Development of Improved Models for Imputing Missing Traffic Counts

Ming Zhong[*,1] and Satish Sharma[2]

[1]*Transportation Group, Department of Civil Engineering, University of New Brunswick, GD-128, Head Hall, 17 Dineen Drive, P.O. Box 4400, Fredericton, N.B., E3B 5A3, Canada*

[2]*Faculty of Engineering, University of Regina, Regina, SK, S4S 0A2, Canada*

Abstract: Estimating missing values is known as data imputation. A literature review indicates that many highway and transportation agencies in North America and Europe use various traditional methods to impute their traffic counts. These methods can be broadly categorized into factor and time series analysis approaches. However, little or no research has been conducted to assess the imputation accuracy. The literature indicates that the current practices are varied, and the methods used by highway agencies are intuitive in nature. Typical imputation methods used by highway agencies are identified and applied to data from six automatic traffic recorders (ATRs) in Alberta, Canada, to evaluate their accuracy. Statistical analysis shows that these traditional methods result in varying accuracy for traffic counts from different types of roads. In some cases, the imputation errors can be unacceptably high. Therefore, improved imputation methods are proposed. Study results indicate that imputation accuracy can be significantly improved by incorporating correction factors and data from both before and after the failure periods into the traditional models. The improved imputations should provide transportation practitioners better information for decision marking purposes.

Keywords: Traffic analysis, data systems, data analysis, statistical analysis.

INTRODUCTION

State and provincial highway agencies maintain ongoing traffic monitoring programs to collect traffic volume, vehicle classification, speed, and weight data. Of them, traffic volume data are most collected ones. The collected traffic volume data are used to generate various traffic parameters, such as annual average daily traffic (AADT) and design hourly volume (DHV). Traffic volume data and the generated parameters are widely used in transportation planning, design, operation, control, and research [1]. However, due to malfunctions of traffic counting devices, data for certain periods may not be properly recorded. Traffic datasets usually contain missing values and outliers. The presence of missing values and outliers in traffic counts makes data analyses and usage difficult.

Missing values are usually represented by zero hourly volumes or blanks in traffic counter files. Counting machine failures usually result in blanks in traffic data records. There are many reasons for the machine failures, including power surges, lighting, loss of battery power, solar panel failure, vandalism, and environmental fatigue, such as storms and frost heaves. Long-range zero hourly volumes indicate that the connections between sensors and counting machines were cut off, while the machines were still working normally. Signals of no traffic were passed to the machines and resulting records were zeros. Lighting is a major reason for this kind of disconnection [2].

*Address correspondence to this author at the Transportation Group, Department of Civil Engineering, University of New Brunswick, GD-128, Head Hall, 17 Dineen Drive, P.O. Box 4400, Fredericton, N.B., E3B 5A3, Canada; Tel: (506) 452-6324; Fax: (506) 453-3568; E-mail: ming@unb.ca

A previous study [3] indicates that traffic datasets usually have significant missing portions. It would be difficult to eliminate these data from the analysis. Due to time and financial constraints, recounting is not always possible. Consequently, highway agencies estimate missing values for their traffic counts. Estimating missing values is known as data imputation. Data imputation has been popular since traffic data program was established in the 1930s [4]. Proper imputations can help maintain data integrity and improve cost-effectiveness of traffic data programs. A literature review carried out for this study shows that many highway agencies in North America and Europe impute their data.

The literature review also indicates that imputation practices are varied, and nearly no research has been conducted to evaluate imputation accuracy. Albright [5] pointed out that it is necessary to evaluate the alternative imputation methods before auditing imputation practice. He also mentioned that, in some instances, the impact of imputation can be negligible, but in other instances the impact could be identified as unacceptable for most data applications. It is evident from the literature that the models used by highway agencies are intuitive in nature. Most of these models use historical data directly as replacements. These models are simple and easy to use and understand. However, it is possible that some of them may result in large estimation errors. Hence, a systematic study is needed to examine the accuracy of imputation practices. The improved imputation methods would also supply more accurate information for decision making purposes.

In this study, a literature review is used to examine data imputation practices in North America and Europe. Typical traditional methods used by highway agencies are identified and applied to traffic volume data from six automatic traffic

1874-4478/09

recorders (ATRs) in Alberta, Canada, to examine their accuracy. Improved models based on correction factor, data from both before and after the failure, and advanced time series analysis technique are also developed and tested.

LITERATURE REVIEW

The imputation principles in North America and the practices of highway agencies in Canada, United States, and Europe are examined in this section.

North American Background

There are increasing concerns about data imputation and Base Data Integrity. The principle of **Base Data Integrity** is an important theme addressed in both the American Society for Testing and Materials (ASTM) Standard Practice E1442, *Highway Traffic Monitoring Standards* [6] and the American Association of State Highway and Transportation Officials (AASHTO) *Guidelines for Traffic Data Programs* [7]. The principle is that traffic measurements must be retained without modification and adjustment. Missing values should not be imputed in the base data. However, this does not prohibit imputing data at analysis stage. In some cases, traffic counts with missing values could be the only data available for certain purpose and data imputation is necessary for further analysis. In such cases, imputation can be applied to missing traffic counts, and imputed data should be stored separately from base data. In accordance with the principle of Truth-in-Data, *AASHTO Guidelines* [7] also recommends highway agencies should document the procedures for editing traffic data.

Canadian Provincial Agencies

Traffic count data analysis and imputation practice of Saskatchewan and Manitoba Highways and Transportation was examined by Sharma *et al*. [2]. For automatic traffic recorder (ATR) data, the highway agencies usually use the last year data from the same site to impute missing hourly volumes. For example, Saskatchewan uses the data from the same day of the week in the same month of the last year as replacements. The maximum imputation period is set to be 21 days. If there are more than 21 days data missing, the counts would be discarded. For short-period traffic data, 48-hour counts with missing values are usually reduced to 24 successive hours. If not possible, the counts will be put into the next rotation for retaken.

Alberta Transportation has implemented an extensive ATR program across the province. Totally 361 pairs of ATRs are used to monitor the traffic over the provincial highway networks. Due to the wide coverage of ATR program, no short-term traffic counts are taken for individual road sections. However, each year more than 400 turning movement counts (9 or 12 hour) are taken at various intersections. These counts are factored into annual average daily traffic (AADT) annually. Alberta Transportation has adopted the principle of Base Data Integrity and stopped imputing missing hourly volumes. However, it was reported that Alberta Transportation's statistics contractor used historical data to estimate monthly average daily traffic (MADT) [2].

US State DOTs

In 1990, the New Mexico State Highway and Transportation Department conducted a survey of traffic monitoring practice in the United States [8]. It was shown that when portable devices failed, 13 states used some procedure to estimate the missing values and complete the data set. When permanent devices failed, 23 states employed some procedure to estimate the missing values [9]. Different methods were used for this purpose [8], including:

- In Alabama, if less than six hours of recorded data was missing, the values are estimated using data directly from the previous year or other data from the month. If more than six hours of recorded data was missing, the day is voided.
- In Delaware, estimates of missing values are determined based on a straight-line interpolation of the data from the months before and after the failure.
- In Indiana, the previous year's data is directly used for estimation and maximum imputation period is one week.
- In Montana, the estimates are calculated from the historical data collected at the same location. If there is no change in the area that would impact traffic patterns, the historical data is used directly. Otherwise, a factor is used to reflect the changes.
- In Oklahoma, missing values are estimated directly for periods up to nine hours based on the data collected on the same day of the week in the same month.
- In South Dakota, missing values are estimated directly from the previous three years' data.
- In Vermont, missing values are directly estimated by the data from the same day of the same month in the previous year.
- In summary, most of these methods simply copy or take the average of historical data as estimates. Growth factors are rarely used for imputation. Most US agencies impute their data manually and no much automation has been applied. Only in Kentucky is a computer program used to estimate and fill in the blanks, although the detailed procedures underlying the model are unknown [8].

European Authorities

In 1997, the Federal Highway Administration (FHWA) conducted a research for traffic monitoring programs and technologies in Europe [10]. It was reported that highway agencies in Netherlands, France, and the United Kingdom used some computer programs for data validation routines. For example, a software system INTENS was used in Netherlands for data analysis and validation. The software used a "smart" linear interpolation process between locations from which data were available to estimate missing traffic volumes. In France, a system MELODIE was used for data validation. Data validation was conducted visually by system operator. Invalid data were replaced with the previous month's data. Several data validation systems were used in the United Kingdom. One of them was used by the Central Transport Group (CTG) to validate permanent recorder data. Invalid data were replaced with data extracted from the valid data of last week collected from the same site. These historical data are multiplied by a factor taken from nearby sites

that did work correctly and used to convert the previous week's traffic volumes to the current week. No research has been found for assessing the accuracy of such imputations.

In England, a survey of practical solutions used by consultancies and local authorities was conducted in 1993 [11]. It was reported that there were two broad categories of solutions. One was "by-eye" method and the other was computerized package. "By-eye" method involved manual estimation of missing values. Most automated practical solutions to patching were based upon simple, moving or exponentially weighted moving average, or their variants. For example, Department of Transportation (DOT) in London employed an exponentially weighted moving average model to update missing values. The process involved validating new traffic count data against old data from the same site collected over the previous weeks at the same time. Following equation was used to estimate missing or rejected data at time t, $\hat{x}_{t,s}$:

$$\hat{x}_{t,s} = (1-\theta)x_{t-1,s} + (1-\theta)\theta x_{t-2,s} + (1-\theta)\theta^2 x_{t-3,s} + ... + (1-\theta)\theta^{n-1} x_{t-n,s} \quad (1)$$

where $x_{t-1,s}$, $x_{t-2,s}$, ..., $x_{t-n,s}$ represent the observations for that particular site and vehicle category s, at the same times for weeks 1, 2, …, n before the current observation; θ is a constant such that $0<\theta<1$. A value of 0.7 was typically used for the parameter θ.

Table **1** summarizes the imputation models used by the above agencies, in terms of model input, estimation function, output, and the maximum imputation period. Table **1** clearly shows that imputation practices are varied and that choosing an imputation method is quite an individual and independent matter. The literature review also indicates that little research has been conducted to evaluate imputation accuracy. It seems that highway agencies just intuitively select some methods to impute their data, and simply assume such methods would provide the imputed data with a satisfactory degree of accuracy.

Missing value imputation has been well researched in the sub-filed of statistics and applied to various fields [12]. For example, multiple imputation has been broadly used to estimate nonresponsive values in surveys [13, 14]. It has been involved as "the state of art" of imputation and applied to various other fields as well. Various advanced techniques have also been used for this matter. For example, Gupta and Lam [15] used neural networks to estimate missing values. Singh and Harmancioglu [16] used entropy theory to estimate hydrologic records. Zhong *et al*. [17] applied geneti-

Table 1. Summary of Imputation Models Used in the Practice

Agency	Model Inputs	Prediction Function and Output	Maximum Imputation Period
Alabama	Hourly volume from the previous year or other part of the month	Directly use input as output	6 hours
Alberta	Do not impute missing hourly volumes, but use historical data to estimate monthly average daily traffic [MADT]	ATRs with missing data will be assigned to ATRs without missing data, and a ratio is used to convert MADTs from ATRs without missing data to those for ATRs with missing data	If less than one week of good data are recorded, no MADT is produced
Central Transport Group, England	Valid data of last week collected from the same site	Historical data are multiplied by a factor taken from nearby sites that did work correctly	N/A
Delaware	Hourly volumes from the same hour of the month before and the month after	Taking average of the two hourly volumes	N/A
France	Hourly volume from the same hour in the previous month	Directly using hourly volume from the previous month	N/A
Indiana	The previous year's data	Directly use input as output	One week
London, England	Historical hourly volumes from the same hours of the same days in the previous weeks	Historical values are averaged with the weights calculated based on their time lags with missing values	N/A
Montana	Historical data from the same location	Expanded by factors if there are changes in the area	N/A
Netherlands	Data from other location	A "smart" linear interpolation	N/A
Oklahoma	Data from the same day of the weeks of the month	Directly use input as output	9 hours
Saskatchewan [Manitoba]	Hourly volume from the same hour on the same day of the week in the last year	Directly use input as output	21 days for Saskatchewan
South Dakota	Hourly volumes from the same hours on the same day of the week in the previous years	Taking average of hourly volumes from previous years as output	N/A
Vermont	Data from the same day of the same month in the previous year	Directly use input as output	N/A

cally designed neural network and locally weighted regression to estimate missing traffic counts. However, personal communications with a few highway agencies indicated that these techniques are unlikely accepted by practitioners because of their complexity. Moreover, the huge amount of data from traffic monitoring programs also "prohibits" implementing complicated analysis procedures to generate each individual imputed value and validate the whole dataset, unless these advanced techniques can be integrated into computer package with a high degree of automation. Therefore, this paper focuses on accessing accuracy of imputation models used in the practice and proposing improved models that is still easy for understanding and implementation.

STUDY DATA

Hierarchical grouping method proposed by Sharma and Werner [18] was used to classify Alberta ATRs into groups. Five groups were obtained to represent study data. Based on the traffic variation information discovered from the ATR data, these groups are labeled as commuter, regional commuter, rural long-distance, summer recreational, and winter recreational groups, as shown in Fig. (**1**). Six ATRs were selected from four of these groups: two from the commuter group, two from the regional commuter group, one from the rural long-distance group, and one from the summer recreational group. Because there are not enough data in the winter recreation group, no ATRs were selected from that group. Table **2** shows ATR sites selected from different trip-pattern groups and functional classes, their AADT values, monitoring counters, and training and test data used in this study. Fig. (**2**) shows the geographic locations of these study sites on a map of the province Alberta. ATR CM1 is on a minor collector commuter road near Red Deer, and ATR CM2 (principal arterial) is on a commuter freeway close to Calgary. ATRs RC1 (minor collector) and RC2 (principal arterial) are located in rural areas and are away from any regional centers. ATR RLD (minor arterial) is on a rural long-distance road section of TransCanada 1 between City of Canmore and Calgary. ATR SR is on a recreational road within Banff National Park.

Table 2. Study ATR Sites from Different Groups and Experimental Data

Trip Pattern Group	ATR	Functional Class	Monitoring Counter	AADT	Training Set	Testing Set
Commuter	CM1	Minor Collector	C011145	4042	1996 – 1999	2000
	CM2	Principal Arterial	C002181	41575	1996 – 1999	2000
Regional Commuter	RC1	Major Collector	C022161	3905	1996 – 1999	2000
	RC2	Minor Collector	C003061	3580	1996 – 1999	2000
Rural Long-distance	RLD	Minor Arterial	C001025	13627	1996 – 1999	2000
Summer Recreation	SR	Major Collector	C093001	2002	1996 – 1998	2000

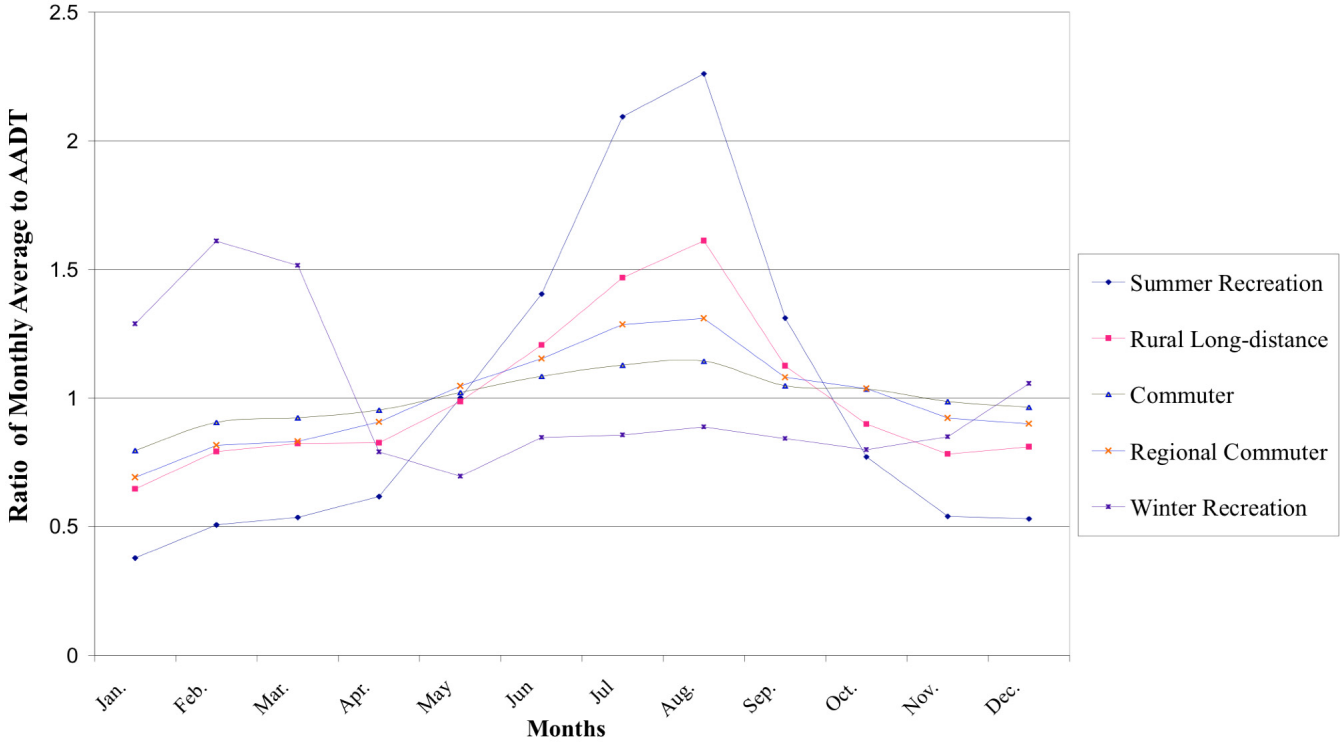

Fig. (1). Grouping results of Alberta ATR sites.

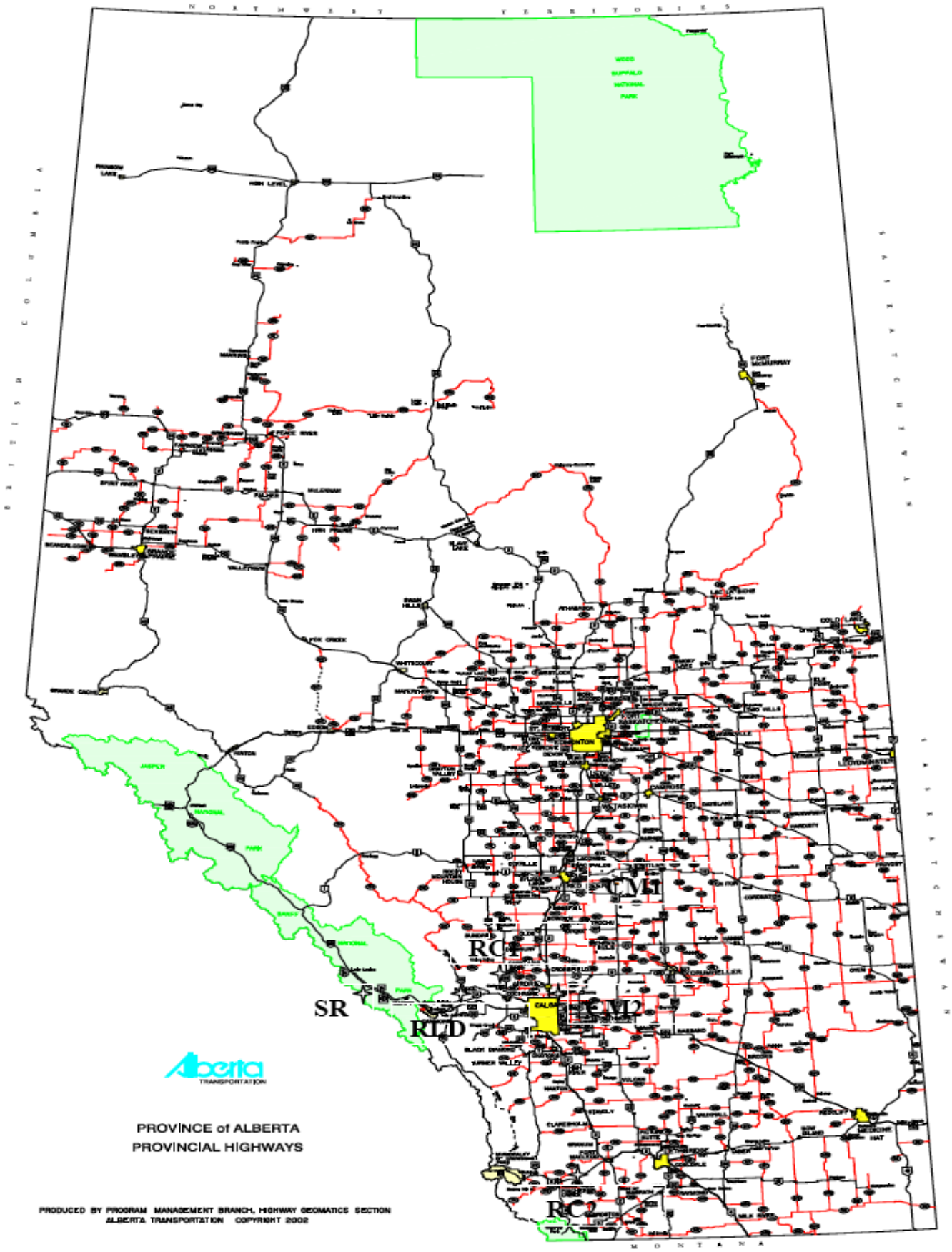

Fig. (2). Geographic locations of the study ATR sites.

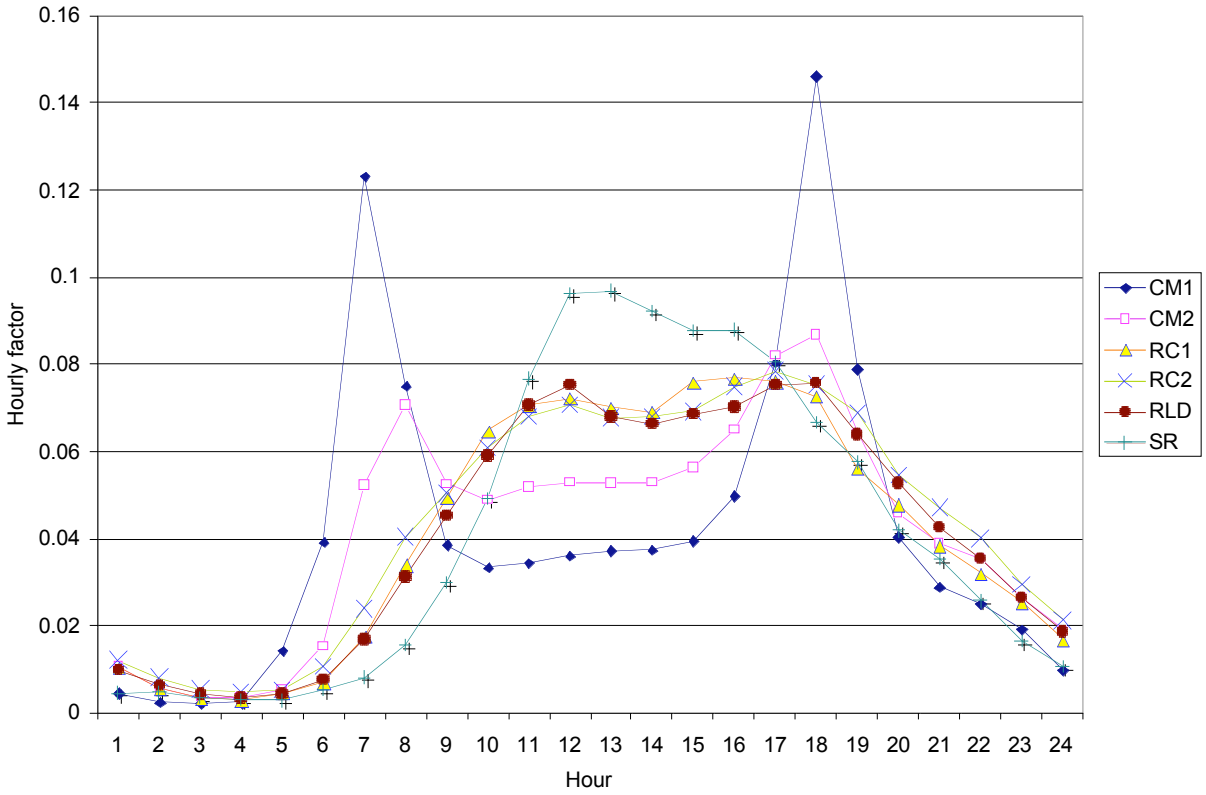

Fig. (3). Hourly patterns of study ATR sites.

Fig. (**3**) shows hourly patterns of these ATRs. For commuter ATRs - CM1 and CM2, there are two peaks in a day: one is in the morning and the other is in the afternoon. The regional commuter ATRs - RC1 and RC2 also have two peaks in a day, but not as remarkable as those of the commuter roads. The pattern of rural long-distance ATR - RLD has two small peaks. However, the first peak occurred nearly at noon, instead of in the early morning. The summer recreational ATR - SR only has one peak occurring nearly at noon. Most recreational travels on the road take place in a few hours in the afternoon and early evening.

Depending on data availability, four or five years data were used in the experiments for each ATR, as shown in Table **2**. These are no missing values in the study data. The data are in the form of hourly traffic volumes for individual travel directions.

STUDY MODELS

The literature review shows that the imputation methods used in the practices can be broadly categorized into factor and time series analysis approach. Factor models represent the mainstream imputation practice. Most of highway agencies in North America and European mainland use factor models. Time series analysis models, especially exponential moving average method, are used by some agencies in United Kingdom. In this study, typical imputation models used by highway agencies are identified and applied to the study data to examine their accuracy. Improved models based on correction factors, data from before and after the failure period, and autoregressive integrated moving average (ARIMA) techniques are also developed.

Traditional Imputation Methods

For traditional factor approach, the methods used by Saskatchewan Highways and Transportation in Canada, South Dakota Department of Transportation (DOT), and Delaware DOT in the United States, and the Highway Administration in the France are presented here. They represent the models using data from the last year, the models using data from the previous years, the models using current-year data from both before and after the failure, and the models only using current-year historical data as replacement values.

In Saskatchewan, missing data are directly imputed with the last year data. The imputing data are usually from the same day of the same week. South Dakota DOT estimated missing values with the average of data from the same periods in the previous three years. For Delaware, estimates of missing data are based on the average of the data from the same periods in the months on either side. French Highway Administration simply uses the previous month data as estimate of missing value.

For traditional time series analysis approach, the method used by DOT in London is selected as example here. It uses an exponentially weighted moving average model to estimate missing values. The estimated values are calculated with Equation (1).

Improved Imputation Methods

The literature review indicates that most imputation models from highway agencies only use historical data. The information available from after the failure period is usually neglected. However, for missing value estimation, usually

data from both before and after the failure periods are available. Incorporating more data into imputation models may provide more accurate estimates.

The literature review also indicates that traditional factor models rarely use correction factors to impute missing data. It is assumed that there are no seasonal differences between the period used for imputation and missing portion. The data from certain periods are usually "copied" to missing parts as estimates. However, traffic time series usually contain trend and seasonality. In some cases, correction factors may have to be used to improve accuracy.

In this study, improved models based on data from both before and after the failure and correction factor are developed and tested. Advanced time series analysis model based on Box-Jenkins technique are also developed in this study. The models are compared with traditional models used by the highway agencies. The improved models are as follows:

1. Monthly factor models: Monthly factors developed from training set are used to convert data from both before and after the failure month into replacement values for the failure month in test set. Such a method is expected to have some improvements over the traditional factor method that directly uses the average of the data from both before and after the failure as the replacement value (e.g., Delaware method), because the applied correction factors can roughly capture seasonal variations in traffic time series. The monthly factor model is as follow:

$$updateValue = \frac{mf_i / mf_{i-1} \times Value_{i-1} + mf_i / mf_{i+1} \times Value_{i+1}}{2} \quad (2)$$

where: mf_i is the average monthly factor of the failure month calculated from training set;

mf_{i-1} and mf_{i+1} are the average monthly factors of the month before and after the failure month, respectively, which are also calculated from training set;

$Value_{i-1}$ and $Value_{i+1}$ are the hourly volumes of the same hour from the same day of the week in the months before and after the failure month respectively in test set.

2. Both-side London method: London exponential moving average model is extended to use data from before and after the failure. The model is in the form as the following:

$$\hat{x}_{t,s} = \frac{1}{2}[(1-\theta)x_{t-1,s} + (1-\theta)\theta x_{t-2,s} + ... + (1-\theta)\theta^{n-1}x_{t-n,s} + (1-\theta)x_{t+1,s} + (1-\theta)\theta x_{t+2,s} + ... + (1-\theta)\theta^{n-1}x_{t+n,s}] \quad (3)$$

where $\hat{x}_{t,s}$ is the estimated replacement value, $x_{t-1,s}$, $x_{t-2,s}$, …, $x_{t-n,s}$ represent the observations for that particular site and vehicle category s, at the same times for weeks 1, 2, …, n before the current observation; $x_{t+1,s}$, $x_{t+2,s}$, …, $x_{t+n,s}$ represent the observations for that particular site and vehicle category s, at the same times for weeks 1, 2, …, n after the current observation; θ is a constant such that $0 < \theta < 1$.

3. Box-Jenkins Forecasting Procedure (ARIMA models): This procedure is based on fitting an autoregressive integrated moving average (ARIMA) model to a given set of data and then taking conditional expectations. A typical multiplicative seasonal ARIMA model is in the form:

$$\phi_p(B)\ \Phi_P(B^S)W_t = \theta_q(B)\Theta_Q(B^S)\alpha_t \quad (4)$$

where B denotes the backward shift operator, $\phi_p, \Phi_P, \theta_q, \Theta_Q$ are polynomials of B with the order p, P, q, Q respectively, and $\{a_t\}$ is the Box-Jenkins notation for a purely random process with mean zero and variance σ_a^2. $W_t = \nabla^d \nabla_S^D X_t$ and $B^S W_t = W_{t-S}$. $\{W_t\}$ is the differenced time series. $\{X_t\}$ is original non-stationary time series. ∇ is differencing operator and d, D is the order of differencing to remove both trend and seasonality. An ARIMA model considering seasonality in the data is often represented by ARIMA(p, d, q)(P, D, Q)$_s$, where p, d, q are the order of autoregressive, differencing, and moving average components; P, D, and Q are the order of seasonal autoregressive, differencing, and moving average components; S is the seasonal periodic component which repeats every S observations [19].

Both the traditional methods and the improved models mentioned above are summarized in Table **3** for clarification purpose. These models are applied to the study data to assess their accuracy. The estimates from these models are compared with true values, and the performance of the models is evaluated with absolute percentage errors (APEs). Depending on the model, the number of patterns or observations varied. The APEs are calculated as:

$$APE = \frac{|actual\ volume - estimated\ volume|}{actual\ volume} \times 100 \quad (5)$$

The key evaluation parameters consist of the average and 95th percentile errors. These statistics give a reasonable idea of the error distributions by including (e.g., when calculating average errors) or excluding (e.g., when calculating the 95th percentile errors) large errors caused by outliers or special events.

RESULTS AND DISCUSSION

By assuming data are missing at random in traffic counts, the traditional models used by Saskatchewan Highways and Transportation, South Dakota DOT, French Highway Administration, Delaware DOT, and DOT in London, and the improved models, including monthly factor model, both-side London method, and ARIMA model, were used to impute missing hourly volumes from various days and seasons of the year for all study sites. For the purpose of presentation and discussion, the imputation results for 12 daytime hours (from 8:00 a.m. to 8:00 p.m.) on Wednesdays in July and August are used in this paper. In general, study results can be distinguished for three groups. One is for those sites with lower traffic variations (e.g., CM2), one for those sites with moderate traffic variations (RC1, RC2, and RLD), and another one for those sites with large traffic variations (e.g.,

Table 3. Summary of Study Models

Model Category	Model Name	Model Inputs	Prediction Function and Output
Traditional models	Saskatchewan	Hourly volume from the same hour on the same day of the week in the last year	Directly use the input as output
	South Dakota	Hourly volumes from the same hours on the same day of the week in the previous 3 years	Taking average of the hourly volumes from the previous 3 years as output
	Delaware	Hourly volumes from the same hour of a month ago and a month after	Taking average of the two hourly volumes
	France	Hourly volume from the same hour in the previous month	Directly using the hourly volume from the previous month as output
	London	Historical hourly volumes from the same hours of the same days in the previous 12 weeks	The historical values are averaged with the weights based on their time lags with missing values (Equation (1))
Improved models	monthly factor	Hourly volumes from the same hours of the same days of the week in the months before and after the failure month	The hourly volumes are multiplied with the ratio of monthly factor for the failure month and those for the months before and after the failure and then averaged (Equation (2))
	Both-side London	Hourly volumes from the same hours of the same days in the 12 weeks from both before and after the failure	The input hourly volumes are averaged with the weights based on their time lags with missing values (Equation (3))
	Autoregressive integrated moving average [ARIMA]	12 hourly volumes [from 8:00 a.m. to 8:00 p.m.] from previous 8 days [same days of the week, totally 12 × 8 = 96 hourly volumes]	12 predicted hourly volumes [from 8:00 a.m. to 8:00 p.m.] based on ARIMA technique (Equation (4))

CM1 and SR). CM2, RLD and SR are used here for the presentation of the study results.

Figs. (**4-6**) show average imputation errors of traditional models (fine lines) and improved models (bold lines) for individual travel directions for CM2, RLD, and SR respectively. It can be seen that the errors for SR are higher than those for RLD, and the errors for RLD are higher than those for CM2. For example, most of average imputation errors for CM2 are less than 10%, and those for RLD are usually less than 15%. Most of average imputation errors for SR are more than 15% and some of them are over 30%.

From Figs. (**4-6**), it can be seen that Saskatchewan, South Dakota, French method, Monthly factor (MF) model and ARIMA model result in varying accuracy for different ATRs. For example, Saskatchewan method has the largest errors at most hours for SR, but it results in the smallest errors for the westbound traffic of RLD. South Dakota method results in the largest imputation errors for CM2 at both travel directions, whereas it results in the smallest errors for SR at most hours. French method usually leads to the largest (RLD) or the second largest errors (CM2 and SR) for these sites. MF models usually result in low to moderate errors. However, for some hours, the errors of MF models are higher than those of Delaware method. It seems that, in some cases, using seasonal correction factors to model hourly variations may not be good enough. ARIMA models result in the smallest errors for CM2 at most hours, but they result in large errors for RLD and SR in a large percent of cases. It is obvious that the pattern stability of ATR sites has considerable influences on the performance of ARIMA models.

Both Delaware and London methods provide consistently good imputations for all study sites, as shown in Figs. (**4-6**). The Delaware method uses a linear interpolation to estimate replacement values for the failure month based on data from either side. Such a method can roughly capture seasonal variations and provide better imputations, compared with the methods only use data before the failure (e.g., French method). The London method uses a large number of historical values as inputs, and weighs them in an exponential way to estimate replacement values. That is, as historical value is far away from estimated value, its influence on the estimation deceases. Such a procedure is useful for taking seasonal variations into estimation. Improved London method based on data from both before and after the failure show even higher accuracy than the Delaware and the traditional London method for all study sites. Using more data as inputs contributes to such achievements. Similar to the Delaware method, another advantage for such an approach is that seasonal variation can be interpolated based on data from both before and after the failure. Such interpolation should be more accurate than extrapolation based on data only from before the failure (e.g., traditional London method).

Fig. (**7**) compares the mean 95th percentile errors of both the traditional imputation methods and the improved methods for CM2, RLD, and SR. The errors shown in these figures are the mean 95th percentile errors for 12-hour (from 8:00 a.m. to 8:00 p.m.) imputations. The solid bars on the left represent the errors for the traditional models whereas the shaded ones on the right represent those for the improved models considered in this study. It is encouraging to find that MF models and both-side London models usually result in the lowest or nearly the lowest mean 95th percentile errors. For example, the mean 95th percentile errors for MF models are lower than the traditional factor models in most cases for these study sites. Both-side London method result in the lowest imputation errors for nearly all cases. It seems that these methods are robust to different traffic patterns, and can provide reasonable replacement values for different study sites, in spite of their differences in regional and traffic characteristics.

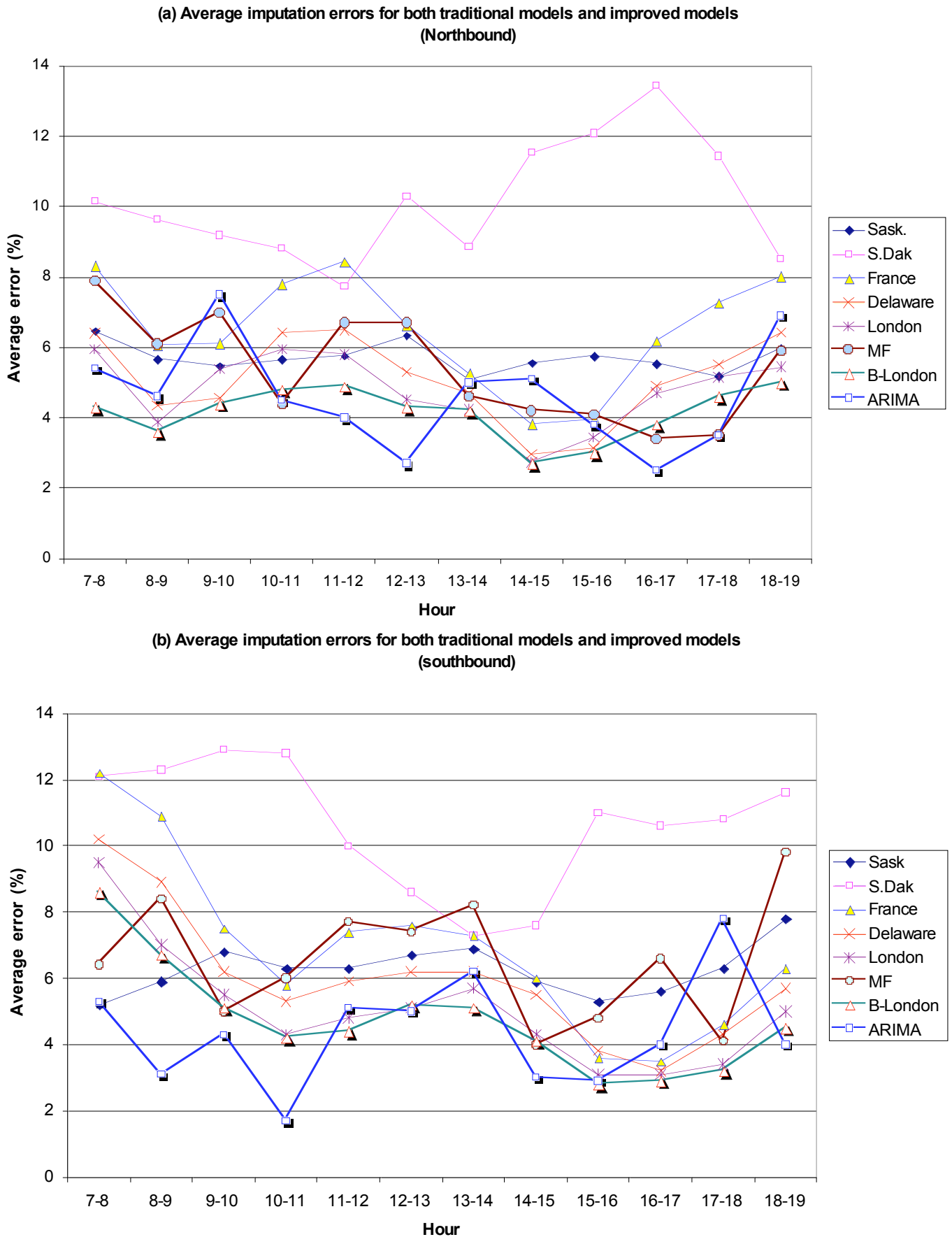

Fig. (4). Average imputation errors for the commuter count CM2.

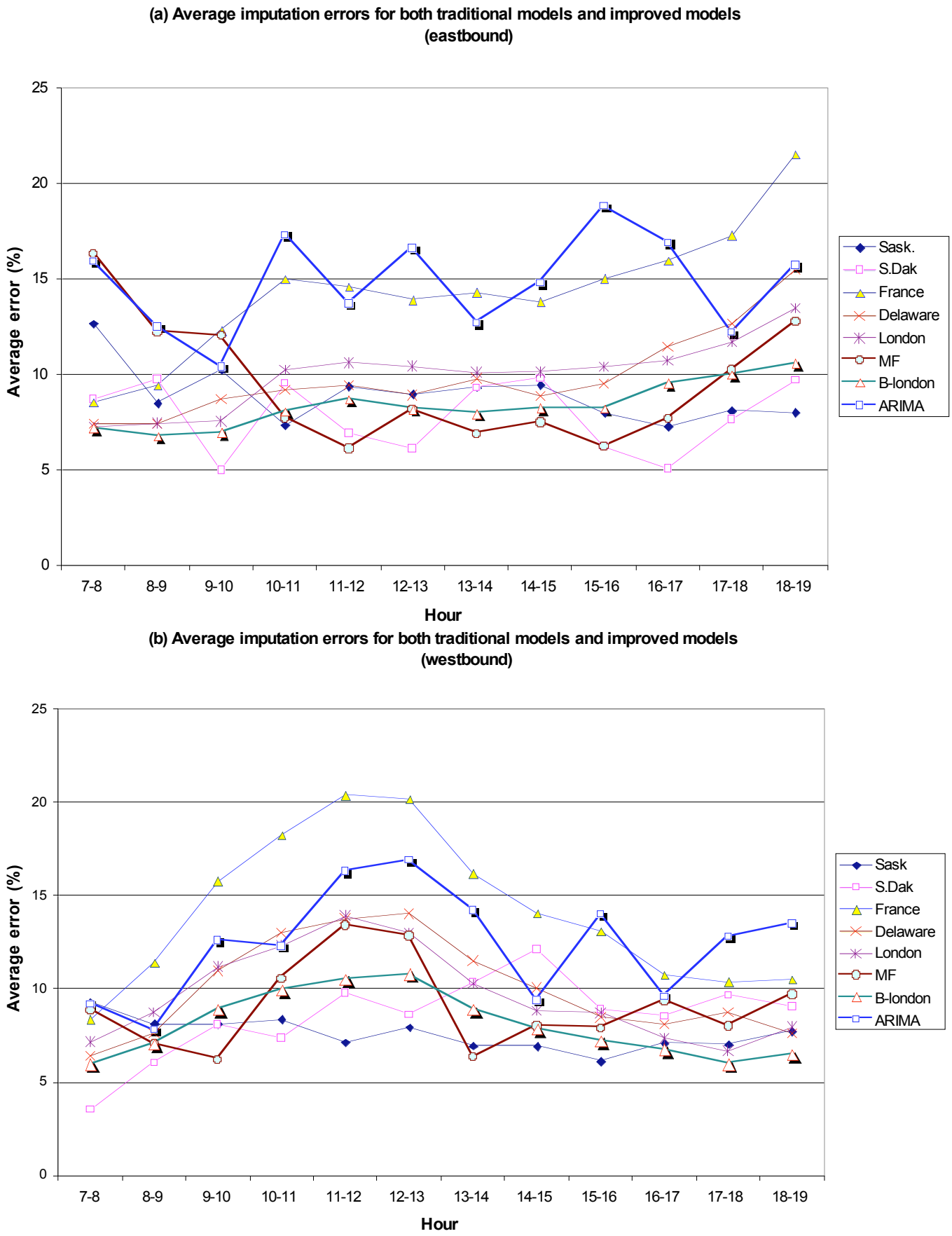

Fig. (5). Average imputation errors for the rural long-distance count RLD.

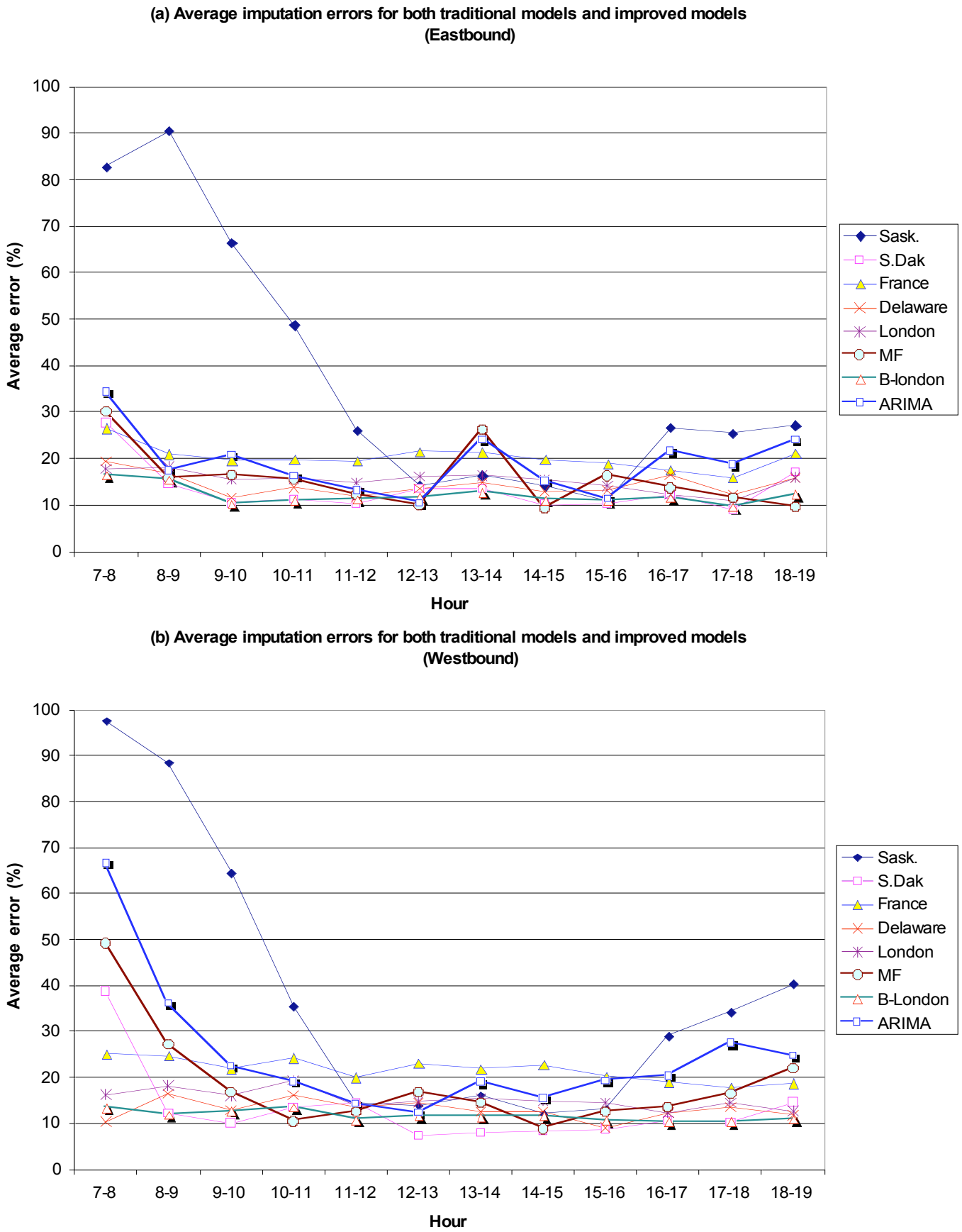

Fig. (6). Average imputation errors for the recreational count SR.

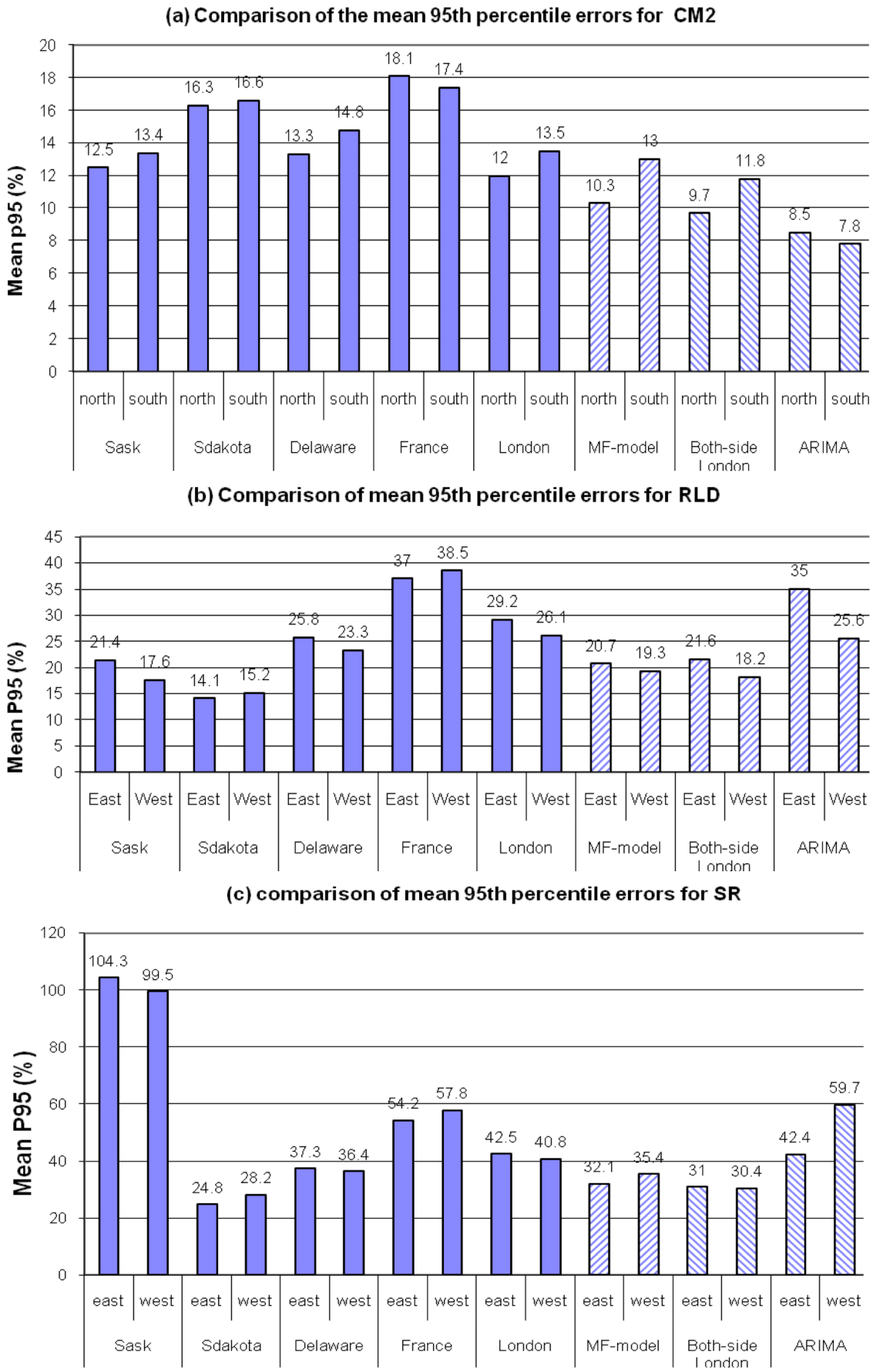

Fig. (7). Comparison of mean 95th percentile errors of all imputation methods for the counts CM2, RLD and SR.

CONCLUDING REMARKS

A previous study [3] shows that traffic data sets usually have significant missing portions. The literature review indicates that highway agencies and transportation authorities employ various methods to impute missing data. The imputation practices are varied and intuitive in nature. No study has been conducted to assess their accuracy.

This study shows that transportation practitioners usually use simple factor and time series analysis models to estimate missing values. In most cases, raw data are directly patched to missing parts as replacements. Correction factors are rarely used. For time series analysis models, the existing literature shows that exponential smoothing method is being used by some agencies in England. For most of these methods, only historical data are used as inputs, and the information available from the data after the failure period is usually neglected.

The imputation accuracy of various traditional models used by transportation practitioners and the improved models proposed in this study are assessed based on traffic volume data from six ATRs in Alberta, Canada. The statistical analyses show that some of these traditional methods could result in large errors. For instance, Saskatchewan method uses the last year data directly as estimates and results in large imputation errors for those sites with considerable yearly variations. When applied to the recreational count SR, such a method resulted in the average errors of over 80% for the morning peak hours. Similar problem exists for South Dakota method because the previous years' data are directly averaged, and no correction factors are used to reflect trend of time series. It resulted in the largest errors for the commuter count CM2 in both travel directions. The French method uses last month data as replacements and results in moderate to high errors. On the contrary, Delaware and London methods are robust to different traffic characteristics and provide good imputations for all study sites. The average errors for these methods range from 4-6% for the counts with stable patterns (e.g., CM2) to 10-15% for the counts with unstable patterns (e.g., SR). Delaware method uses the data from both before and after the failure for imputation. London method uses weighted sum of a large number of historical observations as replacement values. Obviously, using more data for estimation contributes to such robustness.

ARIMA models result in different accuracy for different types of ATRs (e.g., commuter or recreational). The reason may be that it is easy to discover a regular imputation pattern for those sites with stable traffic pattern (e.g., CM2). However, for those sites with unstable traffic pattern (e.g., SR), such a regular pattern may not exist due to a large seasonal, daily, and hourly variation.

The results of this study clearly show that there are possibilities to improve the traditional methods by using correction factors and data before and after the failure. The superior performance of the monthly factor model and both-side London model for ATRs from various types of roads emphasizes this conclusion. For example, the mean 95th percentile errors for these methods are lower than those of traditional methods by 2-6% for the commuter count CM2, as shown in Fig. (**7a**). For the recreational count SR, the improvements range from 5-6% to over 70% (Fig. **7c**). Therefore, it is recommended to highway agencies that the monthly factor and the both-side London method be considered and implemented to improve their imputation accuracy.

The improved models proposed in this study can also be used to estimate replacement values for outliers in traffic data sets after they are identified with some methods. The improved imputations obtained from these models should provide more reliable data for traffic engineers and officials.

This study shows that the models using data from both before and after the failure or more sophisticated techniques (e.g., exponential smoothing) provide more accurate imputations for traffic volume counts. It is believed that the rules discovered here are also applicable to imputations on vehicle classification, speed, and weight data. Future research is going to test the imputation methods on these data.

ACKNOWLEDGMENTS

The authors are grateful towards Natural Science and Engineering Research Council (NSERC), Canada for their financial support. The authors would like to thank Alberta Transportation for the data used in this study. The authors would also like to thank Peter Kilburn at Alberta Transportation, Tom Anderson and Wayne Gienow at Saskatchewan Highways and Transportation for their comments and support.

REFERENCES

[1] FHWA. *Traffic Monitoring Guide*. FHWA-PL-95-031. Federal Highway Administration, U.S. Department of Transportation, Washington, D.C. (3rd ed.), 1995.

[2] S.C. Sharma, M. Zhong, and Z.B. Liu, "Traffic Data Collection and Analysis Practice Survey for Highway agencies in West Canada", Technical Report, Faculty of Engineering, University of Regina, 2003.

[3] M. Zhong, S.C. Sharma, and Z.B. Liu, "Studying the Counting Efficiency of Continuous Traffic Monitoring Programs", In the 32nd Annual Congress of the Canadian Society of Civil Engineering, Saskatoon, 2004.

[4] D. Albright, "History of estimating and evaluating annual traffic volume statistics", *Transportation Research Records 1305*, Transportation Research Board, Washington, D.C., 1991, pp. 103 -107.

[5] D. Albright, "Standards, innovation, and the future of traffic monitoring", *ITE J.*, vol. 63, no. 1, pp. 31- 36, 1993.

[6] ASTM. *Standard Practice E1442, Highway Traffic Monitoring Standards*. American Society for Testing and Materials (ASTM), Philadelphia, PA, 1991.

[7] AASHTO. Guidelines for Traffic Data Programs. American Association of State Highway and Transportation Officials (AASHTO), Washington, D.C., 1992.

[8] NMSHTD. 1990 Survey of Traffic Monitoring Practices among State Transportation Agencies of the United States. New Mexico State Highway and Transportation Department, Santa Fe, New Mexico, 1990.

[9] D. Albright. "An imperative for, and current progress toward, national traffic monitoring standards". *ITE. J.*, vol. 61, no. 6, pp. 23-26, 1991.

[10] FHWA. FHWA Study Tour for European Traffic Monitoring Programs and Technologies. Federal Highway Administration, U.S. Department of Transportation, Washington, D.C., 1997.

[11] E.J. Redfern, S.M. Waston, M.R. Tight and S.D. Clark. "A Comparative Assessment of Current and New Techniques for Detecting Outliers and Estimating Missing Values in Transport Related Time Series Data", in Highways and Planning Summer Annual Meeting. Institute of Science and Technology, University of Manchester, England, 1993, pp. 163-174.

[12] R.A. Little and D.B. Rubin, *Statistical Analysis with Missing Data*, New York: Wiley, 1987.

[13] D. B. Rubin, *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley, 1987.

[14] P.M. Wright. "Filling in the Blanks: Multiple Imputation for Replacing Missing Values in Survey Data". In SAS 18th Annual Conference, New York, 1993.

[15] A. Gupta and M.S. Lam, "Estimating missing values using neural networks", *J. Oper.. Res. Soc.,* vol. 47, no. 2, pp. 229-239, 1996.

[16] V.P. Singh and N.B. Harmancioglu, "estimation of missing values with use of entropy", *NATO Adv. Res. Workshop*, Izmir, Turkey, pp. 267-274, 1996.

[17] M. Zhong, P.J. Lingras and S.C. Sharma, "Estimation of missing traffic counts using factor, genetic, neural, and regression techniques", *Transport. Res. Part C Emerg. Technol.,* no. 12, pp. 139-166, 2004.

[18] S.C. Sharma and A. Werner, "Improved Method of Grouping Provincewide Permanent Traffic Counters", *Transportation Res. Record 815*, Transportation Research Board, Washington D.C.1981, pp. 12-18,

[19] G. Box and J. Jenkins, *Time Series Analysis: Forecasting and Control,* Holden-Day, San Francisco, 1970.

Received: October 10, 2008 Revised: December 23, 2008 Accepted: December 25, 2008