Road Segments Traffic Dependencies Study Using Cross-Correlation

Redouane Benabdallah Benarmas, Kadda Beghdad Bey
Ecole Militaire Polytechnique - Chahid Abderrahmane Taleb (EMP)
PO Box 17, Bordj El Bahri,
Algiers, Algeria

Received: Mai 2021 / Accepted: Oct 2021

Abstract

Traffic Prediction on a urban road network become more complex face to exponential growth in the volume of traffic, it is necessary to study the relationship between road segments before the prediction calculation. The spatial correlation theory has been well developed to interpret the dependency for understanding how time series are related in multivariate model. In large scale road network modeled by Multivariate Time Series, the Spatial-temporal dependencies detection can limit the use of only data collected from points related to a target point to be predicted. This paper present a Cross-Correlation as method to dependency analysis between traffic road segments, Scatterplot of Cross-Correlation is proposed to reveal the dependency, we provide a comparative analysis between a three correlation coefficients sush as Spearman, Kendal and Person to conclude the best one. To illustrate our study, the methodology is applied to the 6th road ring as the most crowded area of Beijing.

Keywords: Intelligent Transportation System, Traffic Forecasting, Time series, Cross-Correlation.

1 Introduction

Traffic flow prediction has become one of the important research fields in Intelligent Transportation System (ITS). The prediction of traffic flow information is important for control and guidance and makes the transport users better informed. The special characteristics of road network, such as their high scale and complex dependencies between segments, make the problem of prediction very challenging. The development of modern's statistical theory and machine learning methods has accelerated the pace of research on a variety of approaches, which usually revolved

[0]R.Benabdallah Benarmas
Ecole Militaire Polytechnique - Chahid Abderrahmane Taleb (EMP)
PO Box 17, Bordj El Bahri Algiers, Algeria
Tel: +213-666-116241
Email: benarmas.it@gmail.com

around classification and regression methods. The earliest traffic prediction methods mainly include Auto Regressive Integrated Moving Averaging (ARIMA)[12], Kalman Filter[13, 14], Support Vector Machine (SVM)[15], Markov chain model [16] and Artificial Neural Network [17, 18, 19]. The first's solutions were provided on simpler contexts, which aim to predict the road flow at a given location, the model is qualified as uni-variate. The classical regression techniques were sufficient to solve the problem. Recently and for industrial needs, the problems are exposed in more complex and varied contexts, such as the prediction of several values in different places by using a data collected from different sources(See Fig. 1).

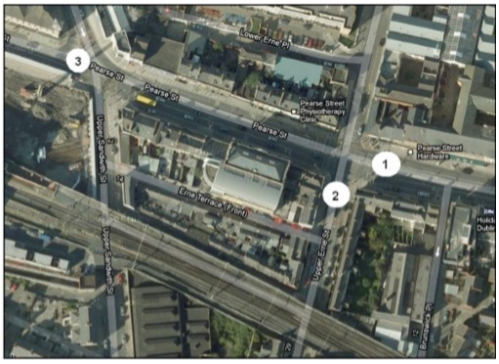

Figure 1: Road Network Point Positions

For this problem, the simple generalization of uni-varied models was insufficient, because on a road network, the flow is a stochastic phenomenon which evolves over time and which has an impact on others flows in neighboring points or located in other places, therefore, the modeling of spatial-temporal dependencies becomes necessary. Furthermore, the reliability and accuracy of prediction method is not based only on the used model but also on the choice and determination of the historical used data. The determination of relevant values used in the calculation aims to characterize in an efficient manner the spatial and temporal dependencies between a given point and different points in the road network. In road traffic prediction, the complexity is assessed in relation to the calculation time, the expertise and the effort required to providing a solution. Flexibility must also be achieved, so as to have a model less sensitive to the sizing of used data. For large scale road network, the detection of dependency between flows evolution captured from different road segments reduces significantly the data used for prediction calculation. The main goal in our work, is to demonstrate that a cross-correlation can interpret the dependency between road traffic segments, and reduce consequentially the data used for prediction calculation for a target point, furthermore, we provide a comparative study relatively to the coefficient used in cross-correlation calculation at second stage.
This paper is organized as follow. Section II present a brief review of related work. Section III consists in the definition of the working context which permit the problem formulation. Section IV is devoted to the description of our dependency detection method. Experiment is performed and results are discussed in Section V, finally we conclude some comments and future work directions.

2 Brief review of related work

Traffic forecasting in large scale road network need an interpretation of the spatial and temporal traffic patterns in each segment and the relationship between each one. Spacial correlation is used to build a GIS system based method [21] to extract the real-time traffic information .These models were insufficient, because on a road network, the flow is a stochastic phenomenon which evolves over time and which has an impact on other flows in neighboring points or located in other places, for this purpose, spatial temporal modeling is necessary where the probability density must be defined in a joint way. Pan et al. Introduced the spatial-temporal correlation to the short-term traffic flow prediction by using the random region transmission framework [20]. Time auto-correlation analysis is carried out by [22] using journey time data collected on London's road network, the analysis was applied for uni-variate model.
Recently, the spatial-temporal correlation theory has been well developed to interpret the dependency for understanding how time series are related in multivariate model. At this stage, traffic data in large scale road network in most of studies are represented by multivariate time series, furthermore, a Cross correlation has been widely used in special analysis and in several contexts such as economics and environment [5], and present a potential for road traffic analysis. Many methods consider the spatial-temporal correlation as basic technique in the research on road traffic. A cross-correlation between network-aggregated density was proposed in [15] as a natural indicator of traffic phases for road networks. [7] Suggest that the method can be used to investigates the relationship between traffic flow series and the spatial distance of the road network sections. [10]Propose a de-trended cross-correlation analysis (DCCA) to measure the relationship between air pollution and traffic congestion in the urban area. The previous works merely consider the spatial-temporal correlation as technique to understand the interactions between different segments on road network, for traffic prediction, [9] use auto-correlation and cross correlation measure to find a seasonal patterns and provides a theoretical assumption for traffic forecasting. Recently [8] propose a new approach to identify the most influential locations, in this work, the captured correlation network between different locations might facilitate future studies on controlling the traffic flows.

3 Context Study and Problem Formulation

The problem formulation in the field of road prediction is based on the definition of a precise context on which the study will be carried out because an approach or a method's choice depends closely to the implementation. In literature, the theory of road traffic prediction is described on three aspects, the precision of the variables used in modeling, the dimension of the quantity measured and the prediction horizon. For the first, a choice between a macroscopic or microscopic modeling must be made to define the precision of the predicted variables, in a macroscopic model, the dynamic behavior of objects in the road is described as a homogeneous flow with a density or a speed. Our work study will be with a macroscopic modeling using a history of measured quantities and coming from different data sources, the prediction will be calculated from a history made up of

several variables.

3.1 Data Collection and Modelling

Whatever the approach and methods adopted for traffic prediction, multi-varied modelling is based on the data definition and the relationships between them, making it possible to capture the dependencies between these data, then an adequate prediction method will be applied to arrive at prediction. There are different ways to get a road traffic data, it can be collected by a network of sensors, namely the detection loops and the traffic counters, another method is to use the GPS on board vehicles or installed in the phones. In any case, data is regularly reported in varying amounts to what is generally called a Traffic Management Center (TMC), using a data transmission network. At this level, traffic is aggregated in observation intervals into three main quantities: flow, speed and volume (density).Big data [4] was heavily used for predicting road traffic and allowed motivation for the adoption of the data-driven models.

3.2 Problem Formulation as a multi-varied time series

A time series $T = t^1, t^2, ..., t^n$ is a sequence of real numbers obtained through repeated measurements over time. We represent the evolution of the variable collected over time and at given point in the road network, as a time series denoted as, $X = (x1, x2, \ldots., xT)$, For each t between 1 and T, x is a vector of dimension D, D represents the dimension of the vector x or the component is valued by the quantity collected in a place X^t belongs to R^D is the observation at time t. The problem is to predict the value $X^h, h = t + t^h$, h denotes the prediction horizon. For large scale road network, the global traffic state agregated in n time intervals can be expressed by a high-order matrix given a large-scale road network with m segments, this traffic state is denoted as Xmn.

$$\mathrm{X}^{mn}\begin{Bmatrix} x^1 1 & ..... & x^1 n \\ x^2 1 & ..... & x^2 n \\ .. & & .. \\ .. & & .. \\ x^m 1 & ..... & x^m n \end{Bmatrix}$$

4 Dependency detection model

In the present work, detecting cross-correlations between traffic recorded in two points is the most usual way to diagnose and understand a complex large - scale road network. The simplest method is the traditional Pearson Cross-Correlation Analysis. In statistics, Pearson Correlation Coefficient (PCC) is one of the most popular statistical measures, and is frequently discussed in almost all fields, such as climatology [3], economics [5], and signal processing. As explained in previous sections, global traffic state in n time interval of large - scale road network can be represented in each point by m time series, then a cross - correlation is calculated

between each time series x and y and stored in cross - correlation matrix denoted Xcc.

$$Xcc = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

This matrix represents mutual traffic segments dependency on the road network, at this stage, the prediction is calculates by using only the data in point which the dependency is strong by means of cross-correlation ,precisely given a parameter θ we consider j, when $Xcc(i, j) > \theta$.

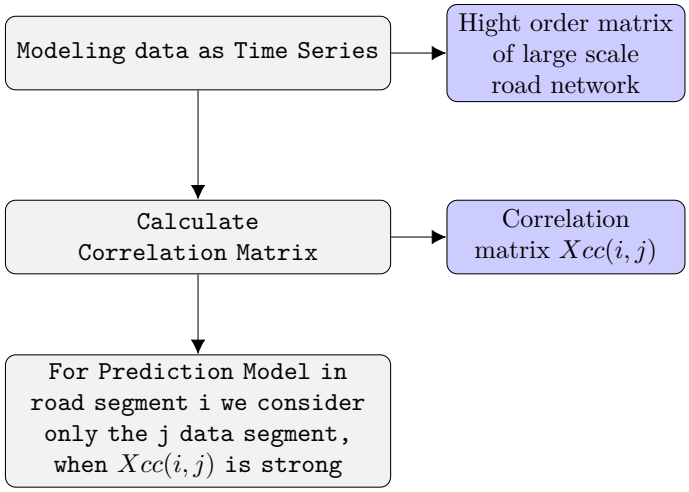

Figure 2: Dependency detection process

5 Experiments

In this experiment, traffic data was obtained from Baidu research Open Access dataset [1]. Baidu is widely used by many researchers in experiments [2]. A large-scale traffic prediction dataset was provided for the 6th ring road (bounded by the lon ,lat box of 116.10, 39.69, 116.71, 40.18), which is the most crowded area of Beijing. Fig. 3 shows the spatial distribution of these road segments.

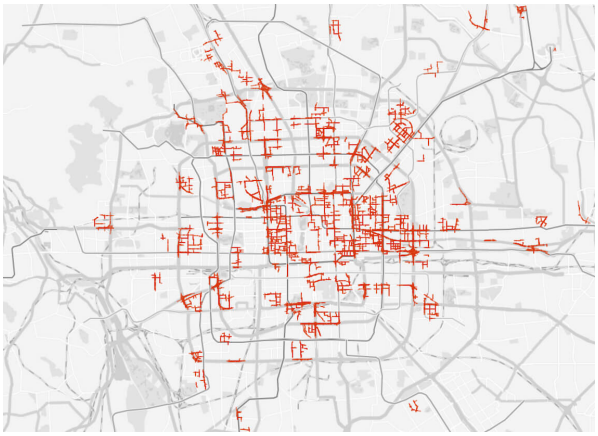

Figure 3: spatial representation of road segments

The traffic speed of 15,073 road segments is recorded per minute, then averaged with a 15-minute time. Thus, there are totally 5856 time steps.

5.1 Experiment setup

The data set used consists of two part, the first consist of the topology description of the road network. Table below shows the fields of the road network sub-dataset.

Field	Type	Description
link id	Char(13)	road segment id
width	Char(3)	width,15 :<3.0m; (30:3.0m,5.0m); 55:(5.5m,13m); 130:<13m
direction	Char(1)	direction,0:unknown,default two-way;1:two-way; 2:single-way,from start node to end node; 3:single-way,from end node to start node.
snodegps	Char(30)	gps coordinate (lon,lat) of start node
enodegps	Char(30)	gps coordinate (lon,lat) of end

The second part is a data dump of speed represented as follow :

15257588940, 0, 42.1175
..., ..., ...
15257588940, 5855, 33.6599
1525758913, 0, 41.2719
..., ..., ...

The data is formatted as line delimited, in total there was 88,267,488 rows in the specific time period, this is 2.5GB of zipped and approximately 8.5 GB unzipped. We limit our study only for the highway (lenght >13 meters) so we read only the line which the segment matched the identifier of the Highway (width field) stored in road network sub-dataset file, at this stage we use a fetch python program to extract the desired data from large dataset,then the result is formatted as large

matrix and stored in CSV file, rows represent time stamp and column segment ID, figure below shows the variation of traffic for three segments.

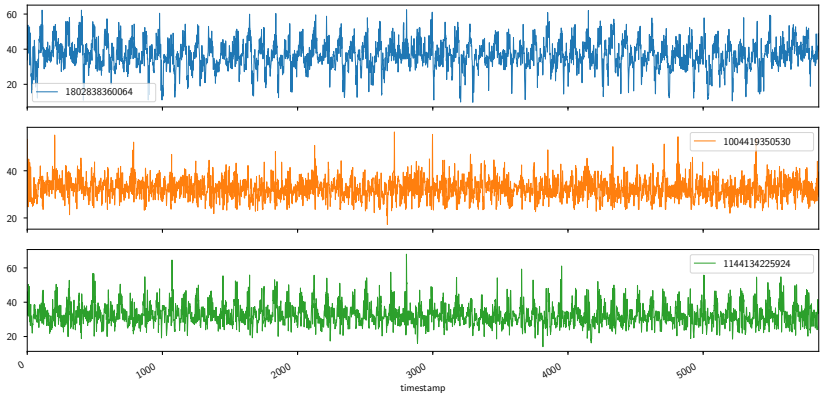

Figure 4: Traffic Averaged in three road points

The reason for formatting a data is to allow the reading in pandas data frame then calculate the cross- correlation with numpy LIB. Finaly, Scatterplots of spatial cross-correlation was been used to visually reveal the causality behind road traffic segments.

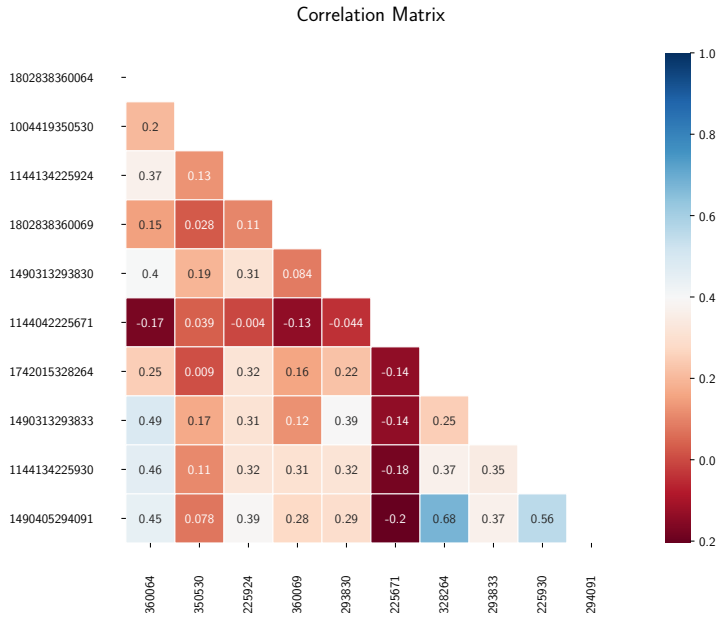

Figure 5: Correlation matrix of 8 observation points

Scatterplots of spatial cross-correlation can be used to reveal the causality between two variables visually.(See Fig. 5) Based on the global cross-correlation coefficient, we can determine the data traffic segments used to predict the traffic in target points. A positive association means that both the variables are moving in same direction. If the coefficient is equal to 0, it does not necessarily mean that there is no relation between the two variables. It means that there is a no linear relationship, but there might be another type of functional relationship, for example, quadratic or exponential. If correlation is plus/mines 0.8 and above, high degree of correlation or the association between the dependent variables are strong. Correlation between plus/mines 0.5 to plus/mines 0.8, sufficient degree of correlation and less than +/-0.5, weak correlation.
To conclude the best coefficient we make a comparison study between the three coefficients cited above [11],ech coefficient extract the data used for prediction which is calculated by using use a legacy prediction method such ARIMA the following four case:
case 1 : by using Person coefficient.
case 3 : by using kendal coefficient.
case 2 : by using Spearman coefficient.
results are evaluated by means of root error.

$$RMSE = \sqrt{(\frac{1}{N})\sum_{i=1}^{n}(t_i - p_i)^2}$$

Where pi = predicted traffic flow; ti = actual traffic flow; N = the number of predictions. Table shows the prediction errors for different models.

5.2 Interpretation

For the three cases, the dependent segments set used in prediction calculation is not same, We consider the segments when $0.2 < \theta < 0.4$. The results are listed in the following table:

Correlation Coefficient	**RMSE**
Person	70.144
Spearman	80.140
Kendal	89.847

It can be seen that the prediction is more accurate when using the dependents road data segments by means of Person cross-correlation. It was also observed that the prediction depends on the choice of used data segments. As shown in Fig. 6, the results are more conclusive if we limit data by considering a strong correlation.

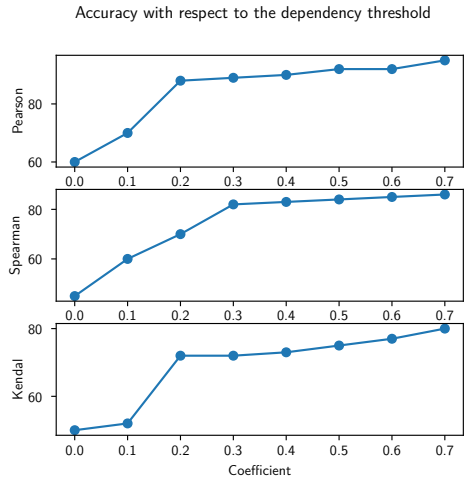

Figure 6: Accuracy with respect to dependency threshold

6 Conclusion

This paper is devoted to laying the foundation for development of spatial cross-correlation theory in Road Traffic Forecasting. The basic measurements and analytical methods are put forward and applied to an urban study of China. Pearson's correlation coefficient and other coefficients can well reflect the relationship between data traffic denoted as dependency Road segments. Finlay, on the basis of experimentation results and empirical analyses, we can conclude that statistical analysis for traffic forecasting can complement other approach such as machine learning methods and reduce data and time processing for the prediction calculation.

References

[1] Baidu research Open–Access dataset. Available online WWW.ai.baidu.com

[2] Binbing Liao, Jingqing Zhang, Chao Wu, Douglas McIlwraith, Tong Chen, Shengwen Yang, Yike Guo, Fei Wu (2018).Deep Sequence Learning with Auxiliary Information for Traffic Prediction.

[3] Yuan, Elena Xoplaki, Congwen Zhu, Juerg Luterbacher(2016).A novel way to detect correlations on multi-time scales, with temporal evolution and for multi-variables Naiming

[4] Kyusoo Chong and Hongki Sung (2015).Prediction of Road Safety Using Road/Traffic Big Data.

[5] Yanguang Chen (2015).A New Methodology of Spatial Cross-Correlation Analysis

[6] Lele Zhang, Callum Stuart, Samithree Rajapaksha, Gentry ,White, Timothy GaroniTimothy Garoni (2017).Study of Cross-Correlations in Traffic Networks with Applications to Perimeter Control.

[7] Qinghua Daxue, Xue bao, Tingting Zhao , Yi Zhang, Yu Zhou, Shumin Feng (2011). Spatial cross correlations of traffic flows on urban road networks.

[8] Shengmin Guo , Dong Zhou, Jingfang Fan , Qingfeng Tong, Tongyu Zhu , Weifeng , Daqing L, and Shlomo Havlin (2019).Identifying the most influential roads based on traffic correlation networks .

[9] Fei Su, Honghui Dong, Limin Jia, Zhao Tianand Xuan Sun(2017).Space–time correlation analysis of traffic flow on road network.

[10] Kai Shi , Baofeng Di, Kaishan Zhang, Chaoyang Feng, Laurence Svirchev (2018).Detrended cross-correlation analysis of urban traffic congestion and NO2 concentrations in Chengdu.

[11] jan hauke , tomasz kossowski (2011).Comparison of values of person's and spearman's correlation coefficients on the same sets of data.

[12] William B.M, Durvasula P.K, Brown D.E(1998). Urban freeway travel prediction: Application of seasonal ARIMA and exponential smoothing models.

[13] Okutani I, Stephanedes Y.J (1984). Dynamic prediction of trafc volume through Kalman ltering theory.

[14] Xie Y, Zhang Y, Ye Z (2007). Short-Term Trafc Volume Forecasting Using Kalman Filter with Discrete Wavelet Decomposition.

[15] Zhang Y,Xie Y (2007). Forecasting of Short-Term Freeway Volume with v-Support Vector Machines.

[16] Yu G, Hu J, Zhang C, Song G(2003). Short-term trafc ow forecasting based on Markov chain model.

[17] Wangyang Wei, Honghai Wu and Huadong Ma (2019). An AutoEncoder and LSTM-Based Traffic Flow Prediction Method.

[18] Xingyuan Dai, Rui Fu, Yilun Lin, Fei-Yue Wang, Fellow, Lili Fellow(2017)DeepTrend: A Deep Hierarchical Neural Network for Traffic Flow Prediction

[19] Chan, K. Y. and Dillon, T. S. (2013). On-road sensor configuration design for traffic flow prediction using fuzzy neural networks and taguchi method.

[20] Pan T.L, Sumalee A, Zhong R.X, Payoong N.I(2013). Short-Term Trafc State Prediction Based on Temporal-Spatial Correlation.

[21] Baofeng DI, Kai S, Kaishan Z, Laurance S and Xiaoxi H (2016). Long-Term Correlations and Multifractality of Traffic Flow Mesured By GIS for Congested and free-Flow Roads.

[22] Cheng T, James H (2011). Spatio-Temporal Autocorrelation of Road Network Data.