International Journal of Engineering Research and Development
e-ISSN: 2278-067X, p-ISSN: 2278-800X, www.ijerd.com
Volume 13, Issue 1 (January 2017), PP.08-16

Application of Cumulative Axle Model To Impute Missing Traffic Data in Defective AVC Stations

Yooseok Jung[1], Jusam Oh[2]
[12]*Korea Institute Of Civil Engineering And Building Technology, 283, Goyangdae-Ro Goyang-Si, Korea.*

Abstract: An automatic vehicle classification (AVC) station is typically composed of three sensors per lane. Instances of data missing from the traffic datasets collected at such stations can occur as a result of issues such as one of the sensors malfunctioning. Although various data imputation methods, such as autoregressive integrated moving average (ARIMA), exponential smoothing, and interpolation, have been proposed to deal with this problem, they are either too complicated or have significant errors. This paper proposes a model, called the "cumulative axle model," that minimizes such errors in traffic volume data resulting from a malfunctioning sensor at AVC stations. Evaluations conducted in which missing traffic volume data imputation was simulated using the proposed cumulative axle model indicate that our method has a mean absolute percentage error (MAPE) of 2.92%. This is significantly more accurate than that of conventional imputation methods, which achieve a MAPE of only 10% on average.

Keywords: Defective AVC, missing traffic data, imputation, cumulative axle model, MAPE

I. INTRODUCTION

Korea introduced nationwide traffic surveys for the first time in 1955. In order to publish the "Annual traffic volume report," the Korea Institute of Construction Technology, commissioned by the Ministry of Land, Infrastructure and Transport, has subsequently been operating automatic vehicle classification (AVC) at 630 locations since 2014. With the equipment gathering traffic data on a daily basis, instances of missing data have been occurring owing to various internal and external factors. The internal factors include controller malfunction, sensor malfunction, and communication problems. The external factors include road construction and traffic lane control [1]. The missing data rate in the Minnesota DOT case was more than 40%; in Alberta, the missing data rate was 50% over a period exceeding seven years; in some years, the rate even increased to 70–90% [2].

When data are missing, ex post facto imputation is needed to utilize the traffic data. In such scenarios, traffic parameters such as annual average daily traffic (AADT) and design hourly volume (DHV), which are the basic data employed when designing and planning roads, may cause severe harm if they are overestimated or underestimated [3]. Hence, missing short-term traffic data were previously estimated using statistical techniques such as regression models, EM, and time series models, and heuristic techniques such as the average of the data before and after the period under consideration. These methods estimate missing data using the past and future traffic data of that particular location. However, this may generate errors in estimating the traffic, as it does not consider special traffic conditions that may have occurred at the time of occurrence of the missing data.

In AVC stations, the sensors most frequently used to investigate regular traffic are loop sensors and piezo sensors, which are usually buried in pavement. These sensors are installed either as a loop-piezo-loop (L-P-L) type or a piezo-loop-piezo (P-L-P) type to gather information not only on the traffic but also on speed, occupancy ratio, number of axles, axle spacing, vehicle classification, and overhang. L-P-L type sensors measure speed and traffic by measuring the time between two loop sensors, whereas piezo sensors classify vehicles by measuring the number of axles. In the P-L-P type configuration, the loop sensor detects the vehicle, and the piezo sensor measures speed through time difference, and can also obtain the number of axles. Consequently, a defective sensor configuration is not able to collect all the relevant information. In such cases, the AVC station will stop the data collection if at least one malfunctioning sensor is detected, even though one or both of the other sensors may still be active.

This paper proposes a method for defective AVC stations that imputes missing traffic volume data using the cumulative number of axles measured in cases where at least one piezo sensor is operational. Further, the proposed method is compared with existing missing data imputation methods. Even though a defective AVC station has limitations, using the reduced sensor configurations to collect data and the proposed "cumulative axle model" to estimate traffic volume via axle counts can provide results that are more accurate than the conventionally used estimation techniques. When the sensor measurements are stored, the cumulative loop sensor value becomes the traffic data, and the cumulative piezo sensor value becomes the cumulative number of axles. Here, cumulative number of axles is related to the traffic and the vehicle class proportion. A cumulative axle model that can estimate the traffic using the cumulative number of axles can be made via linear regression analysis of the cumulative number of axles, and the traffic at a certain period of time can be obtained when the

AVC stations are operating normally. Thus, when this model is applied, traffic can be estimated with very high accuracy even in scenarios where only the data from the piezo sensors can be gathered.

II. BACKGROUND

Traffic surveys using permanent survey equipment sometimes lose data from a certain period because of equipment abnormality or malfunction [4]. Despite rapid developments in traffic-related measuring technology and their commercialization, losses in the quality and accuracy of traffic data as a result of incomplete measurements is still a serious problem [5]. Methods that impute missing data can generally be divided into three groups: prediction methods, interpolation methods, and statistical learning methods [6]. The time series technique is the most frequently used prediction method [5]. Applying the contemporary methods mentioned above in a commuter area results in approximately 10–20% error, while more advanced methods such as genetic algorithms and time delay neural network methods have errors below 6% [7].

In general, data is imputed by considering major factors based on past data. Korea uses the past tendency of an identical survey location, patterns of a nearby location with similar traffic patterns, or changes in traffic patterns under specific weather conditions to impute missing data (Ministry of construction and transportation, 2013). The most representative time series methods are the autoregressive integrated moving average (ARIMA) [8] and exponential smoothing methods. A recent study in Canada imputed missing data with an average median error of 16% using time series smoothing methods and pattern recognition, which uses Euclidean distance, on the traffic data of permanent equipment that had lost 90% of its data. The utilization of these methods requires prudence and, because they cannot be systematically automated, their generalization is difficult [9].

Axle correction factor typically estimates traffic by applying the factor to the number of axles measured by portable traffic counters [10]. The factor is also essential in estimating the AADT of traffic volume survey sites that use pneumatic sensors [11]; [12] by calculating the axle correction factor of permanent traffic survey sites [13]. Hence, different axle correction factors are often provided for each area [14]. Forty states in the U.S. apply axle correction factors to equipment that measure only the cumulative number of axles in traffic surveys (Albright, 1991). It has also been used to show the vehicle model distribution characteristics of a survey site [4] [15].

III. OPERATION OF DEFECTIVE AVC STATION

An AVC station is composed of P-L-P or L-P-L sensors for each lane, as shown in Fig. 1. A loop sensor gathers signals when a vehicle passes over it, whereas a piezo sensor gathers signals via the pressure caused when a vehicle's axle passes over it. The speed of the vehicle, distance between axles, and vehicle body length are calculated by combining the two types of signals, thus allowing for vehicle classification.

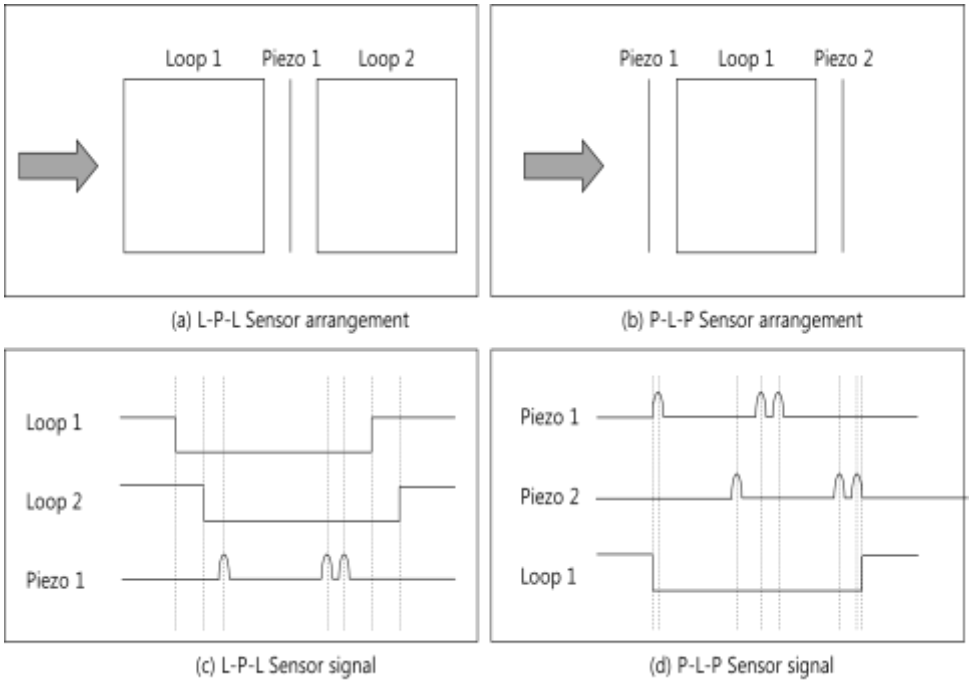


Fig.1: Sensor arrangement and signal diagram for AVC

At present, data collection is suspended when at least one sensor is missing. However, more information can be obtained from a defective AVC station that is in operation. The method proposed in this paper is presented in Table 1. With this proposed cumulative axle model, traffic volume data is produced in every sensor configuration.

Table 1: Comparison of operation plan for Defective AVC

Sensor Configuration	**Present**	**Defective AVC Operation Plan**		
		Traffic Volume	Speed	Vehicle Classification
~~l□l, □l□~~ (P-L-P, L-P,L~~)~~	Available	Available	Available	Available
~~□l, l□~~ (~~(~~L-P, P-l~~)~~	Data Missing	Available	Not Available	Available
~~□ □~~ (~~(~~L-L~~)~~		Available	Available	Impute with ratio
~~l l~~ (~~(~~P-P~~)~~		Axle Cumulative Model	Not Available	Impute with ratio
~~□~~ (~~(~~L~~)~~		Available	Not Available	Impute with ratio
~~L~~ (~~(~~P~~)~~		Axle Cumulative Model	Not Available	Impute with ratio

IV. CUMULATIVE AXLE MODEL

In general, the loop sensor is used to detect traffic; thus, if the loop sensor malfunctions, then the traffic must be estimated using the piezo sensor. The cumulative axle model, created via linear regression analysis between the traffic per unit time and the cumulative number of axles, can be applied in this case.

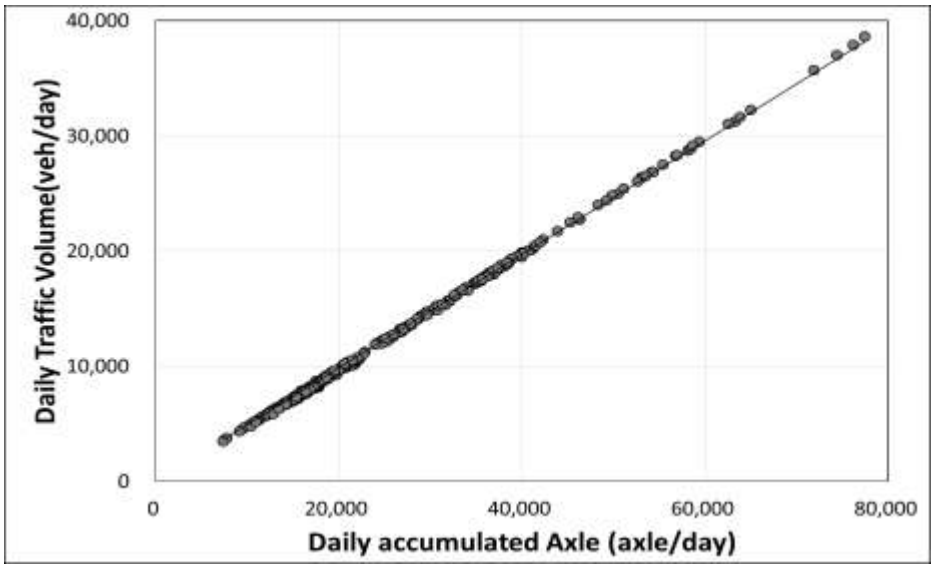


Fig.1: Linear regression analysis example of daily traffic volume versus accumulated axle($y = 0.493x$, $R^2 = 0.99$)

Fig. 2 is an example of the cumulative axle model at ID: 40493 on Route 40. Three hundred and sixty-five units of daily traffic data for each vehicle type and the cumulative number of axles for the year 2014 were used. Linear regression analysis was conducted by considering the traffic as a dependent variable and the cumulative number of axles as an independent variable. The coefficient of determination, R^2 value, shows how well the independent variable can predict the dependent variable, and ranges between zero and one. In this scale, the greater the value, the more appropriate is the regression equation.

The regression coefficient is related to the vehicle type proportion. The cumulative axle model of four sites with different proportions of two-axle vehicles is presented in Table 2 using the same analysis method as above. As expected, the regression coefficient of the model increased as the proportion of two-axle vehicles increased. For instance, the regression coefficient when the proportion of two-axle vehicles was 100%, was 0.500. Further, it was 0.333 when the proportion of three-axle vehicles was 100%, and 0.250 when the proportion of four-axle vehicles was 100%.

Table 2: Axle cumulative model by daily data due to proportion of 2-axle vehicles

ID5	70561	40411	60608	40493
AADT	5,173	11,466	10,468	11,822
Proportion of 2-axle vehicles	0.84	0.89	0.94	0.98
Regression coefficient	0.433	0.449	0.478	0.493
R^2	0.96	0.95	0.98	0.99

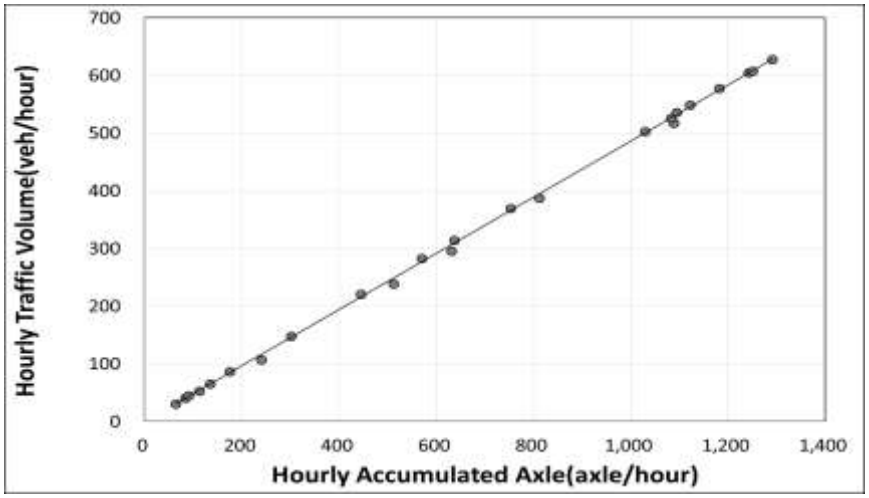


Fig.3: Linear regression analysis example of hourly traffic volume versus accumulated axle ($y = 0.487x$, $R^2 = 0.99$)

A cumulative axle model such as the one presented in Fig. 3 can be created via linear regression analysis of hourly traffic volume and hourly accumulated number of axles by changing the unit time. The example uses 24 units of hourly traffic volume and accumulated number of axles for Wednesday, May 27, 2015, at ID: 40493 of Route 40. Although the amount of data is limited, a precise model with R^2 greater than 0.97 was created.

Table 3 shows the cumulative axle model using hourly traffic volume from four sites with different proportions of two-axle vehicles. This model shows the possibility of imputing missing hourly traffic data, and also reflects the daily variability.

Table 3: Axle cumulative model by hourly data due to proportion of 2-axle vehicles

ID5	70561	40411	60608	40493
AADT	5,173	11,466	10,468	11,822
Proportion of 2-axle vehicles	0.84	0.89	0.94	0.98
Regression coefficient	0.382	0.469	0.456	0.487
R^2	0.97	0.99	0.99	0.99

V. CUMPARATIVE METHOD

A. Historical Approach

Korea's method of imputing missing data is interpolation, which applies a factor approach. This is because the imputing method to obtain daily variability is different for weekdays and weekends. The missing data for Tuesdays, Wednesdays, and Thursdays are substituted with the average data from the previous and next days. As the average of the previous and next days cannot correctly reflect the weekly factor in the case of Mondays, Fridays, Saturdays, and Sundays, the missing data for these days are imputed with the average values from the previous and subsequent weeks.

B. Time Series

The time series method, the representative ARIMA method, and the exponential smoothing method, were compared with the cumulative axle model. The methods used in the estimation can basically be used to impute missing traffic data. The ARIMA model uses the Box-Jenkins methodology, which considers the autoregressive process and moving averages. The ARIMA(p, d, q)$(P, D, Q)_s$ model, which considers seasonality according to the characteristics of the traffic volume, was utilized. Here, P, D, and Q are seasonal autoregressive, differencing, and moving average components, respectively [2].

Exponential smoothing is a time series method that attributes weights to observed values. It has been used to predict traffic volume for decades. The closer the observed value is, the more is the weight given to it. The differently attributed weights are completed by one or several smoothing parameters. It is used in various ways, as it is a method that is not only simple and easy to understand, but is also easily applicable [16]. With the above time series method, traffic volume predictions were obtained using SPSS, and the missing data were imputed.

C. Interpolation

Basic linear Lagrange interpolation and quadratic Lagrange interpolation were applied and compared with the other methods. The advantage of interpolation is that, as a numerical method, its principle is simple, making automation of the method easy. Linear Lagrange interpolation determines the value of the desired location through nearby values and distance weighting, whereas quadratic Lagrange interpolation estimates the value by creating a quadratic equation that passes through three points [17].

VI. RESULTS

Ten AVC sites with various proportions of two-axle vehicles and which have no data missing from the daily traffic volume data of 2014 were chosen randomly. Among the 365 daily traffic volume data of 2014, approximately 15% or 54 data points were randomly assumed to be missing, and the cumulative axle model was applied. The 54 missing data points were determined by generating random numbers using a statistics program, and the same dates were assumed to be missing at all sites. No restrictions on consecutive dates for missing data were applied. Table 4 shows the axle imputing model created using the data for the remaining 311 days and the mean absolute percentage error (MAPE) generated by imputing the traffic data for the 54 missing days using the created model. Fig. 4 is a plot of the traffic volume estimation MAPE versus the proportion of two-axle vehicles. The figure shows that the proportion of two-axle vehicles where the AVC was installed is inversely proportional to the cumulative axle model's accuracy.

Table 4: AADT estimation error (MAPE) by axle cumulative model

ID	**42001**	**20353**	**70561**	**40023**	**40411**	**70012**	**60608**	**21012**	**40493**	**61010**
Proportion of 2-axle vehicles	0.781	0.831	0.842	0.888	0.893	0.923	0.944	0.960	0.977	0.991
Correction Model(α)	0.396	0.430	0.433	0.454	0.449	0.465	0.478	0.486	0.493	0.497
MAPE (%)	4.43	4.99	5.51	3.04	2.87	3.43	1.48	1.47	1.64	0.37

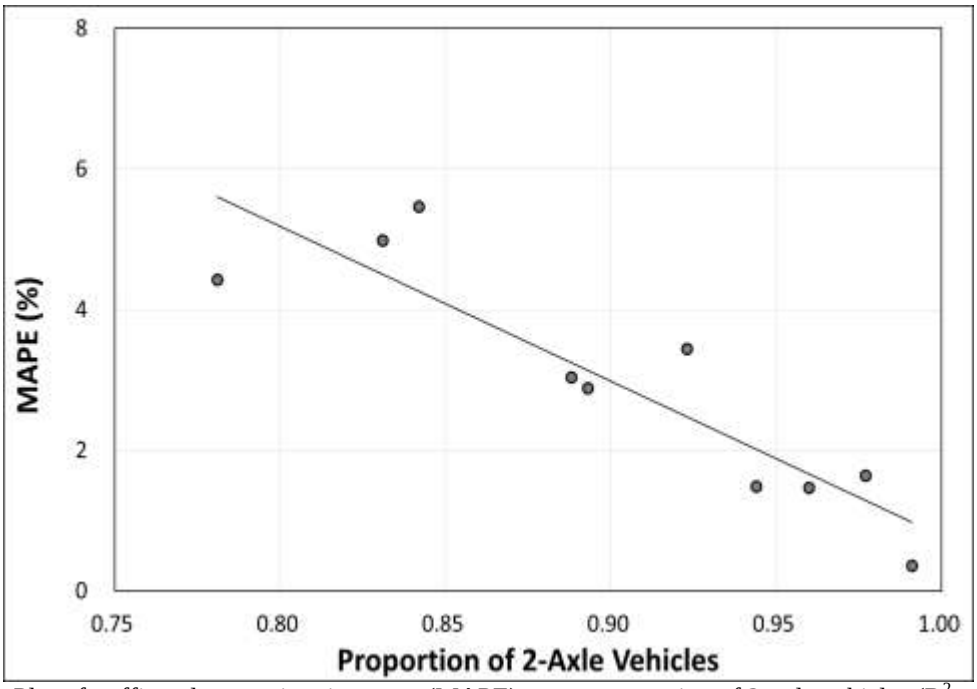
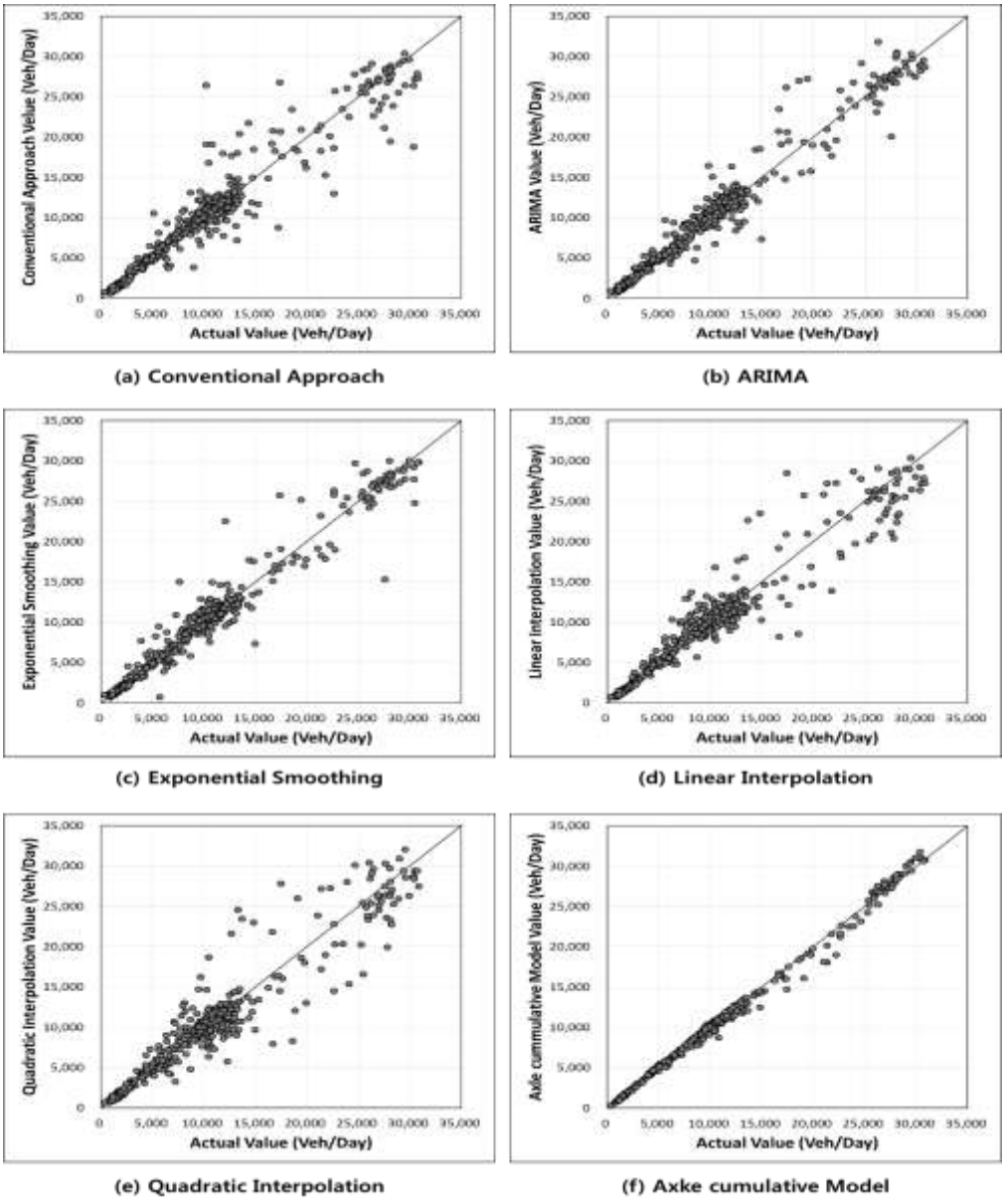


Fig.4: Plot of traffic volume estimation error (MAPE) versus proportion of 2-axle vehicles ($R^2 = 0.81$)

Table 5 gives a histogram of the proportion of two-axle vehicles obtained by analysing the proportion of two-axle vehicles at the 621 AVC sites along the Korean national highway where traffic statistics for 2014 could be obtained. More than 94% of the locations recorded the proportion of two-axle vehicles to be more than 90%. Hence, the cumulative axle model can be deemed to be highly advantageous in terms of accuracy if actually implemented. For reference, according to the reports of the [18] and [19], the proportion of two-axle vehicles at almost all survey sites was more than 90%.

Table 5: Histogram of proportion of 2-axle vehicles

Proportion of 2-axle vehicles (Rank)	Frequency	%
0.75	1	0.16%
0.8	2	0.48%
0.85	9	1.93%
0.9	21	5.31%
0.95	168	32.37%
1	420	100.00%
Total	621	



(a) Conventional Approach (b) ARIMA

(c) Exponential Smoothing (d) Linear Interpolation

(e) Quadratic Interpolation (f) Axke cumulative Model

Fig.5: Plot of actual values versus estimated values

To compare the accuracy of the cumulative axle model, existing imputing methods were applied to the ten previously determined survey sites. Fig. 5 compares the actual daily traffic volumes with the imputed daily traffic volumes for the above missing data imputing methods and the cumulative axle model using graphs. The graphs show that the cumulative axle method is significantly more precise than the other methods. Moreover, the distribution range of ARIMA and exponential smoothing, which are time series analysis methods, is slightly narrower than the existing methods and interpolation methods. In the case of the cumulative axle model, although it is highly accurate, the estimated values are generally slightly greater than the actual values.

Table 6 compares the MAPE obtained using the cumulated axle model to that obtained using the conventional imputation methods. The MAPE obtained for the existing methods does not show any correlation with the proportion of two-axle vehicles. Fig. 6 depicts the graph of MAPE for various imputation methods. The cumulative axle method gives a MAPE of 2.92%, confirming that it is significantly more accurate than the other methods, which have MAPE averaging 10%. This is because the model estimates based on actual data, which enables it to maintain a small error even in the event of sudden changes in traffic volume.

Table 6: Comparison of accuracy between existing imputation methods and axle cumulative model

ID	Proportion of 2-axle vehicles	Historical Approach	ARIMA	Exponential smoothing	Linear Interpolation	Quadratic Interpolation	Axle Cumulative Model
42001	0.781	9.85	6.59	6.80	13.23	14.14	4.43
20353	0.831	8.53	10.12	9.73	11.35	11.95	4.99
70561	0.842	8.88	7.14	9.86	7.07	9.24	5.51
40023	0.888	8.24	7.55	7.05	7.74	9.29	3.04
40411	0.893	6.86	6.73	5.15	8.86	9.73	2.87
70012	0.923	12.26	6.57	6.81	7.44	7.86	3.43
60608	0.944	8.26	7.13	8.03	6.78	7.05	1.48
21012	0.96	11.07	12.46	11.11	8.34	11.39	1.47
40493	0.977	19.94	17.59	16.57	25.11	28.86	1.64
61010	0.991	13.20	13.93	14.42	9.87	10.27	0.37
Average		10.71	9.58	9.55	10.58	11.98	2.92

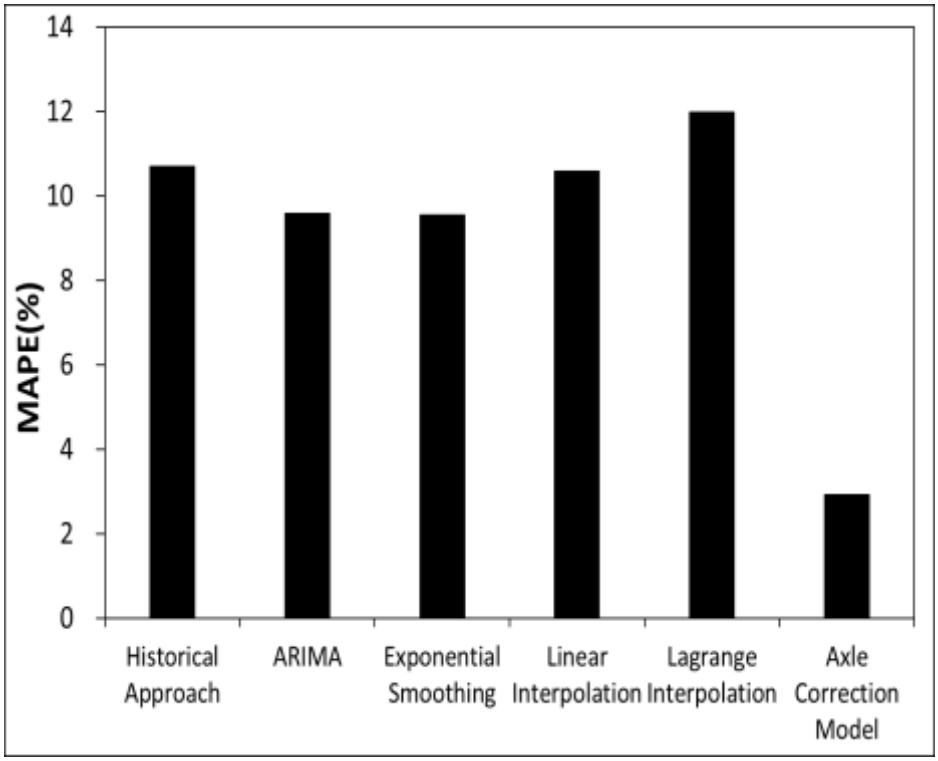


Fig.6: MAPE graph of imputation methods

VII. CONCLUSION

Instances of missing data occur at AVC stations for various reasons, including malfunctioning sensors and controllers, during 365 days of continuous data collection. However, these missing data must subsequently

be imputed for effective utilization of traffic volume data. AVC stations consist of P-L-P or L-P-L sensors for each lane and traffic volume data is gathered by the loop sensor. If this loop sensor malfunctions, the traffic volume can be estimated using the piezo sensor. The cumulative axle model, which creates missing data via linear regression analysis of the traffic volume per unit time and accumulated number of axles, can be applied in such cases.

The daily traffic volume data for 2014, with no missing data, from ten survey sites with various proportions of two-axle vehicles were selected. Subsequently, traffic data for 54 days were assumed to be missing. In the case of the cumulative axle model, the model accuracy increased as the proportion of two-axle vehicles increased. Among the 637 AVC locations along the Korean routes, more than 94% had proportions of two-axle vehicles greater than 90%. The fact that the proportion of two-axle vehicles was more than 90% in the Virginia and Alaska cases also show that it is more advantageous to use the cumulative axle model to estimate missing data. During imputation of the 54 days of missing data, the cumulative axle method recorded a MAPE of 2.92%. This indicates that the method has outstanding accuracy when compared to the 10% error obtained with the existing methods. This is because based on the limited actual data the model maintains a small error even in the event of sudden traffic volume changes.

The accuracy of the proposed operation plans for defective AVC stations using the cumulative axle model and piezo sensors to provide missing traffic volume data has been confirmed. The greatest advantage of this method is its simplicity of logic, which makes it easy to automatically apply this axle imputation model according to sensor operation conditions by implementing it in AVC stations. In future work, the model needs to be improved to consider daily variability and the effect of traffic in different lanes in order to counter increases in error when the proportion of two-axle vehicles decreases. Moreover, the question of whether vehicle type classification is possible using the cumulative axle model needs to be confirmed, and its sensitivity analysis performance obtained.

REFERENCES

[1]. H. Chang, D. Park, Y. Lee, and B. Yoon, "Multiple time period imputation technique for multiple missing traffic variables: nonparametric regression approach," Can. J. Civ. Eng., vol. 39, no. 4, pp. 448–459, 2012.

[2]. M. Zhong, P. Lingras, and S. Sharma, "Estimation of missing traffic counts using factor, genetic, neural, and regression techniques," Transp. Res. Part C Emerg. Technol., vol. 12, no. 2, pp. 139–166, Apr. 2004.

[3]. Jang Jin Hwan and Baik, Nam Chul, "Study on Imputation Technique for Missing Traffic Volume Data" 2005 Korean Society of Civil Engineers Conference, pp. 4060–4066, Oct. 2005.

[4]. FHWA, "Traffic Monitoring Guide." US Department of Transportation, 2013.

[5]. J. Tang, G. Zhang, Y. Wang, H. Wang, and F. Liu, "A hybrid approach to integrate fuzzy C-means based imputation method with genetic algorithm for missing traffic volume data estimation," Transp. Res. Part C Emerg. Technol., vol. 51, pp. 29–40, Feb. 2015.

[6]. L. Li, Y. Li, and Z. Li, "Missing traffic data: comparison of imputation methods," IET Intell. Transp. Syst., vol. 8, no. 1, pp. 51–57, Feb. 2014.

[7]. M. Zhong, S. Sharma, and P. Lingras, "Genetically designed models for accurate imputation of missing traffic counts," Transp. Res. Rec. J. Transp. Res. Board, no. 1879, pp. 71–79, 2004.

[8]. B. M. Williams and L. A. Hoel, "Modeling and forecasting vehicular traffic flow as a seasonal ARIMA process: Theoretical basis and empirical results," J. Transp. Eng., vol. 129, no. 6, pp. 664–672, 2003.

[9]. H. Abdelgawad, T. Abdulazim, B. Abdulhai, A. Hadayeghi, and W. Harrett, "Data imputation and nested seasonality time series modelling for permanent data collection stations: methodology and application to Ontario," Can. J. Civ. Eng., vol. 42, no. 5, pp. 287–302, 2015.

[10]. K. Wiegand, "Traffic Monitoring Program," 2013.

[11]. A. Vandervalk-Ostrander, AASHTO Guidelines for Traffic Data Programs. Aashto, 2009.

[12]. H. Desai, W. Cunagin, K. Cunagin, D. Hoyt, R. L. Reel, and S. Bentz, "Application of Seasonal Adjustment Factors to Subsequent Year Data," Transp. Res. Rec. J. Transp. Res. Board, vol. 2443, no. 1, pp. 143–147, 2014.

[13]. L. J. French, R. W. Eck, J. S. D'Angelo, and W. Virginia, "Use of Permanent Traffic Recorder Data To Develop Factors for Traffic and Truck Variations," West Virginia Department of Transportation, Division of Highways, 2002.

[14]. MassachusettsDOT, 2014 Traffic Counting Report for Cape Cod Massachusetts. 2014.

[15]. Y. Yoo, "An innovative methodology for converting axle counts to vehicle counts," presented at the ARRB Conference, 26th, 2014, Sydney, New South Wales, Australia, 2014.

[16]. Z.-P. LI, H. YU, Y.-C. LIU, and F.-Q. LIU, "An Improved Adaptive Exponential Smoothing Model for Short-term Travel Time Forecasting of Urban Arterial Street," Acta Autom. Sin., vol. 34, no. 11, pp. 1404–1409, Nov. 2008.
[17]. E. Kreyszig, Advanced engineering mathematics. Wiley, 2011.
[18]. VirginiaDOT, Average daily traffic volumes with vehicle classification data on intersatate, arterial and primary routes. 2012.
[19]. AlaskaDOT, Central region traffic volume report 2009-2010-2011. 2011.