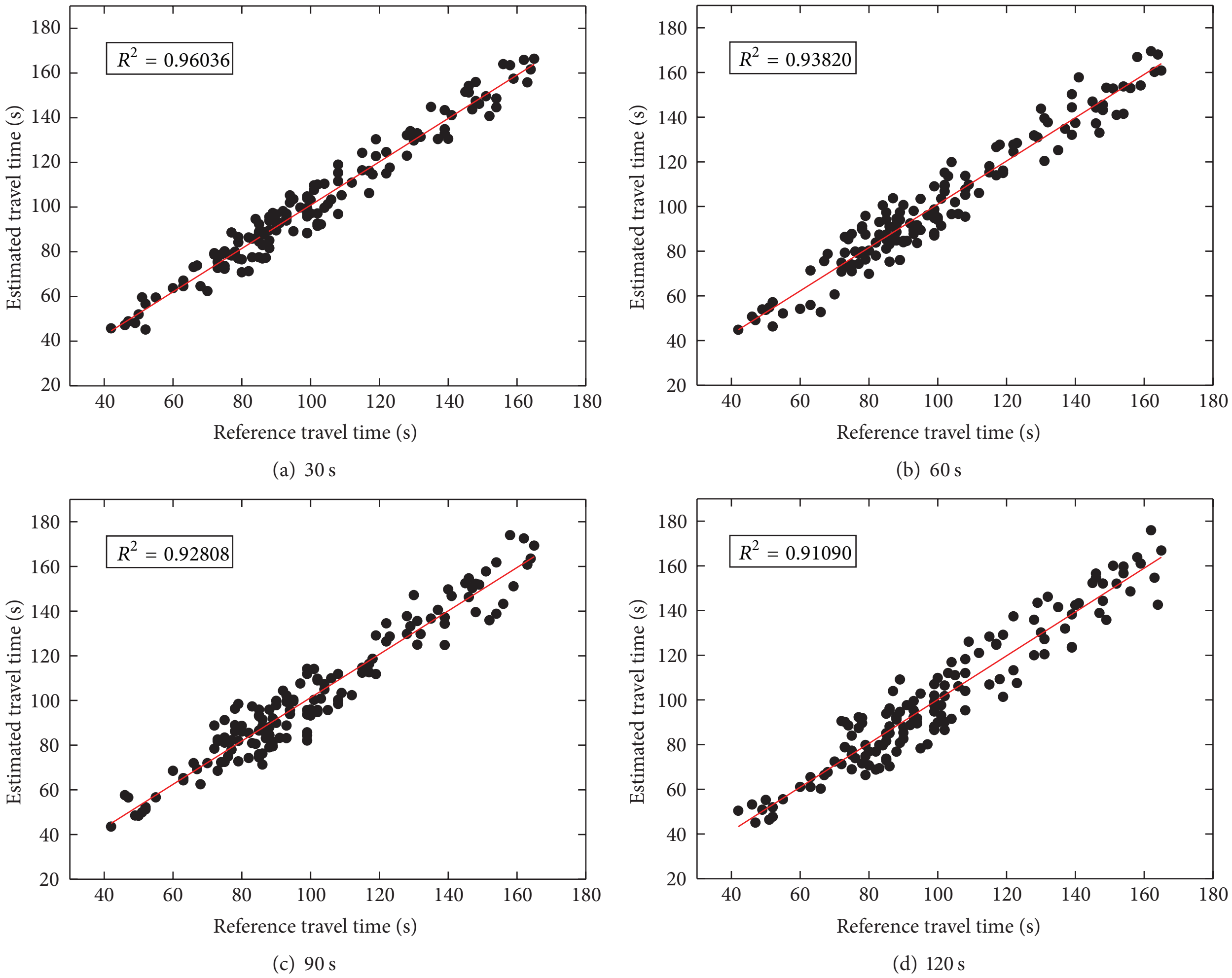
The performance of the three estimation methods, in terms of MAE and RMSE, is indicated in Table 2. Both the MAE and RMSE of the proposed method are less than those found with the other two methods. In addition, when the sampling frequency decreases from one report *per* 30 s to one report *per* 120 s, the MAE and RMSE of the ANN model increase from 6.68 s and 8.24 s to 10.68 s and 12.84 s, respectively; the MAE and RMSE of the spatiotemporal moving average method increase from 8.04 s and 10.46 s to 17.38 s and 20.27 s, respectively, whereas the proposed method still performs well at longer sampling intervals (60, 90, and 120 s), as the MAE and RMSE are always less than 8 s and 10 s, respectively. Therefore, the proposed method returns a higher estimation accuracy than the other two methods under low sampling frequency conditions.

Figure 9 shows the comparison among the three link travel time estimation methods in terms of MAPE. It can be seen that the MAPE of the proposed method not only is less than the other two methods, but also increases marginally as the sampling frequency decreases. The increase in MAPE is always less than 2%. As for the ANN model, the MAPE significantly increases from 7.4% to 11.0% when the sampling frequency decreases from one report *per* 30 s to one report *per* 90 s. As for the spatiotemporal moving average method, the MAPE significantly increases from 8.5% to 16.2% when the sampling frequency decreases from one report *per* 30 s to one report *per* 60 s. Therefore, the proposed method has better robustness than the other two methods under low sampling frequency conditions.

Bayesian method was used to analyze the reliability of the proposed method in the case without observations on CPVs on the target link. The distribution of the absolute estimation error based on the experimental data was selected as the prior distribution. The statistical proportion of the studied case at different error intervals based on the proposed method is shown in Table 3. It can be seen that the proportion of the studied case increases significantly as the sampling frequency decreases; and the estimation error increases as the proportion increases. The posteriori distribution of the absolute error based on the proposed method in the studied case is shown in Figure 10. It can be seen that the distributions are not symmetrical due to the existence of a few large values of the estimation error. This is because the report points of the CPV on the adjacent links could be exceptional due to traffic incidents, which would significantly reduce the accuracy of the most similar HPV recognition and the accuracy of the travel time estimation. The 95% credible interval of the absolute error based on the three estimation methods is indicated in Table 4. It can be seen that the 95% credible interval of the proposed method is the narrowest of the three methods. And the length of the interval based on the proposed method is always less than 17 s, which satisfies the accuracy requirement for the engineering application, whereas the length of the interval based on the ANN method is always more than 15 s and increases to 26.23 s as the sampling interval increases to 120 s. As for the spatiotemporal moving average method, the length of the interval is always more than 20 s and increases to 41.16 s as the sampling interval

Table 3: Proportion of the studied case based on the proposed method (%).

Sampling interval (s)	Interval of the absolute error (s)											
	[0, 2)	[2, 4)	[4, 6)	[6, 8)	[8, 10)	[10, 12)	[12, 14)	[14, 16)	[16, 18)	[18, 20)	[20, 22)	[22, 24]
30	10.7	13.3	18.8	22.9	33.4	52.9	100	100	100	100	\	\
60	18.2	28.5	35.9	40.4	62.2	100	100	100	100	100	100	\
90	27.2	43.8	55.6	66.0	88.6	100	100	100	100	100	100	\
120	40.2	59.8	73.7	92.5	100	100	100	100	100	100	100	100



(a) 30 s

(b) 60 s

(c) 90 s

(d) 120 s

Figure 6: Correlation between estimated link travel times and reference travel times based on the proposed method at different sampling frequencies.

increases to 120 s. Therefore, the reliability of the proposed method is higher than the other two methods in the case without observations on CPVs on the target link.

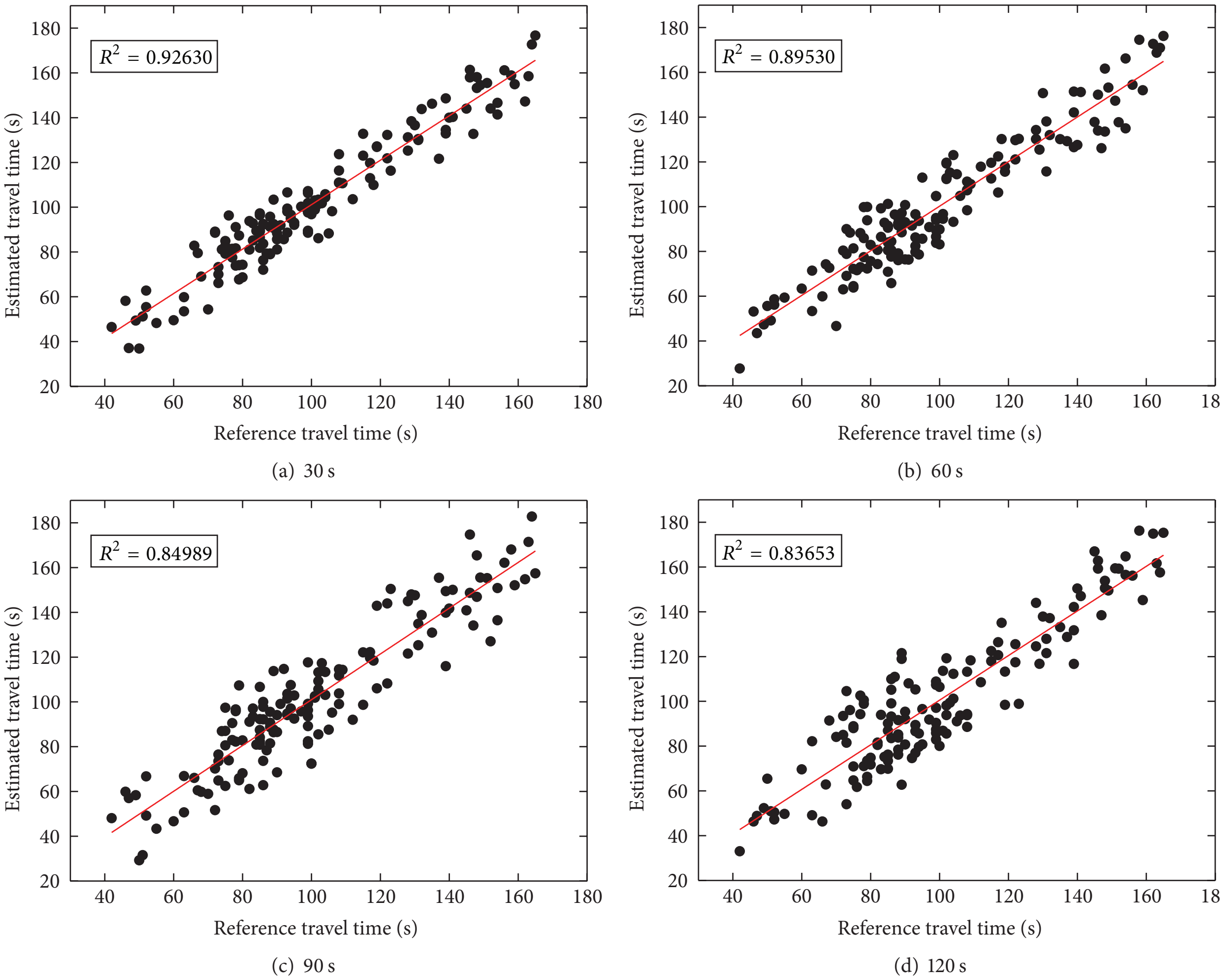
In summary, the proposed method performs better than the ANN model and the spatiotemporal moving average method, under low sampling frequency conditions. This is because it is able to impute missing report points on the target link based on historical data when there is no report point on the target link. When the sampling frequency gets lower, the possibility of no report point on the target link would be higher; then the advantage of the proposed method would be more significant.

4. Discussion and Conclusions

An urban link travel time estimation method using low frequency GPS probe vehicle data was proposed. For the case in which there is no report point from the current probe vehicle on the target link due to the low sampling frequency,

Table 4: 95% credible interval of the absolute error based on the three estimation methods.

Sampling interval (s)	Interval of the absolute error (s)		
	Proposed method	ANN method	Spatiotemporal moving average method
30	[0, 8.82]	[0, 15.85]	[0, 23.78]
60	[0, 11.53]	[0, 19.05]	[0, 32.62]
90	[0, 14.36]	[0, 23.56]	[0, 38.27]
120	[0, 16.44]	[0, 26.23]	[0, 41.16]



(a) 30 s

(b) 60 s

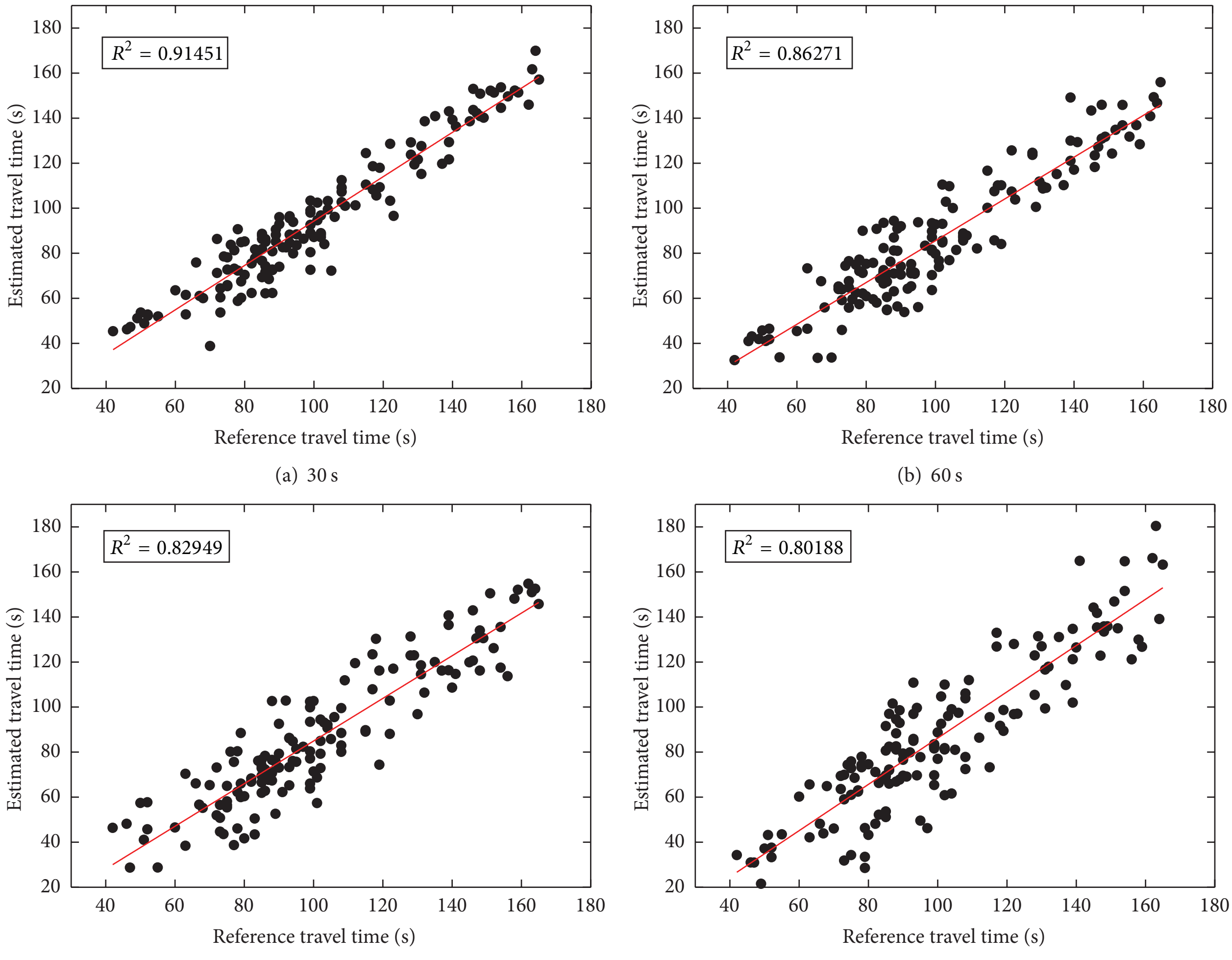
(c) 90 s

(d) 120 s

Figure 7: Correlation between estimated link travel times and reference travel times based on the ANN method at different sampling frequencies.

a K-Nearest Neighbour Rule based model was proposed. The main idea of the model is to recognise the historical vehicle which has the most similar travelling characteristics to the current probe vehicle and uses its report points to create virtual report points for the current vehicle on the target link. Then the virtual report points on the target link were incorporated with the report points on the upstream and downstream links to estimate the link travel time using an improved back propagation neural network model. The proposed methodology was applied to a case study involving an arterial road in Changchun, China, and comparison with the ANN method and the spatiotemporal moving average method was undertaken. Results suggested that the proposed method outperforms the other two methods with higher estimation accuracy and better robustness.

The reliability of the proposed method was validated in the case without report points for CPVs on the target link. When the sampling interval is very long (i.e., 120 s), the 95% credible interval of the absolute error based on the proposed method is $[0, 16.44]$, which still satisfies the accuracy requirement for the engineering application. The accuracy of the proposed method could be influenced by



(a) 30 s

(b) 60 s

(c) 90 s

(d) 120 s

Figure 8: Correlation between estimated link travel times and reference travel times based on the spatiotemporal moving average method at different sampling frequencies.

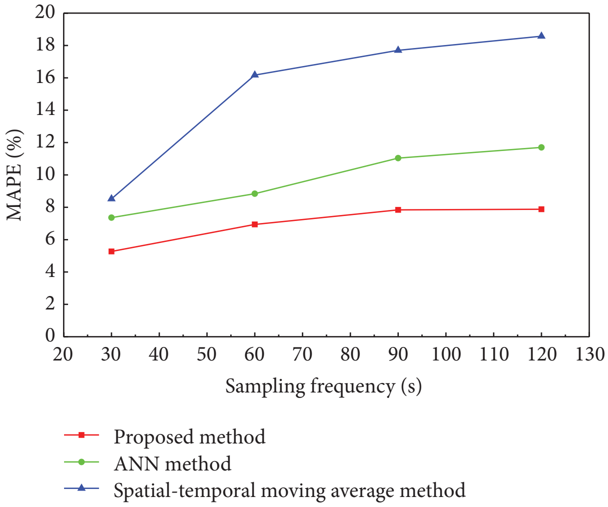


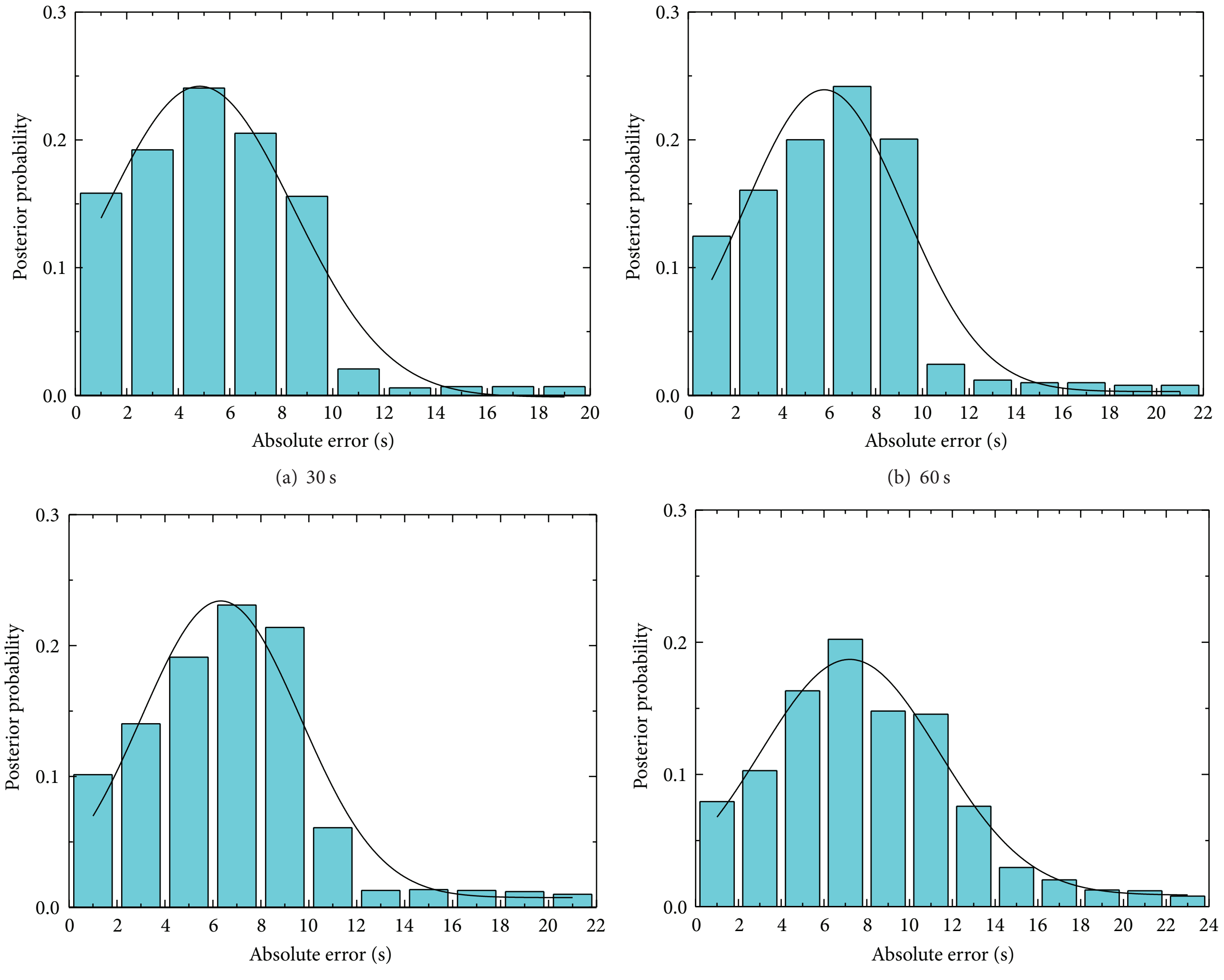
Figure 9: MAPE of the three methods at different sampling frequencies.

the nonrecurring traffic incidents (e.g., disabled vehicles and traffic crashes). This is because the nonrecurring incidents would lead to exceptional observations on CPVs on the adjacent links, which would reduce the accuracy of the similar HPV recognition. Nevertheless, the proposed method is applicable for the general traffic condition.

In future, it is recommended that the virtual report point creation model be improved by constructing a more logical characteristic distance function, which takes the differences among the characteristic parameters into consideration. Besides, the traffic process, the influence of traffic control, and queuing could be considered specifically in urban link travel time estimation. In addition, more efficient algorithms could be introduced to the urban link travel time estimation.

Conflict of Interests

The authors declare that there is no conflict of interests arising from the publication of this paper.



(a) 30 s

(b) 60 s

(c) 90 s

(d) 120 s

Figure 10: A posteriori distribution of the absolute error based on the proposed method at different sampling frequencies.

Acknowledgment

This research has been jointly supported by The National Science & Technology Pillar Programme (Grant no. 2014BAG03B03).

References

[1] J. S. Oh, R. Jayakrishnan, and W. Recker, "Section travel time estimation from point detection data," in *Proceedings of the 82nd Annual Meeting of the Transportation Research Board*, Washington, DC, USA, January 2003.

[2] J. W. C. van Lint and N. J. van der Zijpp, "Improving a travel-time estimation algorithm by using dual loop detectors," *Transportation Research Record*, vol. 1855, no. 1, pp. 41–48, 2003.

[3] J. Kwon and K. Petty, "Travel time prediction algorithm scalable to freeway networks with many nodes with arbitrary travel routes," *Transportation Research Record*, vol. 1935, no. 1, pp. 147–153, 2005.

[4] P. Edara, B. Smith, J. Guo, S. Babiceanu, and C. McGhee, "Methodology to identify optimal placement of point detectors for travel time estimation," *Journal of Transportation Engineering*, vol. 137, no. 3, pp. 155–173, 2010.

[5] P. Chakroborty and S. Kikuchi, "Using bus travel time data to estimate travel times on urban corridors," *Journal of the Transportation Research Board*, vol. 1870, no. 1, pp. 18–25, 2004.

[6] K. Liu, T. Yamamoto, and T. Morikawa, "Feasibility of using taxi dispatch system as probes for collecting traffic Information," *Journal of Intelligent Transportation Systems: Technology, Planning, and Operations*, vol. 13, no. 1, pp. 16–27, 2009.

[7] X. Zhan, S. Hasan, S. V. Ukkusuri, and C. Kamga, "Urban link travel time estimation using large-scale taxi data with partial information," *Transportation Research Part C: Emerging Technologies*, vol. 33, pp. 37–49, 2013.

[8] P. Zheng and M. McDonald, "Estimation of travel time using fuzzy clustering method," *IET Intelligent Transport Systems*, vol. 3, no. 1, pp. 77–86, 2009.

[9] Y. Guessous, M. Aron, N. Bhouri, and S. Cohen, "Estimating travel time distribution under different traffic conditions," *Transportation Research Procedia*, vol. 3, pp. 339–348, 2014.

[10] E. Jenelius and H. N. Koutsopoulos, "Probe vehicle data sampled by time or space: consistent travel time allocation and

estimation," *Transportation Research Part B: Methodological*, vol. 71, pp. 120–137, 2015.

[11] E. Jenelius and H. N. Koutsopoulos, "Travel time estimation for urban road networks using low frequency probe vehicle data," *Transportation Research B: Methodological*, vol. 53, pp. 64–81, 2013.

[12] B. Hellinga, P. Izadpanah, H. Takada, and L. Fu, "Decomposing travel times measured by probe-based traffic monitoring systems to individual road segments," *Transportation Research C: Emerging Technologies*, vol. 16, no. 6, pp. 768–782, 2008.

[13] F. Zheng and H. Van Zuylen, "Urban link travel time estimation based on sparse probe vehicle data," *Transportation Research Part C: Emerging Technologies*, vol. 31, pp. 145–157, 2013.

[14] I. Sanaullah, M. Quddus, and M. Enoch, "Estimating link travel time from low-frequency GPS data," in *Proceedings of the Transportation Research Board 92th Annual Meeting*, Washington, DC, USA, 2013.

[15] C. Diamantini and A. Spalvieri, "Pattern classification by the Bayes machine," *Electronics Letters*, vol. 31, no. 24, pp. 2086–2088, 1995.

[16] E. Kilinc, "Significance of chromatographic and voltammetric data for the classification of green teas in Türkiye: a principle component analysis approach," *Journal of Liquid Chromatography & Related Technologies*, vol. 32, no. 2, pp. 221–241, 2009.

[17] D. Coomans and D. L. Massart, "Potential methods in pattern recognition: part 4. A combination of ALLOC and statistical linear discriminant analysis," *Analytica Chimica Acta*, vol. 132, no. 12, pp. 69–74, 1981.

[18] G. Casalino, N. Del Buono, and C. Mencar, "Subtractive clustering for seeding non-negative matrix factorizations," *Information Sciences*, vol. 257, pp. 369–387, 2014.

[19] Y. Huang, K. B. Englehart, B. Hudgins, and A. D. C. Chan, "A Gaussian mixture model based classification scheme for myoelectric control of powered upper limb prostheses," *IEEE Transactions on Biomedical Engineering*, vol. 52, no. 11, pp. 1801–1811, 2005.

[20] M. Pfeiffer and A. Hohmann, "Applications of neural networks in training science," *Human Movement Science*, vol. 31, no. 2, pp. 344–359, 2012.

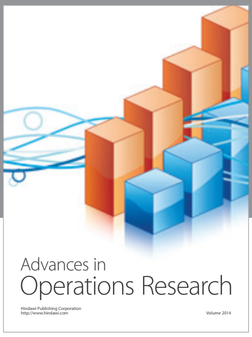
[21] A. D. Dileep and C. Chandra Sekhar, "Class-specific GMM based intermediate matching kernel for classification of varying length patterns of long duration speech using support vector machines," *Speech Communication*, vol. 57, pp. 126–143, 2014.

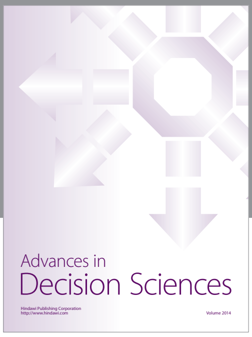
[22] J. Wang, P. Neskovic, and L. N. Cooper, "Neighborhood size selection in the k-nearest-neighbor rule using statistical confidence," *Pattern Recognition*, vol. 39, no. 3, pp. 417–423, 2006.

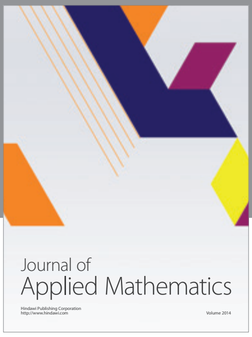
[23] M. Rahmani and H. N. Koutsopoulos, "Path inference from sparse floating car data for urban networks," *Transportation Research Part C: Emerging Technologies*, vol. 30, pp. 41–54, 2013.

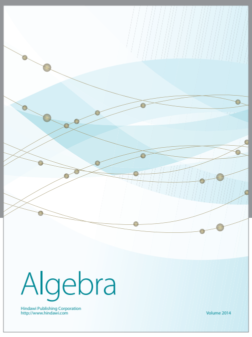
[24] L. Xiao, X. Chen, and X. Zhang, "A joint optimization of momentum item and Levenberg-Marquardt algorithm to level up the BPNN's generalization ability," *Mathematical Problems in Engineering*, vol. 2014, Article ID 653072, 10 pages, 2014.

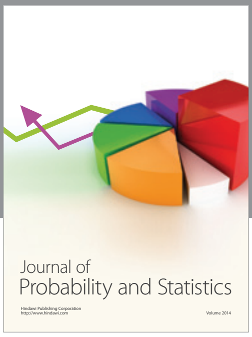
[25] B. Y. Chen, H. Yuan, Q. Li, W. H. K. Lam, S.-L. Shaw, and K. Yan, "Map-matching algorithm for large-scale low-frequency floating car data," *International Journal of Geographical Information Science*, vol. 28, no. 1, pp. 22–38, 2013.


Advances in
Operations Research

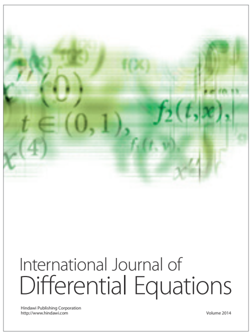

Advances in
Decision Sciences


Journal of
Applied Mathematics

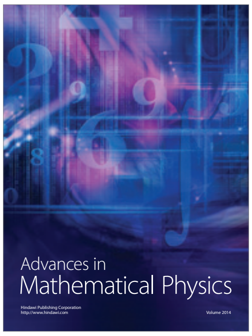

Algebra

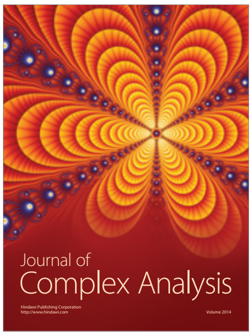

Journal of
Probability and Statistics

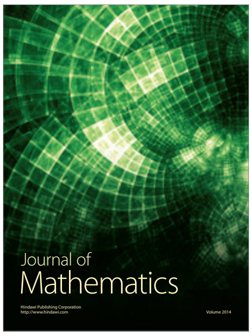
The Scientific
World Journal

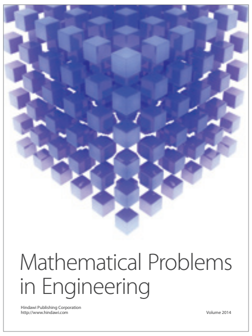

International Journal of
Differential Equations

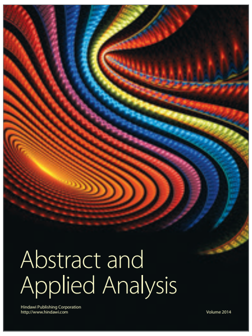
International Journal of
Combinatorics

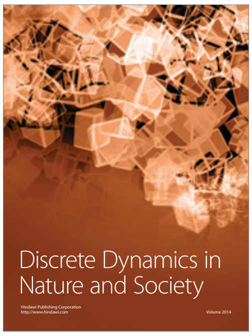

Advances in
Mathematical Physics

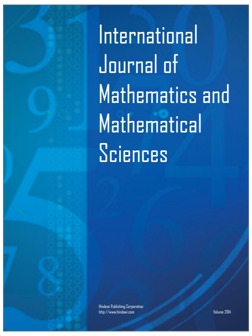

Journal of
Complex Analysis


Journal of
Mathematics


Mathematical Problems
in Engineering

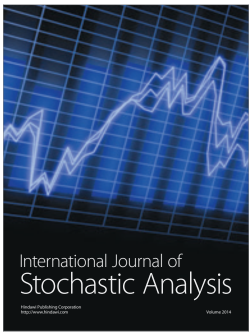

Abstract and
Applied Analysis


Discrete Dynamics in
Nature and Society


International
Journal of
Mathematics and
Mathematical
Sciences

Journal of
Discrete Mathematics

Journal of
Function Spaces


International Journal of
Stochastic Analysis

Journal of
Optimization