University Transportation Research Center - Region 2

Final Report

A Prototype Decision Support System for Optimally Routing Border Crossing Traffic Based on Predicted Border Crossing Times

Performing Organization: University at Buffalo, SUNY

January 2012

Sponsor:
Research and Innovative Technology Administration / USDOT

University Transportation Research Center - Region 2

The Region 2 University Transportation Research Center (UTRC) is one of ten original University Transportation Centers established in 1987 by the U.S. Congress. These Centers were established with the recognition that transportation plays a key role in the nation's economy and the quality of life of its citizens. University faculty members provide a critical link in resolving our national and regional transportation problems while training the professionals who address our transportation systems and their customers on a daily basis.

The UTRC was established in order to support research, education and the transfer of technology in the field of transportation. The theme of the Center is "Planning and Managing Regional Transportation Systems in a Changing World." Presently, under the direction of Dr. Camille Kamga, the UTRC represents USDOT Region II, including New York, New Jersey, Puerto Rico and the U.S. Virgin Islands. Functioning as a consortium of twelve major Universities throughout the region, UTRC is located at the CUNY Institute for Transportation Systems at The City College of New York, the lead institution of the consortium. The Center, through its consortium, an Agency-Industry Council and its Director and Staff, supports research, education, and technology transfer under its theme. UTRC's three main goals are:

Research

The research program objectives are (1) to develop a theme based transportation research program that is responsive to the needs of regional transportation organizations and stakeholders, and (2) to conduct that program in cooperation with the partners. The program includes both studies that are identified with research partners of projects targeted to the theme, and targeted, short-term projects. The program develops competitive proposals, which are evaluated to insure the mostresponsive UTRC team conducts the work. The research program is responsive to the UTRC theme: "Planning and Managing Regional Transportation Systems in a Changing World." The complex transportation system of transit and infrastructure, and the rapidly changing environment impacts the nation's largest city and metropolitan area. The New York/New Jersey Metropolitan has over 19 million people, 600,000 businesses and 9 million workers. The Region's intermodal and multimodal systems must serve all customers and stakeholders within the region and globally.Under the current grant, the new research projects and the ongoing research projects concentrate the program efforts on the categories of Transportation Systems Performance and Information Infrastructure to provide needed services to the New Jersey Department of Transportation, New York City Department of Transportation, New York Metropolitan Transportation Council , New York State Department of Transportation, and the New York State Energy and Research Development Authorityand others, all while enhancing the center's theme.

Education and Workforce Development

The modern professional must combine the technical skills of engineering and planning with knowledge of economics, environmental science, management, finance, and law as well as negotiation skills, psychology and sociology. And, she/he must be computer literate, wired to the web, and knowledgeable about advances in information technology. UTRC's education and training efforts provide a multidisciplinary program of course work and experiential learning to train students and provide advanced training or retraining of practitioners to plan and manage regional transportation systems. UTRC must meet the need to educate the undergraduate and graduate student with a foundation of transportation fundamentals that allows for solving complex problems in a world much more dynamic than even a decade ago. Simultaneously, the demand for continuing education is growing – either because of professional license requirements or because the workplace demands it – and provides the opportunity to combine State of Practice education with tailored ways of delivering content.

Technology Transfer

UTRC's Technology Transfer Program goes beyond what might be considered "traditional" technology transfer activities. Its main objectives are (1) to increase the awareness and level of information concerning transportation issues facing Region 2; (2) to improve the knowledge base and approach to problem solving of the region's transportation workforce, from those operating the systems to those at the most senior level of managing the system; and by doing so, to improve the overall professional capability of the transportation workforce; (3) to stimulate discussion and debate concerning the integration of new technologies into our culture, our work and our transportation systems; (4) to provide the more traditional but extremely important job of disseminating research and project reports, studies, analysis and use of tools to the education, research and practicing community both nationally and internationally; and (5) to provide unbiased information and testimony to decision-makers concerning regional transportation issues consistent with the UTRC theme.

Project No: 49111 -17 - 20

Project Date: January 2012

Project Title: A Prototype Decision Support System for Optimally Routing Border Crossing Traffic Based on Predicted Border Crossing Times

Principal Investigators: Adel W. Sadek and Qian Wang, University at Buffalo, SUNY

Performing Organization: University at Buffalo, SUNY

Sponsors: RITA/USDOT

Disclaimer

To request a hard copy of our final reports, please send us an email at utrc@utrc2.org

Mailing Address:

University Transportation Reserch Center
The City College of New York
Marshak Hall, Suite 910
160 Convent Avenue
New York, NY 10031
Tel: 212-650-8051
Fax: 212-650-8374
Web: www.utrc2.org

Board of Directors

The UTRC Board of Directors consists of one or two members from each Consortium school (each school receives two votes regardless of the number of representatives on the board). The Center Director is an ex-officio member of the Board and The Center management team serves as staff to the Board.

City University of New York
Dr. Hongmian Gong - Geography
Dr. Claire McKnight - Civil Engineering
Dr. Neville A. Parker - Civil Engineering

Clarkson University
Dr. Kerop D. Janoyan - Civil Engineering

Columbia University
Dr. Raimondo Betti - Civil Engineering
Dr. Elliott Sclar - Urban and Regional Planning

Cornell University
Dr. Huaizhu (Oliver) Gao - Civil Engineering
Dr. Mark A. Turnquist - Civil Engineering

Hofstra University
Dr. Dilruba Ozmen-Ertekin - Civil Engineering
Dr. Jean-Paul Rodrigue - Global Studies and Geography

New Jersey Institute of Technology
Dr. Priscilla P. Nelson - Geotechnical Engineering
Dr. Lazar Spasovic - Civil Engineering

New York University
Dr. Mitchell L. Moss - Urban Policy and Planning
Dr. Rae Zimmerman - Planning and Public Administration

Polytechnic Institute of NYU
Dr. John C. Falcocchio - Civil Engineering
Dr. Elena Prassas - Civil Engineering

Rensselaer Polytechnic Institute
Dr. José Holguín-Veras - Civil Engineering
Dr. William "Al" Wallace - Systems Engineering

Rochester Institute of Technology
Dr. James Winebrake

Rowan University
Dr. Yusuf Mehta - Civil Engineering
Dr. Beena Sukumaran - Civil Engineering

Rutgers University
Dr. Robert Noland - Planning and Public Policy
Dr. Kaan Ozbay - Civil Engineering

State University of New York
Michael M. Fancher - Nanoscience
Dr. Catherine T. Lawson - City & Regional Planning
Dr. Adel W. Sadek - Transportation Systems Engineering
Dr. Shmuel Yahalom - Economics

Stevens Institute of Technology
Dr. Sophia Hassiotis - Civil Engineering
Dr. Thomas H. Wakeman III - Civil Engineering

Syracuse University
Dr. Riyad S. Aboutaha - Civil Engineering
Dr. O. Sam Salem - Construction Engineering and Management

The College of New Jersey
Dr. Michael Shenoda - Civil Engineering

University of Puerto Rico - Mayagüez
Dr. Ismael Pagán-Trinidad - Civil Engineering
Dr. Didier M. Valdés-Díaz - Civil Engineering

UTRC Consortium Universities

The following universities/colleges are members of the UTRC consortium.

City University of New York (CUNY)
Clarkson University (Clarkson)
Columbia University (Columbia)
Cornell University (Cornell)
Hofstra University (Hofstra)
New Jersey Institute of Technology (NJIT)
New York University (NYU)
Polytechnic Institute of NYU (Poly)
Rensselaer Polytechnic Institute (RPI)
Rochester Institute of Technology (RIT)
Rowan University (Rowan)
Rutgers University (Rutgers)
State University of New York (SUNY)
Stevens Institute of Technology (Stevens)
Syracuse University (SU)
The College of New Jersey (TCNJ)
University of Puerto Rico - Mayagüez (UPRM)

UTRC Key Staff

Dr. Camille Kamga: *Director, Assistant Professor of Civil Engineering*

Dr. Robert E. Paaswell: *Director Emeritus of UTRC and Distin*guished Professor of Civil Engineering, The City College of New York

Dr. Claire McKnight: *Assistant Director for Education and Training; Associate Professor of Civil Engineering, City College of New York*

Herbert Levinson: *UTRC Icon Mentor, Transportation Consultant and Professor Emeritus of Transportation*

Dr. Ellen Thorson: *Senior Research Fellow, University Transportation Research Center*

Penny Eickemeyer: *Associate Director for Research, UTRC*

Dr. Alison Conway: *Associate Director for New Initiatives and Assistant Professor of Civil Engineering*

Nadia Aslam: *Assistant Director for Technology Transfer*

Dr. Anil Yazici: *Post-doc/ Senior Researcher*

Nathalie Martinez: *Research Associate*

Sundari Prasad: *Graphic Intern*

University Transportation Research Center

Final Report

A Prototype Decision Support System for Optimally Routing Border Crossing Traffic Based on Predicted Border Crossing Times

Prepared by:

Dr. Adel W. Sadek
Associate Professor

Dr. Qian Wang
Assistang Professor

Department of Civil, Structural and Environmental Engineering
University at Buffalo, the State University of New York

January 2012

University at Buffalo The State University of New York

TECHNICAL REPORT STANDARD TITLE PAGE

1. Report No.	2.Government Accession No.	3. Recipient's Catalog No.
4. Title and Subtitle A Prototype Decision Support System for Optimally Routing Border Crossing Traffic Based on Predicted Border Crossing Times		5. Report Date January 2012
		6. Performing Organization Code
7. Author(s) Adel W. Sadek and Qian Wang		8. Performing Organization Report No.
9. Performing Organization Name and Address Department of Civil, Structural & Environmental Engineering University at Buffalo, the State University of New York 212 Ketter Hall Buffalo, NY 14260		10. Work Unit No.
		11. Contract or Grant No. 49111-17-20
12. Sponsoring Agency Name and Address US DOT Region 2 University Transportation Research Center City College of New York Marshak Hall, 910 New York, New York 10031		13. Type of Report and Period Covered Final
		14. Sponsoring Agency Code
15. Supplementary Notes		

16. Abstract

The economic vitality of the "Golden Horseshoe", a densely populated and industrialized region which encompasses Southern Ontario, Canada and parts of New York State including the Buffalo-Niagara Region, is heavily dependent upon the ability to move goods freely and efficiently across the Canadian-US border. This highlights the critical importance of the Niagara Frontier International border crossing, one of North America's busiest portals for travel and trade. This study had two primary objectives. The first objective was to develop a forecasting method for the on-line, short-term prediction of hourly traffic volumes at the Niagara Frontier border crossings. The second objective of the study was to develop queueing models which would use the predicted traffic volume to estimate the future border delay. As a case study, the project considers the Peace Bridge border crossing, which is one of the busiest Niagara Frontier border crossings, serving over 4.76 million cars annually.

For the short-term prediction of hourly volumes at the Peace Bridge, a novel method was developed, which combines forecasts from traditional time series analysis, specifically the Seasonal Autoregressive Integrated Moving Average (SARIMA) model, with forecasts made by Support Vector Regression (SVR). The two models' forecasts are combined using: (1) a simple fixed weight procedure; and (2) the fuzzy adaptive variable weight method, based on the Fresh Degree Function. Based on an analysis of the diurnal distribution of traffic volumes, six separate classes are defined and individual models are developed for weekdays (Monday - Thursday), Fridays, Saturdays, Sundays, holidays and game days. The evaluation results for that part of the study appear to confirm the hypothesis that, while the SARIMA model does a good job capturing the linear characteristics of the data (e.g., seasonality and trend), SVR appears to outperform SARIMA in modeling the data's nonlinear aspects. The study also shows that combining forecasts from the two models, especially using the fuzzy adaptive variable weight method, yields excellent prediction performance, with values for the Mean Absolute Percent Error in the predictions of only about 7%.

For estimating future border delay, the study built and solved an inverse M/M/c queueing model, where it was assumed that arrivals followed a Poisson distribution, service times followed an exponential distribution, and that the system had c service stations open. An inverse model was needed because the hourly volumes predicted from the first part of the study were outgoing volumes (i.e. volumes leaving the border customs and inspection stations), and not incoming volumes. The report includes an example to demonstrate how the inverse queueing model may be solved to estimate the expected average delay.

17. Key Words Time series analysis, Border Crossing, Support Vector Regression, Queueing Models, Peace Bridge	18. Distribution Statement		
19. Security Classif (of this report) Unclassified	20. Security Classif. (of this page) Unclassified	21. No of Pages 29 pages	22. Price

Form DOT F 1700.7 (8-69)

Contents

INTRODUCTION

The economic vitality of the "Golden Horseshoe", a densely populated and industrialized region which encompasses Southern Ontario, Canada and parts of New York State including the Buffalo-Niagara Region, is heavily dependent upon the ability to move goods freely and efficiently across the Canadian-US border. This highlights the critical importance of the Niagara Frontier International border crossing, one of North America's busiest portals for travel and trade, where approximately 30% of the total Canada-US trade crosses, along with millions of immigrants and tourists every year. In recent years, and as a result of the continued increase in travel demand across the border coupled with the need for tighter security and inspection procedures after September 11, border crossing delay has become a critical problem with tremendous economic and social costs.

Given the critical role that the border crossings play in the overall performance of the Niagara Frontier transportation system and hence the region's economy, the Niagara International Transportation Technology Coalition (NITTEC), a coalition of fourteen different agencies in Western New York and Southern Ontario which serves as the Traffic Operations Center (TOC) for the region, presently provides *current* or *instantaneous* border-crossing travel times to the public, based on currently observed estimates of delay at the border. While *current* border-crossing delay estimates are quite valuable in themselves, the ability to provide short-term forecasts of border crossing traffic volumes and delay would provide several additional benefits.

First, for border crossings, *current* travel times may be quite different from the travel times that the drivers would actually experience by the time they arrive at the border (i.e. *experienced* travel times). This is especially the case if there is a significant lag between the time when a traveler receives (or needs to act upon) the information and the time when he/she arrives at the border (at the present time, border delay estimates are typically updated every hour). Second, the ability to predict border crossing traffic would be quite helpful to customs and border protection authorities in terms of determining the needed staffing level to meet the expected travel demand. Third, with predicted border crossing volumes, intelligent routing algorithms could be developed to optimally route border-destined traffic in a fashion that would minimize the overall system travel time.

PURPOSE AND SCOPE

This study had two primary objectives. The first objective was to develop a novel forecasting method for the on-line, short-term prediction of hourly traffic volumes at the Niagara Frontier border crossings. The focus of that part of the study was on first predicting the border crossing volume, rather than predicting the border delay border delay right way. This is because predicting the volume is a pre-requisite for predicting the delay, assuming information about the border crossing staffing level, the number of customs inspection booths open, and the average inspection time is available. Moreover, predicting border crossing volumes have a direct utility for estimating the required border staffing level and for intelligent routing applications as previously mentioned. The second objective of the study was to develop queueing models which would use the predicted traffic volume to estimate the future border delay. As a case study, we consider the Peace Bridge border crossing, which is one of the busiest Niagara Frontier border crossings, serving over 4.76 million cars annually.

It should be noted that while there is currently an extensive literature on the short-term traffic volume prediction problem and delay estimation, the method proposed in this study is unique in terms of its focus on the problem in the context of US-Canadian border crossing traffic which has several unique features (e.g., significant differences between weekday and weekend traffic, sensitivity to special events such as sporting events and national and religious holidays,.. etc.). Moreover, the study proposes a multi-model combined forecast method which combines both traditional time series analysis, specifically the Seasonal Autoregressive Integrated Moving Average (SARIMA) model, with Support Vector Regression (SVR), an approach to solving classification and regression analysis problems recently proposed by the computer science and artificial intelligence (AI) community, in order to improve the quality of the traffic volume predictions. A third unique aspect of the second part of the study involves the development of an inverse M/M/c queneing model to estimate the average delay based on the traffic volume predicted from the first part of the study. Inverse models were needed because the available hourly traffic counts for the Peace Bridge which were used in the study represented outgoing volumes (i.e. number of vehicles processed per hour) rather than incoming volumes.

This report is divided into two parts reflecting the afore-mentioned two objectives of the study. The first part describes the development of the multi-model combined forecasting method used to predict hourly traffic volumes at the Peace Bridge. The second part, on the other hand, discusses the development of the inverse M/M/c queueing model.

PART I A MULTI-MODEL COMBINED FORECASTING METHOD for the ON-LINE PREDICTION of BORDER CROSSING TRAFFIC at the PEACE BRIDGE

LITERATURE REVIEW

Short-term Traffic Volume Prediction Methodologies

As opposed to long-term volume forecasting which provides Annual Average Daily Traffic (AADT) forecasts for relatively long prediction horizons (sometimes reaching up to 20 years), short-term traffic forecasting uses real-time traffic information from roadway sensors to predict the likely changes in traffic flow for much shorter prediction periods, which typically range between 5 minutes to one hour into the future. Short-term forecasting thus provides the functionality needed for on-line transportation system management and control. Among the numerous methods recently proposed in the literature for short-term traffic volume forecasting, two groups of algorithms or approaches have received wide attention. These are: (1) time-series analysis; and (2) AI-based methods such as Neural Networks (NN) and SVR. Given the extensive number of previous studies in this regard, this section will only provide representative examples of each group.

Time Series Analysis

The SARIMA model is one of the most popular and extensively used time series models for short-term traffic volume prediction. Williams and Hoel *(1)* presented the theoretical basis for modeling univariate traffic condition data streams as SARIMA processes. Smith et al. *(2)* showed that SARIMA performs better than nonparametric nearest neighbor regression on the single point short-term traffic flow forecasting problem. In addition, Cools et al. *(3)* used both

ARIMA with explanatory variables (ARIMAX) and SARIMA with explanatory variables (SARIMAX) to predict daily traffic counts.

A limitation of SARIMA with respect to time series forecasting in general, however, is that the model assumes a linear correlation structure among the time series values, and hence may not be able to capture the non-linearity inherent in complex real-world problems. Given this, some researchers have recently proposed combining time series forecasting with other methods such as AI-based algorithms *(4-6)*. In this report, we propose a similar approach based on combining SARIMA with SVR.

AI-based Methods

NNs are among the AI-based methods most widely used for short-term volume forecasting. Many different types of NNs have been proposed including back propagation networks *(7)*, radial basis function networks *(8)*, resource allocating networks *(9),* and wavelet networks *(10)*. Besides NNs, SVR has also recently been used (e.g. *11,12).* Moreover, Huang and Sadek *(13)* recently proposed a novel AI-based forecasting approach inspired by human memory, called Spinning Network (SPN), and used it for short-term traffic volume forecasting. One of the main advantages of AI-based methods for short-term volume forecasting is their ability to capture the non-linear aspects of real-world phenomena.

Traffic Data Classification

Regardless of the method used for short-term volume forecasting, several studies have shown that the prediction results could be significantly improved if the input data is first classified or clustered *(14, 15)*. This is mainly because daily traffic patterns vary significantly based on whether the prediction is for a weekday or a weekend, and whether it is a normal or a special type of a day (e.g., a holiday or a special event day). Besides improving accuracy, dividing the dataset can also help reduce computation cost *(16)*. Given this, preliminary data analysis is often conducted, prior to modeling, to determine if traffic data classification is necessary.

One approach for classifying the traffic data used for volume forecasting is to divide the traffic data directly based on the type of the day to which they belong such as weekdays, Fridays, Saturdays, Sundays or Holidays. We refer to this method as the simple classification method for traffic data. Given the complex nature of the traffic system, however, more sophisticated classification and clustering methods have also been tried in the literature, including clustering methods based on fuzzy set theory.

Fuzzy C-means Method (FCM) is one of the most widely used fuzzy clustering algorithms. In FCM, given a finite set of data $X = \{x_1, x_2, \ldots, x_n\}$, the algorithm returns a list of c cluster centers $C = \{c_1, c_2, \ldots, c_n\}$ and a partition matrix $U = u_{i,j} \in [0,1], i = 1, \ldots n, j = 1, \ldots, c,$ where each element $u_{i,j}$ tells the degree to which element x_i belongs to clusterc_j *(17)*. In the current method, both simple and FCM were examined and the classification results were compared.

Border Crossing Studies

While several previous studies assessed the economic impact of border crossing delay, very few studies focused on modeling traffic flow conditions at border crossings. Moreover, most of those studies have focused on modeling border crossing delay for off-line planning applications, and not on the on-line prediction model which is the focus of this report. Examples include a study by Paselk and Mannering *(18)* which used duration models, and a study by Lin and Lin *(19)* which proposed a delay model for planning applications. More recently, Kam et el. *(20)* describe the development of a NN for predicting delay at border crossing. However, the data used to develop the model came solely from a simulation model, and not from real-world observations.

MODELING METHODOLOGY BACKGROUND

As previously mentioned, this study proposes a multi-model combined forecast method, based on combining forecasts from SARIMA and SVR. A brief description of both SARIMA and SVR is provided below, followed by a description of the method we used to combine the forecasts from the two models.

SARIMA Model

According to Box and Jenkins *(21)*, a time series $\{Z_t | t = 1,2,\dots,k\}$ is generated by $\mathrm{SARIMA}(p,d,q) \times (P,D,Q)_s$ if:

$$\Phi_P(B^s)\varphi_p(B)\nabla_s^D\nabla^d Z_t = \Theta_Q(B^s)\theta_q(B)a_t \qquad \text{[Equation 1]},$$

where p, d, q, P, D, Q are integers and s is the seasonality; $\varphi_p(B) = 1 - \varphi_1 B - \varphi_2 B^2 - \cdots - \varphi_p B^p$ is the nonseasonal autoregressive operator of p order; $\Phi_P(B^s) = 1 - \Phi_{1s}B^{1s} - \Phi_{2s}B^{2s} - \cdots - \Phi_{Ps}B^{ps}$ is the seasonal autoregressive operator of P order; $\theta_q(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q$ is the nonseasonal moving average operator of q order; $\Theta_Q(B^s) = 1 - \Theta_{1s}B^{1s} - \Theta_{2s}B^{2s} - \cdots - \Theta_{Qs}B^{Qs}$ is the seasonal moving average operator of Q order, $\nabla_s^D = (1 - B^s)^D$ is the seasonal differencing operator of order D, $\nabla^d = (1 - B)^d$ is the nonseasonal differencing operator of order d; and a_t is the estimated residual at time *t* – assumed to be distributed as $WN(0, \sigma^2)$.

SARIMA models have been widely and successfully applied to time series forecasting over the last thirty years. By following the three steps of identification, estimation, and diagnostics checking, SARIMA models can be fitted to stationary or weakly stationary time series data.

Support Vector Regression Model

Support Vector Regression (SVR) shares a lot of benefits of the Support Vector Machine (SVM) concept which was proposed by Vapnik in 1995. SVR is based on the structured risk minimization principle. Rather than finding empirical errors, SVR tries to minimize an upper bound of the generalization error. Over the last few years, SVR has been employed to solve several nonlinear regression problems *(22, 23).* Mathematically, the SVR model can be described as follows:

Given a set of data points $\{(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)\} \subset \mathcal{X} \times \mathbb{R}$, where $\mathcal{X}$ denotes the space of the input patterns and m is the total number of training samples, a linear regression function can be stated as $f(x) = <\omega, x> + b$ with $\omega \in \mathcal{X}, b \in \mathbb{R}$ where <,> denotes the dot product in $\mathcal{X}$. Assuming an $\varepsilon -$ insensitive loss function *(24)*, a function intended to allow for ignoring errors that fall within a certain band or distance from the true value, the ω and b are obtained by solving Equation 2 below.

$$\text{minimize } \frac{1}{2}\omega^T\omega + C\sum_{i=1}^{m}(\xi_i^+ + \xi_i^-) \qquad \text{[Equation 2]},$$

$$\text{Subject to} \begin{cases} y_i - < \omega, x_i > -b \leq \varepsilon + \xi_i^+ \\ < \omega, x_i > +b - y_i \leq \varepsilon + \xi_i^-, \\ \xi_i^+, \xi_i^- \geq 0 \end{cases}$$

where,

ε (≥ 0) is the maximum deviation allowed,

C (> 0) is the associated penalty for excess deviation during the training,

and ξ_i^+, ξ_i^- are the slack variables corresponding to the size of this excess deviation for positive and negative deviations, respectively.

In the process of solving this optimization problem, SVR achieves nonlinear regression by mapping the training samples into a high dimensional kernel induced feature space, followed by linear regression in that space. Since the kernel mapping is implicit (depends only on the dot product of the input data vectors), it is possible to map the data to a very high dimension, and still keep the computational cost low. Radial Basis Function (RBF) $K(x_i, x_j) = \exp\left(-\gamma|x_i - x_j|^2\right)$ is one common kernel function. The parameters of the SVR models, namely the penalty factor C and the gamma in the kernel function, are often optimized using the k-fold cross validation method *(25, 26)*.

Combination Forecast Using the Fuzzy Adaptive Variable Weight Method based on Fresh Degree Function

As mentioned above, several researchers have recently proposed combining forecasts from more than one model to improve accuracy. This study used the fuzzy adaptive variable weight method based on fresh degree function to do this. The method is adaptive in the sense that the weights assigned to each model, when combining the forecasts, is a function of how well that particular model performed on recent forecasts. Furthermore, the use of the Fresh Degree Function *(27)*, F(t), which is usually a function of the time series index, t (e.g., t, or t^2 or $\sqrt{t}$), allows one to weigh the performance on the most recent forecast more heavily than prior forecasts. Mathematically, the method proceeds as follows.

The prediction error of model j at time index i is first computed according to Equation 3

$$e_j(i) = y(i) - f_j(i) \qquad \text{[Equation 3]},$$

where,

$e_j(i)$ is the prediction error of model j at time i,
$y(i)$ is the real value,
$f_j(i)$ is the prediction value of model j,

Following this, the method calculates the weighted average absolute prediction error for model j using the Fresh Degree function as shown in Equation 4

$$a_j(i) = \sum_{p=0}^{q} F(i-p)\left|e_j(i-p)\right|\left[\sum_{p=0}^{q} F(i-p)\right]^{-1} \qquad \text{[Equation 4]},$$

where,

$F(t)$, the Fresh Degree Function $=$ t or t^2, or $\sqrt{t}$, $t = 1,2,\dots,s$.

q = the length of the time series (i.e., the number of previous data points) used to calculate the weighted average absolute error, $a_j(i)$ based on the Fresh Degree Function.

Next the method calculates the sum of the absolute prediction error for model j at time step i, $s_j(i)$, from Equation 5. In doing so, the method typically uses a length from the time series, l, which is typically longer than the length q used in equation [4] above.

$$s_j(i) = \sum_{p=0}^{l-1} |e_j(i-p)| \qquad \text{[Equation 5]},$$

The method then calculates the following; (1) $E_j(i)$, which is the ratio of the weighted average absolute prediction error over the past q data points of prediction method j at time i, $a_j(i)$, to the largest $a_j(i)$ of all prediction methods, m (i.e., the max $a_j(i)$); and (2) $E_{jA}(i)$, which is the ratio of the weighted average absolute error difference over the past q data points of prediction method j at time i, , $a_j(i)$, to the largest sum of the l absolute error difference over the past *l* data out of all prediction methods, m (i.e. the max $s_j(i)$), as follows.

$$E_j(i) = a_j(i) / \max_{1\le j\le m}[a_j(i)] \qquad \text{[Equation 6]},$$

$$E_{Aj}(i) = a_j(i) / \max_{1\le j\le m}[s_j(i)] \qquad \text{[Equation 7]},$$

Finally, $E_j(i)$ and $E_{jA}(i)$, are combined using equation 8, which was derived based on a fuzzy set formulation, where α is another weighting parameter reflecting the importance of the more recent forecasts.

$$\tilde{k}_j(i) = 1 - [\alpha E_j(i) + (1-\alpha)E_{Aj}(i)] \qquad \text{[Equation 8]},$$

The $\tilde{k}_j(i)$ are then normalized according to Equation 9, and used to derive the weight, f(i+1), for each prediction method, m, at time step (i+1) according to Equation 10.

$$k_j(i) = \tilde{k}_j(i) / \sum_{j=1}^{m} \tilde{k}_j(i) \qquad \text{[Equation 9]},$$

$$f(i+1) = \sum_{j=1}^{m} k_j(i) f_j(i+1),\ j = 1,2,\dots,m, i = 1,2,\dots \qquad \text{[Equation 10]}$$

MODELING DATASET

Three main International bridges connect western New York and Southern Ontario, Canada over the Niagara River, namely the Queenston-Lewiston Bridge, the Rainbow Bridge and Peace Bridge. In this report, the focus is on predicting the next hour traffic volume on the Peace Bridge, and specifically traffic entering the United States from Canada. Hourly traffic volume and classification counts for the Peace Bridge, since 2003, can be downloaded from the Buffalo and Fort Erie Public Bridge Authority's website. In this study, we mainly used the 2009 and 2010 passenger car traffic data. The data quality appeared to be excellent with very few hourly traffic counts missing (only 9 out of 8760 in 2009 and 7 out of 8760 points in 2010 were missing). To complete the time series, the missing data points were replaced with the average of the count before and the count after the missing point. Given that there were only few data points that were missing, the study did not feel the need to utilize more elaborate imputation algorithms such as those described in *(28).* However, the data was still pre-processed to identify outliers, if any, that can distort the modeling results. This was done using the Frequent Pattern (FP) Tree algorithm as described next.

Data Preprocessing using Frequent Pattern Tree (FP-TREE) Algorithm

Due to various reasons, such as detector malfunctions, transmission distortions, traffic accidents or other possible influence factors, traffic data sets often contain data points that do not comply with the general behavior of the data model. Those data points are often called outliers and need to be detected before the modeling process can start. The outlier detection method used in this project utilized the Frequent Pattern Tree (FP-TREE) algorithm *(29, 30)*. The FP algorithm works by first identifying the frequent patterns in the traffic time series. Patterns are then sorted according to their frequencies, and after sorting, the patterns with the least frequencies are flagged as suspect for containing outliers. The algorithm also helps compress the dataset it is analyzing, which is an attractive feature, given the relatively large dataset considered in this study.

The first step in applying the FP-TREE algorithm to our dataset involved first condensing the time series into a smaller dataset to make the application of the algorithm more practical from a computational cost standpoint. This was done by representing each day-worth of traffic data using only those data points within that day that can capture the overall trend of traffic volume variation over the 24-hour period. Specifically, if the hourly volume met any of the following conditions, it was kept; if not, the hourly volume was dropped:

(1) If the hourly volume belongs to the local maximum or local minimum, keep it *(31)*. Identifying the local maximum or minimum was performed as follows. If there are indices i, j, and m, where $i < m < j$, such that: a_m is the minimum among $a_i, \dots, a_j$, and $a_i/a_m >= R$ and $a_j/a_m >= R$. R is set arbitrarily. Here R is set as 2, if the hourly volume is larger than 100, otherwise, 2.5.

(2) If the hourly volume increases or decreases suddenly and the ratio of this hourly volume to its forward one is greater than 3 times or smaller than 1/3 times, keep it.

(3) If the ratio of the hourly volume to the mean of the same time volume of last week and two weeks earlier is greater than 2 or smaller than 1/2, keep it.

After finding and only keeping the most important points of the day, k-means clustering method was used to group the volumes of the different days corresponding to each hour of the day into 3 clusters, and for each cluster we record the lowest and largest volume value for each hour. In other words, after this step, a 24 by 6 matrix is obtained (the 24 rows correspond to the

24 hours of the day, and the 6 columns correspond to the lowest and the largest volume value of each cluster; because we have 3 clusters for each hour, so for each row of the matrix, there should be 6 numbers).

The next step in applying the FP-TREE algorithm involved determining the number of data points (i.e. the data size) in each of the 72 (24*3) clusters. The clusters were then sorted according to the data size from the cluster having the largest number of data points to the one with the lowest. The ranking of each cluster thus serves as a measure for of how frequent the traffic patterns belonging to that cluster are. The clusters are now order ranked, and for each cluster, a record is kept of the lowest and highest volume range as previously mentioned (i.e. at this point, we would have a 72 * 2 matrix, with the row number of each matrix serving as an indicator of the rank or frequency of the traffic patterns contained within each cluster).

Following this, applying the FP-TREE algorithm involved rewriting the traffic volumes, which were kept for each day, in the following manner. For each day (out of the 730 days corresponding to the two years considered in this study), each of the traffic volume values of the important points retained for that day are checked to identify which cluster in the 72 by 2 matrix it belongs to. The rank order of that cluster in the 72 by 2 matrix is then used to replace the actual traffic volume value. After this is done for all the traffic volume values retained for a given day, the values (i.e. the rank order) are sorted from the smallest to the largest for that day. For example, assume there were three important points or traffic volumes for a given day (say the volumes for the 8:00 am, 11:00 am, and 6:00 pm), and assume that the recorded volumes for those three hours were 500, 600 and 650 vehicles per hour. The recorded volumes were checked against the 72*2 matrix, and it was determined that they belong to the clusters represented by the 32^{nd} row, 15^{th} row, and 20^{th} row, respectively. In that case, we would use '32', '15', and '20' to represent the traffic volume recorded for the 8:00 am, 11:00 am, and 6:00 pm. Finally, after sorting, that day would be represented by the vector [15, 20, 32].

Now we have a new representation for the 730 days. Beginning with the first point, the days which have the same first point are grouped together, and the number of the occurrences is registered as count, the first point is called as root node. For the days which have the same root node, the algorithm proceeds to find the days which have the same second point and the number of occurrences are also counted, and so on. Figure 1 below shows an example. The number before the colon refers to the rank order in the matrix (i.e. the cluster rank which had replaced the traffic volume value for the important hours for each day), while the number after the colon records the occurrence times (i.e. the frequency) of that order. The traffic patterns for those days which only occur 1 or 2 times are finally identified as being suspect for outliers.

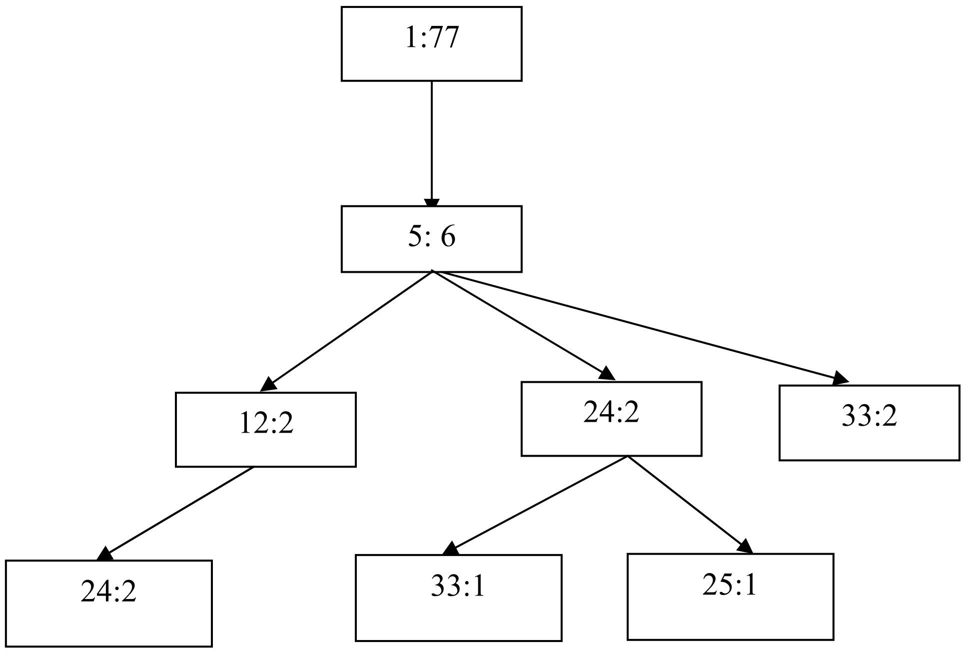

Figure 1 one Example using Frequent Pattern Tree Method

When the FP-TREE algorithm was applied to our dataset, there were a total of 47 days which were flagged as suspect outliers. However, when those days were eliminated, the accuracy of the prediction models did not improve significantly. As a result, it may be concluded that no significant outliers needed to be eliminated from the dataset, and that the quality of the data we were dealing with was very good.

Data Classification

As previously mentioned, there are two ways to divide or classify traffic data before developing the prediction models: a) the simple method based on the type of the day such as weekdays, Fridays, Saturdays and so on; b) more sophisticated methods such as the FCM previously mentioned. Both methods were applied in this study as described below.

Simple Classification Method

Given that daily traffic patterns vary significantly depending upon the type of the day (i.e. weekday vs. weekend vs. a holiday, etc.), the first step in the analysis was to come up with a logical classification scheme that would identify the number of distinct traffic patterns or types of days. For each distinct pattern or day type, we would then develop a separate prediction model. Our approach was to try to tie the classification scheme to easily identifiable properties of a given day in order to facilitate implementation of the prediction process.

To do this, we first considered the group of what may be called "ordinary days", which we defined as consisting of all weekdays, excluding Fridays (i.e. Monday through Thursday). We also excluded the days which, we suspected, might need separate groups (i.e., holidays and game days). We then calculated the mean hourly traffic volume for each hour of the day, and defined an interval of ± 15% of the average hourly volume (15% was chosen based on what may be regarded as acceptable prediction accuracy for the models). The traffic patterns of the "special" days (i.e. holidays and game days) were then compared to the "ordinary days" to determine whether they differed enough to warrant having their own group (i.e., whether they lied within the ± 15% band or not). Figure 2a, Figure 2b and Figure 2c show how two holidays (one Canadian and one US) and one day with a US sporting event all have traffic patterns that differ from "ordinary days" (i.e. they fall outside the ± 15% band).

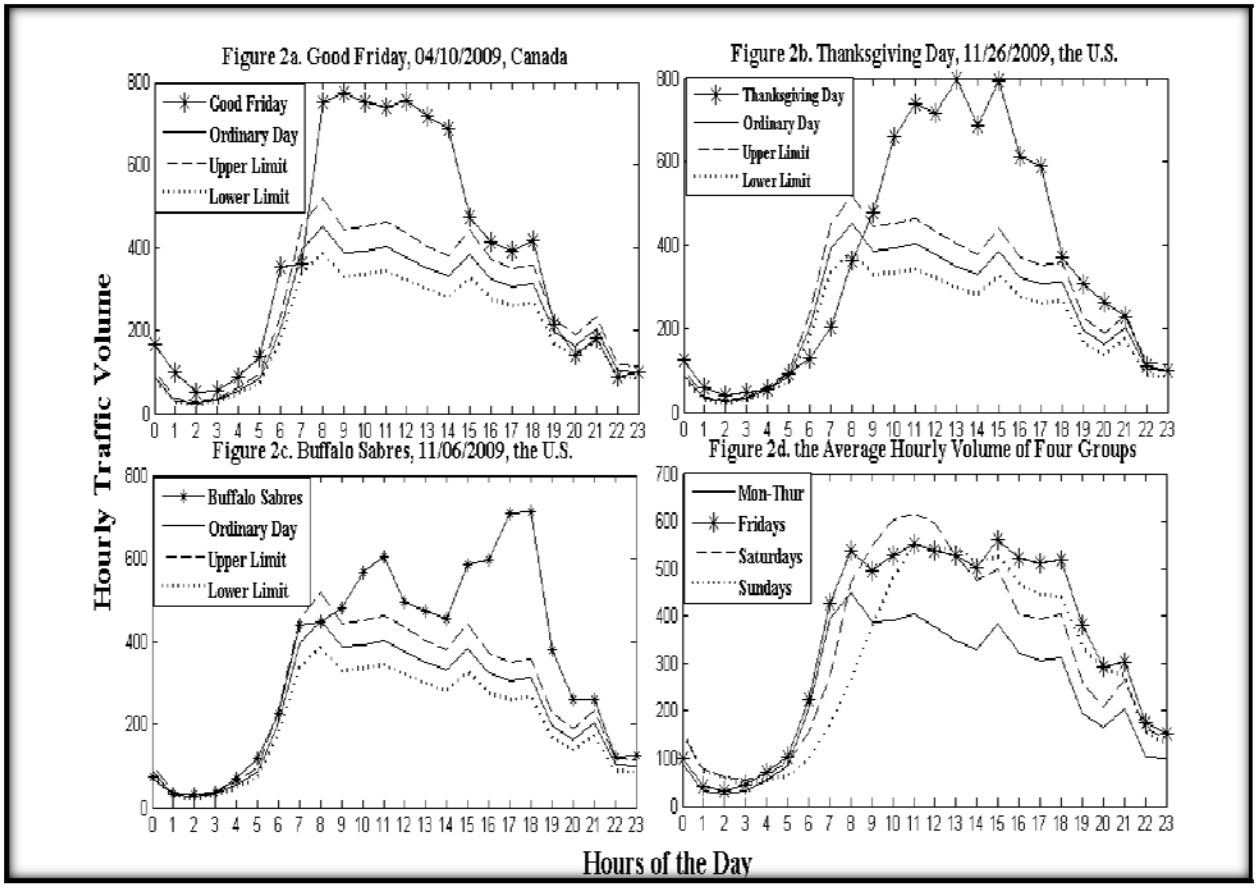

Figure 2 Hourly traffic volume at the Peace Bridge on different days

By following this procedure, the study identified the holidays with traffic patterns that differ significantly from "ordinary days", and a separate group or cluster was defined for those holidays. In total there were 22 such holidays in 2009, 5 of which belonged to Canadian holidays, while the other 17 days including three long weekends: Thanksgiving, Christmas and New Year's holidays) were U.S. or common holidays between the two countries.

For game days, the analysis showed that the days when the Buffalo Sabres' and Buffalo Bills' games were held, had a significantly different trend from ordinary days. As a result, a separate group was defined for those days for which a separate prediction model was developed. In total, there were 50 game days in 2009, 3 of which were also holidays, and 48 game days in 2010.

The study also looked at how traffic patterns generally varied with the day of the week. Figure 2d plots the average hourly traffic volumes for four additional groups (i.e. weekdays excluding Friday, Fridays, Saturdays and Sundays). As can be seen, the diurnal distributions for these four groups differ significantly and warrant defining separate groups.

Based on this, six different groups, for which separate prediction models were developed, were defined: (1) weekdays excluding Fridays (a total of 181 days in 2009 - only 2009 data were used for this group since there were enough data for the analysis); (2) Fridays (35 days in 2009 and 37 days in 2010); (3) Saturdays (with 38 and 41 days in 2009 and 2010, respectively); Sundays (with a total of 83 days in 2009 and 2010); (5) game days (a total of 98 days in 2009 and 2010); and (6) holidays (a total of 22 days in 2009). The following section will describe the process of model development and evaluation for the first five groups (for holidays, given the small data size, a different methodology is being devised based on a case-based reasoning approach). Finally, it needs to mention that the data used in the study included traffic volumes under non-recurrent events (e.g. accidents, emergencies, inclement weathers, etc.). No attempts were made to screen or exclude those data points.

Fuzzy Classification Method (FCM)

The second classification method which the study examined was the Fuzzy Classification Method (FCM). The FCM was applied only to the 2009 on a trial basis, to assess the feasibility of the application of the method and how its results compared to the simple classification method described above. The first step in applying this method was to first put the hourly traffic volumes for the year 2009 in the form of a 21 x 365 matrix as explained below.

The first 15 rows are the hourly volume from the 7:00 am to 9:00 pm (the traffic volume for the other time periods (i.e. from 10:00 pm to 6:00 am) is generally low and therefore prediction is not that important for those periods, and hence were excluded from the analysis). The 16th row expresses the ratio of the total 15 hours' volume for the traffic from the US to Canada to the 15 hours' volume of the opposite direction (i.e. from Canada to USA). The reason for adding this row or variable is the observation that the difference between the two directions traffic volume is significantly larger on holidays and other special days compared to normal days; this variable can thus serve as an indicator for those special events. The 17th row is the mean value of the 15 hours' volume from Canada to USA. The 18th row is the peak hour of the day. The 19th row is to assign a number for different day of the week, for Monday to Thursday, '1' is assigned, for Friday, '2' is assigned, for the weekend, '3' is assigned. The 20th row represents a binary variable to indicate whether there were a Buffalo Sabres or Buffalo Bills game on that day (i.e. it assumes a value of '1' if a game were taking place and '0' otherwise). The 21st row, on the other hand, encodes an indicator variable for days surrounding a holiday. That variable takes the form of a 5 digit binary code, which is used to represent 2 days before and 2 days after the subject day with the 3rd digit representing subject day; holiday-influenced days are coded as 1 and non-holidays as 0.

After the matrix was developed, the FCM was applied to analyze the relationship of the 365 days in 2009. At the end of the process, the method resulted in dividing the 2009 traffic data into a total of five clusters: cluster 1 (144 days), cluster 2 (93 days), cluster 3 (74 days), cluster 4 (32 days), cluster 5 (22 days). It was hoped that the application of the FCM would result in clusters that can be identified by readily available identifiers such as day type (i.e. weekday, weekend, holiday, etc..). However, when we checked the nominal characters of the five clusters, we found that while cluster 1 was relatively pure, with 135 days out of the 144 days contained within it corresponding only to weekdays (Monday to Thursday), there were no clear characteristics to distinguish the other four clusters, which ended up with data from weekdays, Fridays, Saturdays, Sundays, Game days and Holidays mixed together. Given this, the FCM was not pursued any further in this research.

MODEL DEVELOPMENT

Prediction Accuracy Measure

Two measures were utilized in this study to assess the accuracy of the models developed, namely: (1) the Mean Absolute Percent Error (MAPE) as defined in Equation 11; and (2) the Root Mean Square Error (RMSE) as in Equation 12.

$$\mathrm{MAPE} = \frac{1}{n}\sum_{t=1}^{n} \left| \frac{A_t - F_t}{A_t} \right| \qquad \text{[Equation 11]},$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(A_t - F_t)^2} \qquad \text{[Equation 12]},$$

where,

A_t is the actual value and F_t is the forecast value.

It is noted however that in calculating both MAPE and RMSE, only the hourly volume from 7:00 am to 21:00 (i.e. 9 pm) from each day is utilized. This is because the hourly volume values for the other hours (i.e. from 10 pm to 6:00 am) are usually very small (less than 100 vehicles per hour), and hence small errors for those hours may result in artificially high MAPE values.

SARIMA Model

In this study, the Statistical Package for Social Sciences (SPSS) was used to build the SARIMA model. With the data classified into the six groups defined above, the seasonal cycle for each model was set to 24, corresponding to the 24 hours in a day. Because SARIMA models are typically used for off-line modeling of time series datasets, a simple procedure was developed to allow the SARIMA model to be used for on-line prediction. This procedure basically involved fitting the SARIMA model first using a training data set made of the first part of the time series (how the length of the training data was determined will be discussed next), and then recalibrating the SARIMA model to update the model parameters after each prediction. When recalibrating the model, the most recent observation is added to the training data set, and the first or oldest data point in the training time series is dropped to keep the computational burden of model recalibration manageable. This process, which was automated using a syntax file in SPSS, results in what may be viewed as a moving window that updates the part of the time series used for model calibration. It should be noted that although there are more mathematically rigorous methods available to do so (e.g. a state-space representation of the SARIMA model coupled with a Kalman Filter *(32)*, the simple procedure described above was deemed adequate for the purposes of this study, especially since the computation time required for recalibrating SARIMA after each prediction was very short.

To determine the appropriate length of the training dataset, different lengths were tried and the prediction error on a test set consisting of a total of 888 data points (i.e. 37 days or 20% of the total time series for weekdays) was calculated for each length tried. The results are shown in Figure 3a where it can be seen that, for the weekday group, a length of a training dataset corresponding to 960 hours (or 40 days) yielded the best performance.

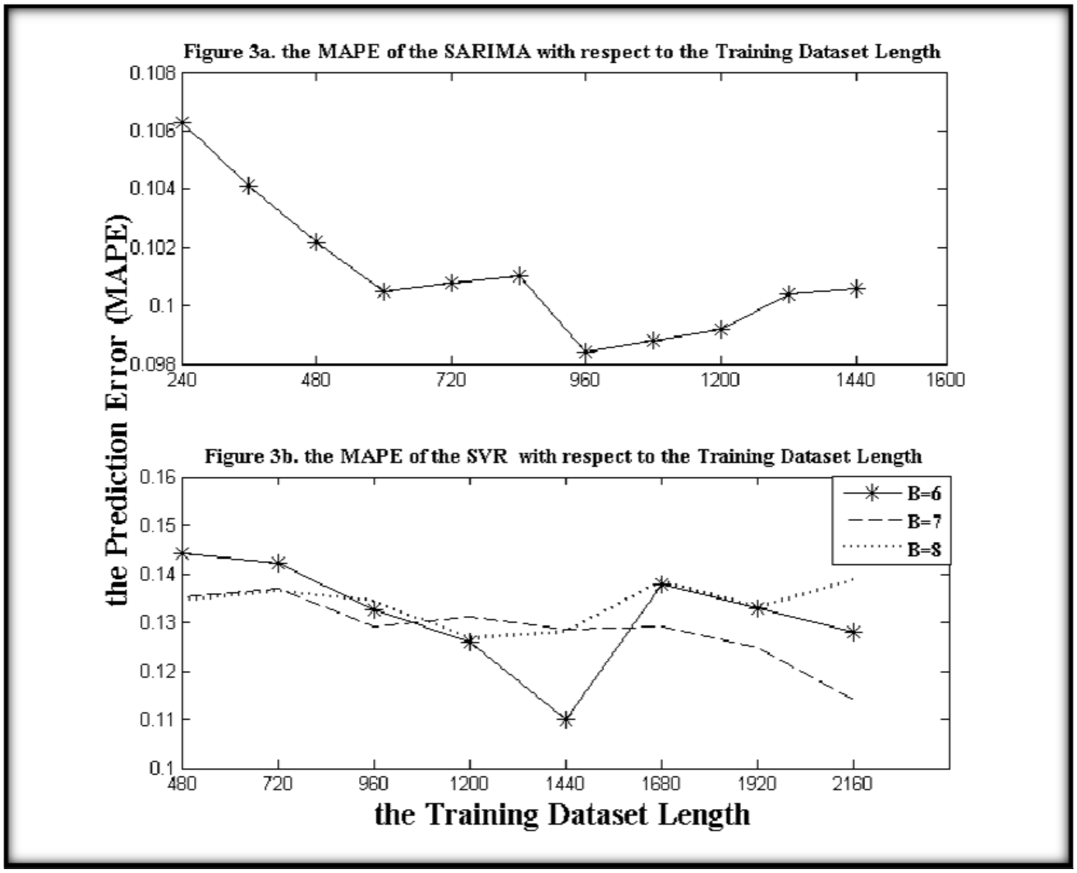

Figure 3 the Parameter Selection of the two Models

The same procedure was followed to develop the other four models for the other classes (i.e. Fridays, Saturdays, Sundays and game days). Table 1 summarizes the results. Specifically, the table lists the size of the training and test datasets for each model, along with the model's MAPE and RMSE. As we can see from Table 1, with the exception of the game days, the MAPE for the models was between 10% and 11%. For game days, the MAPE was slightly higher (around 15%). Also note that for SARIMA, the best performance was achieved when predicting weekday traffic.

Classes	Total Dataset	Training Dataset Length	Test Dataset	MAPE	RMSE
Weekdays (Mon Thru Thurs)	4344 (181 days)	960 (40 days)	3384 (141 days)	9.84%	47.95
Fridays	1728 (35 days in 2009; 37 days in 2010)	960 (40 days)	768 (32 days)	10.28%	64.74
Saturdays	1896 (38 days in 2009; 41 days in 2010)	720 (30 days)	1176 (49 days)	10.8%	57.60
Sundays	1992 (42 days in 2009; 41 days in 2010)	720 (30 days)	1272 (53 days)	11.6%	57.38
Game Days	2352 (50 days in 2009; 48 days in 2010)	1440 (60 days)	912 (38 days)	15.17%	78.91

Table 1 Prediction Performance of the SARIMA Model

SVR Model

For the SVR model, the value of epsilon ε was set as 0.01, and the Radial Basis Function (RBF) was chosen as the kernel function. As opposed to the SARIMA model, SVR can be easily adapted for on-line prediction. Specifically, for this study, the input to the SVR model at a given time step is a vector, X(t), whose length is B, defined as $X(t) = [x(t), \ldots, x(t - B + 1)]^T$. The SVR would then use X(t) to predict the next data point in the series, X(t+1). In other words, the model would always use the most recent B data points or hours to predict the next hour.

For the SVR model, the value of the insensitive loss function (ε)was set as 0.01, and the Radial Basis Function (RBF) was chosen as the kernel function. Before using the SVR model in this fashion, however, it needs to be calibrated by determining the optimal values for the cost factor C and the gamma parameter of the RBF. To do this, a training data set similar to the one used in conjunction with the SARIMA model, needed to be defined; we use the letter O to refer to the length of the training dataset. Moreover, to improve accuracy, we recalibrate the SVR model (i.e., determine new values for C and gamma) every P hours (P=120 hours in this study). The reason the SVR model is only recalibrated every 120 time steps and not after each prediction as was the case with SARIMA, is that the calibration process is computationally extensive and hence cannot be practically performed after each prediction. Once again, to keep the size of the training data set, O, fixed, we adopted a moving window strategy which replaces the first P data points in O with the latest P points added.

To determine appropriate values for both B and O, the values of those two parameters were varied, and the MAPE for a total of 10 days was calculated for each combination. Figure 3b shows the results from these experiments performed on the weekday class. As can be seen, the combination of B=6 and O=1440 appears to yield the smallest prediction error. Given this, these values were the ones adopted for that class.

Table 2 shows the results of the SVR predictions for the five classes, along with the size of the calibration or training dataset (O) and the test dataset adopted for each class. As can be seen, the MAPE for the SVR ranged from 9.42% to 13.62%. A very interesting observation regarding the SVR model performance is that the smallest prediction error for SVR (MAPE of

9.42%) was that corresponding to the game day model. This is in stark opposition to the SARIMA model where the game day SARIMA model had in fact the largest MAPE (15.17%). The RMSE of the game day SVR model is also lower than that in SARIMA model. This in turn tends to support the observation made earlier that traditional time series, given its assumption of linear correlation among the data points, may face difficulties when dealing with the non-linearity of complex patterns such as those observed on game days. For such patterns, the SVR paradigm appears to provide better performance. Again this tends to reinforce the benefits of multi-model combined forecasts as will be discussed in the next section.

Classes	Total Dataset	The Prediction O	Input Dimension B	Test Dataset	MAPE	RMSE
Weekdays (Mon Thru Thurs)	4344 (181 days)	1440	6	2904 (121 days)	10.37%	49.64
Fridays	1728 (35 days in 2009; 37 days in 2010)	1440	7	288 (12 days)	9.87%	58.28
Saturdays	1896 (38 days in 2009; 41 days in 2010)	1440	6	456 (19 days)	12.50%	64.23
Sundays	1992 (42 days in 2009; 41 days in 2010)	1440	6	552 (23 days)	13.62%	57.03
Game Days	2352 (50 days in 2009; 48 days in 2010)	1440	6	912 (38 days)	9.42%	52.17

Table 2 Prediction Performance of the SVR Model

Multi-model Combined Forecasting Method

The comparison of the SARIMA and SVR model results with the observations indicates that each model appears to have its own set of strengths and weaknesses. For example, the SARIMA model is good at handling the linear characteristics of the data, such as seasonality and trend, whereas SVR is capable of capturing the nonlinear characteristics. Given this, this section develops methods to combine forecasts from the two models to improve the quality and accuracy of the predictions.

Specifically, two methods for combining the forecasts from the two models were investigated. The first method, called the simple or fixed weight method, simply compares the performance of each model (i.e. SARIMA and SVR) for predicting a specific hour of the day (e.g. 7:00 am or 8:00 am, etc.), over the whole training dataset. If for example there were more instances in which SARIMA performed better than SVR for the 7:00 am volume prediction, SARIMA is selected for all future predictions for the 7:00 am hour (i.e., the weight assigned for SARIMA in this case would be 1, and for SVR 0). Our analysis of weekday predictions, for example, showed that SARIMA appeared to be the better model for 7 and 8 am, and for 2, 3, 6, 8 and 9 pm. For the remaining hours, SVR was the better model.

The second method is the Fuzzy Adaptive Variable Weight method as previously described. When applying this method, the Fresh Degree Function, F(t), was assumed to be equal to t^2. It should be noted that in this method the performance of the model for predicting a given hour, say 7 am, is evaluated based on the model's performance in predicting that same hour over the past few days and not on the model's performance over the past few hours of that day (i.e., the t index of F(t) is an index referring to the day number or sequence). This is because, as mentioned before, each model outperforms the other for predicting certain hours of the day. With respect to the other parameters of the weight method, q in Equation [4] was set to 3, the moving window length l in Equation [5] was set to 5, and the values of α in Equation [8] were set to 0.84, 0.7, 0.75, 0.84, 0.84 for the weekdays, Fridays, Saturdays, Sundays, and Game days, respectively; those α values were chosen after trying many different values and picking the ones with the best performance.

Table 3 shows the magnitude of the improvement in the quality of the forecasts of the SARIMA and SVR models when their forecasts are combined using both the simple fixed weight method and the fuzzy adaptive weight method (for each class, the first five days were used to calculate the weights for the fuzzy adaptive weight method and performance was evaluated on the remainder of the dataset).

Two important observations can be made with regard to Table 3. First, it is clear that both methods for combining the results appear to improve the quality of the results. Specifically, for all 5 classes and for the two combining methods, the combined multi-model forecast is better than the single model forecast (with the exception of only the fixed weight method when used on the game day group, where the combining method results are better than SARIMA but worse than SVR). The second observation is that the fuzzy adaptive variable method clearly outperforms the simple weight method, and appears to yield a dramatic improvement in the quality of the results. Specifically, with the fuzzy adaptive variable method, the MAPE for all 5 classes is in the range of only 6% or 7%, and the RMSE for all 5 classes is lower than 45.

Criteria / Classes	MAPE				RMSE			
	SARIMA	**SVR**	**Fixed Weight**	**Variable Weight**	**SARIMA**	**SVR**	**Fixed Weight**	**Variable Weight**
Weekdays (35 days)	10.10%	9.37%	7.80%	6.32%	99.60	48.95	74.88	37.49
Fridays (12 days)	10.09%	9.78%	8.36%	7.68%	57.42	51.55	47.49	43.49
Saturdays (19 days)	10.84%	11.80%	9.49%	7.59%	60.00	56.44	77.37	39.07
Sundays (23 days)	11.55%	12.23%	10.59%	7.65%	59.54	52.42	53.59	39.38
Game days (38 days)	15.31%	8.58%	11.93%	7.32%	79.94	48.52	64.60	43.73

Table 3 Models' Prediction Performance Comparison

For a more disaggregate view of the performance of the models, Figure 4 below compares the traffic volume predicted by SARIMA, SVR and the combined forecasting method, against the field observed values for a period spanning a total of 60 hours. As can be seen, for some hours (e.g. those inside circle A), SARIMA outperforms SVR, whereas for other hours (e.g. inside circle B), SVR performs better than SARIMA. By combining the two methods, the final forecast is generally quite closer to the observed values (with naturally a few exceptions such as the hour enclosed by rectangle C).

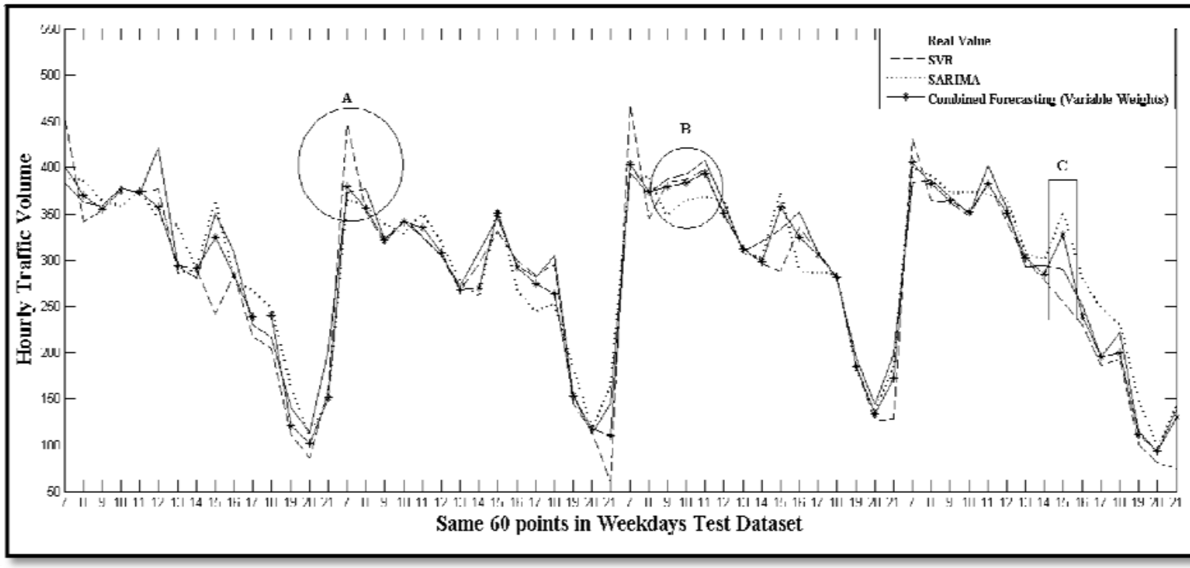

FIGURE 4 Traffic volume predictions of SARIMA, SVR, and the combined forecasting method with variable weight

CONCLUSIONS AND FUTURE WORK

In this part of the study, a multi-model combined forecasting method, combining forecasts from SARIMA and SVR, was proposed and used for the on-line prediction of hourly traffic volumes at the Peace Bridge in Western New York. By combining traditional time series analysis with SVR, the study managed to take advantage of the known strengths of time series analysis, while compensating for its weaknesses in capturing the non-linear aspects of the data, a phenomenon which AI-based methods such as SVR are known to be capable of capturing. Besides that, the study also applied the FP-TREE algorithm to preprocess the data and identify any outliers. We also experimented with the FCM method to classify the data. Among the main conclusions and lessons learned from that part of the study are,

(1) The accuracy of border crossing short-term volume forecasts is improved by classifying the volume data into groups and developing separate prediction model for each group. In this study, a convenient classification scheme involved dividing days into the following six groups: weekdays excluding Fridays, Fridays, Saturdays, Sundays, game days, and holidays.
(2) The SVR method appears to outperform SARIMA when predicting traffic volumes on special days (e.g. game-days), whereas SARIMA performs better than SVR during normal days (e.g. the weekday group). This is consistent with the known strengths and weakness of the two methods, namely the ability of SARIMA, as a linear modeling approach, to capture seasonality and trend versus SVR superior ability to capture nonlinear effects.
(3) While the performance of both SARIMA and SVR methods on the border crossing traffic volume prediction problem appears acceptable, the accuracy is significantly improved by combining the forecasts from the two methods as demonstrated by the results shown in Table 3 and Figure 4.

(4) For combining the SARIMA and SVR forecasts, the Fuzzy Adaptive Variable Weight method appears to outperform the simple fixed weight method as indicated by Table 3.

Several future research directions are suggested by the current work. First, the authors plan to investigate the benefits of incorporating traffic volumes from upstream links into the prediction process to improve accuracy and allow for extending the prediction horizon. We also plan to test the transferability of the approach to other border crossings. A second possible future research area involves dynamically estimating confidence bounds for the volume forecasts, which gives a measure of the reliability of the forecast. Finally, the authors plan to incorporate the prediction models developed into a decision support system for optimally routing border-destined traffic.

PART II AN INVERSE STEADY-STATE $M/M/c$ QUEUEING MODEL to ESTIMATE the DELAY

LITERATURE REVIEW

When vehicles arrive at the border, the process of waiting and going through inspection and customs can be described as a typical queueing system. Given this, the focus of this second of the study was to develop queueing models for estimating the delay which the customers (i.e. the vehicles) in such a queueing system would encounter. Queueing models can be generally represented using Kendall's notation *(33)*: A/B/S/K/N/D,

Where,

A is the inter-arrival time distribution,
B is the service time distribution,
S is the number of the servers,
K is the system capacity,
N is the calling population,
D is the service discipline assumed.

In many cases, it is often assumed that the queueing systems of interest have infinite capacity and that the calling population is also infinite. Moreover, it is often assumed that the queueing systems follow the First-In First-Out service discipline. In such cases, the notation becomes simply A/B/S. Some standard notations for distributions (A or B) are:

$\boldsymbol{M}$ for a Markovian (poisson, exponential) distribution,
$\boldsymbol{E_k}$ for an Erlang distribution with k phases,
$\boldsymbol{D}$ for degenerate (or deterministic) distribution (constant),
$\boldsymbol{G}$ for general distribution (arbitrary),
$\boldsymbol{PH}$ for a phase-type distribution.

A few previous studies could be identified from the literature that applied queueing theory to traffic operations at border crossings and toll plazas. For example, Kim *(34)* proposed an *M/G/1* queueing model for designing a new toll plaza. Kim then solved the model using a non-linear integer programming method, and determined the optimal dynamic lane configuration for operations. Takagi and Wu *(35)* also constructed a queueing model with multi-servers and semi-

Markovian batch arrivals, and derived the distributions of the queue size and the waiting time. Ausin *et al.* (*36*) considered the problems of designing a GI/M/c queueing system. Their goal was to calculate the optimal number of servers so as to minimize an expected cost function which depended on variables such as the number of customers in the queue. Gupta *(37)* estimated the air traffic delays at airports using transient queuing models, including $D(t)/M(t)/1$, $D(t)/D(t)/1$, $M(t)/D(t)/1$, and $M(t)/M(t)/1$, and then compared the estimated delays from different models. Gontijo et al. *(38)* estimated arrival intervals' distribution using Kernel estimation, and developed a $GI^X/M/c/N$ system. Based on this, they calculated the distribution of the number of customers in the system, the average queue size, and the average waiting time. Finally, Park *et al.* derived a queueing model with observable arrival and departure times, but with the number of the service stations unobservable *(39)*.

MODELING METHODOLOGY BACKGROUND

One unique aspect of the current study which we briefly alluded to before, is the fact that the hourly volumes predicted above are outgoing volumes (i.e. volumes leaving the border customs and inspection stations), and not incoming volumes. Given this, we needed to build and solve an inverse M/M/c queueing model in order to estimate the average delay. In that model, it is assumed that arrivals follow a Poisson distribution (however, the average arrival rate $\boldsymbol{\lambda}$ rate is not known). It is also assumed that the system has $\boldsymbol{c}$ service stations whose service times follow exponential distribution with the average service rate as μ. We refer to the model here as an "inverse" model, since what is available for us is the output of the system (in terms of the number of vehicles departing), and not the input (i.e. the arrival rate, $\boldsymbol{\lambda}$). The solution of the inverse queueing model would provide the average delay along with an estimate for the arrival rate as well. The following paragraphs describe how the solution to this problem was derived *(40)*.

First let us assume that $\boldsymbol{N(t)}$ denote the number of customers (i.e. vehicles) in the system (the number in the queue plus the number being served, if any) at time *t* measured from a fixed initial moment (t=0). Given this, the probability of having n customers in the system at certain time moment can be expressed as:

$$P_n(t) = \Pr(N(t) = n)\,, n = 0,1,2, \ldots$$

$$P_i(0) = 1, (P_j(0) = 0, j \neq i) \qquad \text{[Equation 13]},$$

where the number of customers at the initial moment is denoted by $\boldsymbol{i}$ (where $\boldsymbol{i}$ could be 0, 1, 2, …).

For a complete description of the stochastic behavior of the queue-length processes$\{\boldsymbol{N(t), t \geq 0}\}$, the time-dependent solution, $\boldsymbol{P_n(t), n \geq 0}$, needs to be determined. Finding this solution, however, is often quite challenging for transient queues (i.e. systems that consider the time dimension). To simplify the formulation here, we assume steady state conditions and attempt to derive the equilibrium solution which assumes that the system has been in operations for a long time and has reached a steady state. With this assumption, the formulation proceeds as follows *(40)*:

$$\boldsymbol{P_n} = \mathbf{lim}_{t\to\infty} \boldsymbol{P_n(t)} \qquad \text{[Equation 14],}$$

As previously stated, the number of the vehicles in the system under steady state $\boldsymbol{P_n}$ is the value of $\boldsymbol{P_n(t)}$ when $\boldsymbol{t} \to \infty$.

For a queueing system with $\boldsymbol{c}$ servers, in order that steady state exists, the ratio of the arrival to the service rate, denoted by $\boldsymbol{\rho}$, has to be less than 1.0, as expressed by Equation [15] below:

$$\boldsymbol{\rho} = \boldsymbol{\lambda}/\boldsymbol{c}\mu \leq 1.0 \qquad \text{[Equation 15],}$$

Under the steady state, we can derive the probability that the queueing system has no vehicles, $\boldsymbol{P_0}$, as shown below:

$$\boldsymbol{P_0} = [\sum_{n=0}^{c-1} \frac{\left(\frac{\lambda}{\mu}\right)^n}{n!} + \frac{\left(\frac{\lambda}{\mu}\right)^c}{c!\left(1-\frac{\lambda}{c\mu}\right)}]^{-1} \qquad \text{[Equation 16],}$$

With $\boldsymbol{P_0}$ determined, the probability that the queueing system has $\boldsymbol{n}$ vehicles (denoted by$\boldsymbol{P_n}$)and the probability that the queueing system has more than $\boldsymbol{c}$ vehicles ($\boldsymbol{C}$)can be determined as below *(40)*:

$$\boldsymbol{P_n} = \frac{(\lambda/\mu)^n}{n!}\boldsymbol{P_0} \qquad \text{[Equation 17],}$$

$$\boldsymbol{C} = \boldsymbol{Pr(n \geq c)} = \sum_{n=c}^{\infty} \boldsymbol{P_n} = \frac{\left(\frac{\lambda}{\mu}\right)^c}{c!\left(1-\frac{\lambda}{c\mu}\right)}\boldsymbol{P_0} \qquad \text{[Equation 18],}$$

Now, for a steady-state queueing system, we can say that the expected number of vehicles in the queueing system, *E(N)*, is equal to the sum of the expected number of vehicles being served, *E(B)*, and the expected number of vehicles in the queue, *E(Q)*, as shown in equation [19] now. The expected number of vehicles being served, *E(B)*, is equal to $\boldsymbol{c\rho}$, whereas the expected queue length for an M/M/c queueing system can be expressed as $\boldsymbol{\rho C/(1-\rho)}$, where *C* can be estimated from Equation [18] above. In other words, the expected number of vehicles in the queueing system, *E(N)*, is given by Equation [19] below *(40)*:

$$\boldsymbol{E(N) = E(B) + E(Q) = c\rho + \rho C/(1-\rho)} \qquad \text{[Equation 19],}$$

Now, suppose that the queueing system is in steady state at time point $\boldsymbol{t}$, the following equation should exist:

$$\boldsymbol{E(N_H) + V_H = \lambda_H \cdot h + N_t} \qquad \text{[Equation 20],}$$

where,

$\boldsymbol{N_t}$ and $\boldsymbol{N_H}$ refer to the number of vehicles in the system (i.e. in queue or being served) at time $\boldsymbol{t}$ and in the next hour, respectively; $\boldsymbol{V_H}$ is the outgoing traffic volume from the border inspection stations during the next hour that we have predicted; $\boldsymbol{\lambda_H}$ is the arrival rate; $\boldsymbol{h = 1}$ and therefore $\boldsymbol{\lambda_H}.h$ captures the total number of vehicles arriving during the one hour under consideration. The basis of equation [20] is obvious conservation of flow. The number of vehicles at the beginning of the hour, *N(t)*, plus the number of vehicles arriving during that hour .*h,* should be equal to the number of vehicles that departed during that hour, V_H, plus the number of vehicles left in the system at the end of the hour $E(N_H)$. In Equation 20, the only unknown in our case is the arrival rate , and therefore Equation [20] can be used to calculate . Once is calculated, we can calculate the average delay in the queue and the average delay in the system, using the M/M/c appropriate equations as shown below *(40)*:,

[Equation 21],

$-$ [Equation 22],

The difference between and is obviously that the latter also considers the service time.

EXAMPLE

In this part, we will show how to predict the border delay based on the equations derived above, using the "Solver" tool in Excel 2010.

7:00-8:00	output in the next hour (veh)---V_H	500
	number of server station--- c	7
	service time (h)--- μ	0.0125
	vehicle number in the system now--- N_t	200
	expected vehicles in the system after next hour--- $E(N_H)$	257.5199
	number of the arrival rate (veh/h)--- λ_H	557.8
	constraints--- $\rho=\lambda/c\mu <=1$	0.996071
	object--- $E(N_H) + V_H - \lambda_H *1 - N_t = 0$	-0.28006
	average delay in the Queue (h)--- $E(D_{HQ})$	0.449171
	average delay in the system (h)--- $E(D_{HS})$	0.461671

Figure 5 Border Delay Estimation using Solver in Excel

In the Figure 5, the variables highlighted with the yellow color are assumed to be known, the output volume predicted to be processed over the next hour is 500 vehicles (this should have been predicted by the prediction models developed in the first part of this research), the number of the servers or customs and inspection stations is 7, the service time follows an exponential distribution with a mean value of 0.0125 hours, and the vehicle number in the system at the beginning of which would satisfy Equation [20]. In Excel, this is formulated as an objective function which we are seeking to satisfy (object), subject to the constraint that the ratio of arrival to service rate, has to be less than or equal to 1. As can be seen the objective function is basically a transformation of Equation [20].

For this example problem, the arrival rate is estimated to be equal to 557.8 $\boldsymbol{veh/h}$. With this, we then proceed to calculate the average delay in the queue $\boldsymbol{E(D_{HQ}) = 0.45h = 27mins}$ and the average delay in the system $(\boldsymbol{D_{HS}}) = \boldsymbol{0.46h = 27.6mins}$.

CONCLUSIONS AND FUTURE WORK

In part II, we proposed a an inverse M/M/c queueing model to calculate the expected border crossing delay based on the predicted hourly traffic volume in part I. The formulation presented should be regarded as the first step that the authors intend to refine in the future. Specifically, the formulation makes the following three assumptions, which we intend to relax or address in the future:

(1) One key assumption in the formulation above is that the system is under steady-state conditions. In reality, queues at the border are transient or dynamic; this is a research topic we plan to hopefully pursue in the future. Moreover, the constraint that $\rho \leq 1$ may not hold true, especially when short time intervals are considered.

(2) The formulation above assumes that the number of vehicles in the system at the beginning of a given hour is known. From a practical standpoint, that value may not be known, although it could be estimated based on the length of the queue.

(3) The formulation assumes that the distribution of the arrival and service processes exponential. Although such an assumption is very common in queueing theory work, in the future we plan to collect some real-world data to verify that assumption.

ACKNOWLEDGMENTS

Funding for this research was provided by the Region II University Transportation Research Center and by the University at Buffalo, SUNY.

REFERENCES

1. Williams, B.M. and L.A. Hoel. Modeling and Forecasting Vehicular Traffic Flow as a Seasonal ARIMA Process: Theoretical Basis and Empirical Results. *Journal of Transportation Engineering* 129(6), pp. 664-672, 2003.

2. Smith, B.L., B.M. Williams, R.K. Oswald. Comparison of Parametric and Nonparametric Models for Traffic Flow Forecasting. *Transportation Research Part C*, pp. 303-321, 2002.

3. Cools, M., E. Moons, and G. Wets. Investigating the Variability in Daily Traffic Counts through Use of ARIMAX and SARIMAX Models. *Transportation Research Record 2136*, Transportation Research Board, Washington, DC, 2009, pp. 57-66.

4. Zhang, G. P. Time Series Forecasting using a hybrid ARIMA and neural network model. Neurocomputing 50. pp. 159-175, 2003.

5. Chen, K-Y, and C-H Wang. A hybrid SARIMA and support vector machines in forecasting the production values of the machine industry in Taiwan. *Expert Systems with Applications* 32, pp. 254-264, 2007.

6. Tseng, F. Combining neural network model with seasonal time series ARIMA model. *Technological Forecasting and Social Change*, pp. 71-87, 2002.

7. Smith, B. L., and M.J. Demetsky. Short-term traffic flow prediction: neural network approach. Transportation Research Record, 1453, 98-104, 1994.

8. Park, B., C.J. Messer, and T. Urbanik II. Short-term freeway traffic volume forecasting using radial basis function neural network. Transportation Research Record, 1651, 39-47, 1998.

9. Chen, H., and S. Grant-Muller. Use of sequential learning for short-term traffic flow forecasting. Transportation Research Part C 9 319-336, 2000.

10. Xie, Y., and Y. Zhang. A wavelet network model for short-term traffic volume forecasting. Journal of Intelligent Transportation Systems, 10(3):141-150, 2006.

11. Castro-Neto, M., Y-S. Jeong, M-K. Jeong, L.D. Han. Online-SVR for short-term traffic flow prediction under typical and atypical traffic conditions. *Expert Systems with Applications* 36(3), pp. 6164-6173, 2009.

12. Hong, W-C., and Y. Dong. Forecasting urban traffic flow by SVR with continuous ACO. *Applied Mathematical Modeling* 35(3), pp. 1282-1291, 2011.

13. Huang, S. and A.W. Sadek. (2009). A novel forecasting approach inspired by human memory: the example of short-term traffic volume forecasting. *Transportation Research C,* Vol. 17(5): 510-525.

14. Wild, D. Short-term Forecasting based on a Transformation and Classification of Traffic Volume Time Series. *International Journal of Forecasting* 13, pp. 63-72, 1997.

15. Chrobok, R., O. Kaumann, J. Wahle, and M. Schreckenberg. Different Methods of Traffic Forecast based on Real Data. *European Journal of Operational Research* 155, pp. 558-568, 2004.

16. Chung, E. Classification of Traffic Pattern. *Proc. of the 11th World Congress on ITS*, 2004.

17. http://en.wikipedia.org/wiki/Fuzzy_clustering

18. Paselk, T. A. and F.L. Mannering, Use of Duration Models for Predicting Vehicular Delay at a US/Canadian Border Crossing. *Transportation*, Vol. 21, pp. 249-270, 1994.

19. Lin, F.B, and M.W. Lin. Modeling Traffic Delays at Northern New York Border Crossings. *ASCE Journal of Transportation Engineering,* Vol. 127, No. 6, pp. 540-545, 2001.

20. Kam, A., I. Kaysi, B. Abdulhai, and J. Peng. Motorists Guidance Concept for Niagara Peninsula Border Crossings: Delay Prediction Using ANN. *Transportation Research Board Annual Meeting*, Washington D.C., January 2005.

21. Box, G.E.P., and G.M. Jenkins. Time Series Analysis: Forecasting and Control (4th edition). Wiley, 2008.

22. Kim, K-J. Financial Time Series Forecasting Using Support Vector Machines. *Neural Computation* 55. pp. 307-319, 2003.

23. Pai, P-F., and W-C. Hong. Forecasting Regional Electricity Load based on Recurrent Support Vector Machines with Genetic Algorithms. *Electric Power Systems Research* 74. pp. 417-425, 2005.

24. Vapnik, V. *The Nature of Statistical Learning Theory*. Springer, New York, 1995.

25. Stone, M. Cross-validatory Choice and Assessment of Statistical Predictions. *Journal of the Royal Statistical Society* 36(2), pp. 111-147, 1974.

26. Chang, C-C., and C-J. Lin. LIBSVM: a Library for Support Vector Machines. *http://www. csie. ntu. edu. tw/cjlin/libsvm*, 2001.

27. Yongchi, Ma and L. Zhibin. Combination Forecasting Petroleum Production based on Function of Fresh Degree. *Acta Petrolei Sinica* 26(1), pp. 87-95, 2005.

28. Smith, B.L., W. T. Scherer. and J. H. Conklin. Exploring Inputation Techniques for Missing Data in Transportation Management Systems. *Transportation Research Record: Journal of the Transportation Research Board*, No. 1836, pp. 132-142, 2003.

29. Kyung-A Yoon, Doo-Hwan Bae. A pattern-based outlier detection method identifying abnormal attributes in software project data. *Information and Software Technology* 52, pp. 137-151, 2010.

30. Han, J. and J. Pei. Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach. *Data Mining and Knowledge Discovery* 8, pp. 53-87, 2004.

31. Pratt, K.B. Search for Patterns in Compressed Time Series. International Journal of Image and Graphics 2, pp. 89-106, 2002.

32. Shekhar, S. and B.M. Williams. Adaptive Seasonal Time Series Models for Forecasting Short-Term Traffic Flow. *Transportation Research Record: Journal of the Transportation Research Board*, No. 2024, pp. 116-125, 2007.

33. Kendall, D.G. Stochastic Processes Occurring in the Theory of Queues and their Analysis by the Method of the Imbedded Markov Chain. *Annals of Mathematical Statistics* 24(3), pp. 338-354, 1953.

34. Kim, S. The Toll Plaza Optimization problem: Design, operations, and strategies. *Transportation Research Part E* (45), pp. 125-137, 2009.

35. Takagi, H. and De-An Wu. Multiserver Queue with Semi-Markovian Batch Arrivals. *Computer Communications* (27), pp. 549-556, 2004.

36. Ausin, M.C., R.E. Lillo, and M.P. Wiper. Bayesian Control of the Number of Servers in a GI/M/c Queueing system. *Journal of Statistical Planning and Interference* (137), pp. 3043-3057, 2007.

37. Gupta, S. A Framework to Span Airport Delay Estimate Using Transient Queueing Models. Operation Research Center. 2011.

38. Gontijo, G.M., G. S. Atuncar, F. R. B. Cruz, and L. Kerbache. Performance Evaluation and Dimensioning of GI^x/M/c/N Systems through Kernel Estimation. *Mathematical Problems in Engineering*, 2011.

39. Park, J., Y.B. Kim, and T.R. Willemain. Analysis of an Unobservable Queue using Arrival and Departure Times. Computers & Industrial Engineering 61(3), pp. 842-847, 2011.

40. Medhi, J. Stochastic Modeling Queueing Theory. *Academic Press*, pp. 85-93. 2002.

University Transportation Research Center - Region 2
Funded by the U.S. Department of Transportation

Region 2 - University Transportation
Research Center
The City College of New York
Marshak Hall, Suite 910
160 Convent Avenue
New York, NY 10031
Tel: (212) 650-8050
Fax: (212) 650-8374
Website: www.utrc2.org