Hindawi
Complexity
Volume 2020, Article ID 6939328, 14 pages
https://doi.org/10.1155/2020/6939328

Research Article

Integrating Semantic Zoning Information with the Prediction of Road Link Speed Based on Taxi GPS Data

He Bing,[1] **Xu Zhifeng**,[2] **Xu Yangjie**,[1] **Hu Jinxing**,[1] **and Ma Zhanwu**[3]

[1]*Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China*
[2]*Gannan Normal University, School of Geography and Environmental Engineering, Ganzhou 341000, China*
[3]*University of Science and Technology Liaoning, School of Civil Engineering, Anshan 114051, China*

Correspondence should be addressed to Xu Zhifeng; 664838163@qq.com

Received 26 May 2020; Revised 14 October 2020; Accepted 23 October 2020; Published 7 November 2020

Academic Editor: Kailong Liu

Road link speed is one of the important indicators for traffic states. In order to incorporate the spatiotemporal dynamics and correlation characteristics of road links into speed prediction, this paper proposes a method based on LDA and GCN. First, we construct a trajectory dataset from map-matched GPS location data of taxis. Then, we use the LDA algorithm to extract the semantic function vectors of urban zones and quantify the spatial dynamic characteristics of road links based on taxi trajectories. Finally, we add semantic function vectors to the dataset and train a graph convolutional network to learn the spatial and temporal dependencies of road links. The learned model is used to predict the future speed of road links. The proposed method is compared with six baseline models on the same dataset generated by GPS equipped on taxis in Shenzhen, China, and the results show that our method has better prediction performance when semantic zoning information is added. Both composite and single-valued semantic zoning information can improve the performance of graph convolutional networks by 6.46% and 8.35%, respectively, while the baseline machine learning models work only for single-valued semantic zoning information on the experimental dataset.

1. Introduction

With the increasing number of vehicles, traffic congestion in cities is getting more and more serious. Obtaining real-time and future road states is essential for optimizing driver routes, reducing road congestion and developing sustainable urban transport policies [1, 2]. Road states are usually measured by traffic indexes such as volume, speed, and occupancy [2, 3]. With the support of communication and computing technologies, these indexes can be calculated from monitoring data obtained from sensors placed in the road network. Particularly, taxis with location-positioning capabilities are considered to be flexible probes that can obtain real-time, continuous information on vehicle movements, trip origins and destinations, routes, and passenger status. Studies have also shown that the analysis and prediction of urban road states using location data and artificial intelligence methods can be effective in relieving road traffic stress [4–6].

The layout of urban functional zones is the root cause for the generation of traffic demand, the uneven distribution of traffic flow, and the dynamic characteristics of road network. Traditional methods of urban functional zoning use land-uses, satellite images, and questionnaire surveys to statically delineate urban functional areas by clustering or the establishment of indicator systems. However, the static functional delineation cannot reflect the travel patterns exhibited by human activities [7, 8] and their impact on the formation of regional functions. In recent years, trajectory data such as taxi and bus location data have been gradually applied to the classification of land use and the identification of functional zones in road networks to help city managers better understand the relationship between urban functional zones and travelers' activities. On the other hand, the traffic flow between road segments is spatially correlated [9]. Not only do the traffic flows between upstream and downstream road segments influence each other, but there is also a traffic transfer relationship between multiple road intersections.

Studies have shown that the traffic state at one intersection is directly related to the other 100 intersections [10]. However, the use of trajectory data to identify urban functional zones and predict road link speeds is independent of each other. The information on the functional structure of road network is for planning purposes only and is not integrated with speed predictions for road links. In addition, correlations between road segments and intersections were tested only at small spatial scales. Therefore, it would be valuable to include these two factors in road speed prediction.

As socioeconomic activities develop and change, an urban area usually contains multiple functions simultaneously. This in turn affects the temporal and spatial characteristics of each road. But in the traditional method of road network subdivision, each road belongs to only one functional zone [11, 12]. The results of such singular delineation cannot reflect the dynamic nature of the network zones and have poor relevance to human activities. Recently, some studies [13–16] have proposed multifunctional quantification methods based on textual data mining, such as the Latent Dirichlet Allocation (LDA) model [17]. The results of the multifunctional quantitative calculations are expressed as vectors, which are then clustered to form the final functional zoning of road network [18, 19]. Furthermore, urban transportation networks are typically complex networks characterized by small-world, community structures. Current machine learning algorithms for complex networks have become a research hotspot. Several graph representation methods [20, 21] and graph neural networks (GNNs) [22] have been introduced in complex network modeling. In the past years, data-driven machine learning methods were commonly used to predict the state of the road network, such as support vector machines [23] and neural networks [24]. Recently, there are published literatures based on GNNs in the field of road state prediction, such as Multirange Attentive Bicomponent GCN (MRA-BGCN) [25], Multiweighted Traffic Graph Convolutional Network (MW-TGC) [26], DDP-GCN [27], and T-GCN [28]. In these methods, only primitive information in taxi trajectories and structural information in road network were used, such as speed, travel time, traveled distance, speed limit, and flow direction. It was expected that through the analysis of a large amount of data, efficient models could be learned autonomously. However, less research has been devoted to integrating the spatial semantic zoning information with predictive models and examining the validity of zoning information on the intended models.

For the reasons mentioned above, this paper aims to integrate the semantic zoning information with graph convolutional network for road link speed prediction. Firstly, LDA algorithm was used to obtain the stable semantic zoning information of taxi travel network over a certain period of time, and then the extracted functional vectors were added to the training process of spatial-temporal graph convolutional network and baseline machine learning models. Finally, we compared and analyzed the performance of functional vectors in each model. The main contributions of this paper are as follows: (1) due to the difficulty in reflecting the temporal and spatial dynamic dependencies of each road link by assigning it to a single zone, we obtained the composite functional vectors of each road link using semantic zoning based on the text of the taxi trajectories. (2) We proposed the use of spatiotemporal graph convolutional network to fuse information on semantic zoning, historical speed, and network structure. (3) With the proposed method, the large-scale spatial correlation of road links was integrated into the predictive model using LDA-generated semantic vectors; the local spatial correlation and the temporal dependencies of road links were learned by spatial–temporal graph convolutional network.

This paper is organized as follows. We briefly describe the works related to road speed prediction in Section 2. In Section 3, the detailed steps of the proposed method are explained, and theoretical basis of each algorithm used is introduced. The results of comparison experiments with six baseline algorithms are presented in Section 4 and discussed in Section 5. The conclusion and future work are reported in Section 6.

2. Literature Review

The functional layout and structure of the city is the root cause of the generation of traffic demand and the imbalanced distribution of traffic flows in the taxi trip network. This unevenness in traffic flow is often reflected in the spatial and temporal differences in location. For example, traffic demand in commercial areas is high and road congestion is frequent. For electric vehicles, longer charging times tend to cause congestion near charging stations [29]. In contrast, the road network in cultural district tends to experience traffic peaks during commuting and school hours. Identifying zoning and cluster characteristics of taxi travel networks has been the focus of research in urban planning, transportation network planning, and spatiotemporal trajectory mining for taxi operations and management. There are three types of methods to characterize the zoning or clustering of taxi travel networks. The first method is to detect hotspots and identify clustering patterns using a clustering algorithm based on taxi location points. The second approach is to divide the urban space into regular grids of a certain size [8, 30] or traffic zones [31] and then perform density analysis and clustering pattern discovery in the grids or zones. The last one is the semantic analysis method, which extracts trajectories from taxi location points according to spatial and temporal order and then combines them with textual information such as point of interests (POIs) data [32] and street names to identify the semantic functional areas of the travel network [13–16]. Compared to the previous two approaches, the semantic-based approach makes the functional zoning of the travel network more interpretable. In semantic analysis methods, LDA [17] is a widely used method that first appeared in the field of natural language processing (NLP) for semantic topic recognition. The current researches have been extended to the field of trajectory data mining with good results [14, 16]. When analyzing taxi trajectory data using LDA algorithm, the “word-document-topic” relationship in text mining is referenced to extract the road network semantic zones based on “road-trajectory-topic zone.”

Traffic state estimation refers to the analysis of typical quantities, such as traffic speed, travel time, flow volume, and density. Traditional traffic prediction algorithms include support vector machine (SVM) [23], support vector regression (SVR) [2], ARIMA, and neural networks [24]. The early speed collecting approaches primarily adopt loop sensors, radar, cameras, and other sensors, which are mostly used in road traffic monitoring and autonomous vehicle state detection [33]. Compared to the above sensors, the GPS equipped in taxis has a wider coverage and is more useful for recording the driving speed of each road segment. For example, Shan [3] proposed a multivariate linear regression model based on taxi location data to calculate the travel speed of each road segment by fusing information from the previous interval time and adjacent road segments. Oshyani [34] used an estimator based on indirect inference to predict traffic speed. Shan [35] tested three widely used GPS-based traffic speed estimation methods. Deng [36] introduced a path inference process for congested link speeds from low sampling frequency taxi GPS data. Satrinia [2] predicted the traffic speed using support vector regression. Yao [23] proposed a support vector machine model with spatial–temporal parameters for short-term traffic speed prediction, including multitime-step traffic prediction of several road links, and compared the proposed model with ANN, k-NN, historical data-based model, and moving average data-based model. The abovementioned methods in big data generally lack longevity and scalability due to insufficient robustness of the underlying theory. Despite the good performance of the SVM and ANN, they can only provide deterministic point prediction and failed to provide the corresponding uncertainty quantification. In recent years, there have been some state-of-the-art prediction methods that can measure uncertainty in transportation field [37–39], but these methods have not yet been used for road state prediction.

In recent years, research trend in the field of traffic prediction has been towards deep learning and combinatorial models. For example, Ma [40] proposed a convolutional neural network (CNN) to learn traffic from images and predict large-scale, network-wide traffic speed. Liu [41] introduced an attention CNN to predict traffic speed. Kim [42] employed the capsule network on loop sensor data for traffic speed prediction. As an emerging framework, GNNs have been widely promoted and extended in traffic prediction. Zhao [28] presented a temporal graph convolutional network (T-GCN) that combines GCN with gated recursive unit (GRU) for traffic prediction. Guo [43] proposed attention-based spatial-temporal GCN for traffic flow forecasting. Lu [44] designed a graph Long Short-Term Memory (LSTM) framework to capture spatial-temporal representations in road speed prediction. GCNs can be seen as a special case of GNNs [22], whose spectral domain approach aims to introduce the convolutional theory of signal analysis into irregular graphs to extract spatial features similar to those of CNNs on images. GNNs can also implement graphical feature extraction through message passing. However, it is prone to smoothing problems [22]. Combining the spatial zoning characteristics of taxi travel networks at different scales with road speed prediction is beneficial to the optimization and integration of model design. Following this idea, Huang [45] used spectral clustering to classify the traffic conditions into several clusters and implemented predictions for the clusters with less variability of traffic conditions within each cluster. In order to be able to add semantic information to the predictive model and to learn the spatiotemporal dependence between road segment links, we integrated two state-of-the-art algorithms. We used LDA algorithm for semantic zones detection for road links at large scale and adopted a GCN algorithm for speed prediction that considers spatial–temporal dependencies at local scale.

3. Data and Methods

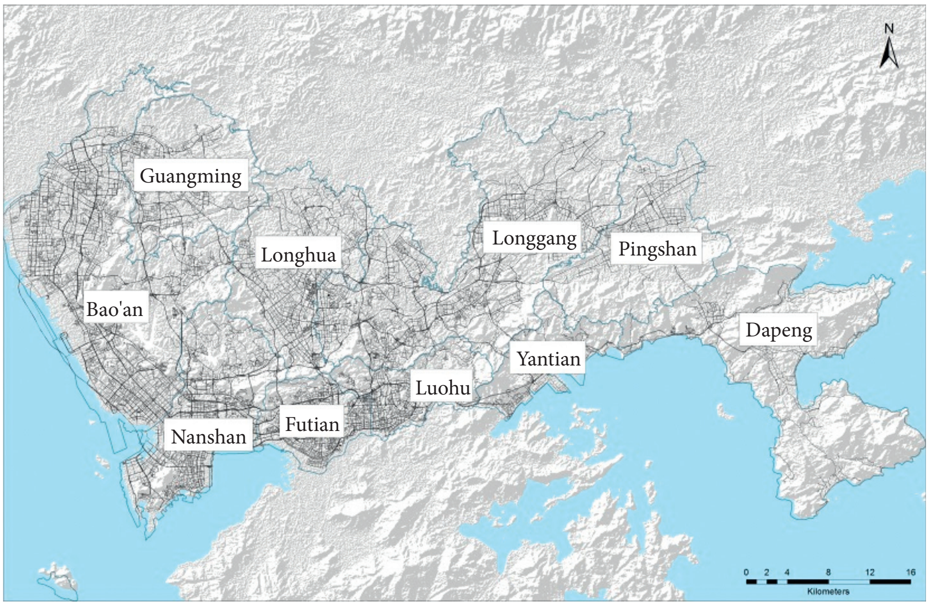
3.1. Data. The data used in this paper are taxi trajectories and road network. Taxi data was collected in Shenzhen, China, from May 1 to May 15, 2015. The raw taxi data was sampled at intervals of about 30 seconds. The road network data was downloaded from OpenStreetMap [46] and manually checked and edited. A sample of taxi data and the road network are shown in Figure 1. Since the taxi data contains some useless information and some errors, we first removed the useless fields from the daily data and saved the following fields: taxi ID, latitude and longitude, timestamp, speed, and operator status. After removing the outliers, the final dataset contains a total of 16,828 taxi trajectories.

3.2. Methods. From the existing literature [14], the temporal and spatial dynamic characteristics of road links can be quantified by trajectory semantic mining methods and represented in the form of functional vectors. Additionally, the characteristics of road links are correlated to their historical states and are influenced by the surrounding road links. In order to incorporate the spatiotemporal dynamic and correlation characteristics of road links into speed prediction, this paper proposes a speed prediction approach for road links based on the integration of LDA and GCN and validates the feasibility of this approach and the effectiveness of the functional vectors of road links through comparative experiments. The detailed flowchart is illustrated in Figure 2.

The proposed approach has five key steps (shown in Figure 2), which are trajectories extraction, map matching, semantic zones detection, semantic zones merging, road link speed prediction, and comparison experiments. The first three steps are used to prepare the experimental data and to discover the semantic zones in taxi travel network. The extracted results are feature vectors of the semantic zones. In order to verify the effectiveness of zoning information in road link speed prediction, we merged the composite semantic zones using modularity [47] in the fourth step and generated two types of features: single-valued zoning features and composite zoning features. In the fifth step, six baseline models were trained using the two types of features and the historical average road link speed, respectively.

Taxi ID	Longitude	Latitude	Time	Speed	Direction	Status
xxxxxx	114.0313	22.67733	2015-05-01 00:05:29	22	7	0
xxxxxx	114.0313	22.67733	2015-05-01 00:08:29	22	217	1
xxxxxx	114.0231	22.68513	2015-05-01 00:12:46	64	283	1
xxxxxx	113.9988	22.69002	2015-05-01 00:14:38	77	291	1
xxxxxx	113.9681	22.6866	2015-05-01 00:18:12	68	256	1
xxxxxx	113.9642	22.685617	2015-05-01 00:18:30	78	254	1

(a)



(b)

Figure 1: The sample data and road network in Shenzhen. (a) Sample location data. (b) Road network in Shenzhen.

Finally, both types of features were added to the prediction process and their effectiveness was compared with six baseline models.

Step 1. Extracting trajectories from taxi location data.

The trajectories in taxi raw data are composed of a collection of discrete points. We extracted the trajectories based on carrier states at the time of taxi positioning. The location where the carrier state changes from 0 to 1 was defined as the origin, and the location where the carrier state changes from 1 to 0 was defined as the destination (see Figure 3). Continuous location points between origin and destination were used as a trajectory. To avoid the impact of searching for passengers on road link speed calculations, we ignored the locations of taxis without passengers.

Step 2. Map matching.

In this paper, the ST-Matching algorithm [48] was used to match all trajectories to the road network. This algorithm takes into account the spatial geometry and topology of the road network as well as the time/velocity constraints of the trajectories. It can handle low-sample-rate localization data within 3 to 5 minutes with excellent matching accuracy and is suitable for the low-frequency data in this paper. The GPS geographic coordinates are converted to planar coordinates, and OpenStreetMap and PostGIS [49] were used to extract the source and target nodes of the road network during the matching process to search for the shortest path. Figure 4(a) shows the prematching trajectory points and Figure 4(b) shows the postmatching trajectory points. It can be seen that all the trajectory points have been correctly aligned to the road network.

Step 3. Semantic zoning based on LDA algorithm.

In order to obtain composite features for semantic zones, LDA algorithm was adopted in this paper. LDA is a semantic topic model proposed by Blei [17] and enables modeling of intertextual semantic topics based on text corpus. In the results obtained by LDA, a topic contains the probability distribution of each word. For each document, it can have multiple topics. When describing the distribution of document topics, a document can be represented as a composite vector of topics, or a topic with the highest probability is used as the topic of a document. For taxi trajectory topic discovery, the former can be used for subsequent machine learning tasks, and the latter can be used to visualize zoning results. The LDA algorithm is defined as follows.

LDA (as shown in Figure 5) assumes that the a priori distribution of the document topic and word is Dirichlet distribution; then for any document d and any topic k, LDA has the following definitions:

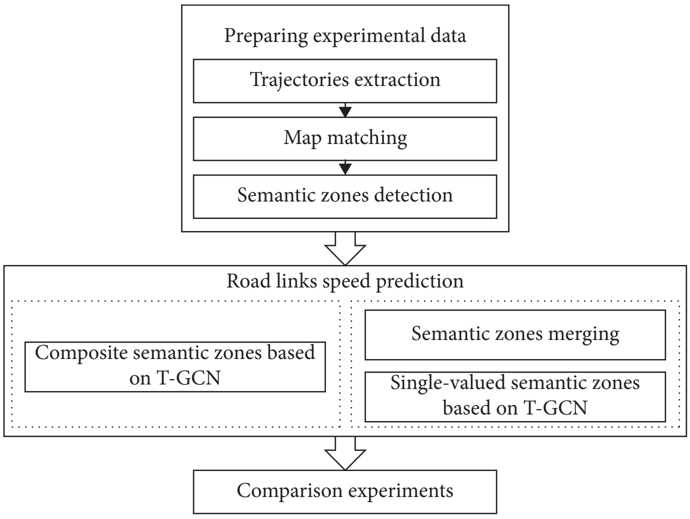

Preparing experimental data
Trajectories extraction
Map matching
Semantic zones detection
Road links speed prediction
Composite semantic zones based on T-GCN
Semantic zones merging
Single-valued semantic zones based on T-GCN
Comparison experiments

Figure 2: The flowchart of proposed approach.

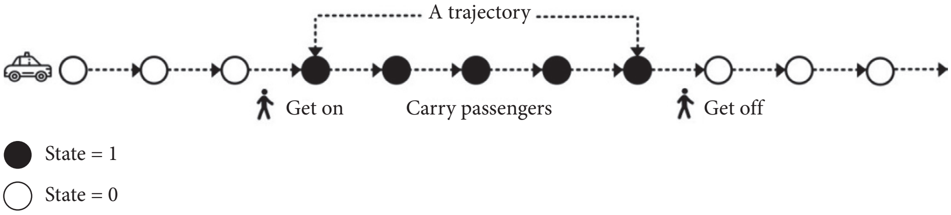


Figure 3: Taxi trajectories extraction.

$$\begin{aligned} \theta_d &= \text{Dirichlet}(\vec{\alpha}),\\ \beta_k &= \text{Dirichlet}(\vec{\eta}),\\ Z_{d,n} &= \text{Multinomial}(\theta_d),\\ W_{d,n} &= \text{Multinomial}\left(\beta_{Z_{d_n}}\right), \end{aligned} \tag{1}$$

θ_d is the probability distribution of each implied topic in the dth document. α is hyperparameter of distribution and a K-dimensional vector. β_k is the probability distribution of the kth topic feature word. η is hyperparameter of distribution and is a V-dimensional vector, and V represents the number of all words in the lexicon. $Z_{d,n}$ is the probability distribution for the nth topic in document d. $W_{d,n}$ is the probability distribution for the nth word in the dth document.

The LDA model is generated as follows:

(1) For each topic k, calculation *Perplexity*, and *JS* Divergence, choose the best number (K) of topic.

(2) For each topic $k \in [1, K]$, draw $\beta_k = \text{Dirichlet}(\vec{\eta})$.

(3) For each document $d \in [1, D]$,draw $\theta_d = \text{Dirichlet}(\vec{\alpha})$.

(4) Then for each word n in document d, draw the topic of the nth word:$Z_{d,n} = \text{Multinomial}(\theta_d)$ and the nth word: $W_{d,n} = \text{Multinomial}(\beta_{Z_{d_n}})$.

When training the LDA model, a common evaluation metric is confusion (as shown in equation (2)) [17]. Smaller perplexity means that the model is a better predictor for new text. Also in this paper, the Jensen–Shannon Divergence (JS Divergence) [50], a method for calculating topics similarity, is used together with perplexity to determine the optimal K.

Perplexity is defined below:

$$\text{perplexity}(T) = \exp\left(-\frac{\sum_{m-1}^{N} \log p(W_m)}{\sum_{m-1}^{N} D_m}\right), \tag{2}$$

where $\sum_{m=1}^{N} D_m$ is the sum of all words in test dataset. T is the test dataset with N documents, D_m represents the number of words in document m, and W_m is the words in document m.

Jensen–Shannon Divergence is defined as follows:

$$\text{TraJS}(\text{topic}) = \frac{\sum_{m-1}^{V} \left(D_{\text{JS}}(\text{topic}_m, \overline{\gamma})\right)^2}{V}, \tag{3}$$

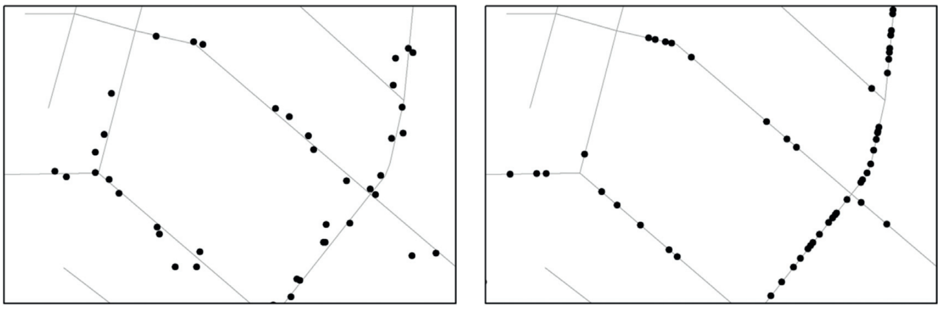
where V is the number of topics, D_{JS} is the JS Divergence of topics, $\text{TraJS}(\text{topic})$ is the variance of topics, topic_m is the mth topic, and $\overline{\gamma}$ is the mean of probability distribution (γ) of topic-word.

The final optimal number of topics is determined by perplexity_TraJS, which is calculated as follows:

$$\text{perplexity\_TraJS} = \frac{\text{perplexi}(D_{\text{test}})}{\text{TraJS}(T_{\text{test}})}, \tag{4}$$

where D_{test} is test dataset.

On the basis of the above definition, we created a topic model for taxi trajectories based on LDA algorithm. When



(a)

(b)

Figure 4: Map matching results. (a) Prematch trajectory points. (b) Postmatch trajectory.

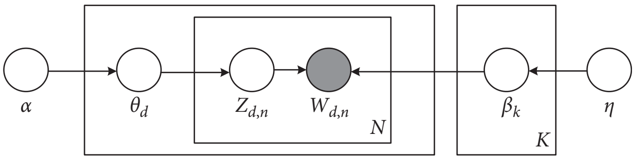


Figure 5: Graphical representation of LDA [17].

building the model, we treated a trajectory as a "document" and each road link number in a trajectory as a "word." All trajectories constitute a word corpus. Then, we used equation (4) to select the optimal number of topics and used the LDA model to extract the topics in trajectories. The obtained topics consist of several road links. Also, each link belongs to multiple traffic topics, which is similar to a document that can contain multiple topics.

Step 4. Merging semantic zones.

After specifying the number of topics, the probability distribution of each road link in each semantic zone is generated by LDA algorithm. The topic for each road link is a vector of probability distribution. This means that each road link will belong to multiple semantic zones, i.e., composite zones. Visualization methods in existing studies will select the topic with the highest probability as the final semantic zone to which a road link belongs. In other words, composite zones will be converted to single-valued zones. However, the resulting zones will be very fragmented. A semantic zone may consist of many small fragments that are scattered in taxi travel network. Therefore, in order to reduce the dispersion at a specific number of zones and to perform comparative experiments between feature vectors of composite and single-valued semantic zones, we used modularity [47] to merge the small fragments and maximize the modularity of road network subdivision. An idealized community division has the highest similarity between nodes within a community and the lowest similarity of nodes between communities. Modularity is commonly used to measure the merits of community subdivision results of complex networks. The higher the quality of the community division, the greater the modularity Q.

The modularity is calculated as follows:

$$Q = \frac{1}{2m}\sum_{vw}\left[A_{vw} - \frac{k_v k_w}{2m}\right]\delta\left(c_v, c_w\right). \tag{5}$$

By using equation (5), we can convert the composite zones obtained by LDA into single-valued zones with larger modularity. In subsequent comparative experiments, we can simultaneously evaluate the effectiveness of composite and single-valued zones in the prediction of road link speed.

Step 5. Building predicting model for road links.

The state of each road link is influenced by the upstream and downstream links. Therefore, incorporating the complex structure and historical state of the road network into model will be beneficial to improve the accuracy of speed prediction for road links and also enables to predict multiple road links at once. GNNs are this kind of methods for learning on a non-Euclidean structure. GNNs introduce the convolution theory from the Euclidean data to the non-Euclidean data to solve the spatial dependencies. In order to address both spatial and temporal dependencies, Zhao [28] proposed the temporal graph convolutional network (T-GCN). The temporal dependencies are obtained by adding the GRU structure to GCN model. In order to compare and analyze the effectiveness of single-valued zoning and composite zoning while maintaining the network topology, we chose to incorporate the results of LDA into T-GCN in the proposed approach. T-GCN is defined as follows:

$$\begin{aligned} u_t &= \sigma\left(W_u\left[f\left(A, X_t\right), h_{t-1}\right] + b_u,\\ r_t &= \sigma\left(W_r\left[f\left(A, X_t\right), h_{t-1}\right] + b_r,\\ c_t &= \tanh\left(W_c\left[f\left(A, X_t\right), \left(r_t \times h_{t-1}\right)\right] + b_c,\\ h_t &= u_t * h_{t-1} + \left(1 - u_t\right) * c_t, \end{aligned} \tag{6}$$

where u_t and r_t are update gate and reset gate at time t. They are used to control the forgetfulness of state information from previous period. c_t is memory contents stored at time t. h_t is output state at time t and h_{t-1} indicates the output of time $t-1$. A represents the adjacency matrix of road network. X_t is feature matrix of each road link at time t.

The training process of T-GCN model is as follows (as shown in Figure 6). Firstly, we calculate the ground truth average speed of each road link for a certain period based on map-matched trajectories. Then, we build training and test

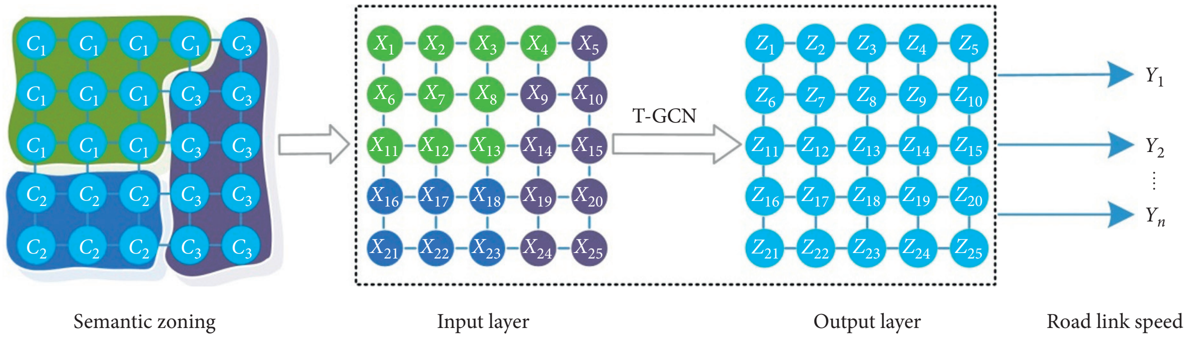

C_1 C_1 C_1 C_1 C_3
C_1 C_1 C_1 C_3 C_3
C_1 C_1 C_1 C_3 C_3
C_2 C_2 C_2 C_3 C_3
C_2 C_2 C_2 C_3 C_3
X_1 X_2 X_3 X_4 X_5
X_6 X_7 X_8 X_9 X_{10}
X_{11} X_{12} X_{13} X_{14} X_{15}
X_{16} X_{17} X_{18} X_{19} X_{20}
X_{21} X_{22} X_{23} X_{24} X_{25}
T-GCN
Z_1 Z_2 Z_3 Z_4 Z_5
Z_6 Z_7 Z_8 Z_9 Z_{10}
Z_{11} Z_{12} Z_{13} Z_{14} Z_{15}
Z_{16} Z_{17} Z_{18} Z_{19} Z_{20}
Z_{21} Z_{22} Z_{23} Z_{24} Z_{25}
Y_1
Y_2
Y_n
Semantic zoning
Input layer
Output layer
Road link speed

Figure 6: The training process of T-GCN.

dataset based on historical data of each road link and extract its semantic zoning information of taxi travel network. Next, we combine with adjacency matrix of road network for T-GCN training. Finally, the speed prediction for each road link is output and compared to the true value to optimize the model.

4. Experiments and Results

4.1. The Results of Semantic Zoning for Taxi Travel Network

4.1.1. Optimal Parameter Selection. LDA model has the following parameters that should be set up firstly: (1) Dirichlet distribution parameter α for trajectory-traffic topic; (2) Dirichlet distribution parameter β for traffic topic-road link; and (3) the number K of traffic topics.

α value affects topics distribution for each trajectory, and β value affects road link distribution for each traffic topic. The greater the two values, the more concentrated the distribution. In order to obtain optimal parameters, the values of $\alpha \in [0.001, 0.01, 0.05, 0.1, 0.25, 0.5]$ and $\beta \in [0.001, 0.01, 0.01, 0.05, 0.1, 0.25]$ were compared. We found that traffic topic model was better differentiated when $\alpha = 0.25$ and $\beta = 0.01$.

Another important parameter is the number of topics K. If K is too large, the topic division is very detailed, and the likelihood of similarity between topics will increase, while a lower value of K may be not able to distinguish topics well. Therefore, tests are needed to determine the number of themes K. In the previous section, we have determined the optimal value of α and β. With $\alpha = 0.25$ and $\beta = 0.01$, we conducted tests by taking the number of topics from 2 to 100 at an interval of 1. Perplexity (equation (2)), Jensen–Shannon Divergence (equation (3)), and the joint index of perplexity and Jensen–Shannon Divergence (equation (4)) were calculated. The results are shown in Figures 7(a)–7(c).

In Figure 7(a), the perplexity value of LDA decreases as the topic number K increases. After $K = 18$, the perplexity value begins to decrease slowly. In Figure 7(b), as the K-values increase, the value of Jensen–Shannon Divergence starts to increase slowly between 15 and 20 and gradually flattens out. In Figure 7(c), the trend is consistent (Figure 7(a)), which starts to slowly decline after $K = 18$. In order to effectively characterize the traffic flow clustering pattern of travel network topics in study area, $K = 18$ was chosen as the topic numbers for LDA algorithm in the experiment.

4.1.2. The Semantic Zones of Taxi Travel Network. After semantic zoning and semantic zones merging based on modularity, stable semantic zones of travel network was generated within 15 days using the constructed trajectories dataset, where the LDA was modeled using the Gensim [51] library. The result semantic zones were visualized using ArcGIS [52] software. We classified zoning information into two categories, namely, composite semantic zones generated by LDA algorithm and single-valued semantic zones merged from LDA results based on modularity. In the former, each road link contains probability belonging to 18 semantic zones. In the latter, each road link belongs only to the semantic zone with maximum probability.

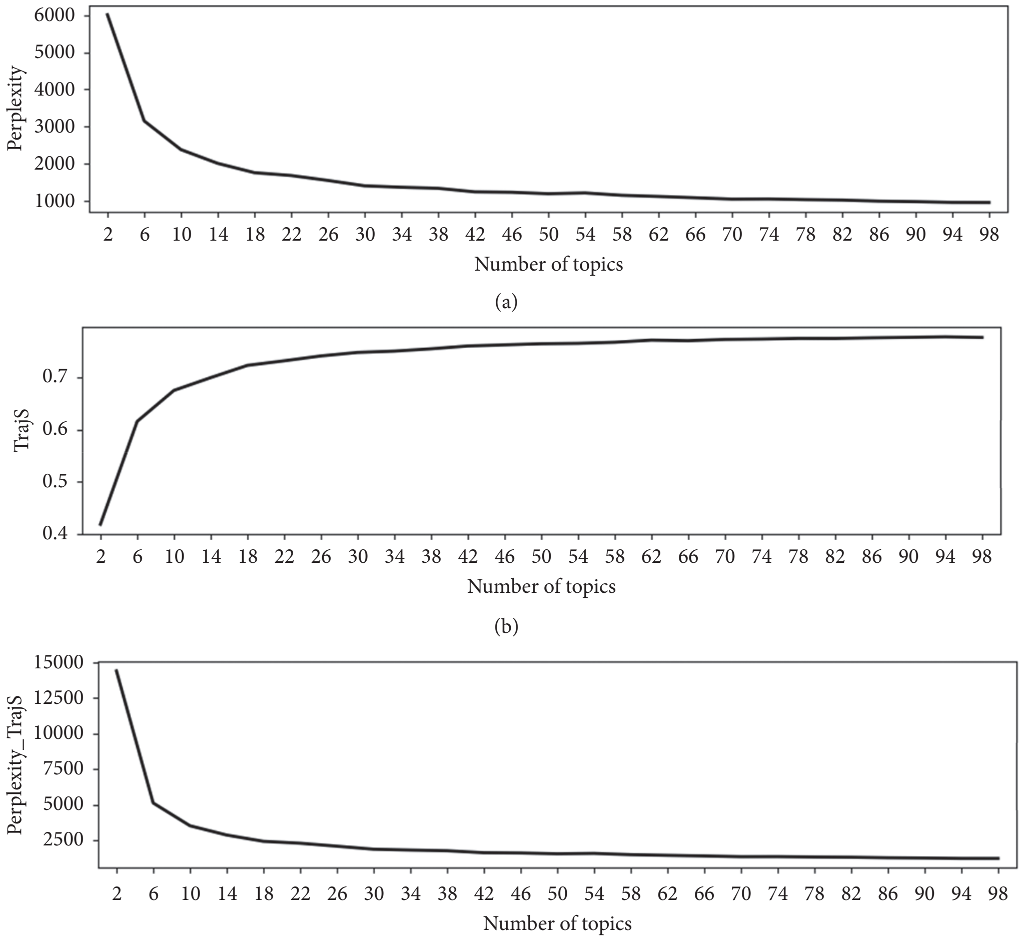
Figure 8 shows the map of composite semantic zones. We used different colors to represent zones. Line widths were set by the probability of a road link belonging to one of the 18 zones. As can be seen from the figure, the distribution of semantic zones is clear across map, but the number of semantic zones varies from district to district (as shown in Table 1). Nanshan, Futian, and Luohu districts have the highest number of topics, and only Longhua district matches the semantic zone nicely. Additionally, we found that traffic topic zones are correlated with land use and are prone to form semantic zones near train stations, airports, residential areas, and commercial areas. On arterial roads, it is also easy to form semantic zones, such as Beihuan Road and Binhai Road. Since the probability that a road belongs to each of the 18 zones is difficult to visualize, some zones have nested road links that belong to other zones (as shown in the upper right corner of Figure 8).

Figure 9 shows the map of single-valued semantic zones. We classified each road link to one semantic zone with maximum probability and used the same color style as Figure 8. As the map depicts, all semantic zones are rendered clearly and overlap disappears. The nested road links between different semantic zones are reduced.

4.2. The Comparison of Prediction Models

4.2.1. Data Preparation and Experimental Setup

(1) Data Preparation. There are 44,609 links in the Shenzhen road network, which has a complex network structure. We



(a)

(b)

(c)

Figure 7: The perplexity (a), Jensen–Shannon Divergence (b), and their joint indexes (c) under different topic numbers.

collected data starting and ending from May 1 to May 15, 2015. Due to the short period of data collection, some of the road links lack taxi track data. As shown in Figure 9, the small amount of data leads to many road links that are not explicitly in a topic. Therefore, in order to compare the effectiveness of single-valued semantic features and composite semantic features in the speed prediction of road links and to examine the learning ability of T-GCN on the spatiotemporal dependencies on road links, an experimental area was selected as an example in this paper. The experimental area contains multiple semantic zones, including composite zoning and single-valued zoning information, which are suitable for the research objectives of this paper. Also, the amount of data in the experimental area is sufficient to make the prediction model fit better. After data processing, the experiment area was selected from composite zoning map and single-valued zoning map extracted by LDA (as shown in the upper right of Figure 10). There were 766 road links in the research area. The data of road link speeds at 7:00–23:00 was selected. We adopted 15 minutes as the time interval, the previous four periods were selected as historical speed features, and the speed of the next one period was used as the prediction value. Also, in the comparison experiment, semantic zoning information will be added to the features. In the splitting of datasets, 80% of all data were used as training sets and 20% as test sets.

(2) Algorithm Selection and Parameter Setting. In this study, Vector Machine Regression (SVR), Random Forest Regression (RFR), Gradient Boost Regression (GBDT), XGBoost, Decision Tree Regression (DTR), and T-GCN were selected for comparative analysis on the impact of semantic zones in road link speed prediction. The first five baseline algorithms are from the Scikit-learn package which is a machine learning library for Python. We used GridSearchCV in Scikit-learn to automatically find the optimal parameters for these five algorithms. The parameters of these five machine learning algorithms are shown in Table 2. The hyperparameters of T-GCN model mainly include batch size, learning rate, training epochs, and number of hidden units. Based on the experience of Zhao [28] and after repeated tests, learn rate was set to 0.01, and catch size was set to 32, and the number of hidden units was set to [8, 16, 32, 64, 100]. We found that T-GCN model has the highest prediction accuracy when the number of hidden layers is 64. In this experiment, Adam optimizer was used for loss

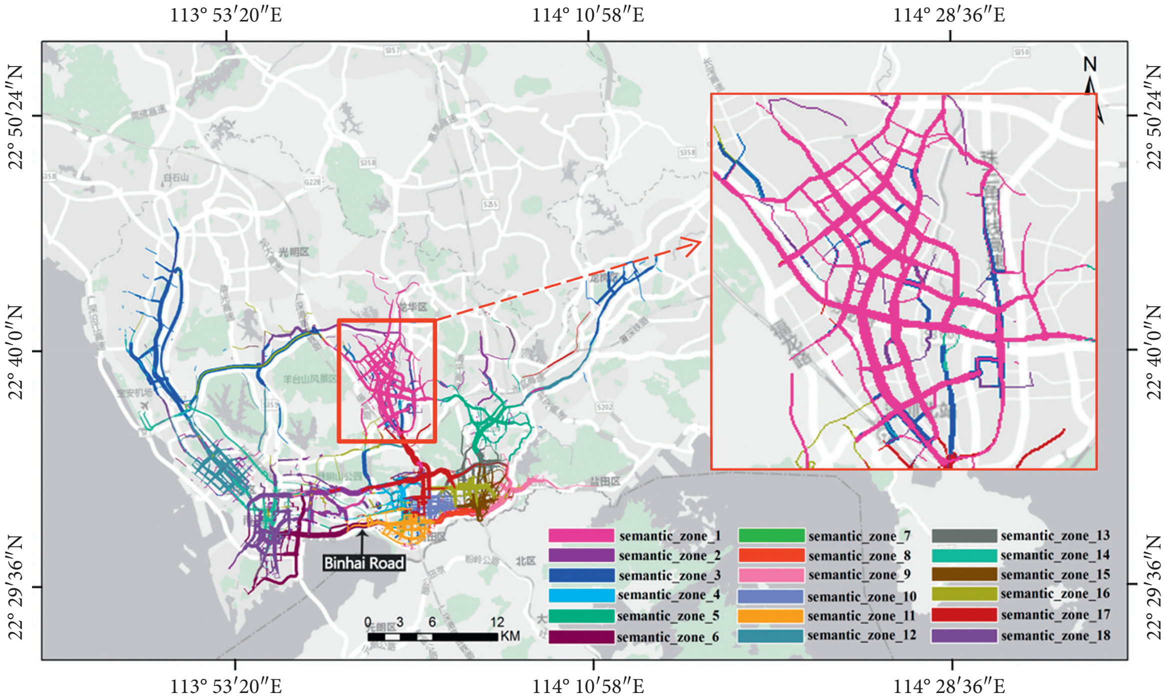


Figure 8: Composite semantic zones of taxi travel network in Shenzhen.

Table 1: Number of semantic zones in each district.

Districts	Number of topic zones
Nanshan	5
Luohu	4
Futian	8
Bao'an	4
Longgang	5
Longhua	3
Yantian	1

calculation and the training loss remained stable when the number of iteration epochs was 100.

(3) Evaluation Metrics. In order to accurately evaluate the performance of prediction models in road link speeds, the following metrics were used in this study. These metrics include Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Coefficient of Determination (R^2), and Accuracy (Acc). RMSE and MAE were used to calculate the error of the model; the smaller the value, the better the model. R^2 was used to test the model's predictive ability on the test dataset; the larger the value, the better the model.

(4) Results. The results were compared and analyzed in three scenarios. The first was only use of raw data to train models and evaluation results; the second was modeled under single-valued zones; and the third was done under composite zones. The accuracy of each model is shown in Table 3.

From Table 3, we can see that the RMSE error of T-GCN decreases gradually with the addition of semantic information (from single-valued zoning to composite zoning), while the RMSE error of the other five machine learning models increases when composite zoning is used instead. For example, the RMSE error of the T-GCN model is about 52.03% smaller than that of SVR model when composite zoning is used. The trend for R^2 is the same as for RMSE error. The ability of the predicted results of T-GCN to represent actual data increases gradually with the addition of semantic zoning information, while the other five models perform poorly under composite zoning. For example, the R^2 of T-GCN is improved about 8.94% compared to that of SVR with the addition of composite partition information. Based on the raw data, the machine learning algorithms were able to achieve average accuracy 83.68%, while the accuracy of T-GCN was only 73.63%. This may be due to the small amount of available data. T-GCN still needs more data to improve its ability to learn spatial dependencies. After adding semantic zoning information, the accuracy of machine learning algorithms was improved by an average of 8.76%. But it showed different performance in composite and single-valued zones, and the accuracy of each machine learning model becomes poor under composite zones. T-GCN showed improved performance with both composite and single-valued zoning information, with a 6.46% improvement in the single-valued zones over raw data only and a 1.88% improvement in the composite zones compared with nonoverlapping zoning information.

5. Discussion

(1) LDA gives the distribution probabilities of semantic zones for each road link. It is a common practice in visualization to select the topic with highest probability as the final semantic zones for road links. In this experiment, we extracted semantic zoning

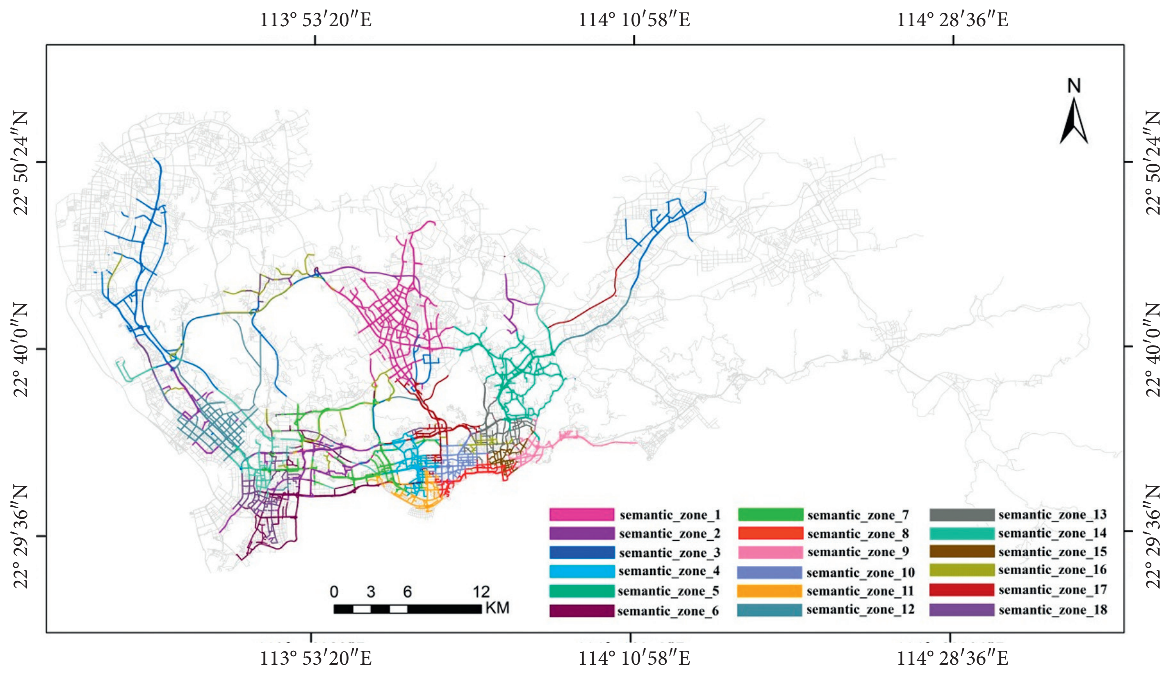


Figure 9: Single-valued topic zones of taxi travel network in Shenzhen.

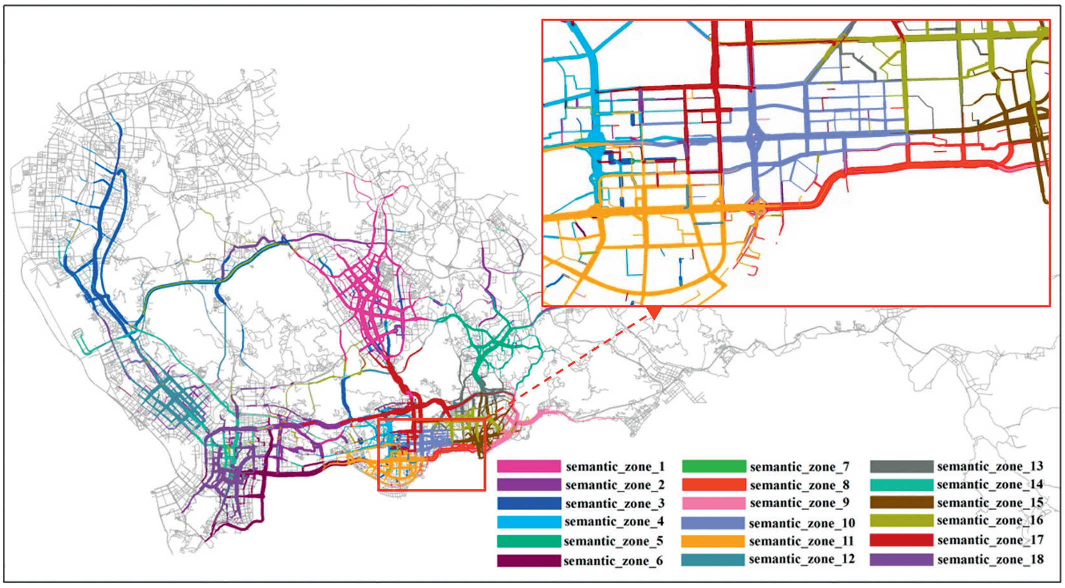


Figure 10: The experiment area.

information according to this idea and used modularity to merge the scattered, nested road links belonging to other zones. We can see that the resulting map becomes tidier and clearer (shown in Figure 9). However, from the predicting results, the accuracy with composite zoning information in T-GCN is improved as compared to the single-valued zones. In addition, LDA is a community discovery algorithm with semantic information, which is different from the traditional division methods targeting on topological relationships for obtaining single-valued communities such as GN [53], FN [54], and FUA [55]. Traffic flow of road link is influenced by dynamic changes at previous times and surrounding road links. The road links will fall in different semantic zones at different times. The probability that one road link belongs to more than one semantic zone in LDA depicts the spatial–temporal ambiguity. Therefore, the composite zones for road links are closer to the actual spatial and temporal characteristics of the transportation network. Ding [56] found that communities identified using topic analysis are more interrelated than communities detected by topological methods, so it

Table 2: Parameter settings.

Model	Parameter				
XGBoost	learning_rate 0.01	max_depth 12	n_estimators 100	Subsample 0.58	Gamma 0.50
GBDT	learning_rate 0.1	max_depth 5	n_estimators 40	Subsample 0.5	min_samples_split 10
DTR	max_depth 5	min_samples_split 50		min_samples_leaf 6	
RFR	n_estimators 100	max_depth 5	min_samples_leaf 10	min_samples_split 50	
SVR	Kernel Linear	C 5		Gamma 0.001	

Table 3: Accuracy evaluation of each algorithm.

Model	Experiment	RMSE	MAE	R^2	ACC
XGBoost	Raw data	6.161418	4.229177	0.837552	0.841541
LDA + XGBoost	Single-valued zoning	4.541672	3.112215	0.883153	0.911551
	Composite zoning	10.154425	7.539924	0.737299	0.553254
GBDT	Raw data	6.440737	4.388805	0.822487	0.834358
LDA + GBDT	Single-valued zoning	4.412345	3.064314	0.885717	0.915466
	Composite zoning	10.375719	7.732406	0.731502	0.533065
DTR	Raw data	6.461293	4.509432	0.821353	0.833829
LDA + DTR	Single-valued zoning	4.244633	2.992957	0.890745	0.922742
	Composite zoning	10.301833	7.658928	0.733414	0.539691
RFR	Raw data	6.278449	4.281197	0.831324	0.838532
LDA + RFR	Single-valued zoning	4.112489	2.829821	0.894147	0.927477
	Composite zoning	10.250276	7.613335	0.734748	0.544287
SVR	Raw data	6.369968	4.186199	0.826567	0.836183
LDA + SVR	Single-valued zoning	3.583199	2.317352	0.907773	0.945007
	Composite zoning	10.856324	7.991883	0.719071	0.489392
T-GCN	Raw data	6.149417	7.8762835	0.7345389	0.736309
LDA + T-GCN (our method)	Single-valued zoning	5.964327	7.7329316	0.7435799	0.8009587
	Composite zoning	5.207459	6.962856	0.7897214	0.81981

would be useful to apply composite zones which are better at characterizing the taxi travel network structure to predict the average speed of road links.

(2) Data augmentation in traffic prediction can be classified into two types: one is to take external information, for example, adding information such as weather and holidays to prediction models. The other is to build features from road network topology and historical time series data or augment data using neural networks [57]. This paper adopted the second approach to augment data for traffic prediction. We used LDA to extract spatial–temporal semantic information from trajectory data and concatenate it with historical average road link speeds to solve the problem of single data source in prediction models. Experiments on predicting average road link speeds were performed on the proposed approach and six baseline models (RFR, DTR, SVR, GDBT, XGBoost, and T-GCN). The experimental results of each model showed that the performance improvement by data augmentation varied obviously. In the experiment, the improvement in prediction accuracy after adding semantic zoning information using T-GCN model is obviously better than other traditional machine learning methods. However, single-valued and composite semantic zoning information can have different effects when added to machine learning algorithms. Adding composite zoning information makes the machine learning algorithms worse. The reason is that each road link may not belong to all semantic zones, and the probability of some road links is zero, so a large number of zero values affect the model fitting. In addition, although the accuracy of machine learning algorithms is higher than T-GCN, but traditional machine learning algorithms such as SVM can only predict the results of one road each time, which is inefficient in practical applications, while T-GCN could predict the average speed of road links all at once. Moreover, as the research on GCNs is deepened, the combination of semantic information with spatial and temporal relationships learned automatically from GCN may improve the prediction performance. It should be noted that the objective of this paper is to verify the validity of

semantic zoning features in predicting the average speed of road links and that data such as weather is difficult to obtain and therefore is ignored.

6. Conclusion and Future Work

This paper proposed a method for predicting average road link speed that integrates the semantic zones of taxi travel network extracted by LDA and the spatial–temporal dependencies learned by T-GCN. Firstly, the taxi location data was preprocessed, and datasets for subsequent tasks were built up after anomaly data filtering, trajectory segmentation, and map matching. Next, we converted the trajectories to a sequence of road numbers and extracted the semantic zones using LDA algorithm. To test the validity of the semantic zoning features, we merged the composite semantic zones obtained from LDA to form single-valued zones using modularity in social network community detection. Finally, we compared the proposed approach with six baseline models. The main findings of this study are summarized below:

(1) Semantic zones of the taxi travel network do exist within a certain period of time. These zones can describe the spatiotemporal dynamic characteristics of the road network.

(2) LDA can be used to quantify the dynamic characteristics of the road network and integrate with the historical state of the network, which helps to improve the accuracy of speed prediction for road links.

(3) Compared with traditional machine learning model, the semantic zoning information has better performance in T-GCN model, which can learn the spatiotemporal dependencies of the travel network simultaneously and can integrate the semantic zones.

In future work, we would like to research on end-to-end algorithms referring to the techniques such as network representation learning and GCNs to reduce the complexity of road link average speed prediction.

Data Availability

As the data forms part of an ongoing study, the raw data needed to reproduce these findings cannot be shared at this time.

Conflicts of Interest

The authors declare that they have no conflicts of interest regarding the publication of this paper.

Acknowledgments

The authors gratefully acknowledge the support of this research from the National Natural Science Foundation of China (41701167) and the Basic Research Project of Shenzhen City (JCYJ20190812171419161).

References

[1] S. Wang, L. Li, W. Ma, and X. Chen, "Trajectory analysis for on-demand services: a survey focusing on spatial-temporal demand and supply patterns," *Transportation Research Part C: Emerging Technologies*, vol. 108, pp. 74–99, 2019.

[2] D. Satrinia and G. A. P. Saptawati, "Traffic speed prediction from GPS data of taxi trip using support vector regression," in *Proceedings of 2017 International Conference on Data and Software Engineering (ICoDSE)*, Palembang, Indonesia, November 2017.

[3] Z. Y. Shan, D. Zhao, and Y. Xia, "Urban road traffic speed estimation for missing probe vehicle data based on multiple linear regression model," in *Proceedings of the 16th International IEEE Annual Conference on Intelligent Transportation Systems (ITSC 2013)*, pp. 118–123, Hague, Netherlands, October 2013.

[4] Y. Y. Li, "Research on traffic condition forecast based on car network big data," *Information and Communications Technologies*, vol. 11, no. 6, pp. 74–78, 2017.

[5] X. Li, Q. Luo, and D. Y. Meng, "Traffic flow-big data forecasting method based on spatial-temporal weight correlation," *Acta Scientiarum Naturalium Universitatis Pekinensis*, vol. 53, no. 4, pp. 775–782, 2017.

[6] X. Li and D. Y. Meng, "Distributed incremental traffic flow big data forecasting method based on road network correlation," *Scientia Geographica Sinica*, vol. 37, no. 2, pp. 209–216, 2017.

[7] H. Cai, X. Zhan, J. Zhu, X. Jia, A. S. F. Chiu, and M. Xu, "Understanding taxi travel patterns," *Physica A: Statistical Mechanics and its Applications*, vol. 457, pp. 590–597, 2016.

[8] L. Xu, D. Xia, X. Zhao et al., "Spatial-temporal travel pattern mining using massive taxi trajectory data," *Physica A: Statistical Mechanics and its Applications*, vol. 501, pp. 24–41, 2018.

[9] W. Chen and L. Wynter, "Real-time road traffic prediction with spatio-temporal correlations," *Transportation Research Part C: Emerging Technologies*, vol. 19, no. 4, pp. 606–616, 2011.

[10] S. Yang, "On feature selection for traffic congestion prediction," *Transportation Research Part C: Emerging Technologies*, vol. 26, pp. 160–169, 2013.

[11] H. Chen and Y. Hu, "Finding community structure and evaluating Hub road section in urban traffic network," *Procedia-Social and Behavioral Sciences*, vol. 96, pp. 1494–1501, 2013.

[12] X. Liu, L. Gong, Y. Gong, and Y. Liu, "Revealing travel patterns and city structure with taxi trip data," *Journal of Transport Geography*, vol. 43, pp. 78–90, 2015.

[13] S. C. Jin, Y. B. Tao, Y. Y. Yan, J. Xu, and H. Lin, "Visual analytics of taxi trajectory data via topical sub-trajectories," in *Proceedings of 2019 IEEE Pacific Visualization Symposium (PacificVis)*, pp. 174–178, Bangkok, Thailand, 2019.

[14] K. Liu, S. Gao, and F. Lu, "Identifying spatial interaction patterns of vehicle movements on urban road networks by topic modelling," *Computers, Environment and Urban Systems*, vol. 74, pp. 50–61, 2019.

[15] J. Bao, P. Liu, X. Qin, and H. Zhou, "Understanding the effects of trip patterns on spatially aggregated crashes with large-scale taxi GPS data," *Accident Analysis & Prevention*, vol. 120, pp. 281–294, 2018.

[16] F. M. Zhang, X. Y. Zhu, W. Guo et al., "Analyzing urban human mobility patterns through a thematic model at a finer scale," *International Journal of Geo-Information*, vol. 5, no. 78, pp. 1–17, 2016.

[17] D. M. Blei, A. Y. Ng, M. I. Jordan, and J. Lafferty, “Latent dirichlet allocation,” *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[18] L. Yu, X. He, and J. Y. Liu, “Discovering urban functional regions based on sematic mining from spatiotemporal data,” *Journal of Sichuan University (Natural Science Edition)*, vol. 56, no. 2, pp. 246–252, 2019.

[19] S. L. Chen, H. Y. Tao, X. L. Liu, and L. Zhou, “Discovering urban functional regions using latent semantic information: spatiotemporal data mining of floating cars GPS data of Guangzhou,” *Acta Geographica Sinica*, vol. 71, no. 3, pp. 471–483, 2015.

[20] M. Kinderkhedia, “Learning representations of graph data-a survey,” 2019, http://arxiv.org/abs//1906.02989.

[21] P. Goyal and E. Ferrara, “Graph embedding techniques, applications, and performance: a survey,” *Knowledge Based Systems*, vol. 151, pp. 78–94, 2017.

[22] Z. H. Wu, S. Pan, F. Chen, and G. D. Long, “A comprehensive survey on graph neural networks,” *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–21, 2019.

[23] B. Yao, C. Chen, Q. Cao et al., “Short-term traffic speed prediction for an urban corridor,” *Computer-Aided Civil and Infrastructure Engineering*, vol. 32, no. 2, pp. 154–169, 2017.

[24] P. Yu, A. Martins, S. Kalakou, and F. Moura, “Spatiotemporal variation of taxi demand,” *Transportation Research Procedia*, vol. 47, pp. 664–671, 2020.

[25] W. Chen, L. Chen, Y. Xie, W. Cao et al., “Multi-range attentive bicomponent graph convolutional network for traffic forecasting,” 2019, http://arxiv.org/abs//1911.1209.

[26] Y. Y. Shin and Y. Yoon, “Incorporating dynamicity of transportation network with multi-weight traffic graph convolution for traffic forecasting,” 2019, http://arxiv.org/abs//1909.07105.

[27] K. Lee and W. Rhee, “DDP-GCN: multi-graph convolutional network for spatiotemporal traffic forecasting,” 2019, http://arxiv.org/abs//1905.12256.

[28] L. Zhao, Y. Song, C. Zhang, Y. Liu et al., “T-GCN: a temporal graph convolutional network for traffic prediction,” *IEEE Transactions on Intelligent Transportation Systems*, http://arxiv.org/abs//1811.05320, 2019.

[29] K. Liu, C. F. Zou, K. Li, and T. Wik, “Charging pattern optimization for lithium-ion batteries with an electrothermal-aging model,” *IEEE Transactions on Industrial Informatics*, vol. 14, no. 12, pp. 5463–5474, 2018.

[30] S. Zhang, J. Tang, H. Wang, Y. Wang, and S. An, “Revealing intra-urban travel patterns and service ranges from taxi trajectories,” *Journal of Transport Geography*, vol. 61, pp. 72–86, 2017.

[31] J. Tang, F. Liu, Y. Wang, and H. Wang, “Uncovering urban human mobility from large scale taxi GPS data,” *Physica A: Statistical Mechanics and its Applications*, vol. 438, no. 15, pp. 140–153, 2015.

[32] L. Tang, D. Cai, Z. Duan, J. Ma, M. Han, and H. Wang, “Discovering travel community for POI recommendation on location-based social networks,” *Complexity*, vol. 2019, Article ID 8503962, 8 pages, 2019.

[33] H. Han, X. Li, W. Wang, L. Han, and C. Xiang, “Markov velocity predictor and radial basis function neural network-based real-time energy management strategy for plug-in hybrid electric vehicles,” *Energy*, vol. 152, pp. 427–444, 2018.

[34] M. F. Oshyani, M. Sundberg, and A. Karlström, “Consistently estimating link speed using sparse gps data with measured errors,” *Procedia-Social and Behavioral Sciences*, vol. 111, pp. 1227–1236, 2014.

[35] Z. Y. Shan, Y. N. Wang, and Q. Q. Zhu, “Feasibility study of urban road traffic state estimation based on taxi GPS data,” in *Proceedings of IEEE 17th International Conference on Intelligent Transportation Systems (ITSC)*, pp. 2188–2193, Qingdao, China, October 2014.

[36] B. Deng, S. Denman, V. Zachariadis, and Y. Jin, “Estimating traffic delays and network speeds from low-frequency GPS taxis traces for urban transport modeling,” *European Journal of Transport and Infrastructure Research*, vol. 15, no. 4, pp. 639–661, 2015.

[37] K. Liu, Y. Shang, Q. Ouyang, and W. D. Widanage, “A data-driven approach with uncertainty quantification for predicting future capacities and remaining useful life of lithium-ion battery,” *IEEE Transactions on Industrial Electronics*, p. 1, 2020.

[38] K. Liu, Y. Li, X. Hu, M. Lucu, and W. D. Widanage, “Gaussian process regression with automatic relevance determination kernel for calendar aging prediction of lithium-ion batteries,” *IEEE Transactions on Industrial Informatics*, vol. 16, no. 6, pp. 3767–3777, 2020.

[39] K. Liu, X. Hu, Z. Wei, Y. Li, and Y. Jiang, “Modified Gaussian process regression models for cyclic capacity prediction of lithium-ion batteries,” *IEEE Transactions on Transportation Electrification*, vol. 5, no. 4, pp. 1225–1236, 2019.

[40] X. L. Ma, Z. Dai, Z. B. He, J. H. Ma et al., “Learning traffic as images: a deep convolutional neural network for large-scale transportation network speed prediction,” *Sensors*, vol. 17, no. 4, pp. 1–10, 2017.

[41] Q. Liu, B. Wang, and Y. Zhu, “Short-term traffic speed forecasting based on attention convolutional neural network for arterials,” *Computer-Aided Civil and Infrastructure Engineering*, vol. 33, no. 11, pp. 999–1016, 2018.

[42] Y. Kim, P. Wang, Y. Zhu, and L. Mihaylova, “A capsule network for traffic speed prediction in complex road networks,” in *Proceedings of 2018 Sensor Data Fusion: Trends, Solutions, Applications*, Bonn, Germany, October 2018.

[43] S. Guo, Y. Lin, N. Feng, C. Song et al., “Attention based spatial-temporal graph convolutional networks for traffic flow forecasting,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 922–929, 2019.

[44] Z. L. Lu, W. F. Lv, Y. B. Cao, Z. P. Xie et al., “LSTM variants meet graph neural networks for road speed prediction,” *Neurocomputing*, vol. 400, 2020.

[45] L. Huang, Y. Yang, X. Zhao, C. Ma, and H. Gao, “Sparse data-based urban road travel speed prediction using probabilistic principal component analysis,” *IEEE Access*, vol. 6, pp. 44022–44035, 2018.

[46] OpenStreetMap, https://www.openstreetmap.org/.

[47] M. E. Yang, “Modularity and community structure in networks,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 10323, pp. 8577–8582, 2006.

[48] Y. Lou, C. Zhang, Y. Zheng, X. Xie et al., “Map-matching for low-sampling-rate GPS trajectories,” in *Proceedings of 17th ACM SIGSPATIAL International Symposium on Advances in Geographic Information Systems*, pp. 352–361, Seattle, Washington, USA, 2009.

[49] PostGIS, http://www.postgis.org/.

[50] J. Lin, “Divergence measures based on the Shannon entropy,” *IEEE Transactions on Information Theory*, vol. 37, no. 1, pp. 145–151, 1991.

[51] Gensim, https://radimrehurek.com/gensim/.

[52] ArcGIS, https://www.esri.com.

[53] M. Girvan and M. E. J. Newman, “Community structure in social and biological networks,” *Proceedings of the National Academy of Sciences*, vol. 99, no. 12, pp. 7821–7826, 2002.

[54] M. E. Newman, “Fast algorithm for detecting community structure in networks,” *Physical Review E*, vol. 69, no. 6, pp. 23–31, 2004.

[55] V. D. Blondel, J. L. Guillaume, R. Lambiotte et al., “Fast unfolding of communities in large networks,” *Journal of Statistical Mechanics: Theory and Experiment*, vol. 38, no. 10, pp. 1–9, 2008.

[56] Y. Ding, “Community detection: topological vs. topical,” *Journal of Informetrics*, vol. 5, no. 4, pp. 498–514, 2011.

[57] A. Koesdwiady and F. Karray, “New results on multi-step traffic flow prediction,” 2018, http://arxiv.org/abs//1803.01365.