View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Scholarship@Western

RESILIENT INFRASTRUCTURE

June 1–4, 2016

RESILIENT INFRASTRUCTURE LONDON 2016

IMPUTATION OF MISSING CLASSIFIED TRAFFIC DATA DURING WINTER SEASON

Hyuk-Jae Roh
City of Regina, Regina, SK, Canada

Satish Sharma
University of Regina, Regina, SK, Canada

Prasanta K. Sahu
Birla Institute of Technology and Science Pilani, Rajasthan, India

ABSTRACT

Highway agencies collect traffic data to calculate traffic parameters such as Annual Average Daily Traffic (AADT), Design Hourly Volume (DHV) and then to use as input in the planning, operation and management of their highway systems. The traffic data are usually collected through traffic monitoring programs. In particular, the Weigh-in-Motion (WIM) system is one of data collection systems to capture configuration patterns of vehicle travelling on the detection area. It is learned from literatures that traffic monitoring devices are prone to be in malfunctioning and, consequently, providing erroneous or missing traffic data due to the adverse weather conditions in which they operate. It is very critical for transportation agencies to be able to estimate classified missing traffic data in high accuracy level because the truck traffic plays a crucial role in developing pavement design and evaluation long term pavement performance. Several imputation methods have been cited in the literature but none of them have been designed to impute classified traffic data missed during severe winter weather conditions. To do this, winter weather model is structured and then calibrated to relate classified traffic volume variation to weather factors (snowfall and temperature) with traffic data collected from WIM stations located on highway network of Alberta, Canada and weather data collected from weather stations nearby WIM stations. Performance of the developed weather model is compared with a nonparametric regression method namely k-Nearest Neighbour (k-NN) method in terms of several error measures. It is concluded that winter weather models show better performance in terms of error measures than k-NN method while imputing the missing classified traffic data.

Keywords: Weather model; Imputation; Missing data; k-NN method; Error measures

1. INTRODUCTION

Highway and transportation agencies implement large-scale traffic monitoring programs to fulfill the planning, operation and management needs of highway systems. The traffic volume data are typically collected by highway and transportation agencies using vehicle data collection techniques equipped with a variety of distinctive detection technology designed according to specific purpose. Typically, harsh weather environments provide additional difficulties in managing traffic counters. In this weather conditions, they are highly prone to malfunctioning and then providing erroneous or missing traffic data. For example, by inspecting the (Permanent Traffic Counters) PTCs datasets from Alberta Transportation, Minnesota Department of Transportation, and Saskatchewan Highways and Transportation, Zhong et al. (2004) found that more than half of the PTC could have missing values in datasets.

Highway agencies in North America and many other parts of the world commit a significant portion of their resources to the collection of traffic volume data. The collected data are used in the planning, design, control, operation, and management of traffic and highway facilities. Collected data could also be used to carry out research. However, the presence of missing values in the collected data due to the malfunctioning of counting devices is inevitable. For example, it is noted in the literature (Datla, 2009) that the percentage of Permanent Traffic Counters (PTCs) sites with missing values for Alberta, Saskatchewan, Minnesota, and Colorado during different years is ranged, in general, from 40% to 60%. Only traffic data without missing or erroneous parts can provide true estimates of traffic parameters. The presence of missing values in the collected data limits the accuracy of such estimates.

Several methods, ranging from simple factor approaches to advanced techniques, such as neural networks and genetic algorithms, have been used in the literature to estimate missing traffic volumes (Datla, 2009). However, none of the past studies have given much consideration to the distorted traffic composition caused by severe weather conditions, while imputing missing traffic data during winter months.

The estimation of missing data is termed imputation. A number of researchers (Smith et al. 2003; Zhong, 2003) have recently shown the feasibility of imputing traffic data. However, these efforts have been aimed mainly at the traffic flow from non-winter season days only for total traffic volumes. Imputation for classified traffic data in particular winter season has never been explored in literature. Nonetheless, extra consideration should be given to impute traffic data in winter days compared to the imputation practice for non-winter season due to another dimension of variations in traffic volumes caused necessarily by severe weather conditions. The objective of this paper is first to discuss briefly the adaptability of available imputation techniques for winter season traffic and then the calibrated winter weather models are applied to estimate the missing traffic volumes during winter months. The model is developed to identify the relationships between weather and classified traffic volumes in the framework of interaction regression modelling using 154 million vehicular records collected from six WIM sites located on Alberta highway networks. The winter weather model developing procedure is not a main topic of this paper and thus only the models with calibrated parameters is utilized for comparison purpose. For detailed description of traffic weather models developing procedures, readers are recommend to visit the works done by Roh et el. (2016). The performance of the traffic-weather models developed for the imputation of missing classified traffic volumes such as total traffic, passenger car traffic, and truck traffic during winter months is compared with another commonly used non-parametric regression method known as k-Nearest Neighborhood (k-NN) method to confirm that they are more accurate than k-NN.

2. METHODS IN PRACTICE FOR TRAFFIC DATA IMPUTATION

Our review of the literature has indicated that available techniques for imputing traffic data can be broadly categorized into four groups: heuristic methods, pattern matching methods, time series methods, and artificial intelligence methods. This section provides a brief discussion of these methods with respect to their adaptability to imputing winter weather traffic.

2.1 Heuristic methods

Heuristic methods tend to exploit some of the inherent properties of traffic data from historical records, and are probably the most common approach to solving the problem of missing data (Smith et al. 2003). According to the study conducted by Zhong et al. (2005), the simplest heuristic method directly employs historical good values as replacements to the missing values. Other methods include taking the average values of data from previous or surrounding time periods. The more sophisticated methods in this approach utilize moving average or weighted moving average over the last few days. Zhong et al. (2005) also assessed the imputation accuracy of the above mentioned methods by using non-holiday (or non-winter) traffic data collected from two Provinces in Canada (Alberta and Saskatchewan). They found the methods directly taking good historical values or simply calculating historical average values as replacements resulted in varying accuracy for different study sites and the mean absolute relative errors (MARE) could reach up to 80%. They indicated that the moving average methods seemed better than other heuristic methods. However, because the moving average values are obtained based on volumes over the last few days, there is an inherent drawback to these methods that they are not able to reflect sudden fluctuations in traffic volumes during abnormal periods such as days having severe snowfall and temperatures.

2.2 Pattern matching methods

Zhong et al. (2006) suggested that by comparing a set of candidate hourly volume patterns (without missing values) with the study curve (with missing values), the values from the candidate pattern that best matches the study curve can be used as a replacement. In their study, they used an example of replacing 12 hourly volumes from 8:00 am to

8:00 pm during the daytime. For this purpose, the pattern matching process was based on the 12 available hours, i.e., eight hours from 1:00 am to 8:00 am and four hours from 9:00 pm to 12:00 am. The corresponding hourly volumes from the best matching candidate curve were used to replace the 12 missing values. They reported that this method performed well when updating the missing values from non-holiday Wednesdays. However, it could be noted that, if the missing data were abnormal traffic volumes during severe weather condition periods, the candidate curve matching well with the early morning and late evening hours will not necessarily reflect the abnormal traffic pattern during inclement weather condition. Hence, this method was not considered appropriate for imputing traffic during severe weather conditions.

2.3 Time series methods

Autoregressive integrated moving average (ARIMA) models are a type of time series model which have been popularly used in traffic forecasting. The ARIMA models can be simply understood as linear estimators regressed on past values of the modeled time series. Detailed discussions regarding the theory of these models can be found in standard references (Fuller, 1996). With respect to the performance of ARIMA models, many researchers have reported successful cases when dealing with normal traffic (Williams & Hoel, 2003). However, Redfern et al. (1993) indicated that changes in seasonal traffic structure due to various reasons would definitely bring problems to the estimation using ARIMA methods. Kirby et al. (1997) found that the ARIMA model did not yield good results when trying to apply it to data with the summer holiday season included. Zhong (2003) reported that when weekend traffic pattern was taken into consideration, the MARE errors resulting from ARIMA models could be as high as 97%.

2.4 Artificial intelligence methods

Genetic algorithms (GAs) and artificial neural networks (ANNs) are some typical artificial intelligence techniques that have been applied in prediction or estimation of traffic data. They are expected to discover useful hidden knowledge from vast amounts of data and to make more accurate predictions. The most successful application of these techniques in traffic data imputation is reported in Zhong et al. (2004) who employed GAs to select final input variables for regression and neural network models. They achieved high accuracy with most of the MARE errors ranging from 1 to 3% for the regression method and 3 to 15% for the neural network models. However, it should be noted that the high estimation accuracy of their study resulted only for the Wednesday traffic in July and August. There was no discussion regarding the imputation of traffic from other time periods, especially winter days. Moreover, the models they proposed had to be redeveloped (with the volumes from different hours as inputs) every time for each individual hour during the imputation process, which would be time consuming in data preparation.

3. A METHODOLOGY FOR IMPUTING CLASSIFIED TRAFFIC VOLUME DURING WINTER WEATHER

The previous section presented a description of the imputation methods in practices. Zhong (2003) reviewed and evaluated different methods to impute missing traffic data during summer months. These methods include simple factor approaches (Garber and Hoel, 2002), autoregressive integrated moving average (ARIMA) models (Redfern et al. 1993), weighted regression analysis, neural networks, genetic algorithms, and genetically designed neural networks (Zhong et al. 2005). His study concluded that genetically designed neural network approaches are superior to other methods when traffic data from summer months are being imputed. However, none of the existing imputation methods considered the variations in traffic volumes due to severe winter conditions. Therefore, the suitability of these methods to impute missing traffic data during winter months is unknown.

Therefore, to compute the missing data in this study, a nonparametric regression method namely k-Nearest Neighbour (k-NN) method has been used. The missing data computations have also been performed by using the winter weather models developed through the extensive modelling works conducted by Roh et al. (2016). The results from both the techniques are compared and it is concluded that winter weather models result in higher level of accuracy while imputing the missing classified traffic data. The following subsection discusses the principles of the k-NN method.

3.1 Principles of k-NN Method for Data Imputation and k Value Determination

The k-NN method relies on memory/instance based learning for large data sets (Liu et al. 2008). It matches the current input variables with similar historical records (Liu and Sharma, 2006). In practice, traffic volume with temporal variation is defined as a state vector at time lags of, $t-1, t-2, \cdots$, etc. Because the k-NN model geometrically attempts to reconstruct a time series (Mulhern and Caprara, 1994), the inclusion of historical averages in the state vector clarifies the position of each observation along with the cyclical flow-time curve and improve forecasting accuracy (Smith et al., 2002). The state vector $x(t)$ used in this study is given in Equation 1.

[1] $$x(t) = [V(t), V(t-1), V(t-2), V_{hist}(t), V_{hist}(t+1)]$$

Where, $V(t)$, $V(t-1)$ *and* $V(t-2)$ are the traffic flows at time intervals t, $t-1$, and $t-2$, respectively. $V_{hist}(t)$ and $V_{hist}(t+1)$ are the historical average volumes at the day-of-month associated with time interval *t* and *t+1*. In cases of imputing classified traffic volumes, the historical average is calculated based on volumes that are from the same day during the same period in the past for the same vehicle class. After the state vector is defined, the k-value is found out based on Euclidean Distance, and the *k* observations with the shortest Euclidean Distances are recognized as neighbors (Liu et al. 2008). The Euclidean distance (*d (p, q)*) from historical record *p* to the current condition *q* can be written as following Equation 2:

[2] $$d(p,q) = \sqrt{\begin{matrix}\left(V_p(t) - V_q(t)\right)^2 + \left(V_p(t-1) - V_q(t-1)\right)^2 + \left(V_p(t-2) - V_q(t-2)\right)^2 \\ \left(V_{hist,p}(t) - V_{hist,q}(t)\right)^2 + \left(V_{hist,p}(t+1) - V_{hist,q}(t+1)\right)^2\end{matrix}}$$

In this study we chose 4 nearest neighbors along with their corresponding output volumes to determine the *k*-value. For example, Figure 1 shows the Euclidean Distances and the corresponding output volumes for the 4 nearest neighbors of one particular day. The day is November 18, 2009 with a volume of 23,324 passenger vehicles that represents the typical traffic condition of this road site with normal winter condition. For this particular day, Figure 1 shows the first three neighbors' volumes are closer to the actual traffic volume. Also, the Euclidean Distances for these neighbors are relatively low. Therefore, the *k* -value is 3 chosen for this case.

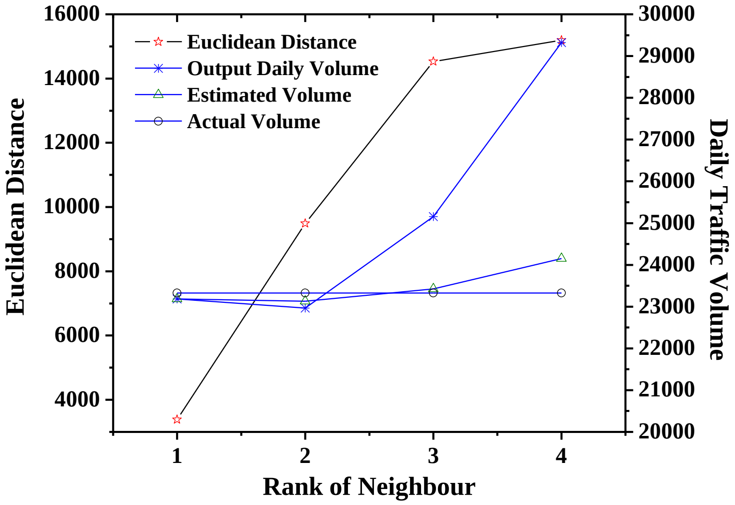

Figure 1: Euclidean Distance and Output Volume of the Closer 4 Neighbors

3.2 Application of k-NN Method and Winter Weather Model for Forecast Generation

With the determination of k=3, the k-NN method was applied to impute the daily traffic volume using Equation 3 for winter days. This subsection provides a sample application of k-NN method to impute 18 winter days for classified traffic volume by using the WIM site data located on Highway 2. The imputation period was chosen carefully to reflect inclement weather conditions starting from the Nov 14 to Dec 14 in 2009 (Figure 2). The daily traffic data from the entire years of 2005, 2006, 2007 and 2008 were treated as the historical database to determine neighbors for the k-NN method.

[3] $$y'(t) = \frac{\sum_{i=1}^{k} \frac{y_i(t)}{d_i}}{\sum_{i=1}^{k} \frac{1}{d_i}}$$

Where, d_i is the Euclidean Distance of the i^{th} neighbor and $y(t)$ is the output volume of i^{th} neighbor. The estimated and actual values for passenger cars and trucks within the imputation period are summarized in Table 1 and shown graphically in Figures 3 and 4 respectively. An important observation can be made here by examining the plots for Friday, December 4 (named as 91204 in x-axis at the top in Figure 2), with a snowfall of 11 cm and the average temperature of -7°C. The k-NN method results in large over estimation for both the passenger cars and truck traffic; whereas winter weather model imputed values that are closer to actual values.

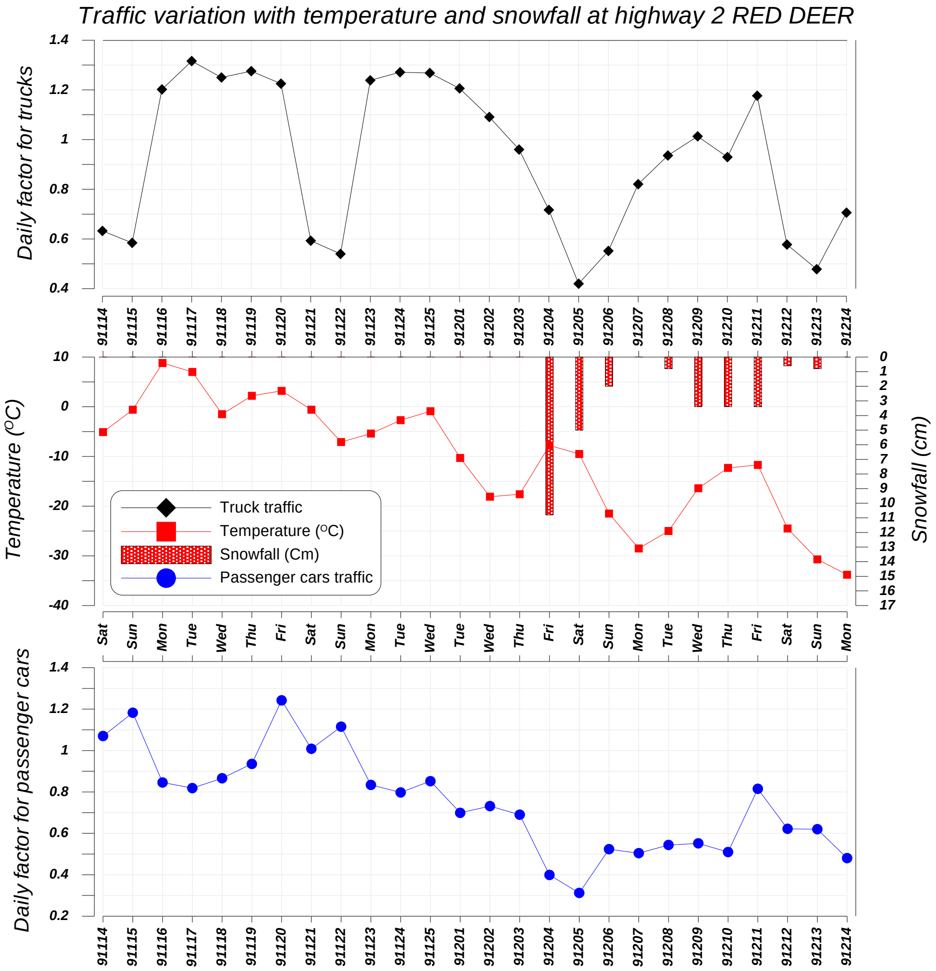

Figure 2: Traffic Variations for Days Assumed to be Missing in the Data Set

Table 1: Estimation Results for k-NN and Winter-Weather Model for Weekdays for Passenger Cars and Trucks

Date	Actual_TT	k-NN_TT	Interaction_TT	Actual_PCT	k-NN_PCT	Interaction_PCT
2009-11-16	5,722	5,442	5,912	22,772	22,383	23,707
2009-11-17	6,264	5,964	6,266	22,042	21,935	23,148
2009-11-18	5,949	5,948	6,079	23,324	23,426	23,023
2009-11-19	6,072	5,841	6,147	25,189	25,158	24,495
2009-11-20	5,830	5,564	5,883	33,450	32,872	32,576
2009-11-23	5,895	5,575	5,545	22,457	22,581	20,795
2009-11-24	6,051	5,695	5,779	21,477	19,546	20,166
2009-11-25	6,036	5,686	5,923	22,942	20,137	21,908
2009-12-1	5,740	2,556	5,883	18,825	17,784	19,509
2009-12-2	5,194	5,137	5,720	19,703	21,585	19,682
2009-12-3	4,570	4,965	5,739	18,586	21,089	21,307
2009-12-4	3,414	5,309	4,674	10,750	22,677	18,050
2009-12-7	3,906	3,820	5,068	13,585	16,904	17,371
2009-12-8	4,455	4,135	5,469	14,649	16,177	17,369
2009-12-9	4,822	5,574	5,510	14,861	22,058	16,677
2009-12-10	4,425	4,260	5,644	13,726	18,786	18,594
2009-12-11	5,600	4,447	5,421	21,953	20,620	22,765
2009-12-14	3,361	3,415	4,873	12,939	15,410	18,085

TT (Truck Traffic), PCT (Passenger Cars Traffic)

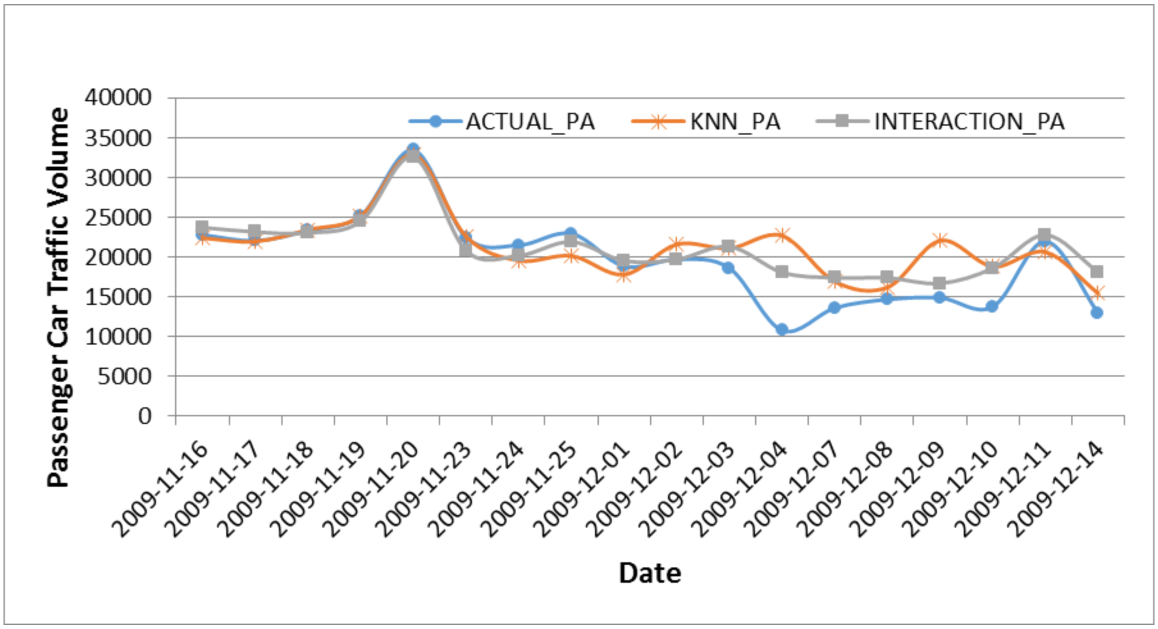

Figure 3: Comparison of the Actual and Estimated Volumes Using k-NN and Winter-Weather (Interaction) Model for Passenger Cars for Weekdays

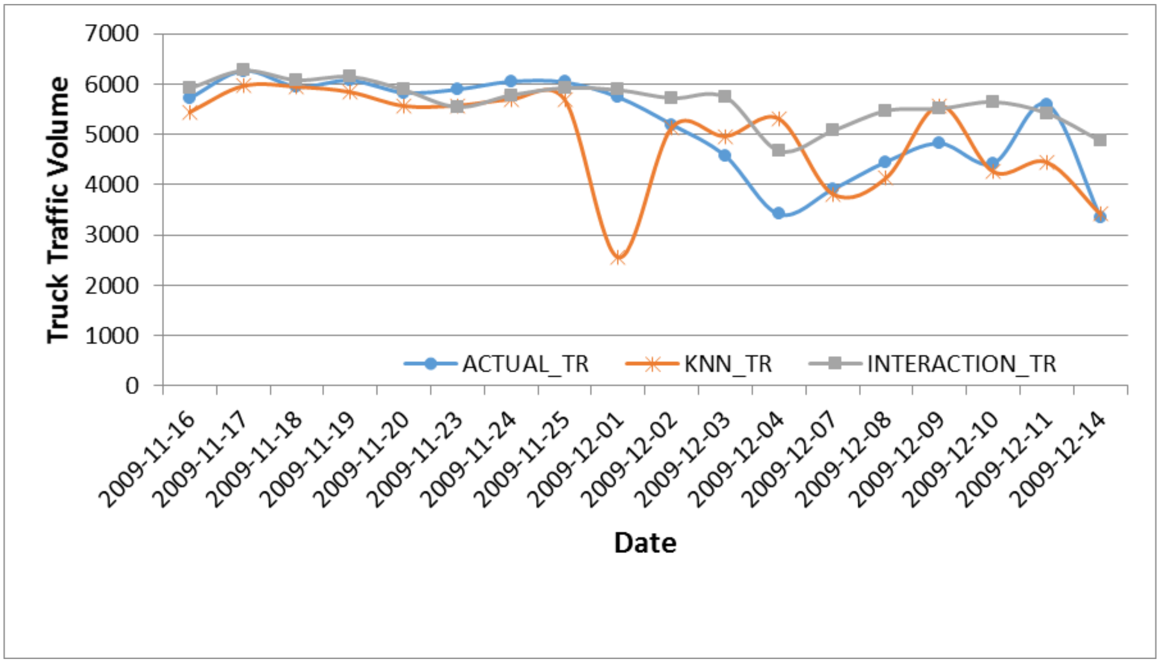

Figure 4: Comparison of the Actual and Estimated Volumes Using k-NN and Winter-Weather (Interaction) Model for Truck Traffic for Weekdays

3.3 Statistical Comparison of k-NN Method and Winter Weather Model Performances

The performance of the k-NN method as compared to the winter-weather model was also evaluated using a set of error measures in terms of forecasting accuracy. Usually, accuracy examines how well the model reproduces the already known data. The error measures used for this purpose are mean absolute percentage error (MAPE) which is a useful measure in order to eliminate the effect of variability observed in data sets. The formulation is as follow, Equation 4:

[4] $$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{X_i - F_i}{X_i} * 100\right|$$

Where, X_i and F_i are the actual and estimated daily traffic volumes, respectively. The minimum absolute percentage error (*MinAPE*, given in Equation 5) and maximum absolute percentage error (*MaxAPE*, given in Equation 6) represent the smallest and the largest error in the results.

[5] $$MinPE = Min\left|\frac{X_i - F_i}{X_i} * 100\right|$$

[6] $$MaxPE = Max\left|\frac{X_i - F_i}{X_i} * 100\right|$$

The 50th percentile error represented by E50 in Table 2 means that 50% of the errors resulting from the estimation are placed below the value of E50. Similar interpretation can be made for E95. Based on estimated error measures, it is clear that winter-weather model results in better imputation results. For truck traffic volume estimation, even though MAPE for k-NN method is less than winter-weather model, the error measure of E95 for winter-weather model (i.e., 38.11) is much lower than the value for the k-NN method (55.48), meaning that more data points are estimated accurately using the winter-weather model. For passenger cars, all error measures show that winter-weather model performance is better than k-NN method. The statistical errors given in Table 2 are also shown graphically in Figures 5 and 6.

Table 2: Imputation Results for k-NN and Winter-Weather Model (Interaction) for Passenger Car and Truck Traffic

Statistics	k-NN_TT	Interaction_TT	k-NN_PCT	Interaction_PCT
MAPE	11.49	13.20	17.28	14.32
MinAPE	0.02	0.03	0.12	0.11
E50	5.16	5.21	9.27	5.56
E95	55.48	38.11	57.81	43.99
MaxAPE	55.52	45.00	110.94	67.91

TT=Truck Traffic; PCT=Passenger Car Traffic

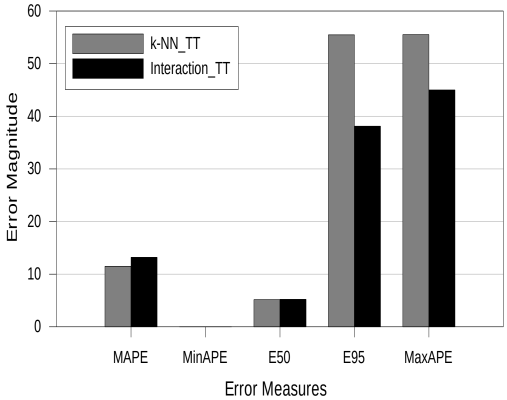

Figure 5: Comparison of Error Measures for k-NN Method and Winter-Weather Model for Trucks

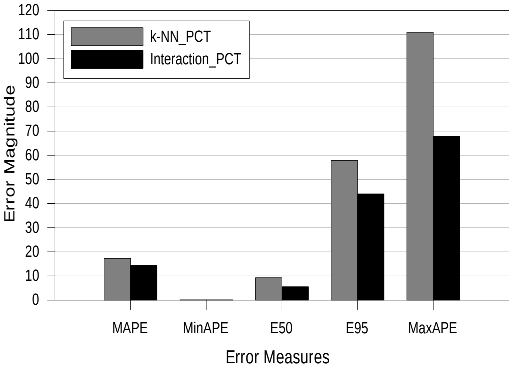

Figure 6: Comparison of Error Measures for k-NN Method and Winter-Weather Model for Passenger Cars

4. CONCLUDING REMARKS

This paper focused mainly on a successful application of the traffic-weather models developed in imputing missing classified traffic volumes such as total traffic, passenger car, and truck traffic during winter severe weather conditions. The estimated missing traffic flow values are compared with the results from another imputation technique known as k-NN method, which provides estimation results based on the historical data base. The comparison of results from both the techniques suggested that the winter weather model results are closer to the actual values.

REFERENCES

Datla, S. 2009. *Development and Application of Traffic Volume - Winter Weather Relationships*, Ph.D. Thesis, University of Regina, Regina, Saskatchewan, Canada.

Fuller, W. A. 1996. *Introduction to Statistical Time Series,* Wiley, New York, USA.

Garber, N. J. and Hoel, L. A. 2002. *Traffic and Highway Engineering,* Brooks/Cole Publishing Company.

Kirby, H. R., Watson, S. M., and Dougherty, M. S. 1997. Should We Use Neural Networks or Statistical Models for Short-Term Motorway Traffic Forecasting? *International Journal of Forecasting*, 13 (1): 43-50.

Liu, Z. and Sharma S. 2006. Statistical Investigations of Statutory Holiday Effects on Traffic Volumes. *Transportation Research Record*, 1945: 40-48.

Liu, Z. B., Sharma, S., and Datla, S. 2008. Imputation of Missing Traffic Data during Holiday Periods. *Transportation Planning and Technology*, 31 (5): 525-544.

Mulhern, F. J. and Caprara R. J., 1994. A Nearest Neighbor Model for Forecasting Market Response. *International Journal of Forecasting*, 10 (2): 191-207.

Redfern, E.J., Watson, S.M., Clark, S.D., Tight, M.R. and Payne, G.A. 1993. *Modeling Outliers and Missing Values in Traffic Count Data Using the ARIMA Model*, ITS Working Paper 395, Institute for Transport Studies, University of Leeds.

Roh, H., Sahu, P., Sharma, S., Datla, S., and Mehran, B. 2016. Statistical Investigations of Snowfall and Temperature Interaction with Passenger Car and Truck Traffic on Primary Highways in Canada. *Journal of Cold Region Engineering*, ASCE, forthcoming.

Smith, B.L. Scherer, W.T. and Conklin, J.H. 2003. Exploring Imputation Techniques for Missing Data in Transportation Management Systems. *Transportation Research Record*, 1836: 132-142.

Smith, B., Williams, B.M., and Oswald, R.K. 2002. Comparison of Parametric and Nonparametric Models for Traffic Flow Forecasting. *Transportation Research Part C*, 10: 303-321.

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

Zhong, M. 2003. *Data Mining Applications for Updating Missing Values of Traffic Counts*, Ph.D. Thesis, University of Regina, Regina, Saskatchewan, Canada.

Zhong, M. Sharma, S., and Lingras, P. 2004. Genetically Designed Models for Accurate Imputations of Missing Traffic Counts. *Transportation Research Records*, 1879: 71-79.

Zhong, M., Sharma, S., Liu Z. 2005. Assessing the Robustness of Imputation Models based on Data from Different Jurisdictions: Alberta and Saskatchewan Examples. *Transportation Research Records*, 1917: 116-126.

Zhong, M., Sharma, S. 2006. Matching Hourly, Daily, and Monthly Traffic Patterns to Estimate Missing Volume Data. *Transportation Research Records*, 1957: 32-42.