Utilizing Artificial Neural Network in GPS-Equipped Probe Vehicles Data-Based Travel Time Estimation

MENGYUN XU¹, KEHUA GUO², JIE FANG¹, AND ZESHAN CHEN¹
¹College of Civil Engineering, Fuzhou University, Fuzhou 350108, China
²School of Information Science and Engineering, Central South University, Changsha 430074, China
Corresponding author: Jie Fang (fangjie@fzu.edu.cn)

ABSTRACT Real-time traffic status information provides good references for urban traffic control and management. Travel time is easy to understand and widely employed in representing traffic status. With significantly improved positioning accuracy and coverage, trajectory data collected from GPS-equipped probe vehicles have great potential for traffic state recognition. This paper presents a machine learning enabled travel time estimation method based on the GPS-equipped probe vehicles data. This research considers the spatial–temporal relevancy while solving the travel time allocation problem: the travel time of target segment might be associated with its previous travel times and/or the traffic states of nearby relevant segments. After data normalization and network clustering, an artificial neural network (ANN) algorithm considering such spatial–temporal relevancy was conducted to infer the travel time distribution among the traveled segments within one path. Furthermore, a weighted summation of the travel time estimation result from various trajectories was calculated to better represent the segment travel time in one time step. The proposed method was evaluated by evaluating the estimation results with automatic vehicle identification obtained ground truth. The experimental results illustrated that by utilizing the ANN to consider the spatial–temporal relevancy, the proposed method is effective and efficient in estimating the travel time.

INDEX TERMS
Artificial neural network (ANN), data clustering, global positioning system (GPS) equipped probe vehicle, travel time estimation, travel time allocation.

I. INTRODUCTION
Real-time traffic conditions estimation on urban arterials is the basis of urban traffic control and congestion management. Traffic conditions estimation enables to capture the variations of the road network and it is the foundation of an accurate and reliable intelligent transportation system (ITS). Multitudinous variables such as flow, speed, occupancy and travel time are very useful in the traffic condition estimation. Compare to others, travel time is a relatively intuitive concept and it can be easily understood by travelers whether professional or not. Moreover, travel time denotes the total time for a vehicle travel through the whole specified route, taking into account the stops, queuing delays, signal delays and delays caused by turning. The overall travel time on network segment is what the traveler cares more about, so it is widely used in Advanced Traveler Information Systems (ATIS). ATIS provides the commuters with the reliable traffic information (e.g. the travel time between origin and destination) to make a reasonable travel plan [1], [2]. Therefore, how to accurately estimate travel time is one of the significant methods to obtain traffic information.

The vast development of the remote sensing and communicating technology has derived many advanced intelligent equipment for traffic information collecting. These equipment fall into two categories, fixed detectors (e.g., loop detectors [3], cameras [4], and magnetic sensors [5]) and mobile sensor devices (e.g., probe vehicles with Global Position System [6], [7] or cell phone [8]). Different techniques bear different advantages and disadvantages. The fixed detectors generally possess better accuracies so that early smart-city engineering projects invested billions into it. Nevertheless,
it suffers from the main disadvantage of low coverage and massive expenditures in both the initial construction and the follow-up maintenance [6]. Obtaining traffic information from cell phone probing may overcome some shortcomings of the fixed devices. However, relatively lower accuracies and the personal privacy problems hinder its development. Different from the aforementioned techniques, GPS-equipped probe vehicle (GPV) is a suitable option that may overcome these problems. For instance, in a medium-sized city such as Zhangzhou, China, over 10000 GPS-equipped commercial probe vehicles which were tracked at all times allow to provide the possibility of generating higher precision level traffic data. Using the data from the GPV can avoid the problem high maintenance cost, and there are no personal privacy problems. Thus, the use of GPV in traffic management and collecting traffic information is growing rapidly.

However, travel time estimation from GPV is facing severe challenges because of low sampling frequencies (less than once per minute due to data transmission costs and bandwidth limitations). One critical challenge is that there are multiple paths in the network between two scarce observed points, and which means different paths may go through different number of network segments [9]. Another challenge involves allocating the observed travel times to each network segment. Allocation and estimation are always performed simultaneously for consistent estimation [10]. These limitations remind us that there is an urgent demand of developing a high-efficiency GPV based algorithm for data processing and travel time estimating.

Various researches aimed at low-frequency GPS-equipped probe vehicles data based travel time estimation method. These studies allocated the travel time between two consecutive points to the traversed network segments or estimated the probability distribution of trip travel time. Therefore, multiple models were proposed, such as statistical regression models [11]–[13], fuzzy logic [14], [15], probabilistic graphical models [16]–[18]. Hellinga et al. [13] presents a heuristic method to decompose the travel time between two observed points to individual segments. A likelihood function considering the stopping probability and congestion probability is maximized to calculate the most likely travel time for each road segment. The results confirm that the proposed method has desirable accuracy while estimating the travel time. Van Hinsbergen et al. [18] proposed a probabilistic graphical modeling framework of Dynamic Bayesian Networks for estimating and predicting travel time distributions, considering the spatial-temporal dependence on the network. An efficient and synthetically statistical regression model for urban travel time estimation was proposed by Jenelius and Koutsopoulos [12]. The method allocates the travel time detailed into every segment travel time and intersection delays, and it makes use of the correlations between different links. Moreover, they further studied [10] the solution of allocating the measured time to traversed components and estimating the travel time distribution parameters based on the different data sampled, time or space. Revealing that the sampling protocol are vital factors in travel time estimation using probe vehicle data. Ramezani and Geroliminis [16] has been built a Markov chain model for analyzing the spatial-temporal evolution of traffic parameters. They integrate the correlation between states of successive link to obtain the arterial route travel time distribution. Besides, Ma et al. [19] proposed a generalized Markov chain approach for estimating the travel time probability distribution based on the correlations in time and space. A heuristic clustering method has been developed to cluster link travel time with homogeneity and underlying traffic conditions.

Nevertheless, for real-time applications, a certain balance is needed in the complexity of the model, calculation accuracy and the computational efficiency of the travel time estimation. Considered this problem, Liu and Ma [20] described a real-time arterial data collection and archival system developed at the University of Minnesota, furthermore, an innovative algorithm considering time-dependent was proposed for arterial real-time travel time estimation applications. Rahmani et al. [21] develops a non-parametric method for route travel time distribution estimation which enables to efficiently calculate and support real time applications. The link travel time allocation is based on the traversed distance on each link as well as prior link travel times. Using information about prior link travel times leads to some potential biases associated with the use of sparse probe vehicle data were reduced. However, except the time-dependence, the traffic condition of target segment may be affected by its upstream and downstream information. Thus, for real-time applications, additionally considering the spatial continuity in non-parametric travel time estimated method may obtain desirable results [17], [22], [23].

With the extensive development of machine learning, some studies applied Artificial Neural Networks (ANN) to infer complex spatial-temporal relevancy of traffic or solving difficult traffic problems. Zeng and Zhang [23] discussed the temporal-spatial interactions of the ANN based travel time prediction method and revealed that is useful in improving modeling accuracy. Differential effects of various inputs (e.g. volume, occupancy, and speed) setting on the ANN was analyzed and it can be found that combining all of various together tend to yield the best results. Hodge et al. [24] introduced a short-term prediction of traffic flow using a binary neural network incorporating both spatial and temporal characteristics into the prediction process. Hoitflener et al. [17] utilize machine learning framework to train static parameters of the roadways, besides, a hybrid approach based flow modeling and machine learning was used to estimate and predict travel times. A Neural Network model was proposed by Zheng and Zuylen [25] to estimate complete link travel time for individual probe vehicle traversing the link. The input to the neural network model is based on non-synthetic data with spatial-temporal dependency, such as collected variables time stamps, speeds and positions. All experimental results showed that the accuracy and computational efficiency of using ANN are desirable.
and reasonable for inferring complex spatial-temporal relevancy [24], [26], [27].

However, using non-synthetic data in travel time estimation may contain some noise data and lead to imprecise outcomes. Besides, the same ANN models were conducted on the different road segment with diverse traffic characteristics, which may produce errors as well. Moreover, most previous travel time estimation methods assume the traffic state of each segment is known and similar as the last time step, which may introduce error due to the vastly changing dynamic nature of the traffic flow. Therefore, this study proposed a travel time estimation method to allocate the observed travel times to each network segment from GPV. Artificial Neural Networks (ANN) were applied to infer complex spatial-temporal relevancy for travel time allocations. Several segments with higher relevancy were clustered into a group for building ANN models and enhancing the accuracy of estimating.

The rest of this paper is organized as follows: Section 2 describes in detail the travel time estimation method based on GPV data, including the path travel time handling, the scaling and allocating methods, Artificial Neural Network models as well as a weighting summation. Section 3 presents the data collection process and discusses the experiment results. Finally, some conclusions and future work are drawn in Section 4.

II. GPS PROBE DATA BASED TRAVEL TIME ESTIMATION METHOD

To describe the procedures of travel time estimation, hereby the authors clarify the definition of three concepts:

Trajectory: The trajectory Traj\textsubscript{a} of vehicle \textit{a} is composed of a sequence of paths between GPS sample points \textit{Traj}_{a} = \{Path\textsubscript{a,1}, Path\textsubscript{a,2}, ..., Path\textsubscript{a,i+k−1}\} (\textit{i} \geq 1), which represents the trajectory connecting the beginning \textit{p}_{a,i} and ending GPS point \textit{p}_{a,i+k} of a trip.

Path: The traveled path Path\textsubscript{a,i} denotes the actual path from GPS point \textit{p}_{i} to its consecutive GPS point \textit{p}_{i+1}. Multiple network segments might be included in one path \textit{Path\textsubscript{a,i}}: \{\textit{S}_{j}, \textit{S}_{j+1}, ..., \textit{S}_{i+n}\} (\textit{n} \geq 0). This research focused on the travel time allocation problem and thus, it is assumed that the actual path of the trip has been acquired. For more information regarding path inference, we refer to [28].

Segment: A segment \textit{S}_{j} is the road in-between two adjacent intersections. Trajectories are connected through various continuous paths, which might contain more than one segment. To support the traveler information system, various combinations of segments with its travel time need to be acquired. Therefore, this paper lay particular emphasis on the estimation of segment travel time \textit{T}_{j}.

This section is organized as follows: Since the traversed path between consecutive sampling points in low sampling frequency probe vehicle data might contain multiple network segments (e.g., \textit{Path\textsubscript{a,1}} and \textit{Path\textsubscript{a,i+1}} in Fig. 1) or cannot experience a completely segment (e.g., \textit{Path\textsubscript{a,k,i+k−1}} in Fig. 1). Section 2.1 presents the different situations that might be encountered while handling the traveled paths.

Section 2.2 presents the proposed artificial intelligence enabled travel time estimation method. Firstly, a scaling procedure is performed to handle the partially travelled segments. Secondly, the GPV data observed path travel times needs to be allocated to each segment. This research considers the spatial-temporal relevancy while solving the travel time allocation problem. The travel time of target segment might be associated with its previous travel times and/or the traffic states of nearby relevant segments. In section 2.2.3, a neural network algorithm considering such spatial-temporal relevancy was conducted to infer the travel time distribution among the travelled segments within one path and thus, the segment travel time for one trajectory can be estimated. Furthermore, a weighted summation of the travel time estimation result from various trajectories was calculated to better represent the segment travel time in one time step.

A. PATH TRAVEL TIME HANDLING

For low sampling frequency probe vehicle, the path travel time allocation problem could have three different scenarios, which is shown in Fig 1.

As shown in the Fig 1, Nodes (blue square) stand for the intersections in the road network; Segments (gray line) represent the road segments connecting the intersections; \textit{P}_{a,i} (red dot) is the observed GPS sampling point after map matching; Path (yellow arrow line) represents the actual travel path between two consecutive sampling points.

Type one: the consecutive sampling points in a trip locates in the same segment, as shown in Fig 1 \textit{Path\textsubscript{a+k,i+k−1}}. For this scenario, the GPV spends the completely sampling interval to traverse partial of one segment. This is typically due to 1) the segment is too long, 2) the segment is congested or

FIGURE 1. Different situation of the traveled path allocation.
3) vehicle is waiting for the signal. The travel time on the untraveled road of the segments is therefore unknown and needs to be inferred based on the collected data. This process of inferring the travel time observed to the complete segments is referred as scaling (more details of scaling were presented in section 2.2.2).

Type two: The consecutive sampling points in a trip are located on the adjacent segments, as shown in Fig 1 Path$_{a,i+1}$. In this case, the collected path travel time includes the time crossing an intersection and needs to be allocated into two segments. Furthermore, the GPV does not traverse each segment entirely so the scaling is also needed.

Type three: As shown in Fig 1 Path$_{a,i}$, the consecutive sampling points in a trip are located on unconnected segments. The travel time needs to be allocated into each segment of the vehicle experienced. Scaling is also required for the first segment and last segment among the traveled segments. For this scenario, the GPV traverses at least two intersections for the specific sampling interval. This is typically due to 1) occasional GPS transmission incontinuity caused larger sampling interval 2) the segment length is too short 3) the intersection signal control contributes little delay to the segment travel time.

B. ARTIFICIAL INTELLIGENCE ENABLED TRAVEL TIME ESTIMATION METHOD

1) SCALING

The GPV traveled path between two consecutive sampling points $p_{a,i}$ and $p_{a,i+1}$ may cover only a fraction of a segment (type one) or partially traverse first and last segments (type two and three). For example, in Fig 2, the traveled distance $x_j$ is part of the total segment length $L_j$. Therefore, the first step is to scale up the collected path travel time $\Delta t_{a,i} = t_{s_{a,i+1}} - t_{s_{a,i}}$ to the segment travel time, in which, $t_{s_{a,i+1}}$ and $t_{s_{a,i}}$ are collected probe vehicle timestamps, $t_j$ is the travel time for the traversed partial of segment $j$, which is equal to path travel time $\Delta t_{a,i}$ in this case ($t_j = \Delta t_{a,i}$). After scaling the travel time to the whole segment, the segment travel time is denoted as $T_j$.

The scaled segment travel time $T_j$ was estimated using a distance scaling factor $\eta$. In this paper, the spatial distribution of travel time along one same segment is assumed to be consistent. Therefore, the distance scaling factor $\eta$ is the ratio of the traveled distance to the segment length. The segment travel time is calculated by:

$$T_j = \eta \times t_j$$  \hspace{1cm} (1) \\
$$\eta_j = \frac{L_j}{x_j}$$  \hspace{1cm} (2)

The segment travel time $T_j$ is more reliable when the segment is mostly covered by the probe ($\eta$ close to 1), and otherwise might contain more bias (when $\eta$ close to 0).

2) ALLOCATING

The traversed path, for low sampling frequencies GPS data, might contain multiple segments and intersections depending on the ever-changing traffic status and collecting frequency, such as illustrated in scenario two or three. As presented in the article [12], the travel time of a trip consists of the running time along segments and delays at intersections and traffic signals (turns). The delays at intersections and traffic signals can hardly be estimated directly, thus, it is not reasonable to allocate travel time among different segments using the distance directly (e.g., A vehicle may spend a long time waiting for a traffic signal, but traveled only a short distance in the segment). Previous studies proved that using information such as distance traversed on the segment or prior segment travel time is expected to reduce the allocation bias [21]. Whereas the prior segment travel time has been proven to be a desirable parameter, the spatial relevancy characteristics can neither be ignored. Besides, most previous travel time estimation method assumes the traffic state of each segment is known and similar as the last time step, which may introduce error due to the vastly changing dynamic nature of the traffic flow. Thus, in this paper, apart from the distance traveled on the segment, a travel time estimation method was proposed to involve an artificial neural network that takes into account both the temporal and spatial relevancy of the traffic network.

As is shown in Fig 3, for the scenarios that a pair of consecutive sampling points $p_{a,i}$ and $p_{a,i+1}$ located on different segments (scenario 2 and 3), the inferred traveled distance in segment $j$ is denoted as $x_j$; $t_j$ is travel time allocated in the traveled segment. Afterwards, the segment travel time $T_j$ can be inferred through scaling $t_j$.

The allocation procedures break the observed path travel time $\Delta t_{a,i}$ into segment travel time ($t_j, t_{j+1}, \ldots, t_{j+n}$), taking into consider that the traffic status could be different among
into different groups in accordance with the similarities, the clustering (HBC) [35], [36]. These methods cluster data grid-based clustering (GBC) [33], [34], and hierarchy-based clustering (HBC) [29], [30], density-based clustering (DBC) [31], [32], used in data mining, such as partition-based clustering (PBC)

traffic information. Several clustering methods were widely high-relevancy segments for extracting and analyzing the traffic information. The HBC method is a more suitable method for clustering the network segments to identify the segments with high-relevant travel time variations, e.g. the peak hour or non-peak hour appeared at almost the same time in a day, which may be identified as having a relatively high-relevancy. Among the above mentioned clustering methods, HBC and DBC are suitable to solve our problem since they cluster the data based on their similarity. Comparing to the HBC, DBC requires the number of clusters to be pre-defined. However, in our cases, it is difficult to set a pre-defined number of clustered segments. Therefore, the HBC would be a more suitable method for clustering the network segments to extract and analyze the traffic information. The HBC method contains the following steps:

1. **Extracting relevant data**
   - Treating each independent dataset as a cluster and calculated the relevancy between each pair of clusters.
   - Merging two or more dataset into a new cluster according to their relevancy.
   - Recalculating the relevancy of the new clusters.
   - Repeating steps 2 and 3 until all categories are merged into one group or there are no pairs of clusters with relevancy above the threshold \( \theta \).

Following the above procedures, the authors firstly cluster the segments in the network according to their relevancy Pearson correlation coefficient (PCC), a linear correlation measurement in statistics, was adopted to measure the relevancy between each pair of segment, which can be calculated by (7) and (8).

\[
\rho_{X,Y} = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y},
\]

\[
\text{cov}(X,Y) = E[(X - \mu_X)(Y - \mu_Y)]
\]

where \( \rho_{X,Y} \) is the correlation coefficient, \( \text{cov}(X,Y) \) is the covariance of \( X \) and \( Y \), and \( E \) is the expected value. \( \mu_X \) and \( \mu_Y \) represent the mean values of \( X \) and \( Y \), and \( \sigma_X \) and \( \sigma_Y \) are the standard deviations of \( X \) and \( Y \). The value of PCC ranges from \(-1\) to \(1\), according to the Cauchy–Schwarz inequality, where \(1\) means total positive linear correlation, \(-1\) means total negative linear correlation, and \(0\) means no linear correlation. In this paper, a relevancy threshold of \( \theta \geq 80\% \) was selected, means that segments with relevancy higher than \(80\% \) (both positively and negatively related) will be categorized into one cluster. Secondly, geographical characteristics were considered by combining the segments.

\[
\Delta_t = t_{p,i+1} - t_{p,i} = \sum_{n=0}^{N} t_{j+n},
\]

where, \( t_{j+n} \) is the allocation coefficient taking into account both the spatial-temporal relevancy and traveled distance, which is calculated as follows:

\[
\varphi_{j+n} = \sum_{n=0}^{N} \frac{t_{est,j+n,t} \times y_{j+n}}{X},
\]

\[
\gamma_{j+n} = \frac{x_{j+n}}{L_{j+n}}
\]

3) **ARTIFICIAL NEURAL NETWORK FOR TRAVEL TIME ESTIMATION**

In order to capture the temporal and spatial relevancy, all segments in the network were tested to examine their relevancy and clustered into groups. After data normalization, a neural network model is conducted to estimate the traffic status of segments in each cluster. Here, the network clustering method, data normalization method and the artificial neural network framework are presented in detail.

\( a \): **NETWORK CLUSTERING**

Data clustering techniques are effective tools to discover high-relevancy segments for extracting and analyzing the traffic information. Several clustering methods were widely used in data mining, such as partition-based clustering (PBC) [29], [30], density-based clustering (DBC) [31], [32], grid-based clustering (GBC) [33], [34], and hierarchy-based clustering (HBC) [35], [36]. These methods cluster data into different groups in accordance with the similarities, the distance between the data and cluster centers or the density of each group.

In this paper, the purpose of data clustering is to identify the segments with high-relevant travel time variations, e.g. the peak hour or non-peak hour appeared at almost the same time in a day, which may be identified as having a relatively high-relevancy. Among the above mentioned clustering methods, HBC and DBC are suitable to solve our problem since they cluster the data based on their similarity. Comparing to the HBC, DBC requires the number of clusters to be pre-defined. However, in our cases, it is difficult to set a pre-defined number of clustered segments. Therefore, the HBC would be a more suitable method for clustering the network segments to extract and analyze the traffic information. The HBC method contains the following steps:

- Repeating steps 2 and 3 until all categories are merged into one group or there are no pairs of clusters with relevancy above the threshold \( \theta \).

Following the above procedures, the authors firstly cluster the segments in the network according to their relevancy Pearson correlation coefficient (PCC), a linear correlation measurement in statistics, was adopted to measure the relevancy between each pair of segment, which can be calculated by (7) and (8).

\[
\rho_{X,Y} = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y},
\]

\[
\text{cov}(X,Y) = E[(X - \mu_X)(Y - \mu_Y)]
\]

where \( \rho_{X,Y} \) is the correlation coefficient, \( \text{cov}(X,Y) \) is the covariance of \( X \) and \( Y \), and \( E \) is the expected value. \( \mu_X \) and \( \mu_Y \) represent the mean values of \( X \) and \( Y \), and \( \sigma_X \) and \( \sigma_Y \) are the standard deviations of \( X \) and \( Y \). The value of PCC ranges from \(-1\) to \(1\), according to the Cauchy–Schwarz inequality, where \(1\) means total positive linear correlation, \(-1\) means total negative linear correlation, and \(0\) means no linear correlation. In this paper, a relevancy threshold of \( \theta \geq 80\% \) was selected, means that segments with relevancy higher than \(80\% \) (both positively and negatively related) will be categorized into one cluster. Secondly, geographical characteristics were considered by combining...
segments with geometric connectivity into the same cluster. Moreover, geometric connectivity will be examined in the clustering procedures so that segments distanced from each other will be not categorized into one same cluster even if their travel time profile happen to appear high relevancy. Besides, in order to ensure the spatial continuity, connecting segments will be incorporated in the same cluster.

b: DATA NORMALIZATION

Segments with different lengths process different travel time associated with their length and other characteristics, even in the free flow condition. Therefore, for better utilizing the neural network to estimate the segment travel time, the travel time data needs to be normalized. Feature scaling, also known as data normalization, is a method used to standardize the range of independent variables or features of data. Therefore, in this paper, the segment travel times in a day were adjusted into a notional common range [0, 1], the normalized function is then

\[ X^* = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}} \]  

(9)

c: ARTIFICIAL NEURAL NETWORK FRAMEWORK

The machine learning method, Artificial Neural Network (ANN), was adopted to identify the spatial-temporal relevancy of travel time in the target network.

It was concluded by the earlier studies that the previous traffic states would have influence on subsequent time steps. Furthermore, the relevancy of the segment state decreases as the time elapses. Therefore, in terms of temporal relevancy, the travel times at previous time steps can be used as the input for the ANN. Apart from temporal relevancy, the segments with spatial relevancy should also be considered as input data. After data clustering, the road network was classified into several groups with high-relevancy segments. The previous travel times on multiple segments in the same cluster were considered as the input to utilize one same neural network, which also reduce the number of ANN required for the network.

The output of the conducted ANN is the estimated travel time parameter of each segment in the current time step. The outputted estimations represent the relevancy in-between different segments, instead of the actual travel time value. This estimation result was then used as the reference for allocating the observed path for more accurate segment travel time estimation results.

• Input layer

The ith cluster input data \( X_i \) is present by

\[
X_i = \begin{bmatrix}
X_{i,j} & X_{i,j+1} & \cdots & X_{i,j+n}
\end{bmatrix}
\]

\[
\begin{bmatrix}
T_{i,j-t-m} & T_{i,j+1,t-m} & \cdots & T_{i,j+n,t-m}
\end{bmatrix}
\]

\[
\begin{bmatrix}
T_{i,j-t-m+1} & T_{i,j+1,t-m+1} & \cdots & T_{i,j+n,t-m+1}
\end{bmatrix}
\]

\[
\vdots
\]

\[
\begin{bmatrix}
T_{i,j-t} & T_{i,j+1,t} & \cdots & T_{i,j+n,t}
\end{bmatrix}
\]

\[
\begin{bmatrix}
T_{i,j+1,t-1} & T_{i,j+2,t-1} & \cdots & T_{i,j+n,t-1}
\end{bmatrix}
\]  

(10)

where \( X_{i,j} \) is the input data of segment \( j \) belong to ith cluster; \( T_{i,j,t-m, j+n} \) is the \((t-m, j+n)\)th input neurons presents the travel time of segment \( j + n \) at previous time step \( t-m \). The number of input neurons \( N_i \) in each neural network is determined by the number of segments in the cluster \( n \) as well as the number of time steps \( m \) involved in estimation.

• Hidden layer

After data clustering, the road network was classified into several groups for establishing the ANN models, which reduces the number of the input neurons. Thus, there is no need to construct an over complicated ANN models with multiple hidden layers. Besides, simply adding the number of hidden layers may result in overfitting or deteriorating the model accuracy. Therefore, following the procedures of existing literatures, the authors tested the model with various number of hidden layers. The experiments demonstrate that the accuracy does not improved significantly with increased number of hidden layers. As a result, to limit the complexity of ANN models and minimize the number of parameters, the design of models in this paper include one hidden layer.

The number of hidden neurons is determined by the rule that reduce the number as much as possible on the premise of satisfying the precision and efficiency. Therefore, the range of hidden neuron numbers was calculated by empirical formulas [37], [38], which are shown as follows,

\[ N_h = \log_2 N_i, \]  

(11)

\[ N_h = 2N_i + 1 \]  

(12)

where \( N_h, N_i, N_o \) is the number of hidden neurons, input neurons and out neurons separately;

In the end, the number of hidden neurons \( N_h \) corresponding to the best performance in terms of Mean Square Error (MSE) and Mean Absolutely Error (MAE) was chosen to build the network.

In which, \( H_{i,h} \) represents the value of the \( h \)th hidden neuron for \( i \)th cluster. \( \omega_{i,j+n,t-m, h} \) represents the weight connecting the \((j+n, t-m, h)\)th input neuron and the \( h \)th hidden neuron. \( b_{i,h} \) denotes bias for \( h \)th hidden neuron. \( \delta \) is an activation function for training the complex non-linear relationship and fully combining data characteristics. Considering the travel time fluctuations during peak hour and non-peak hour, \( \delta \) should be a non-linear function. Several widely used activation functions are logistic sigmoid, hyperbolic tangent (Tanh) and the
rectified linear unit (ReLU) functions. Recently, ReLU has been widely used in machine learning algorithms, as it has less computational load, and may alleviate the problem of over-fitting. Considering the data features and the speed of convergence, the form of ReLU $\delta(x) = \max(0, x)$ was chosen in this study, or

$$
\delta(x) = \begin{cases} 
0, & x \leq 0 \\
x, & x > 0 
\end{cases},
$$

(14)

- output layer

The number of output neurons $N_o$ is equal to the number of segments $n$ in each cluster. The structure of output layer was presented by

$$
Y_i = \begin{bmatrix} 
Y_{i,j+1} \\
Y_{i,j+2} \\
\vdots \\
Y_{i,j+n} 
\end{bmatrix} = \begin{bmatrix} 
\zeta \left( \sum_{h=1}^{h} \omega_{i,h,j+1} h_{i,h} + b_{i,j+1} \right) \\
\zeta \left( \sum_{h=1}^{h} \omega_{i,h,j+2} h_{i,h} + b_{i,j+2} \right) \\
\vdots \\
\zeta \left( \sum_{h=1}^{h} \omega_{i,h,j+n} h_{i,h} + b_{i,j+n} \right) 
\end{bmatrix},
$$

(15)

in which, $Y_{i,j+n}$ represents the value of the $n$th output neuron in $i$th cluster, which is the predicted travel time of segment $j + n$ at current time step. $\omega_{i,h,j+n}$ represents the weight connecting the $h$th hidden neuron and the $(j + n)$th output neuron, $b_{i,j+n}$ denotes bias for $(j + n)$th output neuron. $\zeta$ is an activation function, which is selected from the same function as used for the hidden units $\delta$.

The Artificial Neural Network framework for travel time estimation in the segment was shown in Fig 4.

4) WEIGHTING

Multiple trajectories from different GPVs might cover same road segment. That is, each segment may be covered by multiple trajectories, so various segment travel time estimations are made during a time step. The number of observations is determined by the length of time step and the penetration level of GPV. Each estimated travel time $T_{j,k}$ measured from different trajectories (probes) has different influence on the time step averaged segment travel time $T_{j,\text{average}}$. The time step averaged segment travel time can be obtained through:

$$
T_{j,\text{average}} = \frac{\sum_{k} T_{j,k} \times \omega_{j,k}}{\sum_{k} \omega_{j,k}},
$$

(16)

where $\omega_{j,k}$ is the weight assigned to the $k$th segment travel time estimation $T_{j,k}$ in the time step. The weights $\omega_{j,k}$ is a parameter related to probe coverage ratio in the segment ($\gamma_{j+n}$)

III. EXPERIMENT RESULTS AND ANALYSES

A. EXPERIMENTAL NETWORK AND DATA

This research selected the urban area of Zhangzhou (a medium-scale city in Fujian, China) for field-experiment, as is shown in Fig 5. The experiment area includes several representative locations with heavy traffic generation and attraction, such as School, Hospital, Park, Shopping Mall, Resident district, Government, Metro center (marked as blue square 1 to 7 in Fig 5) and so on. Besides, the experimental area includes roads with various classifications and speed limits, e.g. 40 km/h in the central area (e.g. segment 1) or 60 km/h on the outer expressway (e.g. segment 13). Furthermore, after examining the field data, the coverage of GPV in this experiment area is fair and significant changes in traffic state can be observed.

The selected field-experiment network contains 22 segments. Each segment includes the two directions of road between two adjacent intersections. Average distance of segments is around 600 meters, the shortest segment is around 151 meters (e.g. segment 18), and the longest segment which belongs to expressway is around 1445 meters (e.g. segment 13). Time step is set as 5 min, thus a day of 24 hours was divided into 288 time steps.

1) AUTOMATIC VEHICLE IDENTIFICATION DATA

To verify the accuracy of the proposed method, the results are evaluated using the ground truth data collected by automatic vehicle identification (AVI) techniques. AVI techniques can be of various types, the chosen data in this paper are...
from license plate matching techniques. AVI techniques are another type of interval detector except from probe vehicle, both of these interval detectors enable to calculate the travel time directly between two consequence detectors or sampling points [39]. The AVI detectors were installed after the signalized intersections to acquire entire travel time, including running time, signal delays and turning delays, with approximately 5 minutes collecting interval. Therefore, accurate and adequate ground truth can be collected to evaluate the estimation results of the proposed method.

2) PROBE VEHICLE DATA
The probe vehicle data were collected by around 10,000 GPS-equipped commercial probe vehicles, which mainly consist of commercial fleets and taxicabs, covering most areas of Zhangzhou. The sampling interval ranges from 1-3 minute and the penetration rate of the GPV is almost 2%. An example shows the format of GPS data was demonstrated in Table 1, which includes the vehicle ID (Mdtid), current location (Longitude and Latitude), direction, timestamp and vehicle type.

The collected data might be noisy and includes values that needs to be pre-processed or excluded from the data set due to poor data transmission quality, such as null values, repeating values or outliers. The GPS sample points usually drop off the road due to measuring and computing error. To address the issues, topology-based method was adopted to match the GPS points to the road candidate nodes. The crossing angle between the heading direction of the vehicle and the road direction as well as the nearest neighbor method were major considerations. After map matching, a shortest path between the two candidate nodes was selected to represent the real trajectory of the vehicle in-between the two GPS points. This research focused on the travel time allocation problem and thus, it is assumed that the map matching has

![FIGURE 5. The experimental area in Zhangzhou.](image)

**TABLE 1. Data format of a gps sample point.**

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mdtid</td>
<td>5469114</td>
</tr>
<tr>
<td>Longitude</td>
<td>117.7317350</td>
</tr>
<tr>
<td>Latitude</td>
<td>24.5153490</td>
</tr>
<tr>
<td>Direct</td>
<td>2</td>
</tr>
<tr>
<td>Time</td>
<td>2017/05/04 17:10:00</td>
</tr>
<tr>
<td>Veh-type</td>
<td>41</td>
</tr>
</tbody>
</table>
been accomplished and the actual path of the trip has been acquired. For more detailed information regarding the map match and noisy data processing technique, we refer to the previous work [28].

The collected field data (GPVD and AVI) in Zhangzhou are from March 1, 2017 to September 30, 2017. No significant influence from severe weather and seasonal variations was recorded in the experiment period, such as heavy rain or snows. Thus, there is no need to take the weather into account in this study. In order to avoid influences of weekly variations, the data from Mondays through Fridays other than weekends and holidays (the most representative and cyclical periods “weekday”) are used, which is a total of 152 days.

B. DATA PREPARATION

As aforementioned, the field-experiment contains 22 segments (S\textsubscript{j}, denotes segment ID j = 1,2,…,22), each segment is independent. Thus, each segment travel time can be initially treated as a cluster. For hierarchy-based clustering, the processes of computing the relevancy and merging two categories with the highest relevancy into a new cluster are executed iteratively. The travel time data to be used for determining the relevancy of segments is the ground truth collected through automatic vehicle identification technologies (AVI). The results of data clustering are shown in Fig. 6. The vertical axis depicts 22 initial clusters based on collecting AVI data within 152 days. The horizontal axis is the relevancy between two clusters. Considering the relevancy threshold $\theta > 80\%$, as shown in the example, several segments drop on the left of the red line can be merged into new clusters. Such as segment sets [S2, S8, S4, S11], segment sets [S14, S15, S7, S10, S9, S20, S1], segment sets [S21, S22] can be merged into three new clusters: $C_{2,4,8,11}$, $C_{14,15,7,10,9,20,1}$, $C_{21,22}$. Fig 7 provides the example of travel time comparison on two relevant segments (S2 and S8, S15 and S14) on 4 May. The collected data distributed closely to the line of $y = x$ on both sides, and the value of relevancy calculated by PCC is above 80%.

The lengths of segments in $C_{2,4,8,11}$ are almost the same, besides, they possess similar features in terms of traffic demand. The working hours of Government agencies are the same, this is certified by the high relevancy approximates 90% between S2 and S8 (nearing Government agencies). Hospital and Metro center are locations where traffic volumes are discretely distributed over the whole day. The results show that there is a high relevancy between the segments around the Hospital (S2 & S8) and the Metro center (S4 & S11). In order
to ensure the spatial continuity, S3 (adjacent to S2, S4 and S8) was incorporated into the same cluster. Therefore, the first cluster contains five segments [S2, S8, S4, S11, S3], which is expressed by Cluster I.

Similar procedures were performed to establish Cluster II. The segments in C14,15,7,10,9,20,1 are mainly spread on the sides of Residential and Business districts where have closer commuting relationship and similar traffic demands. The travel time distributions of S21 and S22 are similar to each other even if they are far away. Considering the spatial continuity, the segments (S5, S6, S18, S19, S21, S22) between Residential and Business districts were taken into account Cluster II. In cluster III, some segments S13, S16 or S17 (belonging to expressway, speed limits 60 km/h) show a low relevancy with other segments and were merged according to road classification. As aforementioned spatial continuity, S12 nearby S13 was considered in Cluster III. Therefore, Cluster II includes thirteen segments [S14, S15, S7, S10, S9, S20, S1, S5, S6, S18, S19, S21, S22] and four segments [S2, S8, S4, S11] are merged into Cluster III.

Above all, the paper grouped 22 segments into three clusters according to the relevancy test. The segments with similar road classification or traffic demand characteristics are clustered together.

C. ARTIFICIAL NEURAL NETWORK MODELS

Respective ANN will be built for each cluster. ANN establishing include two mainly procedures: artificial neural network training and results testing. The whole data set (152 workdays) was divided into training subset (half of the total data set, 76 workdays, around 21,888 timestamps and around 4,830,000 trajectories) and testing subset (the rest).

The input neurons of each cluster are previous travel times on each segment in that cluster, whose number \( N_i \) depends on the quantity of selected prior time step \( m \) and the number of segments \( n \). \( m \) can be chosen flexibly according to the data characteristics and practical needs. Considering that the relevancy of previous travel states decreases as the time elapses, segment travel times on previous three time step was selected, i.e. \( m = 3 \). The output neurons are the estimated travel time parameters at present time step, \( N_o \) is determined by the number of segments in the cluster \( (N_o = n) \).

The range of neurons number in the hidden layer is calculated in accordance with the empirical algorithm, and then the optimal number of neurons is determined by comparing the modeling accuracy and effectiveness. Therefore, the range of neurons numbers in the hidden layer was calculated based on (12) to (13) and shown in the Table 2.

Using the same data set and limited on the same number of training epochs, the modeling errors after normalization (mean squared error and mean absolute error) to different number of hidden neurons are shown in the Fig 8. The most desirable testing results in terms of the minimum errors and the fair convergence speed are represented by blue circles in the figure. Therefore, the number of hidden neurons used to build the Artificial Neural Network are chosen to be 30 for Cluster I, 45 for Cluster II and 14 for Cluster III.

The proposed ANN structure does not cause huge computation load and can be easily processed by regular workstations with fair convergence speed. For online applications, the reference time for each 5 min time step are as low as 10 seconds in the experimental study area (roughly 50 segments for both directions in 5 km² area, running on one regular workstation). For city-wide applications, the processing of data will require more computational power, which can be easily handling through utilizing Edge/Fog Computing technologies.

D. EVALUATION RESULT AND DISCUSSION

The research measures the accuracy of the proposed algorithm by the Root Means Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE). Both of accuracy indexes are widely used in traffic state estimation and prediction as they can better demonstrate the difference between the estimations and the ground truth, which are defined by the following equations,

\[
\text{RMSE}_j = \sqrt{\frac{1}{M} \sum_{t=1}^{M} (TT_{\text{estimation},j,t} - TT_{\text{real},j,t})^2}, \quad (17)
\]

\[
\text{MAPE}_j = \frac{1}{M} \sum_{t=1}^{M} \left| \frac{TT_{\text{estimation},j,t} - TT_{\text{real},j,t}}{TT_{\text{real},j,t}} \right| \times 100\% \quad (18)
\]

where \( TT_{\text{estimation},j,t} \) is the estimated segment travel time and \( TT_{\text{real},j,t} \) is the ground truth of segment \( j \) at time step \( t \); \( M \) is total number of time steps in the experiment period.

Segment travel time is estimated using the proposed method and evaluated by comparing with the ground truth data collected by AVI. The overall errors (MAPE and RMSE) over all experimental segments are summarized in Table 3. The estimated results show a good match with the ground truth data during the whole experimental period, the MAPE are less than 20% and the RMSE is around 0.15min. Given the complexity of urban road networks, the estimation result closely matched the ground truth traffic status. In general, estimations in Cluster I has higher accuracy than the other two clusters, this is possibly due to their better spatial relevancy, as demonstrated in the relevancy test (Fig 6).

Meanwhile, according to the results, shorter segment length may lead to a larger error, such as S18, S21 and S22 with segment lengths less than 250 meters. The MAPE over these segments is relatively higher. For shorter segments,
their travel time is more easily affected by the intersections on both sides and it varies significantly with the control signal. However, shorter segments require less travel time, thus a small RMSE can be observed on the results. Among them, S7 and S14 with MAPE over 18% are segments near the staggered intersections. The traffic states in these segments were heavily disturbed by the staggered intersections and changes dramatically during the rush hour. Such unstable traffic status leads to larger MAPE in the estimated travel time.

Fig 9 provides the comparison between the estimated travel time and the ground truth. In general, the estimation results of the proposed method show advantages to capture the daily travel time variability. As shown in the figure, the morning rush hour on S2 and S3 is not significant, but the model successfully caught the congestion happening rapidly in evening rush hour. The estimated travel time is close to the ground truth for Residential and Commercial Districts (e.g. S7, S9 and S19), whereas larger fluctuations can be observed compared to the ground truth data during congestion hour. Such area has more significant peak period, but the traffic status changes more substantially. The estimated values slightly underestimated during rush hour in S9, however, a sharp increasing travel time during early morning was well estimated by the method. For some segments (e.g. S7 and S19), the travel time slightly overestimated the free flow speed before 6 am. The expressway
FIGURE 9: Estimated travel time for six segments in experimental area compared with AVI travel time.

Segment S13 displays more stable travel time along the day. There is no significant peak hours captured as the traffic states are more stable on the expressway. The estimated travel time has similar distribution with ground truth, but with larger fluctuations.

Since the estimated travel time on each segment varies with its segment length, it is unreasonable to compare the traffic states among different segments utilizing segment travel time directly. Space mean speed ($V_{space}$) was defined as total travelled distance divided by their estimated travel time.
For demonstrating the spatial-temporal evolution characteristics among segments, $V_{\text{space}}$ would be a more suitable option. A smaller $V_{\text{space}}$ means that a longer travel time is required to pass through the segment and the traffic state tends to be more congested. Therefore, the comparison of $V_{\text{space}}$ among 13 segments (belonging to Cluster II, westbound and northbound) during the whole day is illustrated in Fig 10. Different traffic congestion levels of morning (9 am - 13 pm) and evening rush hour (17 pm - 20 pm) can be observed among all these segments. Besides, $V_{\text{space}}$ before 6 am are more in line with the free flow speed and close to the speed limit. According to the results in S10-S9-S7-S6, the congestion in evening rush hour gradually formed and spread from the downtown to west as the time elapsed. This coincident with the field data observations that traffic congestions usually propagates towards the direction of suburban area at the evening rush hour (westbound). For S7 and S18, it is observed that the traffic states are congested for most of the day. As mentioned in the previous section, these two segments are relatively short and thus, significantly influenced by the signal controls delays of the intersection at both sides of the segment.

IV. CONCLUSION AND FUTURE WORK

Travel time estimation based on the data collected by GPS equipped probe vehicle is one of the significant applications of interval detectors. An effective machine learning enabled travel time estimation method was proposed to utilize the sparse GPS probe vehicle data. Considering the different scenarios of handling the GPS trajectories, the procedure of the proposed estimation method includes the travel time scaling and allocating. After data normalization and network clustering, an ANN was conducted to infer the complex spatial-temporal relevancy of travel time on the clustered segments, which is a vital factor for travel time allocation. Finally, the paper compares the estimated travel time based on GPV to corresponding ground truth from AVI system. The desirable matching results verify the effectiveness of the proposed ANN enabled travel time estimation method.

For future study, the author world suggests to investigate whether the modeling accuracy can be further improved by breaking the ANN into time-specific networks, e.g. having respective rush hour and off-peak models for each cluster.
Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) models may be selected to evaluate them on solving the travel time estimation problems and compared with the benchmark performance of ANN algorithm proposed in this article. Further, it is also recommended to study the data and considers other traffic parameters in travel time estimation, such as volumes, occupancy and so on. Finally, collecting data from multi-source (the GPV and the ground truth from AVI) can be used to estimate travel time.

REFERENCES


MENGYUN XU received the B.E. degree from the Department of Civil Engineering, Fuzhou University, China, where she is currently pursuing the M.Sc. degree in transportation engineering. Her research interests include intelligent transportation systems and traffic data processing.

KEHUA GUO received the Ph.D. degree in computer science and technology from the Nanjing University of Science and Technology. He is currently a Professor with the School of Computer Science and Engineering, Central South University. His research interests include social computing, ubiquitous computing, big data, and image retrieval.

JIE FANG received the M.S. and Ph.D. degrees from the Civil and Environmental Engineering Department, University of Wisconsin–Madison, USA. He is currently a Professor with the Department of Civil Engineering, Fuzhou University, and the Director of their Transportation Program. His research interests include intelligent transportation systems, proactive traffic control, and the utilizing of artificial intelligence algorithms in the transportation area.

ZESHAN CHEN received the B.E. degree from the Department of Civil Engineering, Fuzhou University, China, where he is currently pursuing the M.Sc. degree in transportation engineering. His research interests include intelligent transportation systems and traffic data processing.

***