# Traffic Flow Data Preprocessing Method Based on Spatio-temporal Similarity

Ran Tian[1,a,*] , Shanwei Li [2,b] and Guoying Yang[1,c]

[1]Lanzhou Petrochemical Polytechnic, Lanzhou 730060, Gansu, China

[2]Gangdong Industry Polytechnic,Guangzhou 510300,Guangdong,China

[a]77580193@qq.com,[b]99970768@qq.com,[c]79546245@qq.com

*Corresponding author

**Keywords:** Spatio-temporal similarity; Traffic flow; Data preprocessing

**Abstract.** In view of the existing traffic collection system, the information transmission system has many problems such as the failure of the detection device and the failure of data transmission in the data acquisition, transmission and storage operations. In this paper, a method of traffic flow data preprocessing based on spatio-temporal similarity is proposed, which can ensure the data quality of traffic flow basic database by effectively processing dirty data before data extraction and application.

## Introduction

As urban traffic pressure continues to rise, intelligent transportation systems play an increasingly important role in urban traffic management and control systems.[1] The correctness, completeness and reliability of the basic traffic flow database are the basic guarantees for the efficient operation of the entire intelligent transportation system. Then, the existing traffic collection system and information transmission system have many problems in data collection, transmission and storage operations and are prone to dirty data such as failure of the detection device, failure of data transmission and the like. These dirty data may lead to failure or even wrong decision making by the intelligent transportation system. Therefore, it is necessary to effectively deal with the dirty data before the application of data extraction to ensure the data quality of the basic database of traffic flow.[2]

In this area, there are already a lot of related research and engineering applications. The traditional traffic flow data cleaning process usually includes data attribute analysis, determine the cleaning program, test the cleaning results, dirty data cleaning and data update five stages, and in order to deal with dirty data travel in different situations, such as data errors, data loss, data drift Etc., need to be dealt with accordingly, the process is cumbersome. At the same time, during the dirty data cleaning, a series of corresponding technical methods and measures have appeared so far, which can be divided into two major categories: the statistical-based prediction model and the intelligent algorithm as the main research methods Prediction models include Kalman filtering model, parametric regression model and time series model.[3] However, these methods only consider the historical factors for data processing, not suitable for time-varying complex systems with low accuracy. The latter method includes state Phase space reconstruction model, wavelet decomposition model, neural network, support vector machine and so on, the method is too complicated, and it is difficult to achieve the real-time processing effect in practical application.

In view of the existing problems and deficiencies mentioned above, this paper proposes a traffic flow data preprocessing method based on spatial and temporal similarities, so as to improve the traditional traffic flow data preprocessing process, while taking into account the complexity of the algorithm and the accuracy of results.

## Mathematical Model

The traffic flow itself is time-varying. It changes according to the road and time, and the traffic flow in the previous period will have an impact on the traffic flow in the later period. Therefore, the time-dependent characteristics of the traffic flow are affected by traffic conditions , The time series of traffic flow have long-range correlation when the traffic flow density is within a certain smaller range under the same local train ratio, that is, the current and future traffic conditions are affected by the historical traffic conditions, and the trends of the traffic flow time series and The trend of historical time series is positively correlated. While the traffic flow density is small and crowded, the interaction between vehicles becomes very small, then the traffic flow time series shows short-range correlation[4]. At the same time, the urban road is an open and complex large system in urban traffic, which makes the traffic flow influenced by many external factors such as weather conditions, traffic accidents, traffic control and so on. The data analysis shows that traffic flow has periodicity and similarity in a long time scale and time-varying, chaos and correlation in a short time. Therefore, this paper deals with the dirty data in traffic flow by mining similarities in traffic flow data. Mainly divided into three steps:

Step 1: traffic flow data for non-dimensional treatment.

Step 2: Fill data through the spatial and temporal similarities of traffic flow data.

Step 3: Modify the abnormal traffic flow data through the fluctuation coefficient of traffic flow data.

**Dimensionless Data.** Establish positive indicator fuzzy quantitative model:

The positive index value is bigger the better, that is, the result of dimensionless processing is a strictly monotonically increasing function of the evaluation result.

$$R_j(x)=\begin{cases}\frac{1}{2}+\frac{1}{2}\sin[\frac{p}{x_{jmax}-x_{jmin}}(x_j\,\frac{x_{jmax}+x_{jmin}}{2})] & ,x_{j\min}<x<x_{j\max}\\ 0\quad , \quad x_{j\min}\,\text{吵}x \text{ or } x\quad x_{j\max} & \end{cases}\tag{1}$$

$x_{jmin}$:For the jth evaluation index evaluation, the maximum score in the system.

$x_{jmax}$:For the jth evaluation index evaluation, the minimum score in the system.

$R_j(x)$:The jth evaluation index after the non-dimensional evaluation of the value.

$x_j$:The original score of the jth evaluation index.

Negative indicator fuzzy quantitative model:

Negative index value is as small as possible, that is, the result of dimensionless processing is a strictly monotonically decreasing function of the evaluation result.

$$R_j\left(x\right)=\begin{cases}\frac{1}{2}-\frac{1}{2}\sin[\frac{p}{x_{jmax}-x_{jmin}}(x_j\,\frac{x_{jmax}+x_{jmin}}{2})] & ,x_{jmin}<x<x_{jmax}\\ 0,x_{jmin}\,\text{吵}x \text{ or } x\quad x_{j\max} & \end{cases}\tag{2}$$

Fixed index fuzzy quantitative model:

$$R_j\left(x\right)=\begin{cases}\frac{1}{2}-\frac{1}{2}\sin[\frac{p}{x_{jmax}-x_{jmin}}(x_j\,\frac{x_{j\mathrm{mod}}+x_{jmin}}{2})], & x_{jmin}<x<x_{j\mathrm{mod}}\\ \frac{1}{2}+\frac{1}{2}\sin[\frac{p}{x_{jmax}-x_{jmin}}(x_j\,\frac{x_{jmax}+x_{j\mathrm{mod}}}{2})], & x_{j\mathrm{mod}}<x<x_{jmax}\\ 0,x_{j\min}\,\text{吵}x \text{ or } x\quad x_{j\max} & \end{cases}\tag{3}$$

$x_{jmod}$:The jth evaluation index evaluation, the score used in the most appropriate value.

**Spatiotemporal Similarity of Traffic Flow Data.** Data similarity in traffic flow generally refers to the law of periodic similarity of traffic flow on the scale of day, week, month and so on. Most of the work is based on the similarity of days, and seldom see the measurement based on long time scales The similarity of the work. Based on this, this paper analyzes the similarity of traffic flow in the week scale by using similarity coefficient and volatility coefficient. This work not only helps to select training samples, but also helps to further forecast traffic flow and provide an effective basis for alleviating urban traffic congestion.

Due to the cyclical nature of work and rest, there is a periodicity of the traffic flow generated by travel, in units of days, weeks, and months, that is, traffic flows of days or weeks or days of the week but with the days of the week exist A certain degree of similarity, in general, uses the following two coefficients to measure their similarity.

Take day by day *k* traffic flow data as a vector:

$$l_i = [x_1, x_2, ..., x_k], (i = [1, n]) \tag{4}$$

Then all the data of n days form a matrix:

$$L = [l_1, l_2, ..., l_n]^{\mathrm{T}} \tag{5}$$

The similarity coefficient *S* is the average of the correlation coefficient between each two vectors, the expression is:

$$S = \frac{\sum_{n \ge i > j \ge l} R(i, j)}{n(n-1)/2}, |s| \le 1 \tag{6}$$

Where *R* is the matrix of correlation coefficients of matrix *L*, the expression is:

$$R(i, j)=R(j, i) = \frac{\operatorname{cov}(l_i, l_j)}{\sqrt{D(l_i)}\bullet\sqrt{D(l_j)}} \tag{7}$$

$\operatorname{cov}(l_i, l_j)$ is the covariance coefficient of $l_i$ and $l_j$. $D(l_i)$ and $D(l_j)$ are the variance of $l_i$ and $l_j$.The larger the value of S, the greater the similarities of traffic flow over the two days.

**Traffic flow data fluctuation coefficient.** Let vector $M = [E(l_1), E(l_2), ..., E(l_n)]$, and each element of *M* represents the average traffic flow for the corresponding day,define the volatility coefficient *T* as:

$$T = \frac{\sqrt{D(M)}}{E(M)} \tag{8}$$

Where $D(M)$ is the variance of M; $E(M)$ is the mean value of *M*.The smaller the value of *T*, the smaller the change of traffic flow in each day. Abnormal traffic flow data can be found by the fluctuation coefficient so that it can be processed in time.

## Summary

For the existing traffic collection system, the information transmission system has many problems such as the failure of the detection device and the data transmission failure in the data collection, transmission and storage operations. In this paper, a method of traffic flow data cleaning based on spatio-temporal similarity is proposed, which includes non-dimensional processing of traffic flow data, filling of traffic flow loss data based on spatio-temporal similarity and discovery of abnormal traffic flow data based on fluctuation coefficient. The method can be convenient, fast and accurate for traffic flow data cleaning. It has practical application value in simplifying the process of cleaning traffic data and providing the cleaning quality of traffic data.

## Acknowledgement

This research is support in part by the Scientific research project of Gansu higher education institutions(2016B-133),the Scientific research project of Lanzhou Petrochemical Polytechnic (D2016-01) and the Innovation and Entrepreneurship of college students in Gansu Province-college

graduated(2050305-1435).

## References

[1] Chan K Y, Dillon T S, Singh J, et al. Neural-network-based models for short-term traffic flow forecasting using a hybrid exponential smoothing and Levenberg–Marquardt algorithm[J]. IEEE Transactions on Intelligent Transportation Systems, 2012, 13(2): 644-654.

[2] Zhang L, Liu Q, Yang W, et al. An improved k-nearest neighbor model for short-term traffic flow prediction[J]. Procedia-Social and Behavioral Sciences, 2013, 96: 653-662.

[3] Guo J, Huang W, Williams B M. Adaptive Kalman filter approach for stochastic short-term traffic flow rate prediction and uncertainty quantification[J]. Transportation Research Part C: Emerging Technologies, 2014, 43: 50-64.

[4] Wang Z, Lu M, Yuan X, et al. Visual traffic jam analysis based on trajectory data[J]. IEEE Transactions on Visualization and Computer Graphics, 2013, 19(12): 2159-2168.