Transportation Research Forum

Developing a Strategy for Imputing Missing Traffic Volume Data
Author(s): Dr. Mei Chen, Jingxin Xia, and Dr. Rongfang (Rachel) Liu
Source: *Journal of the Transportation Research Forum*, Vol. 45, No. 3 (Fall 2006), pp. 57-75
Published by: Transportation Research Forum
Stable URL: http://www.trforum.org/journal

Developing a Strategy for Imputing Missing Traffic Volume Data

by Dr. Mei Chen, Jingxin Xia, and Dr. Rongfang (Rachel) Liu

Archived ITS-generated data can provide a potential resource for many long-term transportation applications. However, missing and suspicious data are inevitable due to detector and communication malfunctions. This paper presents a comparative analysis of various techniques for imputing missing traffic volume data in the archived data management system in Kentucky. The applicability of the techniques, as well as their reliability in terms of data requirement, is also discussed. An implementation strategy for the Kentucky archive data management system is then developed based on the performance and the applicability/reliability analyses.

INTRODUCTION

As part of an intelligent transportation system, various traffic monitoring devices collect large amounts of traffic data to fulfill the operational and management needs of the highway system. Such data may also have great potential for long-term applications. In order to help realize the benefit of such data, an archived data user service has been incorporated into the National ITS Architecture. A number of archived data management systems (ADMS) have been developed in the past decade, aiming to systematically maintain and re-use this transportation data. However, hardware, software, and communication problems with traffic surveillance devices may cause large amounts of missing or suspicious data. Such discrepancies raise concerns about the reliability of such data. To maximize the cost effectiveness of the data collection infrastructure, it is necessary to impute and replace the missing and suspicious data, as recommended in recent works such as Margiotta (1998).

Imputation techniques developed thus far can be classified into three major categories: regression, nearest neighbor and deck replacement, and classification. Their applications in traffic data imputation have been presented by Smith et al. (2003), Al-Deek and Chandra (2004), and Gold et al. (2001) for archived data management systems in Virginia, Florida, and Texas. In addition, Zhong et al. (2004) discussed the application of neural networks and a genetic algorithm for imputing missing traffic data from permanent count stations. The imputation techniques presented by these works vary in complexity and accuracy. Most of these studies focus on the performance of each imputation algorithm.

The objective of this study is to develop an implementation strategy for the archived data management system in Kentucky through comparative analysis of various imputation algorithms for traffic volumes. The following sections of the paper include a brief description of the archived ITS data used in the study and several imputation techniques and their performance comparison. A system-wide imputation strategy is then developed based on the algorithms' individual performance, applicability, and reliability of the data sources. Finally, conclusions and recommendations are presented.

KENTUCKY ADMS

There are two regional ITS deployments in Kentucky, ARTIMIS (Advanced Regional Traffic Interactive Management & Information System), serving Cincinnati/Northern Kentucky, and TRIMARC (Traffic Response and Incident Management Assisting the River Cities), serving Louisville/Southern Indiana. Various traffic detectors (such as loops and radar) collect large amounts of traffic data each day. There are more than 100 detectors under these two systems that

are located in Kentucky. Traffic flow, speed, and lane occupancy data (i.e., percentage time the detector is occupied by vehicles) are available at 15-minute intervals for each of them. An archived data management system (http://adms.uky.edu) has been developed to archive and disseminate information derived from this data to users. The discussion in this paper is based on data from the TRMIARC system.

Due to hardware and communication problems, the data completeness index (total available data as a percentage of total expected data) in TRIMARC was about 75.4% in 2002, and 12.6% of this available data was flagged as suspicious/erroneous based on quality screening rules developed by Lomax et al. (2002). These rules use basic relationships among traffic variables to perform multivariate consistency checks on the data. The majority of the flagged data records bear the signature of device/communication malfunction, such as a zero vehicle count or repeating exact same counts over extended periods of time. Table 1 shows a breakdown of the results summarizing the quality control test for TRIMARC volume data in 2002. The task of imputation is to estimate the values that are missing or labeled as suspicious/erroneous.

Table 1: Overall Quality of 15-Minute Volume Data

Traffic Data Quality Screening Rules	Number of Records Flagged
No vehicle present (Volume = 0, Speed = 0, Occupancy = 0)	10,532
Consistency of elapsed time between records	0
Invalid date, time	1,194
Maximum occupancy (80%)	385
8 Consecutive identical volume	19,553
Muliti-variate consistency (Speed = 0, Volume = 0, Occupancy < > 0)	0
Truncated occupancy values of zero (Occurred when software truncates or rounds to integer value)	11
Muliti-variate consistency (Volume = 0, Speed > 0)	1,494
Maximum volume (750 vehs per 15-min)	10
Total number of records flagged as suspicious for volume	23,396

IMPUTATION ALGORITHMS

A review of existing works by researchers indicates that various heuristic methods could perform well in imputing missing values. Smith et al. (2003) analyzed such methods as average of surrounding detectors, average of surrounding time periods, historic average, and factor up. More sophisticated statistical procedures, such as expectation maximization and data augmentation, were also examined. These algorithms provide accurate estimates, but their high complexity and dependency on input data make them less attractive.

Preliminary analyses of various imputation algorithms were performed using Kentucky data. Initial screening of the algorithms was conducted based on their applicability, accuracy, and complexity. Several algorithms, including historical average, spatial and temporal interpolations, artificial neural network, expectation maximization, as well as a hybrid algorithm combining historical profile and time series analysis, were tested. It was determined that the following techniques should be further evaluated for implementation in Kentucky ADMS.

Historical Average (HIST)

Historical average assumes that traffic volumes tend to be stable over time. In this study, the average volume of the same time-of-day and day-of-week in a historical database was used to impute missing volume. For example, if the k-th time period of a Tuesday was missing in May, then that volume was imputed as the average of all the k-th time periods of all the Tuesdays in the historical database.

Temporal Interpolation (TpI)

The temporal interpolation algorithm imputes a missing value by replacing it with the average of its preceding and succeeding values at the same site. For example, if the volume for the k-th time period of a Tuesday in May was missing, then it was imputed as the average of volumes for the (k-1)-th time period and the (k+1)-th time period, if these two items are available. This algorithm is relatively accurate when there is no significant change in traffic over a short period. However, it is not applicable to cases where volume data is missing for continuous periods of time.

Spatial Interpolation (SpI)

Recognizing the correlation between volumes at adjacent detector stations, spatial interpolation imputes missing traffic volume at a detector station by replacing it with the average between the volumes measured at its upstream and downstream stations. Detectors are installed along a highway at various locations, and each detector station contains a group of detectors installed to monitor traffic conditions in all lanes at that location (mile point). There may be a number of detector stations along a stretch of a highway. For a detector station on an eastbound (northbound) highway, its upstream station is that immediately to its west (north), and its downstream station is that immediately to its east (south). Whether to use the readings from the same time period or not depends upon the distances between stations. In this study, since the distances between adjacent stations are rather short – around 0.6 mile in most cases, traffic volume data for the same time period was used. This algorithm generally works well if the stations are close to each other with no entrance or exit ramps in between. However, its performance is highly dependent upon the accuracy of the data collected at the adjacent detectors, and it is only applicable when traffic volume data are available from both adjacent stations.

Hybrid Algorithm (Hybrid)

A hybrid algorithm integrating the historical average algorithm and time series analysis was developed to impute missing volume data. This algorithm reflects the temporal variation of flow by time-of-day and day-of-week through a triple exponential smoothing model on top of the historical average procedure introduced earlier. The triple exponential smoothing procedure constructs three statistically related series: the smoothed data series, the seasonal index, and the trend series to address the temporal features of the data. Details of the triple exponential smoothing can be found in texts on time series modeling. The basic equations for the triple exponential smoothing model are given as:

(1) $S_t = \alpha \frac{y_t}{I_{t-L}} + (1-\alpha)(S_{t-1} + b_{t-1})$ Overall smoothing

(2) $b_t = \gamma(S_t - S_{t-1}) + (1-\lambda)b_{t-1}$ Trend smoothing

(3) $I_t = \beta \frac{y_t}{S_t} + (1-\beta) I_{t-L}$ Seasonal smoothing

(4) $F_{t+m} = (S_t + mb_t) I_{t-l+m}$ Forecast

where *y* is the observation, *S* is the smoothed observation, *b* is a trend factor, *I* is a seasonal index, *F* is the forecast at m-th period ahead, *L* is the number of periods in a complete season's data, *t* is an index denoting time period, and α, β and γ are constants whose values are obtained through minimizing the mean squared error between the estimates and the observations.

Since traffic flow shows variations over both time-of-day and day-of-week, the period of one week was selected as a season. The hybrid model starts with filling the missing volumes in the time series with the historical averages as described earlier. It then runs the triple exponential smoothing procedure on the time series.

Artificial Neural Network (ANN)

Artificial neural network models can also be used to impute missing values. An artificial neural network is an information processing paradigm which is inspired by the human nervous system. A typical ANN architecture consists of input layers, hidden layers, and output layers. Information from input layers is passed on to hidden layers where the training of the network takes place, after which the output is shown in the output layer. Various ANN models have been developed for transportation applications in recent years, such as highway traffic data analysis and traffic flow or travel time prediction (e.g., Nelson and Palacharla 1993, Pourmollem et al. 1997, Park et al. 1998, and Jiang and Adeli 2006) and transit operation analysis (Chen and Liu 2005 and Chen et al. 2004).

Based on data availability and the missing data pattern, two models of ANN were developed in this analysis. Preliminary tests were performed to select the appropriate neural network structure and learning algorithm. The results showed that a multilayer perceptron network with back propagation learning mechanism had the best performance. The multilayer perceptron is a commonly used neural network with layers of perceptrons (or neurons) that estimate output based on linear combinations of input values. A nonlinear activation function can be used to map nonlinear relationships between the input and output. Back propagation is a training algorithm that involves repeated presentations of input data to the neural network. With each iteration, the difference between the model output and the desired output provides feedback to the neural network for weight adjustment in the next iteration. The goal of the training process is to maximize the match between the model and desired outputs.

Two multilayer perceptron models were developed: ANN Model I was designed to have time-of-day, speed, and occupancy at the current station, and traffic volumes at the adjacent upstream and downstream stations as input variables. This was based on the presumption that traffic flow rates at adjacent stations are correlated. However, considering that sometimes the data from adjacent stations may also be missing, ANN Model II was designed to take inputs from the data collected at the current station, such as time-of-day, speed, and occupancy. Each of the ANN models was trained and validated using the data collected at detector stations. The missing volume was then imputed by feeding the input variables to the neural networks. It should be noted that the ANN model generally requires extensive computing resources in the development and implementation process.

PERFORMANCE EVALUATION

Data Samples

A site at mile-point 6.7 on eastbound I-64 was chosen to illustrate the performance evaluation procedure for each of the imputation algorithms. Traffic flow, speed, and occupancy data for this site was archived at 15-minute intervals. This site was chosen because the detector functioned very well and most of its data records passed quality screening as defined by Lomax et al. (2002). Two other detectors located at mile-points 6.3 and 7.0 were considered as the upstream and downstream sites, respectively. Since the imputation performance needs to be assessed on the same set of data, a common set of test data on which all imputation algorithms were applicable was selected from the set of data that passed the quality screening.

Performance Indices

To evaluate the performance of the imputation algorithms, error indices were calculated based on the deviation of the imputed values from the observed values. The two major indices used were Mean Absolute Percentage Error (MAPE) and Root Mean Square Error (RMSE), which can be estimated as follows.

$$(5)\quad MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{Y_i - Y_i^*}{Y_i}\right| \times 100$$

$$(6)\quad RMSE = \frac{1}{n}\sqrt{\sum_{i=1}^{n}(Y_i - Y_i^*)^2}$$

where, n is the number of imputed values, Y_i is the observation value, and Y_i^* is the imputed value.

Initial Screening of Imputation Techniques

Initial screening of all the algorithms based on 2002 data indicated that all the algorithms performed well except the ANN model II. The MAPE and RMSE indices for ANN model II were 47.4% and 188.5 vehicles, respectively. For the other imputation algorithms, the MAPE index ranged from 7.4% to 12.9%, while the RMSE index ranged from 39.6 to 61.3 vehicles. Therefore, ANN model II was removed from the list of imputation algorithms to be considered further.

Performance Comparison

The performance of each imputation algorithm was evaluated and compared based upon missing data pattern by time-of-day and day-of-week. This was done to account for variation in performance.

Performance by Missing Data Percentage. Preliminary analyses indicated that there were different percentages of missing data at various data collection sites. These percentages ranged generally from 15% to 45%. To account for this variation, the performance indices (i.e., MAPE and RMSE) were used to evaluate each algorithm based on datasets with 20%, 30%, and 40% missing records. This was achieved by randomly extracting these percentages from the data set.

According to Table 2, there is no significant difference in the MAPE and RMSE indices among the three missing data percentages for any algorithm. To further verify this, analyses of variance (ANOVA) were performed to test the significance of missing data in each imputation algorithm in

terms of percentage error (defined as the difference between the imputed and the measured volumes as a percentage of the measured volume). The p-values for these tests ranged from 0.46 to 0.94 as shown in Table 3, indicating an insignificant impact of missing data at a 5% significance level.

Table 2: Imputation Performance by Missing Data Percentage

	20% Missing Volumes		30% Missing Volumes		40% Missing Volumes	
	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE
HIST	0.129	57.495	0.127	57.526	0.129	57.339
Hybrid	0.094	51.255	0.094	50.857	0.095	52.150
SpI	0.095	61.208	0.095	61.679	0.095	61.136
TpI	0.074	38.843	0.073	39.269	0.074	39.760
ANN I	0.100	42.024	0.099	42.403	0.100	43.093

Table 3: ANOVA p-Values with Respect to Missing Percentage, Day-of-Week, and Time-of-Day

Algorithm	HIST	Hybrid	SpI	TpI	ANN I
Missing Percentage	0.57	0.85	0.94	0.46	0.87
Day-of-week	0.13	0.02	0.38	0.27	0.84
Time-of-day	0.00	0.00	0.00	0.00	0.00

For each missing percentage, TpI outperformed other algorithms in both the MAPE and RMSE indices. Additionally, the hybrid algorithm had a slightly better performance than the SpI algorithm, while the performance of the ANN I algorithm in terms of the MAPE index was comparable to that of SpI. The HIST algorithm trailed the other algorithms in performance, and there was no change in the performance ranking across the three missing percentages.

Performance by Day-of-Week and Missing Data Percentage. Table 3 also shows that for most algorithms the day-of-week factor did not have a significant impact on imputation accuracy. The p-values with respect to the day-of-week factor ranged from 0.02 (for Hybrid) to 0.84 (for ANN I) except for Hybrid. This may be attributable to variation in traffic flow patterns between weekday and weekend and the ability of each algorithm to adapt to this change.

Further performance comparison was conducted to examine the impact of the time-of-day factor. Table 4 shows the values of the MAPE and RMSE indices by imputation algorithm, missing data percentage, and day-of-week. It can be observed that, within each day-of-week category, there were no significant differences in performance measures among the three levels of missing data percentages. Both MAPE and RMSE varied within a very small range for each imputation algorithm within either weekday or weekend. Analysis of variance tests using data from each category supported this inference.

Table 5 shows that at a 5% significance level missing data percentage did not have significant impacts on imputation accuracy of an algorithm for both weekdays and weekends. On the other hand, time-of-day was a significant factor in imputation accuracy during weekends in all algorithms except Hybrid and TpI. This exception can be explained by the fact that traffic volume is less variable during weekends and also that both algorithms already have built-in considerations of the time-of-day factor.

Table 4: Imputation Performance by Day-of-Week and Missing Data Percentage

Day of Week	Algorithm	20%		30%		40%	
		MAPE	RMSE	MAPE	RMSE	MAPE	RMSE
Weekday	HIST	0.119	58.247	0.122	58.909	0.118	58.336
	Hybrid	0.090	54.156	0.092	52.324	0.091	55.089
	SpI	0.101	68.681	0.098	64.741	0.101	68.542
	TpI	0.070	41.577	0.072	40.610	0.072	42.857
	ANN I	0.096	45.411	0.095	43.826	0.097	46.802
Weekend	HIST	0.158	55.430	0.155	53.018	0.157	54.571
	Hybrid	0.106	42.515	0.107	41.282	0.107	43.281
	SpI	0.078	33.826	0.079	33.603	0.079	34.050
	TpI	0.083	30.321	0.080	28.960	0.080	29.899
	ANN I	0.110	31.181	0.111	30.010	0.111	31.015

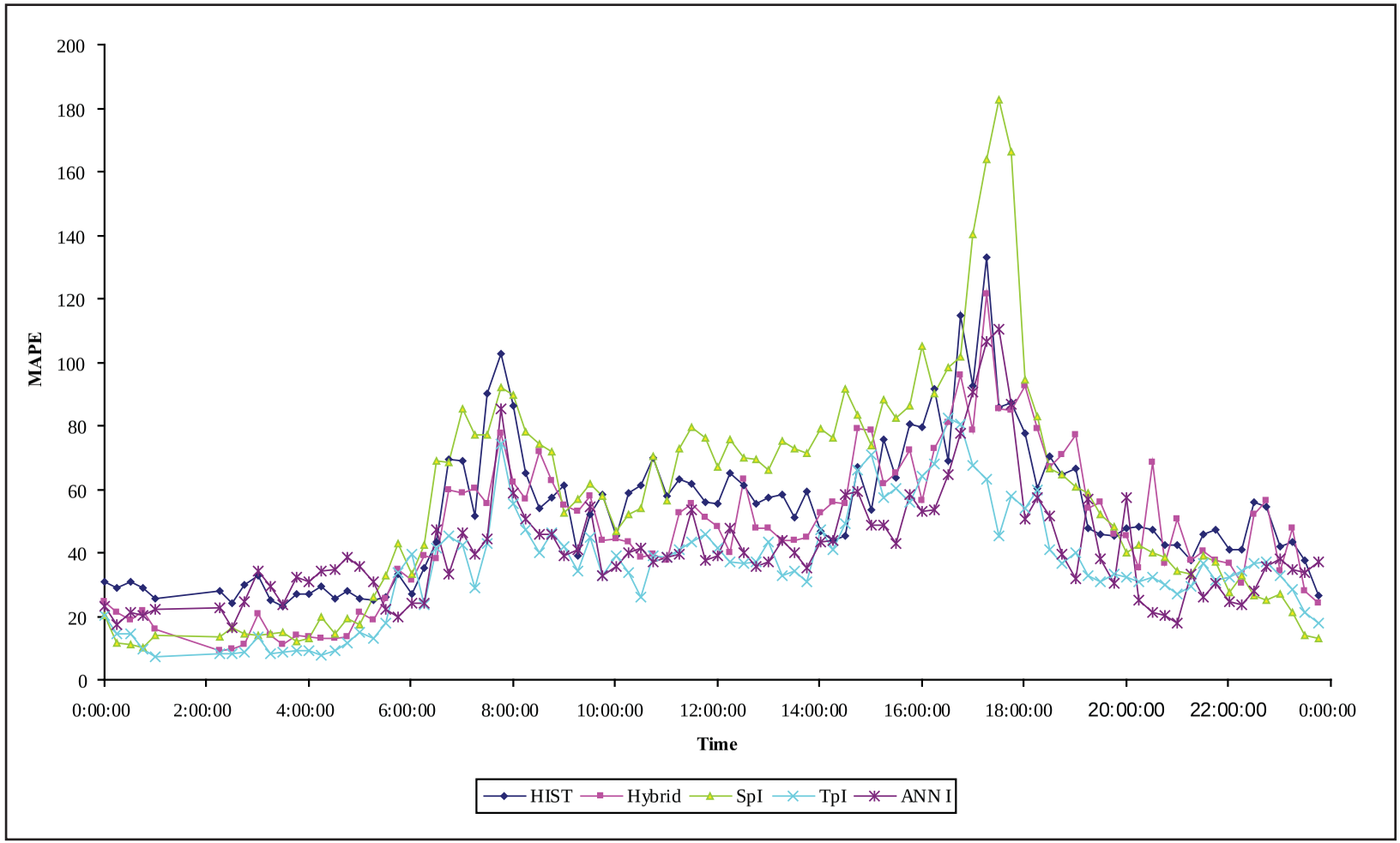
Table 5: ANOVA p-Values with Respect to Missing Percentage and Time-of-Day

Algorithm	HIST	Hybrid	SpI	TpI	ANN I
Weekday					
Missing Percentage	0.95	0.72	0.92	0.37	0.91
Time-of-day	0.00	0.00	0.00	0.00	0.00
Weekend					
Missing Percentage	0.26	0.81	0.69	0.27	0.84
Time-of-day	0.00	0.07	0.29	0.00	0.00

Performance by Time-of-Day, Day-of-Week, and Missing Data Percentage. To further account for the potential impact of time-of-day variation of traffic flow on performance, the MAPE and RMSE indices were calculated by time-of-day on both weekdays and weekends. By analyzing the flow patterns by time-of-day in TRIMARC, a weekday was divided into five periods: early morning (EM) from midnight to 6:00 a.m., morning peak (MP) from 6:00 a m. to 10:00 a m., mid-day (MD) from 10:00 a m. to 4:00 p.m., afternoon peak (AP) from 4:00 p m. to 8:00 p m., and late night (LN) from 8:00 p.m. to midnight, respectively. A typical weekend day was divided into two periods: daytime from 6:00 a m. to 8:00 p.m. and nighttime from midnight to 8:00 a.m. and from 8:00 p.m. to midnight.

Figure 1 shows comparisons of the MAPE and RMSE indices for all five algorithms on a weekday with 20% missing data. The MAPE and RMSE indices by time-of-day are shown in Table 6(a). Both measures showed that the TpI algorithm outperformed the others during the early morning. The Hybrid and SpI indices had similar performance and seemed to outperform the ANN I and HIST algorithms. During late night, both the MAPE and RMSE indices favored the TpI and SpI algorithms, and both algorithms outperformed the other three. For morning peak, afternoon peak, and mid-day, TpI and ANN I had comparable performance and both seemed to provide more accurate results than the others. A paired t-test indicated that, under a 5% significance level, TpI had a slightly better performance than ANN I. Hybrid was the next in line for these periods. Similar

Figure 1(a): Performance Comparison on Weekday with 20% Missing Data - MAPE



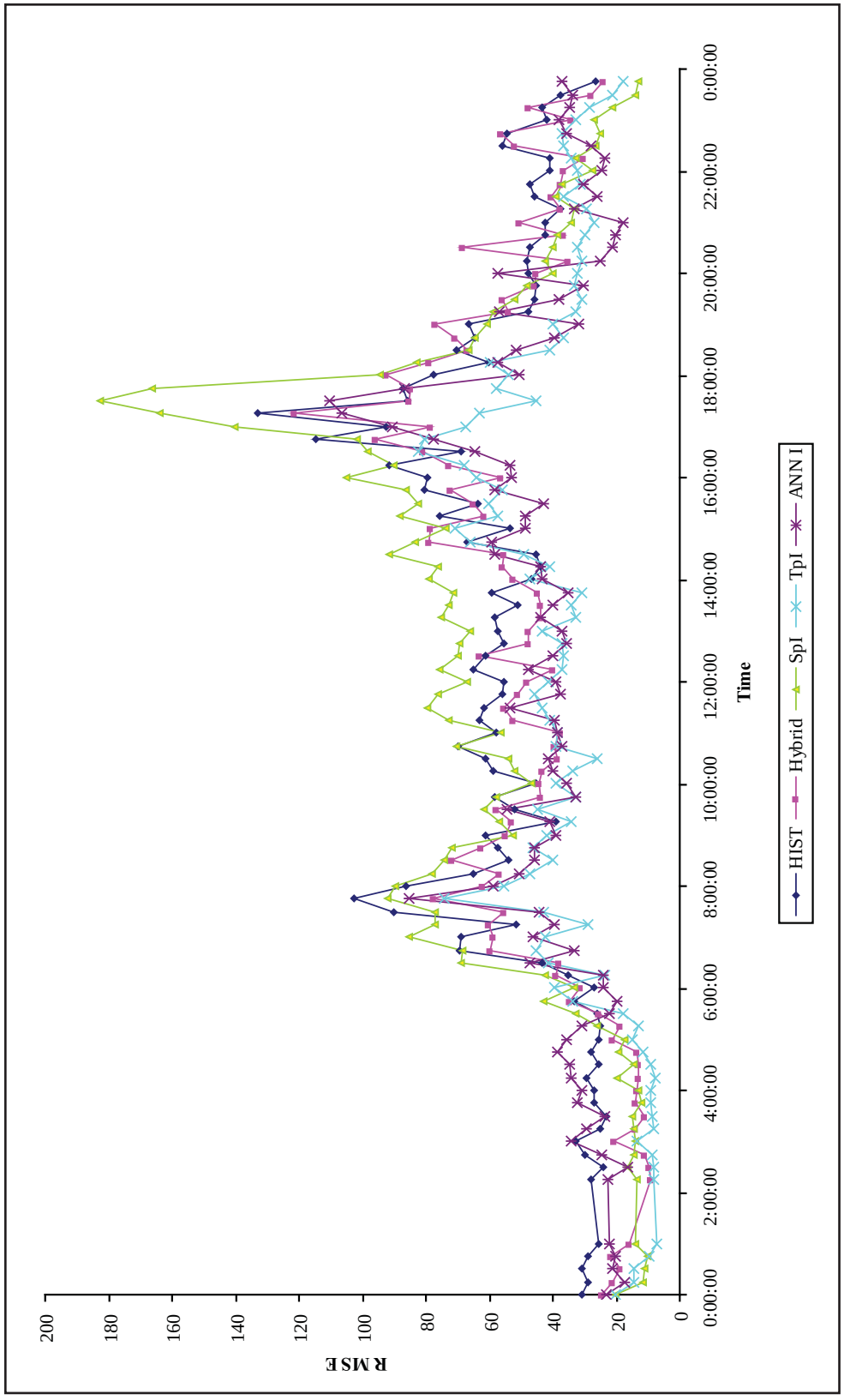


Figure 1(b): Performance Comparison on Weekday with 20% Missing Data - RSME

comparisons were conducted for weekend days on which the performance indices were evaluated for daytime and nighttime, as illustrated in Table 6(b). During the daytime, ANN I, TpI, and SpI were in the first tier in terms of their performance, followed by Hybrid. For nighttime, TpI and SpI were the leaders followed by Hybrid and ANN I.

The performance of the algorithms was also tested using datasets with 30% and 40% missing records. Although differences existed between the values of MAPE and RMSE by day-of-week and time-of-day, the general performance ranking was the same as observed with 20% missing data.

Additional Tests

To generalize the results, two additional sets of data were chosen from the TRIMARC system. One set was collected from Interstate 71 northbound at mile-point 3.0, with its upstream and downstream stations at mile-points 2.3 and 3.8 in 2003; the other was from Interstate 65 northbound at mile-points 134.6, with its upstream station at mile-point 134.2 and downstream station at mile-point 134.8 in 2004. The same tests and performance comparisons were conducted on these sites. Similar to what was observed previously, the impact of missing data percentage on the performance of the algorithms was not statistically significant for either set of data.

The performances of the algorithms by time-of-day and day-of-week are shown in Table 7 and Table 8 for the I-71 and I-65 data, respectively. It can be observed that TpI continued to produce the highest imputation accuracy for all time periods and all days using the MAPE and RMSE indices. The second best was usually among ANN I, SpI, and Hybrid. For most times, HIST appeared to be the least accurate among the algorithms. While the values of the MAPE index did not appear to have an obvious trend, the RMSE values were usually larger during peak periods. The I-64 and I-71 sites, of which 2002 and 2003 data were used, are both located outbound from the city of Louisville. The largest RMSE values were observed during the afternoon peak when both sites experienced their heaviest traffic flows of a typical work day. The I-65 site, of which 2004 data was used in the analysis, is located on the inbound direction; thus the largest RMSE values were observed during the morning peak when traffic was usually the heaviest for a weekday. For weekends the algorithms tended to perform better during night time.

The performance evaluation by time-of-day, day-of-week, and missing data percentage helped to establish the final implementation strategies. The results showed that missing data percentage did not have a significant impact on the rank order of a particular algorithm. Therefore, it was not considered in developing an implementation strategy. However, time-of-day and day-of-week factors affected the rank of an algorithm.

Table 6: Imputation Performance by Time-of-Day with 20% Missing Data for 2002

a) Weekday

Index	Imputation Algorithm	EM	MP	MD	AP	LN
MAPE	HIST	0.257	0.094	0.076	0.084	0.133
	Hybrid	0.134	0.081	0.064	0.081	0.108
	SpI	0.136	0.106	0.090	0.101	0.085
	TpI	0.097	0.063	0.055	0.058	0.086
	ANN I	0.257	0.062	0.051	0.064	0.104
RMSE	HIST	27.7	61.2	60.3	75.0	43.6
	Hybrid	17.9	56.0	53.1	75.4	41.1
	SpI	18.4	68.8	74.7	94.5	30.1
	TpI	13.6	42.5	44.7	51.6	30.5
	ANN I	26.5	45.3	44.3	62.7	28.7

(b) Weekend

Index	Imputation Algorithm	Nighttime	Daytime
MAPE	HIST	0.234	0.117
	Hybrid	0.137	0.093
	SpI	0.090	0.073
	TpI	0.105	0.072
	ANN I	0.191	0.070
RMSE	HIST	49.2	50.6
	Hybrid	33.5	41.7
	SpI	23.8	34.8
	TpI	22.5	31.8
	ANN I	31.8	28.0

Table 7. Imputation Performance by Time-of-Day with 20% Missing Data for 2003

(a) Weekday

Index	Imputation Algorithm	EM	MP	MD	AP	LN
MAPE	HIST	0.121	0.147	0.090	0.124	0.136
	Hybrid	0.107	0.104	0.086	0.091	0.123
	SpI	0.064	0.106	0.067	0.072	0.093
	TpI	0.068	0.077	0.043	0.059	0.069
	ANN I	0.311	0.123	0.085	0.087	0.162
RMSE	HIST	32.727	66.126	58.102	96.356	46.373
	Hybrid	27.621	47.401	53.133	80.313	39.453
	SpI	29.101	67.316	54.611	78.337	45.310
	TpI	20.569	31.006	27.701	44.641	24.047
	ANN I	41.636	67.086	55.953	69.662	53.946

(b) Weekend

Index	Imputation Algorithm	Nighttime	Daytime
MAPE	HIST	0.166	0.133
	Hybrid	0.127	0.110
	SpI	0.087	0.083
	TpI	0.062	0.055
	ANN I	0.241	0.153
RMSE	HIST	44.693	49.934
	Hybrid	34.288	41.877
	SpI	38.552	43.604
	TpI	15.795	19.990
	ANN I	47.384	50.214

Table 8: Imputation Performance by Time-of-Day with 20% Missing Data for 2004

(a) Weekday

Index	Imputation Algorithm	EM	MP	MD	AP	LN
MAPE	HIST	0.080	0.140	0.074	0.083	0.107
	Hybrid	0.079	0.081	0.060	0.070	0.093
	SpI	0.083	0.060	0.053	0.047	0.060
	TpI	0.051	0.059	0.028	0.039	0.046
	ANN I	0.120	0.058	0.045	0.048	0.090
RMSE	HIST	30.180	145.054	87.690	93.060	66.800
	Hybrid	27.404	103.901	81.002	81.753	72.379
	SpI	24.204	71.054	59.937	51.510	36.451
	TpI	21.654	70.866	35.216	42.427	31.625
	ANN I	32.809	70.095	50.702	52.625	49.799

(b) Weekend

Index	Imputation Algorithm	Nighttime	Daytime
MAPE	HIST	0.122	0.108
	Hybrid	0.091	0.073
	SpI	0.101	0.067
	TpI	0.043	0.035
	ANN I	0.124	0.064
RMSE	HIST	61.853	86.873
	Hybrid	46.068	61.230
	SpI	35.295	44.579
	TpI	20.778	25.279
	ANN I	44.209	43.365

IMPLEMENTATION ISSUES

While the performances of the imputation techniques are extremely important in developing an implementation strategy, the applicability of each algorithm needs to be analyzed. In addition, the reliability of the algorithms will be discussed in terms of data requirement.

Applicability

The applicability of an imputation algorithm can be measured by the amount of data it can impute for a given site. Using a total of 725,981 missing data records (which are approximately 35% of the total expected amount) in the TRIMARC system during 2002, the hybrid algorithm and the historical average algorithm are able to impute all missing values provided that a historical database is available. Meanwhile, only about 21% of these missing data can be imputed by TpI, and 22% by SpI and ANN I.

Reliability

The reliability of an imputation algorithm depends on the reliability of the data used. As previously discussed, algorithms, such as HIST, Hybrid, and TpI, only need volume data collected at the current site. Thus, the accuracy of these imputation algorithms depends on the quality of this data or the reliability of the detector that records the data. On the other hand, SpI and ANN I both require traffic volumes from the adjacent sites in addition to those from the current site. Therefore, their performance is tied to the reliability of the adjacent detectors as well.

To illustrate the data reliability issue, we compared the detector data with those collected by the automatic traffic recorders (ATR) at the same site during a 48-hour period. Figure 2(a) displays the comparison of hourly traffic volumes recorded by these two devices. A very good match between the readings from the two devices can be observed, with a mean absolute percentage difference between them of approximately 2%. At other sites, however, the reliability of the collected data was rather low. Figure 2(b) shows a similar comparison at a site on I-65 at mile-point 131.2, indicating that the traffic flow recorded by the TRIMARC device was systematically lower than that recorded by ATR. This could be caused by system errors that might require recalibration of the devices.

Based on this information, a tradeoff between performance and reliability needs to be considered in developing an implementation strategy for the imputation algorithms. To achieve this goal, understanding the data required for implementing each algorithm is necessary. Table 9 summarizes the information required by each of these algorithms. It shows that the rank order of the algorithms from the least demanding to the most demanding in terms of input data is HIST, Hybrid, TpI, SpI, and ANN I. This should be factored into consideration during the development of the implementation strategy.

Table 9: Data Requirements for Implementation

Imputation Algorithms	Data Required
HIST	Historic volumes by time-of-day and day-of-week at current site
TpI	Volumes collected in currounding time periods at current site
SpI	Volumes collected at surrounding sites within the same period
Hybrid	Historic volumes by time-of-day and day-of-week at current site
ANN I	Time series volumes, speeds, and occupancies at current site; Time series volumes collected at both surrounding sites

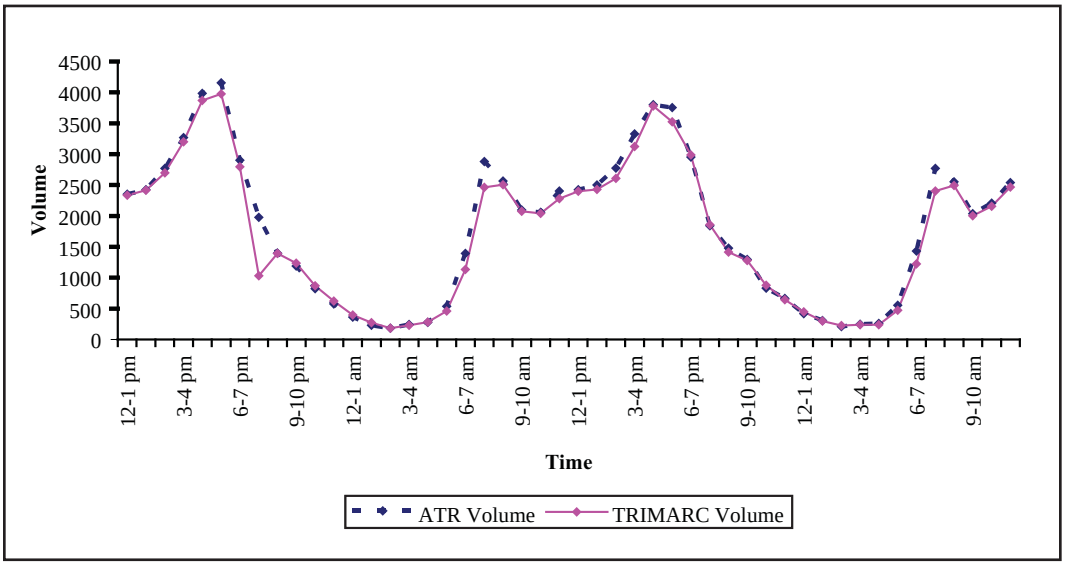
Implementation Strategy

Based on the performance evaluation and applicability/reliability analyses, an implementation strategy was developed for the archived data collected by detectors in the TRIMARC system. Since TpI always had the highest imputation accuracy under all circumstances tested, it should always be the first choice if applicable. The results showed that approximately 21% of all the missing records can be imputed by the TpI algorithm. Other algorithms may be considered when TpI is not applicable.

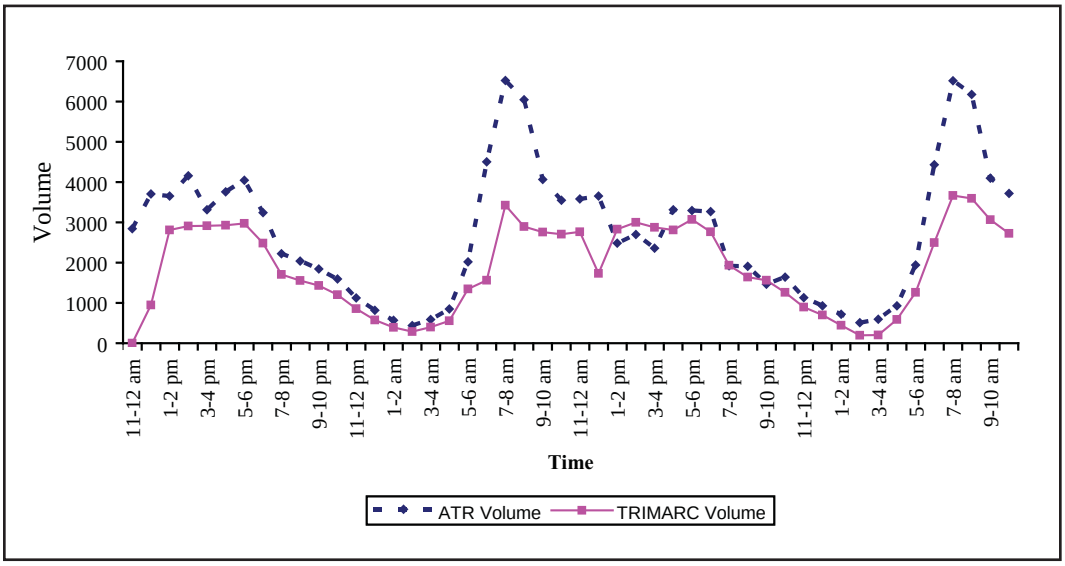
During the early morning of a weekday, Hybrid and SpI had similar performance in terms of MAPE and RMSE, both trailing TpI. A series of paired t-tests were conducted at a 0.05 significance level. For both I-64 and I-65 data sets, the tests did not find enough evidence to differentiate the performances of the two algorithms. For the I-71 dataset, SpI had a very slight advantage over Hybrid; its result was on average 3.8 vehicles more accurate than that of Hybrid during a 15-minute period. Based on their applicability and data requirement, Hybrid was chosen over the SpI since it

Figure 2: Comparison of Volumes Collected by Different Devices

(a) At I-64 eastbound mile-point 6.7



(b) At I-65 northbound mile-point 131.2



requires less data and is applicable to all missing data. Other algorithms will not be considered for this time period since, again, Hybrid is able to impute all missing data.

During the late night, SpI appeared to generate more accurate results than the other algorithms except TpI. ANN I was not considered further since it is more data-demanding than SpI. Further, paired t-tests on imputation differences between SpI and Hybrid confirmed that the former was statistically better (at a 5% significance level), as shown by the MAPE and RMSE indices. Therefore, SpI is the second choice and Hybrid is ranked third.

During the three daytime periods (e.g., morning peak, mid-day, and afternoon peak), ANN I appeared to be more accurate than Hybrid and SpI in most cases. In fact, ANN I seemed to perform better during the daytime compared with early morning and late night periods. For other cases, ANN I at least had similar performance compared with SpI. Therefore, ANN I was chosen to be implemented when TpI was not applicable. The hybrid algorithm should be the next in line since it had comparable performance and can be applied to all missing records. TpI was not further considered since it has similar data requirements as ANN I.

For weekends, the selection and priority of imputation algorithm was determined similarly based on the performance indices as well as the applicability/reliability of each algorithm. During both daytime and nighttime, TpI remained the most accurate algorithm. SpI had comparable performance to Hybrid, however, the performance statistics showed that SpI offered a slightly higher level of accuracy. Therefore, implementation should be in the order of TpI, SpI, and Hybrid. Figure 3 is a graphic representation of the implementation strategy, with "V_TpI" denoting the volume imputed by TpI, and so on. Depending on the procedure outlined in the chart, the imputed value by the proper algorithm can be selected to populate the final imputation column in a data table.

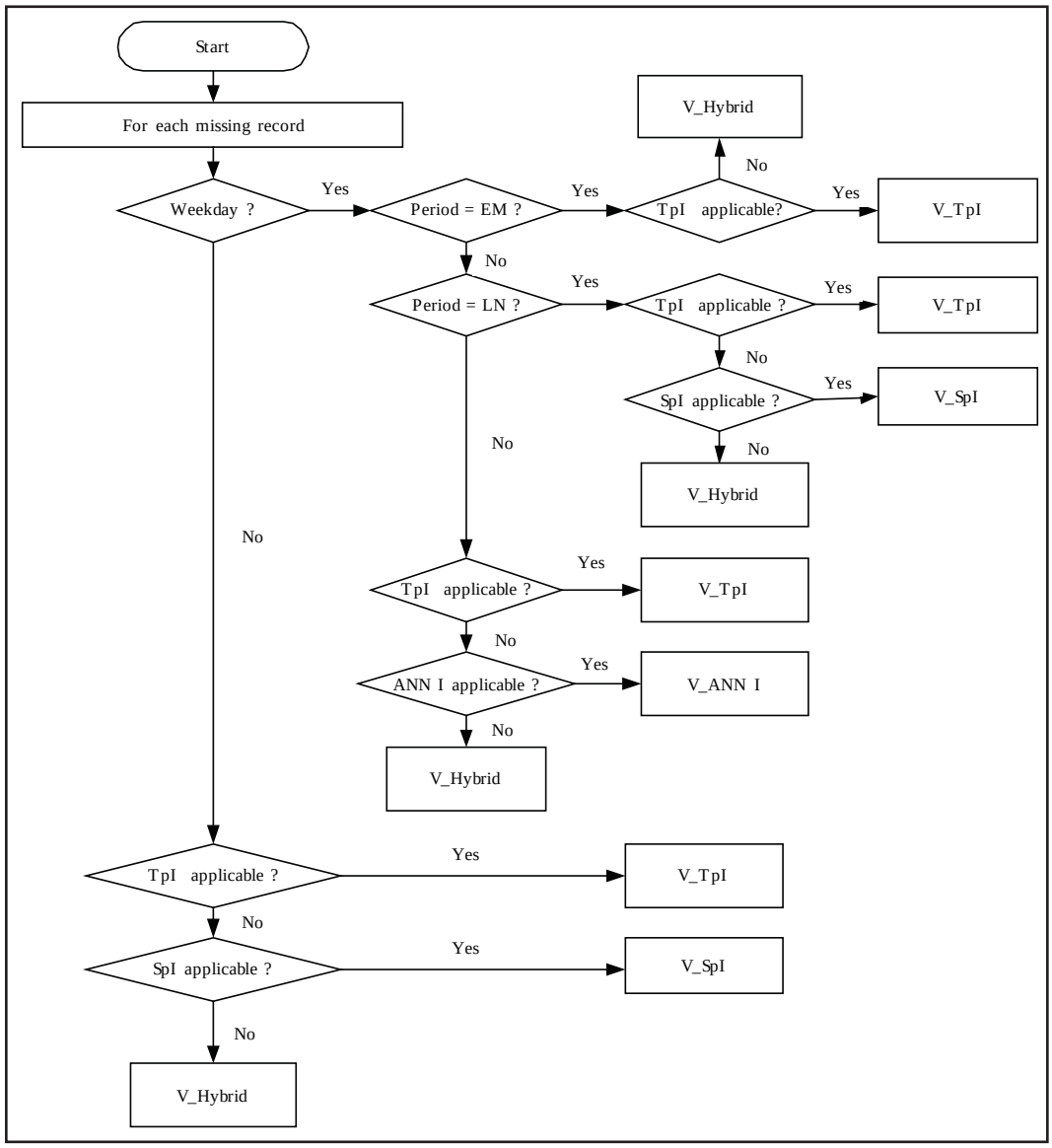
CONCLUSIONS

Archived ITS-generated data can provide a potential resource for many long-term transportation applications. Missing and suspicious data are inevitable due to various hardware, software, and communication reasons. Based on the data collected by the TRIMARC system in Kentucky, several imputation algorithms were developed and evaluated. The methodology presented in this paper is simple to understand and implement. By simulating the missing data patterns, the performances of all imputation algorithms by time-of-day, day-of-week, and different percentages of missing volumes were analyzed. The results showed that the rank of an imputation algorithm was affected by time-of-day and day-of-week factors.

The applicability of the algorithms and their reliability in terms of data requirement were also discussed. An implementation strategy for Kentucky ADMS was then developed based on performance, applicability, and reliability analyses. The imputation was designed to be carried out off-line. The implementation of the imputation strategy does not require substantial computational effort. For the 92 detector stations in TRIMARC, it took approximately 90 hours on a PC with 3GHz processor and 1GB memory to run the data quality screening criteria on traffic data aggregated at 15-minute intervals, to generate imputed results for all the algorithms, and to select the appropriate value for each missing/erroneous record according to the implementation strategy.

Each imputation algorithm discussed in this paper has its own strengths and weaknesses, and their applicability varies with missing data patterns. For data quality control, it is advisable to check the detector maintenance record to gain a better understanding of the sources of data discrepancies. For example, roadway maintenance may damage detectors or communication equipment, which causes complete or partial loss of data at a station. While the periods with a total loss of data are readily identifiable, it could be quite difficult to recognize the records associated with partial data loss when data is aggregated at the station level. Therefore, a more sophisticated data quality assurance procedure should incorporate the information from the device maintenance record.

Figure 3: Implementation Procedure



It should be noted that the hybrid algorithm developed in this study has very good potential to become a powerful imputation algorithm. This algorithm can be applied regardless of missing data patterns as long as a historical data profile is available. It is less demanding on input data than most of the algorithms, while it can produce relatively good estimates. Though the current form of the hybrid algorithm is rather straightforward, a more sophisticated formulation that combines the features of a time series and historical trend is certainly worth exploring.

References

Al-Deek, H. and C. Chandra. "New Algorithms for Filtering and Imputation of Real-Time and Archived Dual-Loop Detector Data in I–4 Data Warehouse." *Transportation Research Record: Journal of the Transportation Research Board* 1867, (2004): 116-126.

Chen, M. and X. Liu. "Using Neural Network to Analyze Passenger Activity and Its Impact on Bus Dwell Time and Travel Time." *Journal of Transportation Research Forum* 44 (3), (2005): 131-141.

Chen, M., X. Liu, J. Xia, and S. Chien. "A Dynamic Bus Arrival Time Prediction Model Based on APC Data." *Journal of Computer-Aided Civil and Infrastructure Engineering* 19 (5), (2004): 364-376.

Gold, D., S. Turner, B. Gajewski, and C. Spiegelman. "Imputing Missing Values in ITS Data Archives for Intervals under 5 Minutes." paper presented at the 80th Annual Meeting of Transportation Research Board, January 7-11, 2001, Washington, D.C.

Jiang, X. and H. Adeli. "Dynamic Wavelet Neural Network Model for Traffic Flow Forecasting." *Journal of Transportation Engineering* 131 (10), (2006): 771-779.

Lomax, T., S. Turner, and R. Margiotta. *Monitoring Urban Roadways in 2002: Using Archived Operations Data for Reliability and Mobility Measurement.* FHWA-HOP-04-011, Federal Highway Administration, Washington, D.C., 2002.

Margiotta, R. *ITS as A Data Resource: Preliminary Requirements for a User Service*. Federal Highway Administration, http://www fhwa.dot.gov/ohim/its/itspage htm, 1998. (last Accessed: November 2005).

Nelson, P. and P. Palacharla. "A Neural Network Model for Data Fusion in ADVANCE." *Proceedings of Pacific Rim Transportation Technology Conference*, American Society of Civil Engineers, Seattle, WA, Vol. 1, (1993): 237-243.

Park, D., R. Rilett, and G. Han. "Forecasting Multiple-Period Freeway Link Travel Times using Neural Networks with Expanded Input Nodes." *Proceedings of the 5th International Conference on Applications of Advanced Technology in Transportation Engineering*, ASCE, (1998): 325-332.

Pourmollem, N., T. Nakatsuji, and A. Kawamura. "A Neural-Kalman Filtering Method for Estimating Traffic States on Freeway." *Journal of Infrastructure Planning and Management* 36 (569), (1997): 105-114.

Smith, B., W. Scherer, and J. Conklin. "Exploring Imputation Techniques for Missing Data in Transportation Management Systems." *Transportation Research Record: Journal of Transportation Research Board* 1836, (2003): 132-142.

Zhong, M., S.Sharma, and P. Lingras. "Genetically Designed Models for Accurate Imputation of Missing Traffic Counts." *Transportation Research Record: Journal of the Transportation Research Board* 1879, (2004): 71-79.

Acknowledgment

This study is funded by the Kentucky Transportation Cabinet and the Federal Highway Administration. The views presented in the paper are those of the authors alone.

Mei Chen, Ph.D.*, is an assistant professor in the Department of Civil Engineering at University of Kentucky. Her research interests include transportation network analysis, travel demand forecasting, simulation and forecasting, and public transit. She received her Ph.D. in Transportation from New Jersey Institute of Technology in 1999.*

Jingxin Xia *is a Ph.D. candidate in the Department of Civil Engineering at University of Kentucky. He has worked on a number of research projects involving analyses of traffic data. He obtained a Master's degree on Transportation Engineering from Southeastern University in China.*

Rongfang (Rachel) Liu, Ph.D., *is an associate professor in the Department of Civil and Environmental Engineering at New Jersey Institute of Technology (NJIT). She is also affiliated with the transportation research centers in NJIT. Prior to joining NJIT, Liu worked in Parsons Brinkerhoff, Inc., as a project manager. She has been involved in a number of transportation planning, programming, and management projects. Her research interests are in the areas of travel behavior and demand forecasting modeling, intermodal transportation planning, operations research and network simulations, economic and environmental impact analysis. Rachel Liu is a professional engineer (PE) as well as a certified planner (AICP).*