Using Incomplete Archived ITS Data to Calculate Annual Average Traffic Statistics:

How much missing data is too much?

Shawn Turner (corresponding author)
Texas Transportation Institute
405C CE/TTI Bldg, 3135 TAMU
College Station, TX 77843-3135
V (979) 845-8829 ~ F (979) 845-6008
shawn-turner@tamu.edu

Eun Sug Park
Texas Transportation Institute
405G CE/TTI Bldg, 3135 TAMU
College Station, TX 77843-3135
V (979) 845-9942 ~ F (979) 845-6008
e-park@ttimail.tamu.edu

Abstract. Some transportation planning and traffic operations/ITS groups are cooperating on data sharing initiatives, including the use of archived ITS data for planning statistics. However, conventional procedures for the calculation of annual average traffic statistics need to be updated to accommodate the incomplete nature of archived ITS data. This paper tested the effects of various missing data patterns on several existing and modified annual average statistic calculation procedures. Five locations were chosen for the study: two urban interstate highways with significant commuter traffic, one urban parkway with commuter traffic and recreational trips, and two rural roads with pronounced seasonal patterns.

A complete year of data was obtained for each study location, then missing data patterns (as identified from an empirical study) were simulated by randomly or systematically removing data. At urban locations in which commuter traffic dampens seasonality patterns, a significant amount of missing data (up to 8 months of consecutive missing data) can be tolerated with little to no effect on annual average traffic statistics. All of the calculation procedures provided similar results at the urban locations. For the two rural locations, it appears that a month of missing data still results in tolerable error for most procedures, whereas two months of missing data is close to or exceeds tolerable error levels for most procedures. The authors concluded that modifications of conventional calculation procedures (that account for small gaps in data on a daily basis) are better suited to archived ITS data.

Paper prepared for presentation and publication
Transportation Research Board 87th Annual Meeting
January 13-17, 2008 ~ Washington, DC

Word Count: 7,344 equivalent words

4,844 words in manuscript body + 5 tables (1,250 words) + 5 figures (1,250 words)

INTRODUCTION

Traffic monitoring data that have been collected for traffic operations and management functions can be archived and used for a variety of analytical purposes in transportation planning, design, and operations. Many traffic management centers have installed permanent traffic monitoring sensors on freeway and arterial streets in urban areas. Since the incorporation of the Archived Data User Service in the National ITS Architecture in 1999, many traffic management centers also have begun to archive this traffic monitoring data for a variety of uses and users, with the most common applications to date being:

- Calculation of traffic count statistics and adjustment factors;
- Development of congestion or performance monitoring programs; and,
- Decision support for operations and maintenance (e.g., incident detection algorithms, ramp meter timing, work zone planning, etc.).

In most transportation agencies, the transportation planning division is responsible for maintaining traffic monitoring programs and publishing official traffic and roadway statistics. These planning groups typically install a limited number of permanent monitoring sites on a sample of roadways in their jurisdiction, and then perform short-duration data collection on a cyclical basis on the remaining roadways. The short-duration traffic data is then adjusted by day-of-week and seasonal adjustment factors to estimate an annual average traffic statistic, such as the annual average daily traffic (AADT) volume or the annual average weekday traffic (AAWDT) volume.

Since both traffic operations and planning staff have limited resources as well as a need for traffic data, these groups have begun to cooperate on data sharing initiatives, including the use of archived ITS data for planning statistics. In reviewing the archived ITS data, some data users have noted data quality issues such as incomplete, invalid, and inaccurate data. This paper addresses the issue of using incomplete archived ITS data in planning statistics. Archived ITS data can be incomplete for a variety of reasons, including communication interruptions, sensor malfunctions, equipment maintenance, road construction, and software or hardware failures.

The conventional procedures for the calculation of annual average planning statistics (as presented in AASHTO's 1992 *Guidelines for Traffic Data Programs*) require a certain level of data completeness. For example, the 1992 AASHTO *Guidelines* require that a full day of traffic data be available for each day of the week and month of the year in the calculation of annual averages. That is, 84 daily traffic counts (monthly average days of the week, MADW) must be 100 percent complete, with no imputation permitted for incomplete days.

This data completeness requirement is not often met by archived ITS data, because the nature of archived ITS data is that small periods of data are typically missing throughout nearly all days. For example, consider an archived ITS data set in which five percent of the data is missing for each day of the year. In sum total, the data set is 95 percent complete for the entire year. According to conventional procedures, however, none of the data would be acceptable because there are no days that are 100 percent complete and no imputation is permitted.

It is the authors' assertion that conventional procedures for the calculation of annual average traffic statistics need to be updated to accommodate the incomplete nature of archived ITS data. Clearly, data that is 95 percent complete (from ITS data) should be acceptable if data that is 23 percent complete (i.e., 84 days / 365 days of the year) from traditional planning-based data is also acceptable. The most pressing questions are:

1. How much missing data (from ITS) is too much? For example, what is the minimum completeness that is acceptable for the calculation of annual average traffic statistics?

2. Does the minimum acceptable completeness vary based on the pattern of missing data? For example, does randomly missing data have a lower acceptable value for minimum completeness than systematically missing data?

The research described in this paper addresses these questions in the following sections:

- **Background –** reviews the literature and practices that have been proposed, developed, or used.
- **Study Methodology –** describes the experimental design and data sets used in this study.
- **Summary of Findings –** presents the findings of the data analysis and provides an interpretation.
- **Conclusions–** presents the conclusions from the study.

BACKGROUND

This section reviews the literature and practices that have been proposed, developed, or used. The literature contains several relevant articles that address missing data in the context of calculating annual average statistics. Recent guidance from an update to the AASTO *Guidelines for Traffic Data Programs* is also summarized.

Literature Review

Perhaps the most relevant study identified in the literature review was from a report authored by Wright et al. (*1*). The primary objectives were to: 1) study and characterize the variability in traffic data (volume, classification, and weight) from continuously monitored road segments; and, 2) study the extent to which this variability is transferred to and affects the precision of data produced from road segments that are monitored only one or two days each year.

The report by Wright et al. developed several conclusions that are particularly relevant for this study:

- There are five different methods for estimating average traffic characteristics, ranging from simple to complex. Their study results showed that almost all five methods produce estimates of traffic characteristics that are within five percent of each other. As a result, Wright et al. recommend the straight forward averaging method because of its simplicity.

- The effects of randomly missing data on annual traffic estimates are negligible. However, the more missing data, the more unreliable the estimates. In many cases, Wright et al. found that the loss of reliability is tolerable. The authors indicated that most of the missing data more closely resembled random patterns than systematic patterns.

Sharma, Lingras, and Zhong have published several papers on the topic of missing data and annual average traffic statistics (*2*,*3*,*4*). These papers pursue a rigorous statistical approach to imputing missing values when calculating annual average traffic statistics. The authors also present the error that is produced when missing values are estimated in the calculation of annual averages. The authors report that the error for simple imputation techniques (such as historical averages or factors) often exceeds ten percent or more, which may be unacceptable in many cases. More sophisticated approaches, such as autoregressive integrated moving average (ARIMA) models and genetically designed regression and neural network models typically had error rates much lower, on the order of less than five percent and sometimes less than one percent.

More recent work from Zhong and Sharma (*5*) considered pattern matching algorithms to impute missing data. The algorithms use historical data to develop a series of candidate patterns and compare them to patterns with missing data. Data from the best-fit pattern is then used to estimate the missing data. The algorithms show improved accuracy over traditional models used in practice and appear suitable to impute missing hourly, daily, and monthly traffic volume data.

Several other papers have addressed the issue of data imputation in ITS data archives in a general sense, and not in relation to the calculation of annual average statistics. Conklin and Scherer (*6*) evaluated five imputation techniques: one approach based on historical averages, two heuristics approaches designed specifically for missing traffic data, and two statistical approaches developed for other application domains. The authors concluded that statistical algorithms are capable of generating the most accurate imputations for a single value of interest; however, for simpler cases where aggregate values are desired, a simple historical average approach may perform adequately.

Chen, Xia, and Liu (*7*) present a comparative analysis of various techniques for imputing missing volume data in the archived data management system in Kentucky, including simple historical averages, time series, and hybrid models. The authors concluded that the hybrid model (combines historical and time series approaches) has the best potential for implementation.

Finally, Al-Deek and Chandra (*8*) presented algorithms used to filter and impute real-time and archived ITS data in the I-4 (Orlando) data warehouse. Regression techniques were developed to fill the gaps formed by invalid or missing data values. Various pairwise regression models were developed and described, and their performance on the archived data from January and February 2003 was analyzed. The pairwise quadratic regression model with selective median was identified as the best model, and is currently being used to impute missing detector data in real time.

State of the Practice

Several groups have developed procedures that enable the use of archived ITS data in planning-based traffic databases. These procedures have taken similar approaches in dealing with incomplete data in the calculation of annual average traffic statistics.

The University of Minnesota at Duluth has developed imputation and calculation procedures for the Minnesota Department of Transportation (MnDOT) so that their extensive ITS data archive can be used for annual average traffic statistics (*9*). In previous work, MnDOT had been selecting 48 to 72 consecutive hours of traffic counts from incomplete traffic data and then factoring the short-term counts by day-of-week and month-of-year factors. The procedure that MnDOT has now implemented using spatial and temporal imputation techniques (i.e., estimating missing data using similar locations or time periods). With these imputation techniques, most locations now have more than 300 days of complete data for the calculation of annual average traffic statistics.

The Arizona Department of Transportation uses temporal imputation procedures for using their ITS data archive for planning statistics (*10*). These procedures will only estimate data for a maximum of 15 consecutive minutes. Once these lower-level time periods have been imputed, conventional calculation procedures are then used for annual average traffic statistics.

Other ITS data archives, such as PeMS in California and ADMS Virginia, have developed similar imputation procedures to fill in small gaps of missing data. This imputation is done for general purposes and not specifically for the calculation of annual traffic statistics. Once the permissible imputation procedures have been performed, conventional calculation procedures are used to compute annual traffic statistics. Caltrans has been investigating tolerable error levels that are present when significant imputation has been performed.

An update to the 1992 AASHTO *Guidelines for Traffic Data Programs* should be available in late 2007 or early 2008. In addition to numerous enhancements on a variety of data topics, the updated *Guidelines* included "provisional" guidance on the calculation of annual average traffic statistics from archived ITS data as follows:

> *Provisional procedures for summarizing ITS data have been developed, but there is some concern that they might be downwardly biased, and, in any event, they require further investigation.*
>
> *For the purpose of using these procedures, the standard time period for an ITS count is set at a single hour (rather than 24 consecutive hours). With this convention, all ITS data are summarized by hour, and modified versions of the procedures for producing monthly and hourly summaries are used. These procedures distinguish the 168 hours of the week in the same way as the conventional procedures described above distinguish the 7 days of the week.*
>
> *In the case of monthly summaries of ITS data for a particular location, the first step is to average all counts for each month and each of the 168 hours of the week to produce a monthly average hour of the week (MAHW) value. For each month and each day of the week, the 24 MAHW values (corresponding to the hours of the day) are summed to produce an MADW value. The MADWs are then combined to produce the various monthly and annual summary statistics for the site using the procedures described above.*

This guidance was considered "provisional" by the *Guidelines* authors because there was not sufficient use of this practice, nor was there significant analytical evidence to support firm recommendations. The research described in this paper is an attempt to develop analytical evidence to bolster these "provisional" guidelines and provide additional details.

STUDY METHODOLOGY

This section describes the experimental design and data sets used in this study. The experimental design for this study was relatively straightforward. The researchers conducted the following steps:

1. Identify several study locations and obtain or develop 100 percent complete data sets;
2. Calculate the reference annual average value based on the complete data;
3. Develop several modified procedures for calculating annual average values based on incomplete data;

4. Remove data from the complete data sets to simulate incomplete data in random and systematic patterns;
5. Use the modified procedures to calculate annual average estimates;
6. Quantify the error for each modified procedure by comparing the annual average estimate to the reference value;
7. Identify the modified procedure(s) with the lowest error, and identify the maximum amount of missing data that still meets minimum acceptable error levels.

These steps are described in additional detail in the sections that follow.

Study Locations and Reference Data Sets

The study locations were selected to include urban and rural roads as well as a variety of seasonal traffic patterns. The five study locations were all permanent counter locations and include:

1. I-10 Eastbound (ITS Station 40) in Phoenix, Arizona – monitors five lanes in the eastbound direction of I-10 near 24th Street;

2. I-94 Eastbound (ITS Station 1120) northwest of Minneapolis, Minnesota – monitors three lanes in the eastbound direction of I-94 near 101st Avenue;

3. George Washington Memorial Parkway in Arlington, Virginia – monitors the Parkway (southbound and northbound) in the vicinity of Theodore Roosevelt Island;

4. US 191/US 287 in Yellowstone National Park, Wyoming – monitors the West Entrance Road (eastbound and westbound) near the intersection with US 89 (Madison Junction); and,

5. State Highway 120 in Yosemite National Park, California – monitors the Big Oak Flat Entrance Station (eastbound only) near the main road entrance to Yosemite National Park.

Figures 1 through 5 illustrate the seasonal patterns at these five locations by showing the monthly average daily traffic (MADT), monthly average weekday traffic (MAWDT), and the monthly average weekend traffic (MAWET) volumes. The first two locations in Phoenix and Minneapolis were chosen to be typical of urban conditions. These locations have consistent commuter traffic as well as significant through traffic. The third location, George Washington Memorial Parkway near Washington, DC was chosen as an urban location that could have more seasonality due to recreational travel. The final two locations in Yosemite and Yellowstone National Parks were chosen to be typical of rural conditions in which recreational travel is a large component of the overall traffic. In these rural locations, the seasonality patterns are much more pronounced than in the three urban locations.

A year of traffic data were obtained from the appropriate government agencies for each of these five locations. In nearly all cases, the year of traffic data was relatively complete with little missing data. If missing data was present, standard statistical techniques were used to estimate missing values so that a 100 percent complete data set was available for the experiment. The reference AADT values were then based on these 100 percent complete data sets.

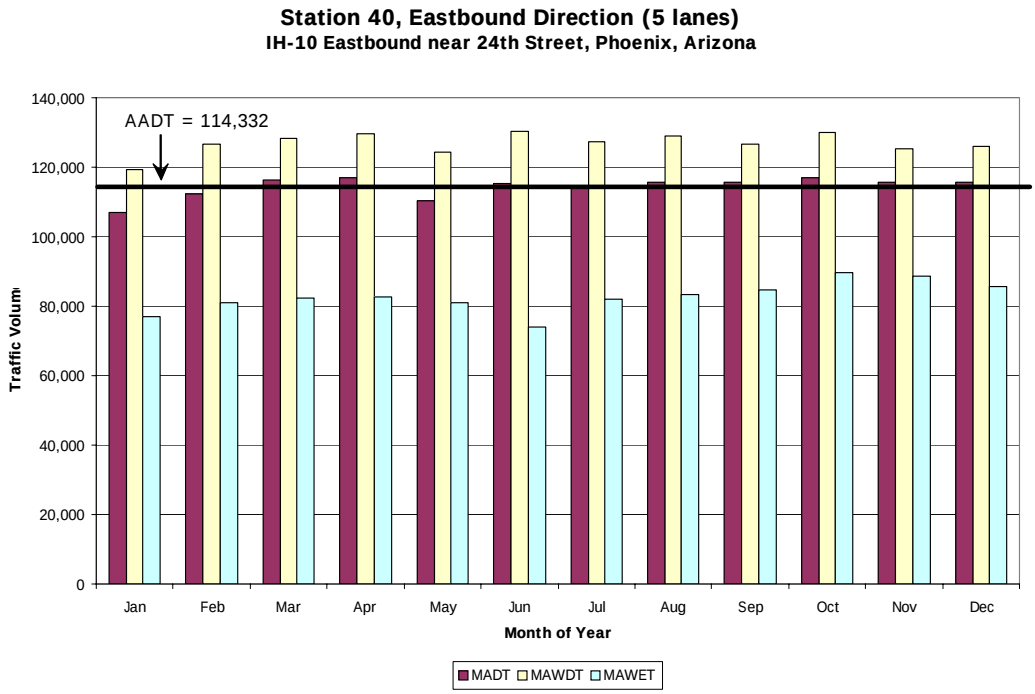

FIGURE 1 Monthly variation in traffic volume: I-10 Eastbound in Phoenix, Arizona

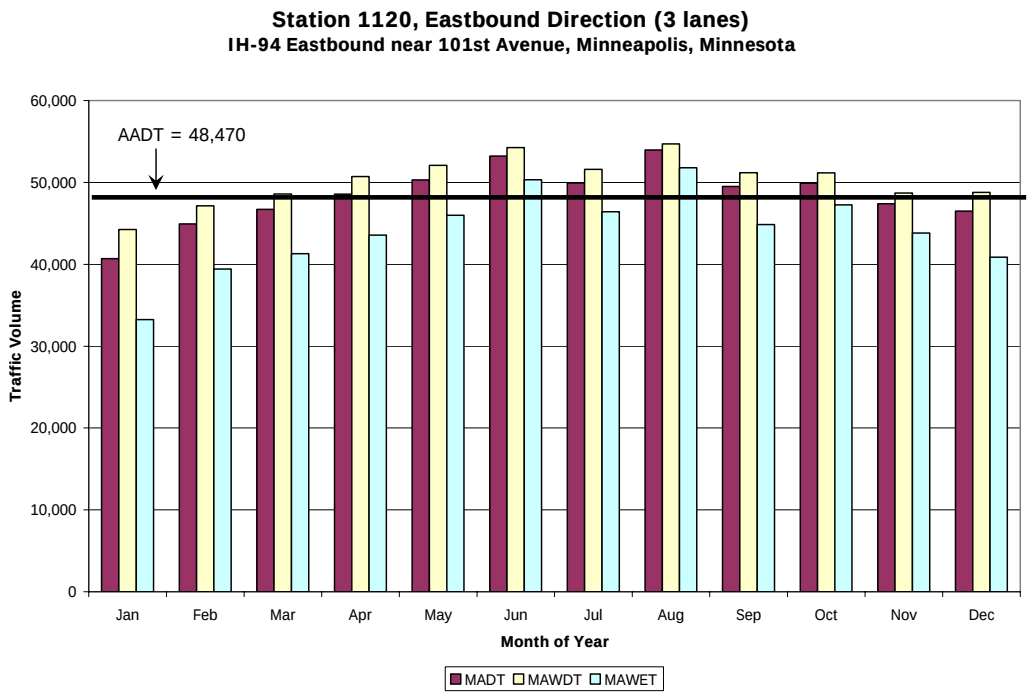

FIGURE 2 Monthly variation in traffic volume: I-94 Eastbound in Minneapolis, Minnesota

FIGURE 3 Monthly variation in traffic volume: George Washington Memorial Parkway in Arlington, Virginia

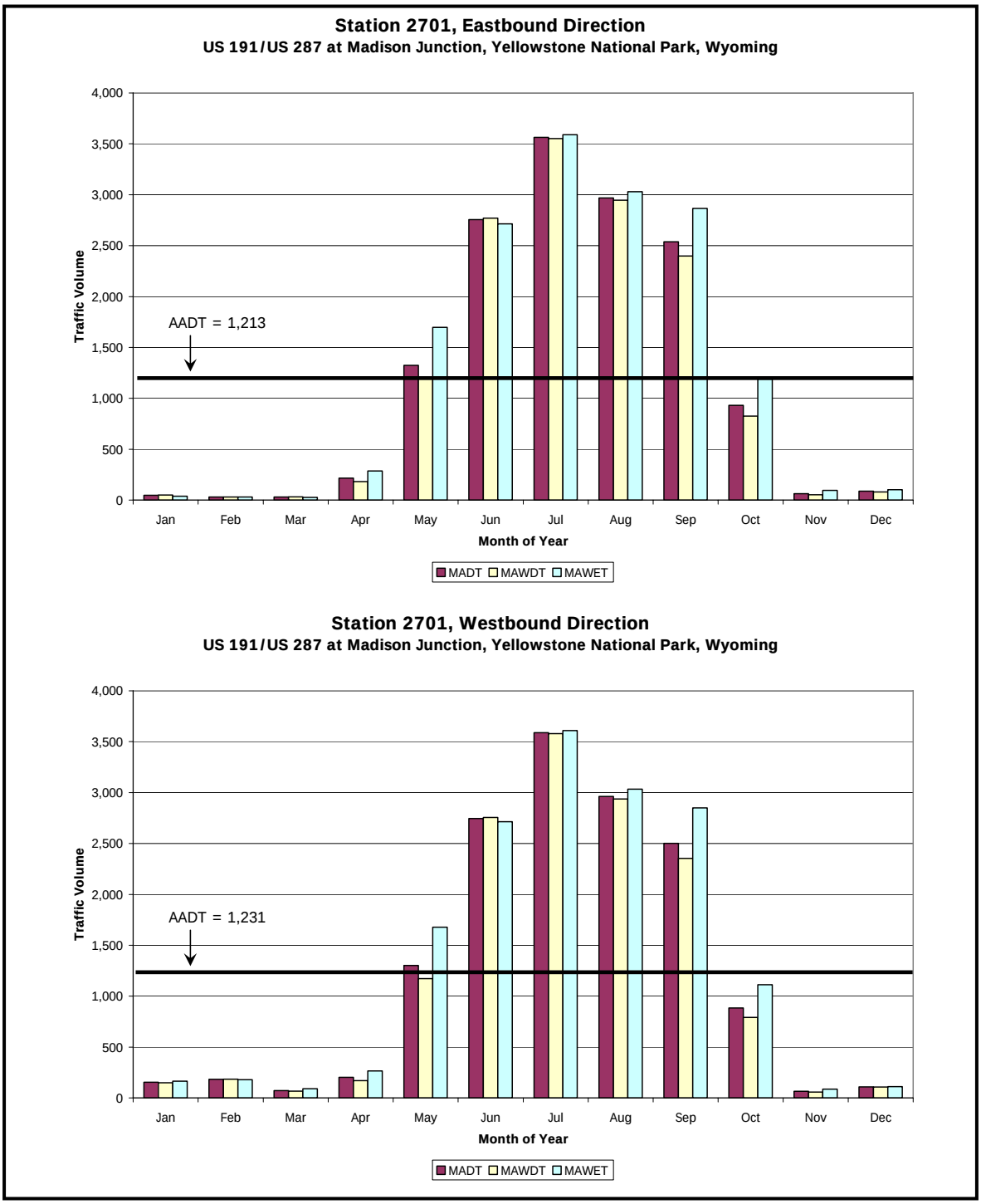

FIGURE 4 Monthly variation in traffic volume: US 191/US 287 in Yellowstone National Park, Wyoming

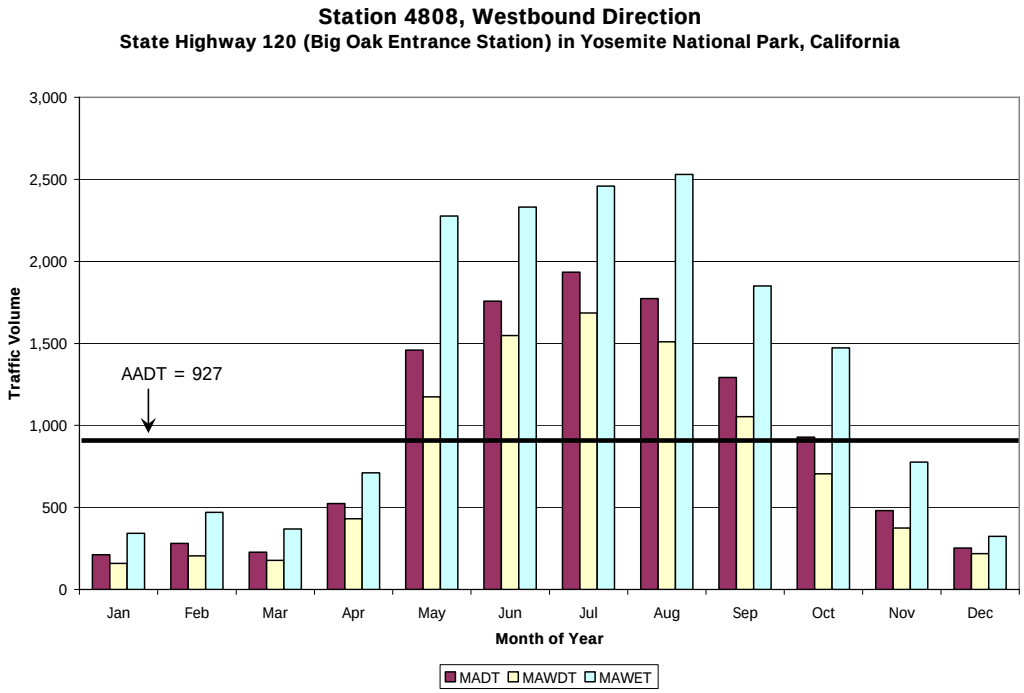

FIGURE 5 Monthly variation in traffic volume: State Highway 120 in Yosemite National Park, California

Modified Procedures for Calculating Annual Average Values

The conventional procedure for calculating annual average daily traffic (AADT) volume (as defined in the 1992 AASHTO *Guidelines* as well as the updated 2007/2008 AASHTO *Guidelines*) is as follows:

- All complete daily traffic volumes are averaged for each day of the week and month of the year, yielding seven values for each month, or 84 total Monthly Average Days of the Week (MADW) values. The 1992 *Guidelines* explicitly recommend that "Missing data should not be imputed." Therefore, each day used in the calculation of MADW values must be 100 percent complete.

- The MADW values are averaged across all twelve months, yielding seven Annual Average Days of the Week (AADW) values.

- To calculate AADT, all seven AADW values are averaged. If annual average weekday traffic (AAWDT) or annual average weekend traffic (AAWET) values are desired, the respective weekdays or weekend days are averaged together.

This conventional procedure effectively divides the calendar year into twelve months of seven daily means, and assumes that calendar days should be weighted equally. The procedure was designed to accommodate the nature of missing data from continuous permanent counter stations. That is, when the equipment was operational, the counters had no missing data. When various outages would occur throughout the year, the problems could be identified and fixed such that at least one good week of data was collected in each month.

The patterns of missing data from traffic operations-based sensors are different and require different annual average statistic calculation procedures. The researchers identified or developed several other modified calculation procedures to test in this study. The calculation procedures that were evaluated in this study include the following:

1. Arithmetic Mean – The AADT value is calculated as the arithmetic mean of all daily traffic volumes. This corresponds to a statistical definition of annual average.

2. ASTM Standard 1442 (1994) – The 84 MADW values are calculated by from all available complete daily traffic volumes. The Monthly Average Daily Traffic (MADT) is calculated for each month as the average of non-missing MADW values for that month. The AADT value is then calculated as the average of all MADT values. Missing data is not imputed; therefore, incomplete days are not used in the calculation of MADW values. Note that one or more of the seven weekdays could be missing when computing the MADT for each month although none of the MADT values are allowed to be missing when computing AADT. This standard was last approved in 1994, but its renewal as a formal standard was not approved in 1999.

3. Conventional AASHTO Procedures (1992 and 2007/2008) – As described above, the 84 MADW values are calculated by from all available complete daily traffic volumes. The Annual Average Days of the Week (AADW) are calculated for each weekday as the average across all months. The AADT value is then calculated as the average of all AADW values. Missing data is not imputed; therefore, incomplete days are not used in the calculation of MADW values.

4. Provisional AASHTO Procedures (2007/2008) – Hourly traffic volumes are used as the basic foundation instead of daily traffic volumes. That is, the 168 average hours of the week (24 hours × 7 days) is calculated from available data and is called the monthly average hours of the week (MAHW). For each month and day of the week, the 24 MAHW values are summed to produce a MADW value. The MADW values are average across all twelve months to produce seven AADW values, then the AADW values are averaged to calculate an AADT value.

5. Sum of the 24 Annual Average Hourly Traffic Volumes – For each hour of the day, hourly traffic volumes are averaged across all months and days (365 possible data values, one for each day of the year), yielding 24 annual average hourly values. The AADT value is calculated as the sum of the 24 hourly values.

6. Modified ASTM Standard (allow missing MADW values) – This procedure is a modified version of the ASTM Standard, with the exception that not all 12 MADT values are required for AADT calculation. For example, one of the 12 MADT values could be missing when computing the AADT.

7. Modified Conventional AASHTO Procedures (allow missing MADW values) – This procedure is a modified version of the conventional AASHTO procedures, with the exception that not all 84 MADW values are required for AADT calculation. For example, one of the twelve months could be missing when computing an AADW value. Similarly, one of the seven AADW values could be missing when computing the AADT.

8. Modified Provisional AASHTO Procedures (allow missing hourly MADW values) – This procedure is a modified version of the provisional AASHTO procedures, with the exception that not all 2,016 hourly MADW values (84 MADW values × 24 hours of the day) are required for AADT calculation. For example, one of the twelve months could be missing when computing an AADW value. Similarly, one of the AADW values could be missing when computing the AADT.

Simulation of Missing Data

Once the study locations were identified and the 100 percent complete data sets were generated, a reference AADT value was calculated using the arithmetic mean formula (procedure 1 as described above). Then, various patterns of missing data were simulated by removing data and recalculating AADT values using the seven other calculation procedures described in the previous section of this paper.

The missing data patterns used in this study were modeled after actual patterns of missing data from an archived ITS data set which was considered to be typical. The archived ITS data were from Phoenix freeways for the entire year of 2004 and were obtained from the Arizona Department of Transportation (ftp://ftp.az511.com/pub/traffic/5min/). The analysis of missing data patterns was performed on a location-by-location basis, with most locations falling into one or more of eight possible categories. These eight basic categories were then modeled to create eight missing data patterns for this study:

Randomly Missing Patterns

1. Less than 1 percent of data randomly missing – 0.03 to 0.5 percent of the data are missing, with at least one missing value in between 9 and 43 different days.

2. Twenty percent of hourly data randomly missing – On average, about four to five missing hours per day, with the number of days with at least one missing value is 363 to as high as 365.

Systematically Missing Patterns

3. The first time period of the day is missing from every day – this models server maintenance that disables data archiving functions during the early morning hours, from January 1 through July 1 or 2. In sum, 0.5 to 2 percent of the data are missing, but 50 percent of the annual days are affected.

4. Three consecutive summer weeks of missing data – 22 consecutive missing days (6 percent of all days), from July 15 through August 5.

5. Seven consecutive weeks of missing data in the beginning and end of the year – 50 consecutive missing days (14 percent of all days), from January 1 through January 27 and December 9 through December 31.

6. Eight consecutive weeks of missing data in the middle of the year – 57 consecutive missing days (16 percent of all days), from August 12 through October 7.

7. Fourteen consecutive weeks of missing data in the beginning of the year – 99 consecutive days (27 percent of all days), from January 15 through April 22.

8. About eight consecutive months of data missing in the beginning of the year – 253 consecutive days (69 percent of all days), January 1 through September 9.

Error Formulation

Once the missing data patterns were applied and the AADT values were calculated, the error between each of the eight alternative procedures were compared to the reference AADT value (calculated as an arithmetic mean). The error formulation used in this study was the absolute relative error (Equation 1). The findings of the study are presented in the next section and include error results for these various combinations of calculation procedures and missing data patterns at each of the study locations.

Equation 1

$$Absolute\ Relative\ Error\ (\%) = \frac{\left| AADT_{Alternative} - AADT_{Reference} \right|}{AADT_{Reference}} \times 100$$

SUMMARY OF FINDINGS

The results of the analysis are shown in Tables 1 through 5. These tables show the absolute relative error for AADT estimates for the various statistic calculation procedures and patterns of missing data. The findings are summarized as follows:

- For certain calculation procedures, AADT estimates could not be calculated for some missing data patterns. This was expected because certain missing data patterns result in the inability to calculate all 84 MADW values, especially when imputation is not permitted. For other calculation procedures, the AADT estimates could be calculated because these procedures remove or modify the restriction on having 84 complete MADW values.

- For the three urban locations (e.g., IH-10 in Phoenix; IH-94 in Minneapolis; George Washington Memorial Parkway in Arlington), all calculation procedures and missing data patterns had errors less than 5 percent. Even with more than eight months of missing data, both interstate roadway locations had errors of 2 percent.

- The highest error in the urban locations was 5 percent on George Washington Memorial Parkway. This maximum error occurred for the scenario with 20 percent of the data randomly missing. This was an unexpected result, as the authors had hypothesized that the maximum error would occur when the missing data pattern was systematic and more substantial than 20 percent missing data.

- For the urban locations, all of the calculation procedures produced similar error in estimating AADT. That is, all AADT calculation procedures produced similar results at the urban locations with less distinct seasonality.

- As expected, the two rural locations with more pronounced seasonality patterns had greater error, particularly in cases where large amounts of data were systematically missing. The error rates for 2 months (7-8 weeks) of consecutive missing data were typically less than 10 percent for most calculation procedures. The error rates for 14 weeks of data ranged from 16 to 35 percent, clearly greater than acceptable error. The error rates were proportionately higher for 36 weeks of missing data, ranging from 23 to 41 percent.

- At these two rural locations with greater seasonality, certain calculation procedures had lower error rates for different patterns of missing data. For example, with 20 percent randomly missing data, the modified ASTM and conventional AASHTO procedures both performed poorly (error between 76 and 90 percent), whereas the modified provisional AASHTO procedures performed much better (error one percent or less). When there are many incomplete days (e.g., the first time period of each day is missing for six months), the procedures developed based on the smaller time intervals (such as Sum of 24 annual average hourly volumes and the modified provisional AASHTO procedures) perform considerably better (error rates are less than 1 percent) than the other procedures based on the complete days' volumes only (error rates are 18 to 38 percent). Most calculation procedures performed similarly when entire days' volumes are missing although the modified ASTM, conventional AASHTO, and modified provisional AASHTO procedures still perform better than the other approaches.

TABLE 1 AADT Estimates for Various Calculation Procedures and Different Patterns of Missing Data: IH-10 EB, Phoenix, Arizona.

AADT Calculation Procedure	100% complete	Less than 0.5% randomly missing	20% randomly missing	First 15 minute of day (0.03%)	3 summer weeks (6%)	7 winter weeks (14%)	8 summer weeks (16%)	14 weeks (27%)	36 weeks (69%)
1. Arithmetic mean	114,326 (reference)	113,943 (<1%)		115,535 (1%)	114,075 (<1%)	115,046 (<1%)	113,978 (<1%)	114,323 (<1%)	116,326 (2%)
2. 1994 ASTM Standard	114,269 (<1%)	114,204 (<1%)			113,859 (<1%)	115,234 (<1%)			
3. Conventional AASHTO	114,269 (<1%)	114,204 (<1%)			113,859 (<1%)				
4. Provisional AASHTO	114,269 (<1%)	114,366 (<1%)			113,859 (<1%)				
5. Sum of 24 Annual Avg Hourly Vol.	114,326 (0%)	114,450 (<1%)	114,609 (<1%)	114,393 (<1%)	114,075 (<1%)	115,046 (<1%)	113,978 (<1%)	114,323 (<1%)	116,325 (2%)
6. Modified ASTM Standard	114,269 (<1%)	114,204 (<1%)		115,402 (<1%)	113,859 (<1%)	115,234 (<1%)	114,108 (<1%)	113,704 (<1%)	116,265 (2%)
7. Modified Conventional AASHTO	114,269 (<1%)	114,204 (<1%)		115,402 (<1%)	113,859 (<1%)	114,870 (<1%)	114,108 (<1%)	113,704 (<1%)	116,265 (2%)
8. Modified Provisional AASHTO	114,269 (<1%)	114,366 (<1%)	114,135 (<1%)	114,089 (<1%)	113,859 (<1%)	114,870 (<1%)	114,108 (<1%)	113,704 (<1%)	116,265 (2%)

Note: Empty cells indicate that AADT estimates cannot be calculated because imputation is not permitted (i.e., the 84 monthly average days of the week (MADW) volumes are incomplete and cannot be calculated).

TABLE 2 AADT Estimates for Various Calculation Procedures and Different Patterns of Missing Data: IH-94 Eastbound, Minneapolis, Minnesota.

AADT Calculation Procedure	100% complete	Less than 0.5% randomly missing	20% randomly missing	First 15 minute of day (0.38%)	3 summer weeks (6%)	7 winter weeks (14%)	8 summer weeks (16%)	14 weeks (27%)	36 weeks (69%)
1. Arithmetic mean	48,486 (reference)	48,524 (<1%)		49,618 (2%)	48,383 (<1%)	49,304 (2%)	48,025 (<1%)	49,554 (2%)	48,273 (<1%)
2. 1994 ASTM Standard	48,476 (<1%)	48,493 (<1%)			48,587 (<1%)	48,521 (<1%)			
3. Conventional AASHTO	48,476 (<1%)	48,493 (<1%)			48,587 (<1%)				
4. Provisional AASHTO	48,476 (<1%)	48,476 (<1%)			48,587 (<1%)				
5. Sum of 24 Annual Avg Hourly Vol.	48,486 (0%)	48,449 (<1%)	48,365 (<1%)	48,499 (<1%)	48,383 (<1%)	49,304 (2%)	48,025 (<1%)	49,554 (2%)	48,273 (<1%)
6. Modified ASTM Standard	48,476 (<1%)	48,493 (<1%)		49,593 (2%)	48,587 (<1%)	48,521 (<1%)	48,454 (<1%)	48,831 (<1%)	48,417 (<1%)
7. Modified Conventional AASHTO	48,476 (<1%)	48,493 (<1%)		49,593 (2%)	48,587 (<1%)	48,787 (<1%)	48,454 (<1%)	49,078 (1%)	48,417 (<1%)
8. Modified Provisional AASHTO	48,476 (<1%)	48,476 (<1%)	48,567 (<1%)	49,540 (2%)	48,587 (<1%)	48,787 (<1%)	48,454 (<1%)	49,078 (1%)	48,417 (<1%)

Note: Empty cells indicate that AADT estimates cannot be calculated because imputation is not permitted (i.e., the 84 monthly average days of the week (MADW) volumes are incomplete and cannot be calculated).

TABLE 3 AADT Estimates for Various Calculation Procedures and Different Patterns of Missing Data: GW Parkway, Arlington, Virginia.

AADT Calculation Procedure	100% complete	Less than 0.5% randomly missing	20% randomly missing	First hour of day (2%)	3 summer weeks (6%)	7 winter weeks (14%)	8 summer weeks (16%)	14 weeks (27%)	36 weeks (69%)
Northbound									
1. Arithmetic mean	37,062 (reference)	37,136 (<1%)	38,172 (3%)	36,641 (1%)	37,036 (<1%)	37,480 (1%)	37,180 (<1%)	37,032 (<1%)	36,770 (<1%)
2. 1994 ASTM Standard	37,103 (<1%)	37,202 (<1%)			36,938 (<1%)	37,127 (<1%)			
3. Conventional AASHTO	37,103 (<1%)	37,202 (<1%)			36,938 (<1%)				
4. Provisional AASHTO	37,103 (<1%)	37,113 (<1%)			36,938 (<1%)				
5. Sum of 24 Annual Avg Hourly Vol.	37,062 (0%)	37,077 (<1%)	37,262 (<1%)	37,047 (<1%)	37,036 (<1%)	37,480 (1%)	37,180 (<1%)	37,032 (<1%)	36,770 (<1%)
6. Modified ASTM Standard	37,103 (<1%)	37,202 (<1%)	38,172 (3%)	36,642 (1%)	36,938 (<1%)	37,127 (<1%)	37,220 (<1%)	37,051 (<1%)	36,979 (<1%)
7. Modified Conventional AASHTO	37,103 (<1%)	37,202 (<1%)	38,172 (3%)	36,642 (1%)	36,938 (<1%)	37,496 (1%)	37,220 (<1%)	37,051 (<1%)	36,979 (<1%)
8. Modified Provisional AASHTO	37,103 (<1%)	37,113 (<1%)	37,110 (<1%)	36,906 (<1%)	36,938 (<1%)	37,496 (1%)	37,220 (<1%)	37,051 (<1%)	36,979 (<1%)
Southbound									
1. Arithmetic mean	36,330 (reference)	36,393 (<1%)	38,279 (5%)	36,113 (<1%)	36,293 (<1%)	36,682 (<1%)	36,410 (<1%)	36,305 (<1%)	36,123 (<1%)
2. 1994 ASTM Standard	36,375 (<1%)	36,468 (<1%)			36,240 (<1%)	36,341 (<1%)			
3. Conventional AASHTO	36,375 (<1%)	36,468 (<1%)			36,240 (<1%)				
4. Provisional AASHTO	36,375 (<1%)	36,380 (<1%)			36,240 (<1%)				
5. Sum of 24 Annual Avg Hourly Vol.	36,330 (0%)	36,349 (<1%)	36,475 (<1%)	36,323 (<1%)	36,293 (<1%)	36,682 (<1%)	36,410 (<1%)	36,305 (<1%)	36,123 (<1%)
6. Modified ASTM Standard	36,375 (<1%)	36,468 (<1%)	38,279 (5%)	36,116 (<1%)	36,240 (<1%)	36,341 (<1%)	36,446 (<1%)	36,291 (<1%)	36,317 (<1%)
7. Modified Conventional AASHTO	36,375 (<1%)	36,468 (<1%)	38,279 (5%)	36,116 (<1%)	36,240 (<1%)	36,721 (1%)	36,446 (<1%)	36,291 (<1%)	36,317 (<1%)
8. Modified Provisional AASHTO	36,375 (<1%)	36,380 (<1%)	36,345 (<1%)	36,243 (<1%)	36,240 (<1%)	36,721 (1%)	36,446 (<1%)	36,291 (<1%)	36,317 (<1%)

Note: Empty cells indicate that AADT estimates cannot be calculated because imputation is not permitted (i.e., the 84 MADW values are incomplete).

TABLE 4 AADT Estimates for Various Calculation Procedures and Different Patterns of Missing Data: US 191/US 287, Yellowstone NP, Wyoming.

AADT Calculation Procedure	100% complete	Less than 0.5% randomly missing	20% randomly missing	First hour of day (2%)	3 summer weeks (6%)	7 winter weeks (14%)	8 summer weeks (16%)	14 weeks (27%)	36 weeks (69%)
Eastbound									
1. Arithmetic mean	1,221 (reference)	1,235 (1%)	123 (90%)	1,671 (37%)	1,072 (12%)	1,402 (15%)	988 (19%)	1,651 (35%)	744 (39%)
2. 1994 ASTM Standard	1,214 (<1%)	1,206 (1%)			1,207 (1%)	1,206 (1%)			
3. Conventional AASHTO	1,214 (<1%)	1,206 (1%)			1,207 (1%)				
4. Provisional AASHTO	1,214 (<1%)	1,213 (<1%)			1,207 (1%)				
5. Sum of 24 Annual Avg Hourly Vol.	1,221 (0%)	1,221 (0%)	1210 (<1%)	1221 (0%)	1,072 (12%)	1402 (15%)	988 (19%)	1,651 (35%)	744 (39%)
6. Modified ASTM Standard	1,214 (<1%)	1,206 (1%)	123 (90%)	1,687 (38%)	1,207 (1%)	1,206 (1%)	1,123 (8%)	1,480 (21%)	874 (28%)
7. Modified Conventional AASHTO	1,214 (<1%)	1,206 (1%)	123 (90%)	1,687 (38%)	1,207 (1%)	1,251 (2%)	1,123 (8%)	1,480 (21%)	874 (28%)
8. Modified Provisional AASHTO	1,214 (<1%)	1,213 (<1%)	1,206 (1%)	1,214 (<1%)	1,207 (1%)	1,251 (2%)	1,123 (8%)	1,480 (21%)	874 (28%)
Westbound									
1. Arithmetic mean	1,238 (reference)	1,251 (1%)	250 (80%)	1,664 (34%)	1,089 (12%)	1,411 (14%)	1,016 (18%)	1,646 (33%)	729 (41%)
2. 1994 ASTM Standard	1,233 (<1%)	1,222 (1%)			1,227 (<1%)	1,225 (1%)			
3. Conventional AASHTO	1,233 (<1%)	1,222 (1%)			1,227 (<1%)				
4. Provisional AASHTO	1,233 (<1%)	1,232 (<1%)			1,227 (<1%)				
5. Sum of 24 Annual Avg Hourly Vol.	1,238 (0%)	1,237 (<1%)	1,226 (<1%)	1,238 (0%)	1,089 (12%)	1,411 (14%)	1,016 (18%)	1,646 (33%)	729 (41%)
6. Modified ASTM Standard	1,233 (<1%)	1,222 (1%)	250 (80%)	1,681 (36%)	1,227 (<1%)	1,225 (1%)	1,151 (7%)	1,479 (19%)	856 (31%)
7. Modified Conventional AASHTO	1,233 (<1%)	1,222 (1%)	250 (80%)	1,681 (36%)	1,227 (<1%)	1,265 (2%)	1,151 (7%)	1,479 (19%)	856 (31%)
8. Modified Provisional AASHTO	1,233 (<1%)	1,232 (<1%)	1,231 (<1%)	1,233 (<1%)	1,227 (<1%)	1,265 (2%)	1,151 (7%)	1,479 (19%)	856 (31%)

Note: Empty cells indicate that AADT estimates cannot be calculated because imputation is not permitted (i.e., the 84 MADW values are incomplete).

TABLE 5 AADT Estimates for Various Calculation Procedures and Different Patterns of Missing Data: State Highway 120 in Yosemite National Park, California.

AADT Calculation Procedure	100% complete	Less than 0.5% randomly missing	20% randomly missing	First hour of day (2%)	3 summer weeks (6%)	7 winter weeks (14%)	8 summer weeks (16%)	14 weeks (27%)	36 weeks (69%)
1. Arithmetic mean	931 (reference)	933 (<1%)	225 (76%)	1,096 (18%)	869 (7%)	1,041 (12%)	844 (9%)	1,166 (25%)	677 (27%)
2. 1994 ASTM Standard	929 (<1%)	920 (1%)			931 (0%)	927 (<1%)			
3. Conventional AASHTO	929 (<1%)	920 (1%)			931 (0%)				
4. Provisional AASHTO	929 (<1%)	928 (<1%)			931 (0%)				
5. Sum of 24 Annual Avg Hourly Vol.	931 (0%)	931 (0%)	922 (<1%)	931 (0%)	869 (7%)	1,041 (12%)	844 (9%)	1,166 (25%)	677 (27%)
6. Modified ASTM Standard	929 (<1%)	920 (1%)	225 (76%)	1,107 (19%)	931 (0%)	927 (<1%)	915 (2%)	1,078 (16%)	716 (23%)
7. Modified Conventional AASHTO	929 (<1%)	920 (1%)	225 (76%)	1,107 (19%)	931 (0%)	948 (2%)	915 (2%)	1,078 (16%)	716 (23%)
8. Modified Provisional AASHTO	929 (<1%)	928 (<1%)	926 (<1%)	928 (<1%)	931 (0%)	948 (2%)	915 (2%)	1,079 (16%)	716 (23%)

Note: Empty cells indicate that AADT estimates cannot be calculated because imputation is not permitted (i.e., the 84 MADW values are incomplete).

CONCLUSIONS

The authors concluded the following after a thorough analysis of the traffic data:

- At urban locations in which commuter traffic significantly dampens seasonality patterns, a significant amount of missing data (up to 8 months of consecutive missing data) can be tolerated with little to no effect on annual average traffic statistics. All of the calculation procedures provided similar results at the urban locations. Therefore, the authors concluded that the critical case for selecting the most suitable calculation procedure will be for rural locations with greater seasonality.

- At rural locations in which recreational traffic produces pronounced seasonality patterns, the amount of missing data is less tolerable. For the two locations examined in this study, it appears that a month of missing data still results in tolerable error (less than 1 percent except for the arithmetic mean and sum of 24 annual average hourly volumes procedures) whereas two months of missing data is close to or exceeds tolerable error levels (up to 10 percent except for the arithmetic mean and sum of 24 annual average hourly volumes procedures).

- At the rural locations, a modified version of the provisional AASHTO procedures provided consistently lower AADT estimation errors for the critical missing data patterns. The only change in this modified provisional AASHTO procedure is that not all 2,016 hourly MADW values (84 MADW values × 24 hours of the day) are required for AADT calculation. The authors recommend that this finding be confirmed at more locations and by other researchers before changes are considered to the provisional AASHTO procedures.

- The authors conclude that systematically missing data up to 50 to 75 percent of all days from operations-based detectors in urban locations with weekday commuter traffic is not fatal flaw for the purposes of calculating annual average traffic statistics. That is, up to 75 percent of all days can be missing data at urban locations, and the errors in AADT estimates will still be less than 5 percent.

- The authors conclude that missing data up to 1 to 2 months (6 to 15 percent) from operations-based detectors in rural areas with significant seasonality patterns is tolerable (errors less than 10 percent with most procedures considered in this study). When the amount of missing data approaches 2 consecutive months or there are many incomplete days whether the missing pattern is random or systematic (e.g., the first time period of each day is missing), the calculation procedures should be selected to minimize estimation error. In this study, the authors found that modified versions of the provisional AASHTO procedures produced the lowest error rates for those situations.

ACKNOWLEDGEMENTS

The authors acknowledge the following agencies for providing traffic data to support this analysis: Arizona Department of Transportation, Minnesota Department of Transportation, and the National Park Service. The opinions and conclusions expressed in this paper are those of the authors and do not necessarily reflect the opinions or conclusions of these agencies.

The authors acknowledge the previous empirical work of Dr. Rich Margiotta of Cambridge Systematics, Inc. and Dr. Joe Wilkinson of Chaparral Systems, Inc. These two friends and data analysts provided the insight and impetus to perform this work based on their previous analyses.

REFERENCES

1. Wright, T., P.S. Hu, J. Young, and A. Lu. *Variability in Traffic Monitoring Data: Final Summary Report.* August 1997, http://www.fhwa.dot.gov/ohim/flawash.pdf, Accessed July 20, 2007.

2. Sharma, S., P. Lingras, and M. Zhong. Effect of Missing Values Estimations on Traffic Parameters. In *Transportation Planning and Technology*, Vol. 27, No. 2, 2004, pp. 119–144.

3. Zhong, M., P. Lingras, and S. Sharma. Estimation of Missing Traffic Counts Using Factor, Genetic, Neural, And Regression Techniques. *In Transportation Research Part C*, Vol. 12, 2004, pp. 139–166.

4. Zhong, M., S. Sharma, and P. Lingras. Genetically Designed Models for Accurate Imputation of Missing Traffic Counts. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1879*, TRB, National Research Council, Washington, D.C., 2004, pp. 71–79.

5. Zhong, M. and S. Sharma. Matching Hourly, Daily, and Monthly Traffic Patterns to Estimate Missing Volume Data. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1957*, TRB, National Research Council, Washington, D.C., 2006, pp. 32–42.

6. Conklin, J.H. and W.T. Scherer. *Data Imputation Strategies for Transportation Management Systems*. Report No. UVACTS, Center for Transportation Studies at the University of Virginia, May 2003, http://cts.virginia.edu/docs/UVACTS-13-0-80.pdf, Accessed July 20, 2007.

7. Chen, M., J. Xia, and R. Liu. *Developing a Strategy for Imputing Missing Volume Data*. Paper #06-2431 submitted for presentation at the Transportation Research Board Annual Meeting, January 2006.

8. Al-Deek, H.M. and C.V.S.R. Chandra. New Algorithms for Filtering and Imputation of Real-Time and Archived Dual-Loop Detector Data in I-4 Data Warehouse. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1867*, TRB, National Research Council, Washington, D.C., 2004, pp. 116–126.

9. Kwon, T.K. TMC *Traffic Data Automation for Mn/DOT's Traffic Monitoring Program*. University of Minnesota-Duluth and Minnesota Department of Transportation, Report No. 2004-29, July 2004.

10. Oz Engineering. AZTech™ *Regional Archived Data System Traffic Conditions Data Aggregation Specification*. unpublished.