CTRF 51 Annual Conference

CTRF-TRF Joint Conference

North American Transport Challenges in an Era of Change
Les défis des transports en Amérique du Nord à une aire de changement

Toronto, Ontario
May 1-4, 2016

Platinum Sponsor

UNIVERSITY OF TORONTO
FACULTY OF APPLIED SCIENCE & ENGINEERING
Transportation Research Institute

UTTRI

TIME SERIES BASED HOURLY TRAFFIC FLOW PREDICTION ON THE GTA FREEWAYS USING TS-TVEC MODEL

Tao Ma, Civil Engineering, University of Toronto
Zhou Zhou, Statistical Sciences, University of Toronto
Baher Abdulhai, Civil Engineering, University of Toronto

Introduction

Short term traffic flow prediction is a cornerstone for intelligent traffic operations, management, policy making, and strategy formulation. It is an essential instrument to support ATMS (advanced traffic management system) implementation and ATIS (advanced traveller information system) service such as congestion mitigation, ramp metering, road pricing, and route guidance, etc. among others. Traffic flow prediction is to forecast macroscopic traffic quantities including traffic volume, speed, and occupancy (i.e. analogy to density) three major indicators of traffic state for a short time future horizon. This research focuses on hourly traffic flow prediction based on advanced time series techniques.

There is a great body of literature in the domain of traffic flow prediction. In our view, the methodologies can be classified into two streams including traffic state estimation and prediction. Traffic state estimation refers to methods that are mainly based on traffic flow theory to approximate the traffic situation at the middle section of a road stretch from boundary conditions. Cell transmission model (CTM) (Daganzo 1994, 1995a) is a major approach to traffic state estimation using discrete second order macroscopic traffic flow models with a supply-demand method. The application of CTM is not only restricted by the first and second Courant–Friedrichs–Lévy conditions where its time interval is limited up to a few minutes, but also suffers from discretization errors, convergence issues, and numerical instabilities (Daganzo 1995b, Treiber and Kesting 2013). On the other hand, traffic state prediction refers to methods that forecast the traffic situation at a fixed location of a freeway stretch based on historical time series data from the loop detector at the location of prediction and its adjacent upstream and downstream locations. The methodologies in this stream include, but not limited to, Neural Networks (Abdi et al., 2010, Zheng et al., 2006, Abdulhai et al., 2002, Qiao et al., 2001), Support Vector Regression (Wu et al., 2004), Kalman Filter (Antoniou et al., 2005, Wang and Papageorgiou 2005), ARIMA (Williams et al., 2003), and nonparametric regression (Chang et al., 2012, Smith et al., 2002), etc. Those methods are developed on the basis of statistical error theory using gradient descent optimization, kernel regression, state-space structure, or time series technique, etc. Their advantages and limitations can be referenced to (Ma 2016) for interested readers.

Forecasting macroscopic traffic quantities is a nonlinear multivariate problem. The challenges lie in complex statistical characteristics of stochastic traffic processes as well as autonomous and interactive dynamics within and between traffic variables. Traffic time series usually exhibit large variation over time and space. According to the analysis to the hundreds of traffic time series, statistical characteristics that often exhibit in traffic time series include seasonality, non-stationarity, serial and cross-sectional dependence over time and space, cointegration, and unknown structural break; from a perspective of traffic flow theory, the dynamics include the fundamental relation between traffic variables, and multiple traffic states. In order to model and forecast traffic state more accurately, all those factors have to be taken into account. However, many of them have not been thoroughly and systematically addressed in the existing approaches to traffic forecast. While many modelling and forecasting methods aforementioned have been developed during the last two decades in this domain, there is no single time series model available in the literature that can incorporate these factors all at once. It is therefore natural and intuitive to seek and devise a model structure that is able to concurrently take care of all those factors for network-wide application. Following this train of thought, the time-space threshold vector error correction (TS-TVEC) model is proposed and developed for short term (hourly) traffic state prediction (Ma et al., 2015).

Methodology

The new statistical model TS-TVEC is designed to concurrently use the information of multivariate traffic time series and their interactive dynamics to improve the accuracy of traffic prediction. This model is established on cointegration and error correction techniques. As well, it incorporates spatial information from upstream and downstream locations. The concept of cointegration is introduced by Granger (1981), and Granger and Weiss (1983), and is precisely defined in Engle and Granger (1987) as follows:

> **Definition:** *The components of the vector* x_t *are said to be cointegrated of order d, b, denoted* $x_t \sim CI\left(d,b\right)$*, if (i) all components of* x_t *are* $I\left(d\right)$; *(ii) there exists a vector* $\alpha\left(\neq 0\right)$ *so that* $z_t = \alpha' x_t \sim I(d-b),\;\; b>0$. *The vector* α *is called the cointegrating vector.*

In other words, elements of a vector time series are said to be cointegrated if their linear combination achieves stationarity by taking difference less number of times than each individual element of the vector.

The relationship between cointegration and error correction model is known as the Granger Representation Theorem that has been proven in Engle and Granger (1987). The most relevant part, statement (4), of the Granger Representation Theorem is cited in this paper for the purpose of this research. A full version of the theorem can be found in Engle and Granger (1987, pp 255) for interested readers.

> **Granger Representation Theorem**: *If the components of* $N\times 1$ *vector* x_t *are cointegrated with* $d=1,\; b=1$ *and with cointegrating rank* r*, then: Statement (4) There exists an error correction representation with* $z_t = \alpha' x_t$ *that is a* $r\times 1$ *vector of stationary random variables:*
>
> $$A(B)(1\text{-}B)x_t = \text{-}\Phi z_{t-1} + \varepsilon_t \qquad (1)$$
>
> *with* $A\left(0\right) = I_N$

where $A\left(B\right)$ is p^{th}-degree polynomials and B is a backshift operator. z_{t-1} are error correction term. $\boldsymbol{\varepsilon}_t$ are zero mean white noise with standard deviation σ. The Granger representation theorem indicates that an error correction model must exist if non-stationary time series are cointegrated, and vice versa. According to the Granger Representation Theorem, the TS-TVEC model expands the prototype error correction model in Eq.1 to accommodate time-space correlations of traffic time series within a regime-switching structure to fit the need for traffic multi-state prediction. The mathematical form of the TS-TVEC model is given by Eq.2.

$$\nabla Y_{0,t} = \begin{cases} \Phi_0^1 + \Phi^1 ECT_{t-1} + \sum_{i=1}^{p} \Gamma_i^1 \nabla Y_{0,t-i} + \sum_{j=1}^{l}\sum_{s=1}^{m} \Pi_{j,s}^1 (Y_{0,t-j} - Y_{s,t-j}) + \varepsilon_t & \text{if } ECT_{t-1} \le \theta_1 \\ \Phi_0^k + \Phi^k ECT_{t-1} + \sum_{i=1}^{p} \Gamma_i^k \nabla Y_{0,t-i} + \sum_{j=1}^{l}\sum_{s=1}^{m} \Pi_{j,s}^k (Y_{0,t-j} - Y_{s,t-j}) + \varepsilon_t & \text{if } \theta_{k-1} < ECT_{t-1} \le \theta_k \\ \Phi_0^{n+1} + \Phi^{n+1} ECT_{t-1} + \sum_{i=1}^{p} \Gamma_i^{n+1} \nabla Y_{0,t-i} + \sum_{j=1}^{l}\sum_{s=1}^{m} \Pi_{j,s}^{n+1} (Y_{0,t-j} - Y_{s,t-j}) + \varepsilon_t & \text{if } \theta_n < ECT_{t-1} \end{cases} \qquad (2)$$

where $Y_{s,t} = \left(x_{s1,t} \quad\quad x_{sn,t} \right)'$ and $x_{sr,t}$ denotes the r^{th} $(r = 0,1, \quad ,n)$ traffic variable at the location s at time t. s denotes the location, s =0 represents the location of prediction, $s \in [1,m]$ denotes the neighborhood site spatially correlated to the location of prediction. ECT_{t-1} is the error correction term defined by $ECT_{t-1} = x_{01,t-1} - \sum_{r=2}^{n} \beta_{r-1} x_{0r,t-1}$; β_{r-1} are the coefficients of linear combination of $x_{0r,t}$. i and j denote time lag. θ_k is the threshold parameter, Φ_0^k, Φ^k, Γ_i^k and $\Pi_{j,s}^k$ are the parameters at the k^{th} $\left(k = 1,2, \quad ,n+1\right)$regime when the ECT_{t-1} is between the threshold θ_{k-1} and θ_k. ECT_{t-1} as a transition variable indicates the deviation from the long run equilibrium. The deviation can be positive or negative where asymmetry may occur and jog among regimes. The coefficient Φ^k indicates adjustment speed for error correction. Different regimes have different adjustment speeds for error correction. The sign of Φ^k is expected to be opposite to the sign of ECT_{t-1}. $\Theta = (\theta_1, \quad ,\theta_n)'$ is the threshold vector that determines the number of regimes and tells when regime switching occurs.

Large scale application

Data source

The TS-TVEC model is investigated with large scale experimentation on the GTA freeway system. Traffic time series are collected from loop detectors deployed on the GTA 400 series of freeways including Highway 400, 401, and 404. Fig.1 shows the locations of loop detectors, the sites of prediction, and related upstream and downstream sites on Highway 401 Exp./Col. east and west bound, and Highway 400 and 404 north and south bound. The data collection stations on the freeways are managed and maintained by the Ministry of Transportation Ontario (MTO), and the traffic data are provided to the ITS lab at University of Toronto for purposes of research.

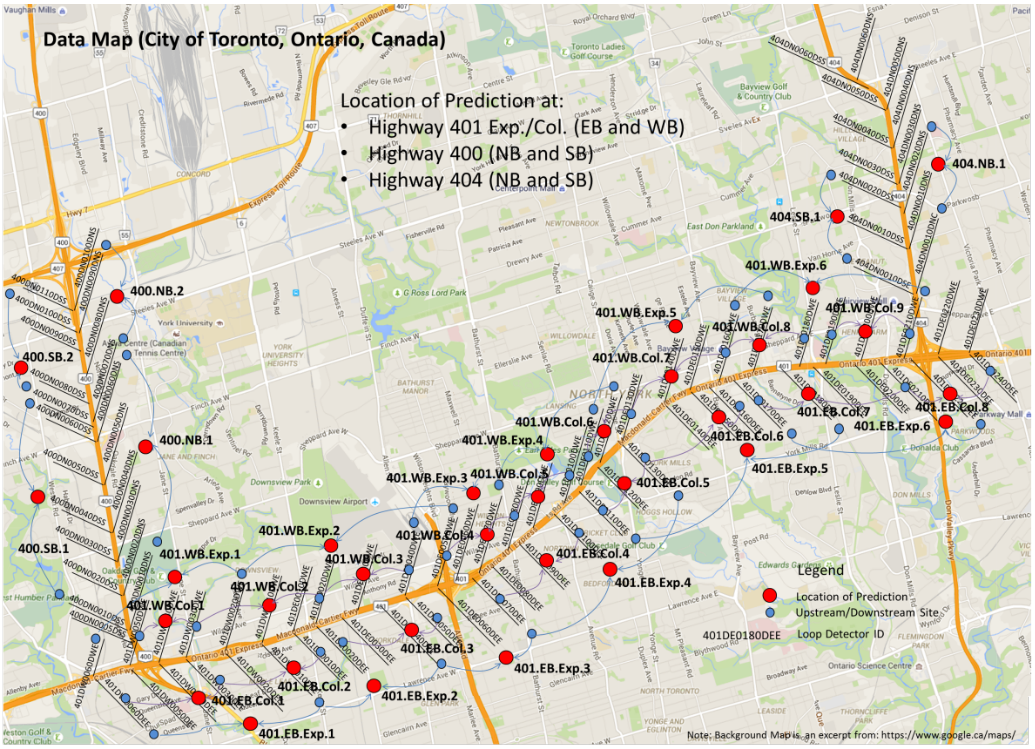

Fig. 1 the GTA 400 series of freeways and data collection stations

In total, the large scale applications of TS-TVEC model are performed at 35 freeway locations with approximately 315 time series. Each site of prediction is denoted by a big red dot on the data map. The locations of prediction selected on Highway 401 focus on core sections between Highway 400 and 404 that assume heavy traffic volumes in, out, and passing through the City of Toronto on a daily basis. These chosen locations cover 12 stretches on the Highway 401expressway and 17 stretches on the collector road. Four locations of prediction are selected on Highway 400 north and south bound between Highway 401 and 407. Two locations are selected on Highway 404 north and south bound between Highway 401 and Finch Avenue.

Statistical test

TS-TVEC model is designed for dealing with complex multivariate time series environment where threshold cointegration effect exists among the time series. Typical hourly traffic time series including volume, speed, and occupancy are chosen from data sets for statistical test. They represent most of traffic time series in the data sets with some variations. Four statistical tests are performed to verify the existence of cointegration and threshold effect among traffic variables. Table 1 shows the results of the Phillips-Ouliaris cointegration test (Phillips and Ouliaris, 1990). The null hypothesis of no cointegration is rejected at 5% significance. The existence of cointegration between traffic variables is further verified by results of Johansen cointegration test (eigen) (Johansen, 1995) shown in Table 2.

Table 1 Phillips-Ouliaris cointegration test

Pair of Traffic Variables	Type	Test Statistic	p-value	Critical Values		
				10%	5%	1%
Volume vs. Speed	Pu	486.69	p < 1%	27.85	33.71	48.00
(q-v)	Pz	515.47	p < 1%	47.59	55.22	71.93
Volume vs. Occupancy	Pu	660.25	p < 1%	27.85	33.71	48.00
(q-o)	Pz	873.49	p < 1%	47.59	55.22	71.93
Speed vs. Occupancy	Pu	729.52	p < 1%	27.85	33.71	48.00
(v-o)	Pz	568.67	p < 1%	47.59	55.22	71.93

Table 2 Johansen cointegration test (eigenvalue)

Pair of Traffic Variables	Rank	Test Statistic	p-value	Critical Values		
				10%	5%	1%
Volume vs. Speed	r<=1	157.53	p < 1%	6.5	8.18	11.65
(q-v)	r=0	211.21	p < 1%	12.91	14.9	19.19
Volume vs. Occupancy	r<=1	201.43	p < 1%	6.5	8.18	11.65
(q-o)	r=0	235.00	p < 1%	12.91	14.9	19.19
Speed vs. Occupancy	r<=1	155.51	p < 1%	6.5	8.18	11.65
(v-o)	r=0	169.20	p < 1%	12.91	14.9	19.19

Hansen and Seo test (Hansen and Seo, 2002) results in Table 3 reject the null hypothesis of linear cointegration at 5% significance and favor the threshold cointegration effect. Bootstrap method is used for estimating asymptotic distribution of the density of test statistics and its critical value p (Hansen, 1999). The Zivot-Andrew unit root test (Zivot and Andrews, 1992) is also performed to verify the structural break in the speed and occupancy time series. Fig.2 shows that the null hypothesis of a unit root process with drift that excludes exogenous structural change is rejected at 5% significance.

Table 3 Hansen and Seo threshold cointegration test

Multivariate	Test Statistic	P-Value	Bootstrap Critical Values		
			0.90	0.95	0.99
Volume vs. Speed	151.93	0.00	13.88	15.32	17.57
Volume vs. Occupancy	126.19	0.00	10.39	12.00	15.11
Speed vs. Occupancy	92.19	0.00	18.78	21.03	24.00

The results of those statistical tests indicate that the threshold cointegration effect exists among traffic volume, speed and occupancy. The necessity of the TS-TVEC model is justified. In addition, the spatial time series that is used as the exogenous term of the model are chosen from the upstream and downstream site where the traffic flow converges or diverges from the traffic flow at the site of prediction.

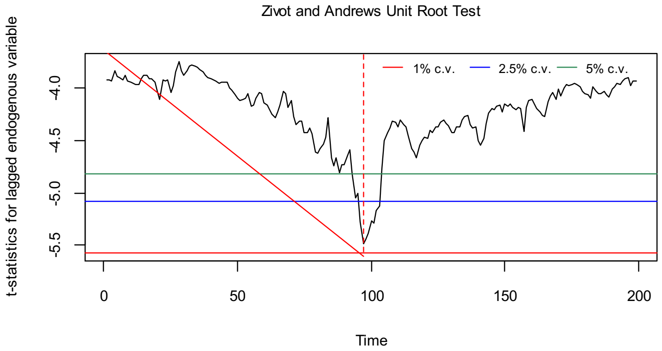

Fig. 2 Structural break test with 200 points of speed series

Model determination and forecast

There are 105 models that are estimated at 35 freeway locations. Model selection concurrently takes into account the MSE, number of regimes, and lags. The cross validation method in conjunction with parsimonious principle is used for model selection. As the TS-TVEC model is a data driven model, its lag order p is varying for different data sets at different locations. The value of p reflects the best model selection given the data set at each location of prediction. The conditional least squares and grid search techniques (Narzo et al., 2014) are employed in this research to identify the model parameters. Statistical diagnostic tests are performed to examine the significance of the model coefficients, normality, and whiteness of noise in the process of model estimation.

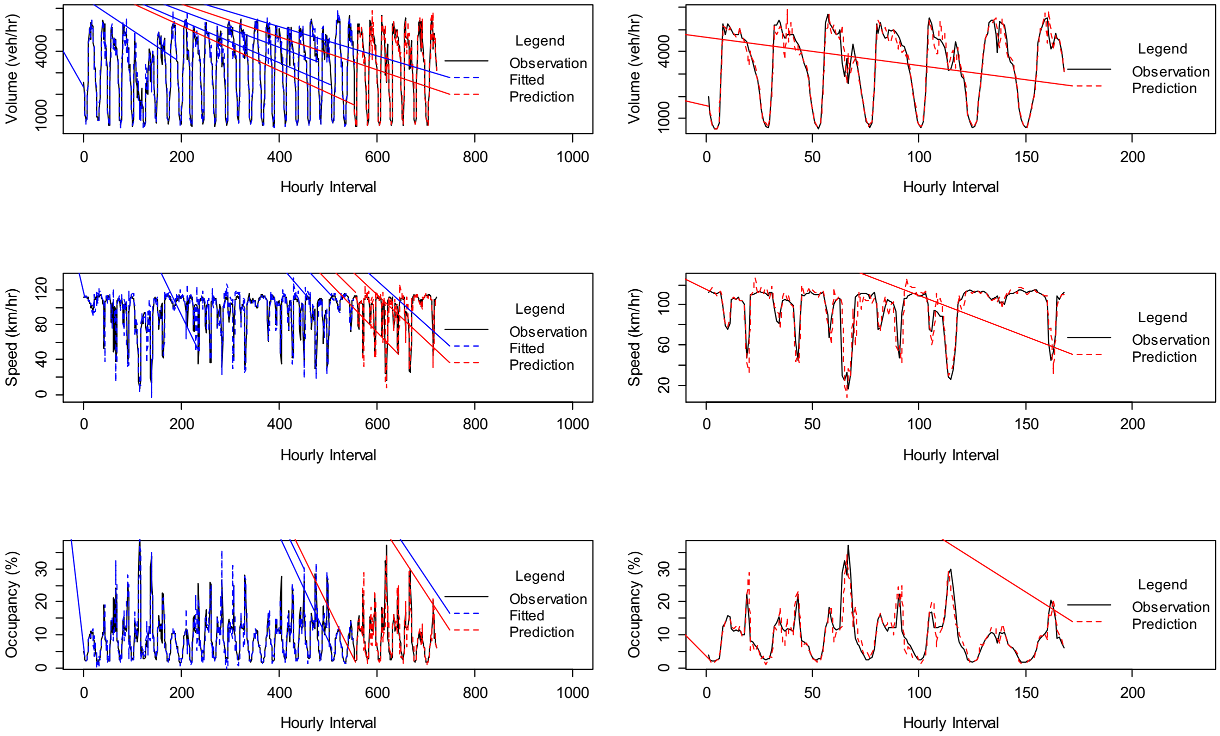

Fig. 3 Model fit and one-step ahead rolling prediction for 7 days
(401EB.Exp.1, Hwy400 ¬ Basket Wave)

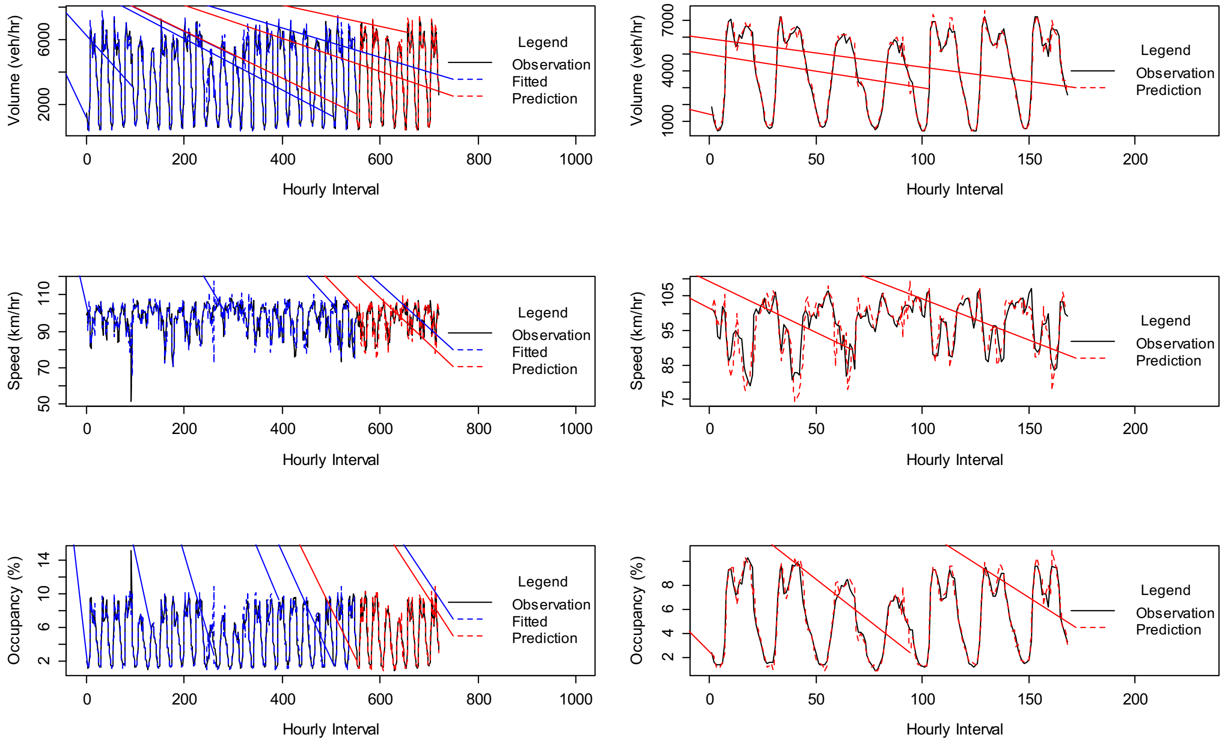

Fig. 4 Model fit and one-step ahead rolling prediction for 7 days
(401WB.Col.5, Bathurst St. ¬ Avenue Rd)

The hourly interval one-step-ahead rolling prediction is performed with the selected TS-TVEC model for seven days in a row for each traffic variable. The prediction includes 168 points of time. Prediction accuracy is assessed by the MSE (mean squared error), Coefficient of Variation, and MAPE (mean absolute percentage error). Exhibitions of model fitness and 168 hourly predictions for each traffic variable from 2 representative sites are shown in Fig.3 and Fig.4 respectively. The complete list of exhibitions of model fitness and 168 hourly predictions from 35 sites can be referenced to (Ma 2016).

Performance of TS-TVEC

According to 35 locations of prediction with 315 time series, Table 4 summarizes prediction accuracy of TS-TVEC model.

Table 4 Summary of model prediction accuracy at 35 locations

Statistics	Volume				Speed				Occupancy			
	MSE	Coefficient of Variation	MAPE	Std Deviation	MSE	Coefficient of Variation	MAPE	Std Deviation	MSE	Coefficient of Variation	MAPE	Std Deviation
min	18864.7	5.87%	4.66%	137.35	16.14	4.16%	3.10%	4.02	0.4	11.65%	9.24%	0.63
max	296926.8	41.29%	23.03%	544.91	284.64	20.43%	16.87%	16.87	41.89	97.40%	62.63%	6.47
median	109632.3	9.29%	7.45%	331.11	142.46	12.93%	9.84%	11.94	8.71	37.87%	20.72%	2.95

The MAPE in traffic volume prediction is between 4.66% and 23.03% with median 7.45%, the coefficient of variation is between 5.87% and 41.29% with median 9.29%, and the standard deviation is between 137.35 and 544.91 with median 331.11. Similarly, the MAPE in traffic speed prediction is between 3.1% and 16.87% with median 9.84%, the coefficient of variation is between 4.16% and 20.43% with median 12.93%, and the standard deviation is between 4.02 and 16.87 with median 11.94. The MAPE in traffic occupancy prediction is between 9.24% and 62.63% with median 20.72%, the coefficient of variation is between 11.65% and 97.40% with median 37.87%, and the standard deviation is between 0.63 and 6.47

with median 2.95. In other words, prediction accuracy of TS-TVEC model is approximately 92.55% for traffic volume, 90.16% for speed, and 79.28% for occupancy at 35 locations of prediction.

The TS-TVEC model shows its advantage in need of a modest data size for model estimation. For instance, one-month hourly data is adequate for a TS-TVEC model to be estimated. One-month hourly data prior to the prediction time is able to provide sufficient information on the most recent dynamics of traffic flow including daily and weekly cyclic patterns. A characteristic of a stochastic process is known as ergodicity that refers to asymptotic independence. Loosely speaking, it means that the further apart two realizations of a time series are with respect to time, the closer to independence they become [Pfaff 2008]. Hence, long time series is not necessarily helpful for improving model estimation and prediction. Traffic time series are short memory time series where observations separated by a long time span exhibit asymptotic independence. Therefore, one-month data is sufficient for the TS-TVEC model to identify the model structure. In practice, TS-TVEC model estimation should be a periodically rolling update process. The most recent one-month data prior to the prediction time should be always used for model estimation.

Conclusion

This research contributes to literature in a few aspects. (1) It discovered the existence of cointegration effect among macroscopic traffic variables. This is beneficial to better understanding the mechanism of the traffic data generating process, thus improving the prediction accuracy; (2) established a vector error correction model for traffic state prediction according to the Granger Representation Theorem; (3) introduced the regime switching structure to capture structural break in traffic time series and reflect multi-states of traffic situation; (4) incorporated spatially correlated information from upstream or/and downstream of the location of prediction into the model to enhance the accuracy of prediction;

Large scale experiments show consistent effectiveness and robustness of the TS-TVEC model. It is our belief that TS-TVEC is a theoretically sound, powerful and competitive method suitable for modelling and forecasting complex multivariate traffic time series where threshold cointegration effect is non-trivial. The model is able to provide accurate predictions, and potentially applicable to a wide variety of traffic circumstances and real time traffic state forecasting.

References

Abdi, J., Moshiri, B., Sedigh A. K., 2010. Comparison of RBF and MLP Neural Networks in Short-Term Traffic Flow Forecasting. IEEE. DOI: 10.1109/ICPCES.2010.5698623

Abdulhai, B., Porwal, H., Recker, W., 2002. Short-term traffic flow prediction using neuro-genetic algorithms. ITS J. 7, 3–41.

Antoniou, C., Ben-Akiva, M., Koutsopoulos, H.N., 2005. Online calibration of traffic prediction models. J. Transp. Res. Rec. 1934, 235–245.

Chang, H., Lee, Y., Yoon, B., Baek, S., 2012. Dynamic near-term traffic flow prediction: system oriented approach based on past experiences. IET Intelligent Transport Systems, 6 (3), 292–305.

Daganzo, C.F., 1994. The cell transmission model: a dynamic representation of highway traffic consistent with the hydrodynamic theory. Transp. Res. Part B 28 (4), 269–287.

Daganzo, C.F., 1995a. The cell transmission model, Part II: Network traffic. Transp. Res. Part B 29 (2), 79–93.

Daganzo, C.F., 1995b. Requiem for second-order fluid approximation of traffic flow, Transportation Res. Part B, 29 (4), pp. 277–286.

Engle, R.F., Granger, C.W.J., 1987. Co-integration and error correction: representation, estimation, and testing. Econometrica 55 (2), 251–276.

Granger, C.W.J., 1981. Some properties of time series data and their use in econometric model specification. J. Econometrics 16, 121–130.

Granger, C.W.J., Weiss, A.A., 1983. Time series analysis of error-correcting models. In: Studies in Econometrics, Time Series and Multivariate Statistics. Academic Press, New York, pp. 255–278.

Hansen, B. E., 1999. Testing for linearity. J. Econ. Surv. 13 (5).

Hansen, B. E., Seo, B., 2002. Testing for two-regime threshold cointegration in vector error correction models. J. Econometrics 110, 293–318.

Johansen, S., 1995. Likelihood-Based Inference in Cointegrated Vector Autoregressive Models. Oxford University Press.

Ma, T., Zhou, Z., Abdulhai, B. 2015. Nonlinear multivariate time–space threshold vector error correction model for short term traffic state prediction. Transp. Res. Part B 76, 27–47

Ma, T. 2016. Nonlinear multivariate time–space threshold vector error correction model for short term traffic state prediction. Ph.D. Dissertation, University of Toronto.

Narzo, A. F. D., Aznarte, J. L., Stigler, M., 2014. Manual of tsDyn Package in R v0.9-41.

Pfaff, B., 2008. Analysis of Integrated and Cointegrated Time Series with R. Springer-Verlag.

Phillips, P.C.B., Ouliaris, S., 1990. Asymptotic properties of residual based tests for cointegration. Econometrica 58, 73–93.

Qiao, F., Yang, H., Lam, W. H. K., 2001. Intelligent simulation and prediction of traffic flow dispersion. Transp. Res. Part B 35, 843–863.

Smith, B. L., Williams, B. M., Oswald, R. K., 2002. Comparison of parametric and nonparametric models for traffic flow forecasting. Transp. Res. Part C 10(4), 303–321.

Treiber, M., Kesting, A., 2013. Traffic Flow Dynamics, Data, Models and Simulation. Springer-Verlag.

Wu, C., Ho, J., Lee, D., 2004. Travel-time prediction with support vector regression. IEEE Transp. Intell. Transp. Syst. 5 (4), 276–281.

Wang, Y., Papageorgiou, M., 2005. Real-time freeway traffic state estimation based on extended Kalman filter: a general approach. Transp. Res. Part B 39, 141–167.

Williams, B. M., Hoel, L. A., 2003. Modeling and forecasting vehicular traffic flow as a seasonal ARIMA process: a theoretical basis and empirical results. J. Transp. Eng. (ASCE) 129, 664–672.

Zheng, W.Z., Lee, D. H., Shi, Q. X., 2006. Short-term freeway traffic flow prediction: Bayesian combined neural network approach. J. Transp. Eng. ASCE 132 (2), 114–121.

Zivot, E., Andrews, D. W.K., 1992. Further Evidence on the Great Crash, the Oil-Price Shock, and the Unit-Root Hypothesis. Journal of Business & Economic Statistics, 10(3), 251–270.