Research on Bus Data Pre-processing for Traffic Flow Prediction

Shuchen Gao* Hongye Yang* Jiaqi Zhang* and Zituo Li *
*Inner Mongolia University of Technology, China

Abstract— **Traffic flow prediction is an important basis for intelligent transportation systems, and its accuracy directly affects traffic control and induced effects. In order to improve the accuracy of the prediction model and reduce the complexity of the model and lessen the training time, this article studies the pre-processing method of bus GPS data. First, the data correlation analysis shows that the key influencing factors are obtained, which lays the foundation for the selection of multiple input factors. Next, on the basis of ensuring the rationality of the data, the fusion algorithm based on SVDD and isolation forest is used to discriminate the outliers. Finally, a combined denoising algorithm based on Hanning+Symlet4 wavelet is proposed. Square error, signal noise ratio and smoothness are used as performance indicators to verify the effective optimization of the data, which provides powerful data support for further research.**

I. INTRODUCTION

The rapid growth of motor vehicle provides new opportunities for the rapid development of China's social economy, but at the same time, it also brings a series of complications to the city development, such as traffic congestion, environmental pollution and so on. Intelligent Transportation System (ITS) has become the main way to solve urban traffic problems, and traffic flow prediction is one of the core research contents of ITS [1]. According to the accurate prediction results of traffic flow, the traffic management department can conduct effective traffic guidance in advance, and travelers can effectively avoid the congested section through road condition information and save travel time [2]. Therefore, it is of great significance to predict traffic flow accurately.

In recent years, domestic and foreign scholars have done a lot of work on traffic flow prediction. Due to the strong real-time and wide coverage of GPS data, the usage of large data advantage to study the traffic field has become an inevitable trend [3]. Early prediction methods mainly include linear and non-linear regression, filtering and so on [4]. In recent years, with the popularity of neural network, new methods have more advantages than traditional methods and can better adapt to the characteristics of non-linearity and instability of traffic data. Domestic and foreign studies have shown that the quality of the input sample has a very important impact on the prediction effect of the prediction model. Therefore, pre-processing input samples of the predictive model is of great significance.

In general, if there are abundant kinds of traffic flow parameters, the traffic state can be directly determined and the prediction will be very accurate. However, if the kinds of traffic flow parameters are not suffcient, the prediction accuracy will decrease [5]. Therefore, the correlation analysis of bus GPS data can provide a reference for the diversified construction of traffic flow parameters. The input sample can be selected efficiently and multiple training of the model can be avoided, which greatly reduces the training time. After that, this paper uses the GPS data of Hohhot bus as the original data. Due to equipment failure or unstable transmission, data will inevitably appear problems such as data loss or mutation. If the traffic flow prediction model is trained directly with original data, it will inevitably produce wrong or meaningless results. Therefore, in view of different data problems, data are modified to ensure that the processed data are reliable and reflect the real traffic state characteristics as far as possible, which provides ideas for the application of high-precision traffic flow prediction in practical engineering.

II. CORRELATION ANALYSIS

There are 22 kinds of data parameters collected in the Bus-borne GPS system. Diversified data support can effectively reflect the complexity, non-linearity and uncertainty of traffic flow. However, if all parameters are not screened as input factors, the complexity and training of the prediction model will have a huge load. Therefore, on the basis of the speed data, the parameters that can positively influence the traffic flow prediction can be selected, which can further improve the traffic flow prediction accuracy. The bus GPS data parameters are shown in Table I.

Correlation coefficient is show in (1). On the basis of speed, correlation test was carried out with other 21 parameters, and blank rows were removed. From Table II, it can be concluded that the bus speed has an extremely strong correlation with the Station Name and the Real Time Status; it has strong correlation with the Cursor Over Ground, the Odometer and the Lat; the correlation with the Line Type, the Lng and the Gather Time is weak; it has a very weak correlation with the Pos Is In Stations and the Relative Location. The experimental results are shown in Table II.

$$r(X,Y)=\frac{Cov(X,Y)}{\sqrt{Var[X]Var[Y]}} \tag{1}$$

Cov (*X, Y*) is the covariance of *X* and *Y*, *Var* [*X*] is the variance of *X* and *Var* [*Y*] is the variance of *Y*.

TABLE I
THE PARAMETER OF BUS GPS DATA

The Parameter	Means	The Parameter	Means
Lng	Bus GPS longitude	Lat	Bus GPS latitude
Bus Speed	The bus speed	Relative Location	The relative position
Real Time Status	The bus real time location	Driver Uuid	The driver number
Pos Uuid	Bus equipment number	Cursor Over Ground	The bus turning Angle
Pos Is In Station	Bus in or out	Station Name	The bus station name
Distance To Pre Position	The distance to next bus station	Line Uuid	The line number
Drv Ic Card	Driver IC card number	Line Type	The line type
Is Off set	Whether the bus leave the station	Driver Name	The driver's name
Gather Time	GPS data acquisition time	All Alarms	Number of warnings in per trip
Bus Uuid	The vehicle number	Odometer	Bus odometer reading
Dev Uuid	The equipment serial number	Sation Uuid	The bus station ID

TABLE II
CORRELATION COMPARISON TABLE OF SPEED AND OTHER DATA PARAMETERS

Rank	Parameter	Absolute correlation coefficient
1	Station Name	0.3059
2	Real Time Status	0.21768
3	Cursor Over Ground	0.15764
4	Odometer	0.15485
5	Lat	0.1546
6	Line Type	0.12554
7	Lng	0.12465
8	Gather Time	0.09123
9	Pos Is In Station	0.01254
10	Relative Location	0.01245

III. DATA PREPROCESSING

The data acquisition interval of Bus-borne GPS system is short and the time span is long. GPS data contains a lot of missing, abnormal and noisy data. Direct application of GPS data to traffic flow prediction will lead to slow model training and model under-fitting. Therefore, it is necessary to carry out high-quality pre-processing of bus GPS data to provide high-precision data support for further research. The data pre-processing steps are shown in Fig.1.

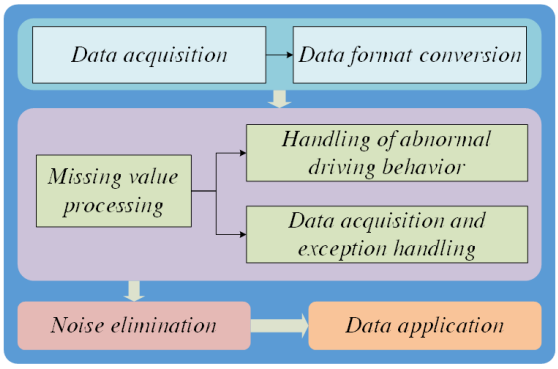

Fig.1 Data Preprocessing Steps.

A. Processing of Missing Values

External factors such as equipment failure and weak acquisition signal will lead to missing value. Bus speed data is a continuous time series, and the range of missing values in the collection process is small, generally 2-3 points of intermittent vacant values, so the time-related point interpolation method is adopted to fill the speed data. The method of time series interpolation is shown in (2).

$$y^t = \frac{1}{n+2}\sum_{i=t-2}^{i=t+n} y^i \tag{2}$$

Where y^t denotes the missing point and n denotes the number of data collected after the missing value. The interpolation of time series is mainly based on summing up several adjacent data and averaging the final results. It is suitable for a small number of time series with fewer breakpoints. If there are other requirements, interpolate function can be called in Python and filled by setting different parameters. The velocity data curve after interpolation of relevant time points is shown in Fig. 2.

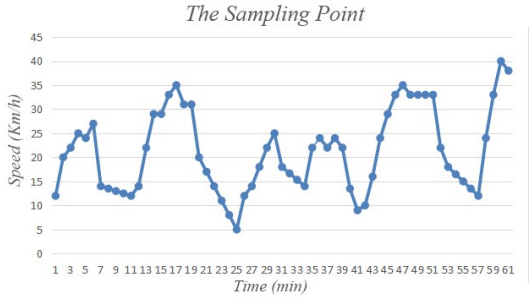

Fig.2 The Velocity Curve after Filling.

B. Processing of Outliers

Outliers are values that differ from the general behavioral characteristics, structures or correlations due to system errors, human errors or variations in intrinsic data. It is too much

work to remove outliers simply by manual filtering, and there is a risk that outliers will be missed. In order to ensure the rationality of the data, this paper divides the abnormal values into two cases: driving behavior abnormality and data abnormality, and proposes an abnormal value discrimination method based on the fusion of two unsupervised algorithms: SVDD and isolation forest.

1. Driving Behavior Abnormal

Buses stop at the platform to facilitate passengers getting on and off. This situation has no effect to the road traffic conditions. The GPS data generated when the bus enters the station does not contain useful information. In order to avoid the interference of inbound data on traffic status prediction and recognition results, it is necessary to eliminate the inbound data of those buses. The bus speed curve before removing the abnormal driving behavior is shown in Fig. 3.

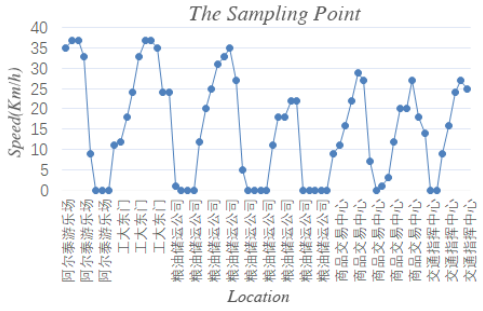

Fig.3 Shows the Data Image of the Docking Station.

When the bus arrives at each station, the bus speed will drop from 20-30 km/h to 0 km/h to wait for passengers to get on and off. Although these data will exist at each station, these stagnation data are not caused by congestion parking behavior, and the identification of traffic flow state under this condition will become abnormal, so the data should be eliminated. However, when the bus is running on the road, the bus speed decreases due to the coming car, pedestrians and car-following operation belonging to the daily bus driving data. This kind of data is beneficial for the model to learn the complexity of the traffic condition. Therefore, this part of data is selected and saved in this paper, and the result after deletion is shown in Fig.4.

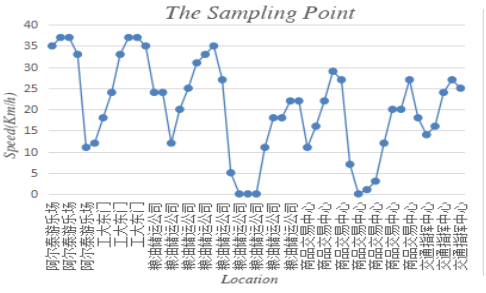

Fig.4 Shows The Data Image without Bus Stop Station.

2. Data Anomaly

Data anomalies is generally related to equipment failure, drivers' driving habits, abnormal road emergencies and other factors. The unsupervised learning method can automatically learn the structure of data and deduce the network model. It is gradually applied to the detection of outliers [6]. Among them, isolation forest and SVDD are more widely used. When using isolation forest or SVDD algorithm individually to deal with outliers, normal points with smaller values are often removed as outliers. Therefore, considering the above two algorithms, the outliers of the data are determined. The specific steps are as follows:

Sept 1. Importing and judging Data according to SVDD and isolated forest;

Sept 2. Marking judgment results, Normal data is in the latter column marker 1. Abnormal data is in the latter column marker -1;

Sept 3. Judging data based on Python. Through operation, the row marked with -1 in the column of both methods are discarded, and the rest of the data is retained;

Sept 4. Saving the processed data.

Because isolation forests cannot learn the traffic complexity, some of the smaller and fewer occurrence points are judged as outliers, which can be reduced by combination method. The mean value and standard deviation of the data before and after the experiment are shown in Table III.

TABLE III
COMPARISON VALUES BEFORE AND AFTER THE REMOVAL OF OUTLIERS

	Before outlier processing	After outliers processing
The mean	15.88	14.32
The variance	12.11	10.00
The quartile	12.00	14.00
The median	18.00	19.00
The maximum	47.00	42.00

C. Processing of Noise

The condition of urban traffic is complex and changeable. Even after filling the missing values and screening the abnormal values, there will still be a lot of noise in the data due to the floating bus's own equipment and various disturbances. If it is directly applied to traffic flow prediction, the model training time will increase, the memory occupied become larger, and the prediction curve will become a basically flat straight line. Through experiments, it is found that the wavelet domain denoising method has a better fitting degree for the data trend with the increase of the number of layers. However, if the layer number of wavelet basis is increased blindly, excessive denoising will occur, leading to the increase of data credibility and mean square error. Therefore, Hanning is introduced to further smooth the data. As a kind of ascending cosine window, Hanning can well

suppress high-frequency noise and smooth data due to its main lobe widening and side lobe reduction [7] and [8]. The window function of Hanning is shown in (3).

$$w(n) = \frac{1}{2}[1 - cos(\frac{2\pi n}{N-1})] \tag{3}$$

Compared with the four-layer symlet wavelet, using Hanning denoising alone can make the curve more smoother and reflect the direction of traffic flow, but it can not retain many data details, which is not conducive for future research. In view of that situation, this paper combines the Hanning and the wavelet denoising method, and chooses root mean squared error (RMSE), signal noise ratio (SNR) and smoothness (R) as the evaluation indexes to measure the denoising effect [9] and [10]. As follows:

RMSE: Square root of variance of original signal and estimated signal after denoising, as shown in (4).

$$RMSE = \sqrt{\frac{[f(n) - F(n)]^2}{n}} \tag{4}$$

SNR: The ratio of useful information to noise in the signal, as shown in (5).

$$SNR = 10lg(\frac{mean[f(n)]^2}{Var[f(n)]}) \tag{5}$$

R: The ratio of the variance root of the difference fraction of the signal after wavelet denoising to the variance root of the difference fraction of the original signal, as shown in (6).

$$R = \frac{\sum_{n=1}^{N}[F(n+1) - F(n)]^2}{\sum_{n=1}^{N}[f(n+1) - f(n)]^2} \tag{6}$$

Where, $f(n)$ is the original signal and $F(n)$ is the signal after denoising. These three kinds of evaluation criteria can represent the information retention, denoising quality and high frequency retention of time series respectively, so as to ensure the availability of denoised data. Three evaluation criteria were used to compare the denoising experiment results, as shown in Table IV.

TABLE IV
COMPARISON OF DENOISING EXPERIMENT RESULTS

	Symlet4 wavelet	Symlet6 wavelet	Hanning	Combinatorial algorithm
RMSE	9.0220	11.2615	15.1450	8.8021
SNR	6.1316	12.2229	10.2564	10.5681
R	0.07630	0.0540	0.00486	0.00953

From Table IV, it can be seen that Symlet6 wavelet only increases SNR by 1.6368 compared with the combination algorithm on the basis of discarding a lot of useful information, and the RMSE of the combination algorithm decreases by 6.3429 compared with the Hanning denoising. Therefore, the denoising effect of Hanning+Symlet4 wavelet combination algorithm is obviously better than that of single algorithm. It has certain improvement in RMSE, SNR and R, which verifies the effectiveness of the combination algorithm.

As shown in Fig.5, the data processed by Hanning+Symlet4 wavelet combination algorithm shows the direction of basic traffic flow and the curve is smoother, and there is no excessive denoising phenomenon, leaving more details.

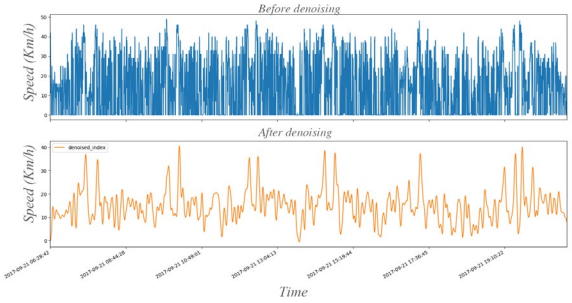

Fig.5 Hanning+Symlet4 Wavelet Combination Denoising Result Diagram.

IV. CONCLUSIONS

In order to improve the accuracy of the prediction model, this paper explores the deep learning in bus GPS data pre-processing. Firstly, on the basis of velocity data analysis, through correlation analysis, the parameters which can play a positive role in traffic flow prediction are selected. Then, this paper completes the cleaning of Bus-borne GPS data. In view of the unique mode of bus driving, on the basis of retaining the daily low-speed and zero-value data of buses due to traffic congestion, the data of pedestrians up and down at bus stops are removed to ensure that the data used can truly reflect the direction of traffic flow. The other outliers of the data were removed by the two unsupervised test methods−SVDD and isolation forest under their joint decision. Finally, in the denoising stage of data, the Hanning is added based on the Symlet4 wavelet for denoising smoothing. The experimental results show that the combined denoising algorithm based on Hanning+Symlet4 wavelet can significantly improve RMSE, SNR and R, and increase the recognition degree of data. In this paper, the method of bus GPS big data pre-processing is deeply analyzed, which provides reliable data source for further data mining, so as to provide the best information service for traffic flow forecasting application.

REFERENCES

[1] Garima Dhawan S N, "An Overview and Evolution of the Intelligent Transportation System as VANETs," *International Journal of Advanced Trends in Computer Science & Engineering*, 2016.

[2] Lv Y, Duan Y, Kang W, "Traffic Flow Prediction With Big Data: A Deep Learning Approach," IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, Vol. 16 No. 2, pp. 865-873, 2015.

[3] Shan Gong, *Research on Vehicle Speed Prediction Model Based on Floating Vehicle GPS Data,* Beijing Jiaotong University, 2009.

[4] Abouaïssa H, Fliess M, Join, Cédri, "On short-term traffic flow forecasting and its reliability," *IFAC-Papers Online* Vol. 49 No.12, pp. 111-116, 2016.

[5] Sładkowski A, Pamuła W, *Intelligent Transportation Systems-Problems and Perspectives*, Springer International Publishing, 2016.

[6] Fei Tony Liu, Kai Ming Ting, Zhihua Zhou, "Isolation Forest Data Mining," Eighth IEEE International Conference, Vol.1 No.170 pp. 413-422, 2008.

[7] Qiang Chen, "An Evaluation Indicator of Wavelet Denoising," *Journal of Geomatics*, Vol.33 No.5, pp. 13-14, 2008.

[8] Yuanyuan Peng, *Application of Wavelet Analysis in One-dimensional Signal Denoising,* Beijing University of Posts and Telecommunications, 2011.

[9] Wufeng Liu, Xuchun He, Yang Xu, Weimin Qiao, "A Program of Arithmetic to Process Data," *Control & Management,* Vol. 09, pp. 209-210+187, 2007.

[10] Fei Kou, *Short-term Traffic Flow Prediction Based on Adaptive Artificial Fish Swarm Algorithm BP Recurrent Neural Net Work,* Beijing Jiaotong University, 2018.