SENSITIVITY ANALYSIS OF AADT ESTIMATION ACCURACY BASED ON PARTIAL AND IMPUTED PTC DATA

Ming Zhong, University of New Brunswick
Satich C. Sharma, Satyasrinivas Gadidasu and Sandeep Delta,
University of Regina

INTRODUCTION

Highway agencies at all levels invest a significant portion of financial and human resources for maintaining a strong permanent traffic counting program. Permanent traffic counts (PTCs) are of importance in understanding geographic, temporal, and long-term traffic variations. In particular, PTC data are used to establish road pattern groups and develop expansion factors to extrapolate short-period traffic counts (SPTC) into estimated AADTs for the majority of road segments without permanent traffic counters (PTCRs).

However, due to various reasons, such as equipment failure (caused by vandalism or environmental fatigue), equipment maintenance, and road construction, there are significant missing portions in collected records. Analyses to PTC data from a few highway agencies indicated that a significant portion of permanent traffic counts contain missing data (Zhong et al. 2006). AASHTO recommended practice is to use only those PTCs that meet the requirements of minimum data quantity in each month of the year (AASHTO 1992). However, an examination of missing data patterns shows many of them are continuous blocks and up to 25% to 30% percent of PTCs have to be excluded from the analysis (Zhong et al.

2006). Previous studies (NMSHTD 1990; Zhong and Sharma 2005) indicate that the practice of handling missing PTC data varies from jurisdiction to jurisdiction. Each agency seems have its own minimum data quantity requirement, imputation method, maximum imputation limit, and AADT estimation approach (Zhong and Sharma 2005). No research has been conducted to investigate the sensitivity of AADT estimation accuracy based on partial or imputed PTC data.

In this study, 11 years of PTC data from Alberta are used to investigate the sensitivity of AADT estimation accuracy based on partial or imputed PTC data. Complete PTCs from different road pattern groups are selected and used in the simulations. First, part of these PTCs (e.g., one-month or two-month) are assumed missing and the rest records are used to estimate AADTs. The estimated AADTs are then compared with the true values and estimation errors are calculated. The study then proceeds to use k-Nearest Neighbours (K-NN) methods to impute the assumed missing part and re-estimate AADTs based on the imputed data. The accuracy of AADT estimation is quantified with various statistics (e.g., average and the 95^{th} percentile errors) and the implications of study results to current AASHTO standards are discussed.

This paper is organized as the following: A Literature Review of minimum PTC data requirements, related research and K-nearest neighbors (K-NN) method is right after the Introduction; then the Study Data and Method section introduces the data and method used; the accuracy of AADT estimation based on partial and imputed data is then presented in the Study Results; and finally the Conclusions of this study are given.

LITERATURE REVIEW

Traffic engineers strategically (or historically in some cases) pick up sample roads and set up PTCRs to monitor traffic continuously. It is assumed that the sample roads are typical for each road group and selected randomly from these groups. Hence, it is statistically valid to use data collected from PTCRs to portrait traffic distribution on the rest roads without PTCRs. However, the rigorousness of such statistical analysis is based on the assumption that PTCRs perform well and hence provide enough data to do so. Literature review

indicates that little research has been found for studying the performance of permanent traffic counting programs in terms of the quality of collected data (Zhong et al. 2006). Practicing traffic engineers and data analysts may have a rough idea about how much missing data exists in collected records, however, detailed statistical analysis is not available.

Federal Highway Administration (FHWA 1985) recommended that each state or province maintain 40 to 60 permanent traffic counters (PTCRs) (or automatic traffic recorders (ATRs)). The new Guide (FHWA 2001) recognizes that most jurisdictions have more PTCRs than recommended and therefore does not specify any particular number. It is recommended that these PTCRs be assigned into 5 or 6 groups based on their functional classes and temporal variations (FHWA 1985, 2001). The data from PTCRs are used not only to study traffic characteristics and trend of the sites being continuously monitored, but also to convert short-period counting data into important traffic summary statistics, such as annual average daily traffic (AADT), on those roads without PTCRs. However, it is not clear if the recommendation from the old Guide (1985) considers the breakdown of counting equipment and how this should be taken into account when designing and operating a traffic monitoring program. Zhong et al. (2006) show that most of PTCRs are actually not able to run perfectly throughout a year, and malfunctions and shutdowns of counting machines result in "holes" or missing values in collected records. The collected data may not be a good base for a rigorous statistical analysis.

For PTC data, the American Association of State Highway and Transportation Officials (AASHTO) *Guidelines for Traffic Data Programs* (AASHTO 1992) recommends that highway agencies adopt two-day minimum of edit-accepted data for each day of the week, each month of the year for calculating annual traffic summary statistics, before the automatic data edition is implemented (hereafter referred as the *two-day minimum standard*). That is, only those PTCs with at least two-day complete data for each day of the week (from Sunday to Saturday), for each month of the year (from January to December) can be used to calculate annual traffic summary statistics. It recommends using one-day minimum of edit-acceptable data

(hereafter referred as the *one-day minimum standard*) for producing annual traffic summary statistics after the implementation of automatic data edition. *AASHTO Guidelines* (1992) also suggests "highway agencies should ensure that the adjustment factors calculated from permanent counter sites are based on the designed minimum of 5 sites" (page 4). However, strictly following these standards may challenge the validity of the recommendations of FHWA and AASHTO because missing data could result in less number of PTCs than recommended in certain groups. Therefore, highway agencies may have to either set up more PTCRs or impute collected records to maintain minimum data integrity.

Zhong et al. (2006) found that more than 25% of PTCs do not meet the two-day minimum requirement and over 30% of them do not meet the one-day minimum requirement of the AASHTO Guidelines. Pattern analysis of missing data indicated many of them are large blocks, rather than sporadic short periods. This is usually because of the nature of traffic monitoring program: due to widely distributed locations of PTCRs, it takes time to get the equipment fixed once it is malfunctioned.

Davis and Nihan (1991) were one of the first transportation engineers to use K-NN method to estimate short-term traffic flows. Their study focused on estimating the development from a free flow state of traffic to a congested state. They concluded that "The nearest neighbor nonparametric regression approach replaces the problem of selecting a class of models and then estimating parameters with the problem of maintaining and sorting an adequately large learning sample". Smith and Demetsky (1997) compared the k-NN method with the historical, time series, and artificial neural network models (ANN). They identified the k-NN method as the one that produced the lowest traffic forecasting errors. An important merit of the k-NN method, as reported by Oswald et al. (2000), is that it maintains every observation making it particularly useful in modeling unusual situations. The k-NN method basically involves the following procedures: (1) determining State Vector; (2) determination of the k Nearest Neighbors; and (3) generation of the Forecast. For a detailed description of the method, please see Hardle (1990).

STUDY DATA AND METHOD

The study uses 11 years of Alberta PTC data from 1995 to 2005, since they are "perfect base data" without "pollutions" from any imputing activities. The data are in the form of hourly volumes. The number of available counters and the file formats vary from year to year. Each counter file contains 8760 (365×24) hourly volume observations for non-leap years or 8784 (366×24) observations for leap years.

14 PTCRs were selected from the following three trip pattern groups: five from the commuter group, six from the rural long-distance group, and three from the recreational group. Due to insufficient data, no counts were selected from the high recreational group. Table 1 shows the trip pattern groups, functional classes and AADT values of the selected PTCRs. It can be seen the PTCRs are located on high functional-class roads (arterials and above) and their AADTs vary significantly, ranging from 550 to 51,300.

Figure 1 shows the locations of the selected PTCRsin the middle and south part of the province of Alberta. It should be noted that these counters are located on ten major highways throughout Alberta and labelled with the "Site No." listed in Table 1. The commuter counters were located close to the major cities in the province: one is PTCR 002321 (Site 12), which is near the provincial capital ─ the city of Edmonton International Airport and the two are adjacent to the city of Calgary PTCR 002181 and 011010. The average remote area is covered by six rural long distance sites. For example, PTCR 016065 and 003061 are located at a moderate distance with Edmonton and Lethbridge respectively. The three recreational sites are scattered in the province of Alberta: PTCR 093001 is near the scenic Lake Louise within the Banff National Park; PTCR 006041 is located close to Waterton National Park; and PTCR 001061 is located west of Cochrane on the old Trans-Canada Highway 1, which has evolved as a recreational road to Rocky Mountains since the new Trans-Canada was built.

The above PTCRs are all having 100% complete records over a period of eleven years. It is assumed that certain portions of PTC data were missing (from one day up to two months) and modified records

are directly used or imputed to re-estimate AADT. The estimated AADT will then be compared with the true values calculated based

Table 1 detailed information of the selected PTC sites

Road Type	Site No.	PTC Site	Functional Class	AADT (Veh/day)
Rural long distance	1	003061	Multilane Arterial	3,739
	2	009045	Major Arterial	1,871
	3	016065	Multilane Arterial	7,090
	4	016121	Expressway	12,549
	5	028085	Major Arterial	2,593
	6	028101	Major Arterial	2,462
Recreational	7	001061	Freeway	16,801
	8	006041	Major Arterial	542
	9	093001	Minor Arterial	2,195
Commuter	10	011010	Expressway	13,376
	11	002181	Expressway	45,606
	12	002321	Freeway	51,301
	13	003081	Expressway	14,971
	14	011145	Major Arterial	5,000

on the original records. The estimation accuracy is quantified with percent error (PE) or absolute percent error (APE) using the following formulae:

$$PE = \frac{true\ AADT - estimated\ AADT}{true\ AADT} \times 100 \qquad (1)$$

$$APE = \frac{|\,true\ AADT - estimated\ \ AADT\,|}{true\ AADT} \times 100 \qquad (2)$$

The key evaluation parameters consist of the average and 95th absolute percentile errors. These statistics give a clear profile of model's error distribution by including (average) and excluding (the 95th percentile) large errors.

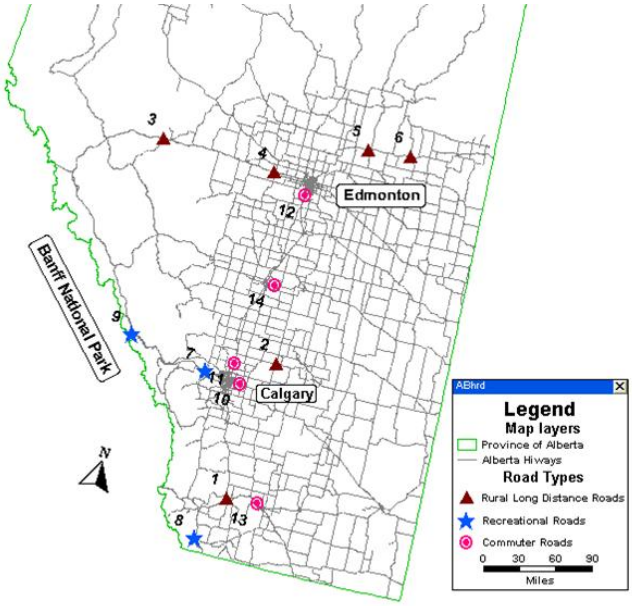

Figure 1 the locations of selected permanent traffic counters

STUDY RESULTS

Analyses were made by assuming that certain range of values was missing in the original records. The missing durations considered in this study include one-day (weekday or weekend day), two-day (weekday or weekend day), four-day (weekday), seven day (a week), 2 weeks, 1 month and up to 2 months. Hourly volume data from specific periods were removed and AADTs were re-estimated based on either partial or imputed records. The estimated AADTs were

compared with true AADTs to quantify the estimation errors. Due to limited space here and the fact that the large missing durations are more significant to traffic data programs, only the results for 1 month and 2 month missing durations are presented.

First, partial PTC data by removing one or two-month observations from the complete records is used to estimate AADTs and compared with the true values. The percent errors are then calculated using Equation (1). The simulations were run for each one or two successive months during a year and the estimation errors for corresponding errors are shown in Figure 1. Figure 1(a) shows AADT estimation errors after removing one-month complete records. It can be seen from the figure that the commuter roads result in the lowest errors, whereas the recreational roads result in the highest errors. The rural long-distance roads show errors between. The AADT estimation errors for the commuter roads are all less than 2% indicating that a month missing data hardly have any impact on estimated AADTs. For example, applying 1% or 2% over- or under-estimation to the five PTCs from the commuter group results in a traffic difference of 100 vehicles/day for PTC 011145 with an AADT of 5000 and about 1,000 vehicles/day for PTC 002321 with an AADT of over 50,000. The rural long-distance roads show marginal increases in estimation errors. In contrast, the recreational roads show significant errors of about 10% underestimation during the summer months and about 5% overestimation errors for winter months. This is consistent to what expected because of their summer recreational nature, that is, they carry relatively very high traffic for summer months, but quite low volumes for the rest seasons. Missing data occurring during summer months has more significant impact on AADT estimations.

There are obvious seasonal effects on estimation accuracy. It can be found from Figure 1(a) that one-month data missing during April, May, September or October have little impact on AADT estimation. However, the other months show significant effects, especially summer months like July and August.

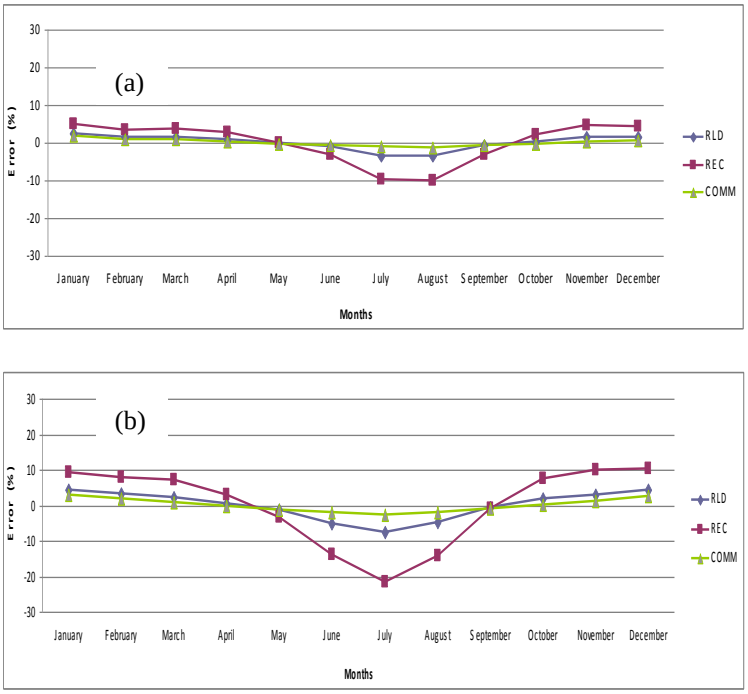

Figure 2 average AADT estimation errors based on (a) one-month missing (b) two-month missing

Figure 1(b) show AADT estimation errors based on removal of two-month data records. Compared to Figure 1(a), the errors are consistently higher. Similar rules can be observed as Figure 1(a) as the commuter roads result in the lowest errors and the recreational roads result in the highest. Two-month missing data result in a maximum error of 7-8% for rural long-distance roads and up to over 20% for the recreational roads. It appears that AADT estimations on the commuter roads are less sensitive to missing data. Two-month missing data has very little impact on estimated AADT (less than 3%). Again, seasonal effect can be observed. For those cases with missing data over the summer months, AADT are underestimated; whereas for the other seasons, AADT are overestimated.

On account of the past successes, the k-NN method is used to impute assumed missing data. As mentioned before, the procedure involved determining state vector, selecting an appropriate number of nearest neighbors (k value), determination of the k nearest neighbors, and generation of the forecast using weighted average values. In this study, a hybrid state vector similar to that from Smith et al. (2002) is defined as the following:

$$x(t) = V(t), V(t-1), V(t-2), V_{hist}(t), V_{hist}(t+1) \quad \ldots\ldots (3)$$

Where, *x(t)* is the state vector for predicting traffic volume factor at time *t+1*, *V(t)*, *V(t-1)*, and *V(t-2)* are the traffic volume factors at time interval t, t-1, and t-2, respectively. $V_{hist}(t)$ and $V_{hist}(t+1)$ are the historical average volume factors associated with time interval t and t+1.

It is found that K-NN method works better with monthly factors rather than monthly volumes. Therefore, average daily traffic (ADT) is first calculated based on partial PTC data and used to develop monthly factors for those months with available data. The monthly factor patterns from candidate PTCs within each pattern group are also computed to prepare for imputation. In cases of imputing monthly factors, the historical average is calculated based on the monthly factors from the same month and the same PTC site in the past.

Once the state vector is determined, the next step is to search for the k nearest neighbours that will be used in prediction. The 'closeness' between observations in a multivariate space can be commonly measured using the Euclidean Distance, and then the k observations with the shortest Euclidean Distances are recognized as neighbours. A series of experiments are carried out for different type of roads and a k value of 4 is selected (Results are not shown here due to limited space).

Once the k value is determined, K-NN methods are used to search for the closest neighbors, in this case, 4 closest monthly factors. Then a weighted average of the obtained monthly factors according to the equation used by smith et al. (2002) is used to impute the missing data. The AADT is estimated by multiplying

ADT with the obtained monthly factor vector and divided with 12. The estimated AADTs are then compared with the true values and the estimation errors are calculated. Figure 2 shows the average and the 95th absolute percentile errors (APEs) for AADT estimations based on the imputed one-month (2(a)) and two-month (2(b)) data using K-NN method. It can be found from both Figure 2(a) and 2(b) that the estimation errors are much lower than those based on the partial data. The average errors for the estimations based on one-month imputed data are all less than 0.5%, regardless which pattern groups they are from. Most of the 95th percentile errors are less than 1% and only a few months from the recreational roads have a slightly higher error. Figure 2(b) shows the errors based on two-month imputed data. Similar to the one-month imputed data cases, all average errors are less than 1% and most of the 95th percentile errors are lower than 1.5%. One in one case of recreation road, K-NN results in the 95th percentile error of more than 2.5% where the whole July and August are assumed missing and imputed. Since the error magnitudes are so small across all cases, it is safe to conclude that no significant effects can be expected.

The low estimation errors resulted from simple imputation methods like K-NN in this study clear show missing data up to two months does not have any significant impact on estimated AADTs and probably group expansion factors. Therefore, it is evident that the AASHTO minimum data requirements, such as one-day minimum and two-day minimum, are too strict. A previous study (Zhong et al. 2006) shows that up to 25% to 30% of PTCs cannot be used to compute AADT and develop group expansion factors because they do not meet the AASHTO requirements. This study clearly shows that even with one or two complete month missing data, simple imputation can still result in very accurate AADT estimates. Study results clearly support that the AASHTO minimum data requirements should be revisited in light of this and future similar studies.

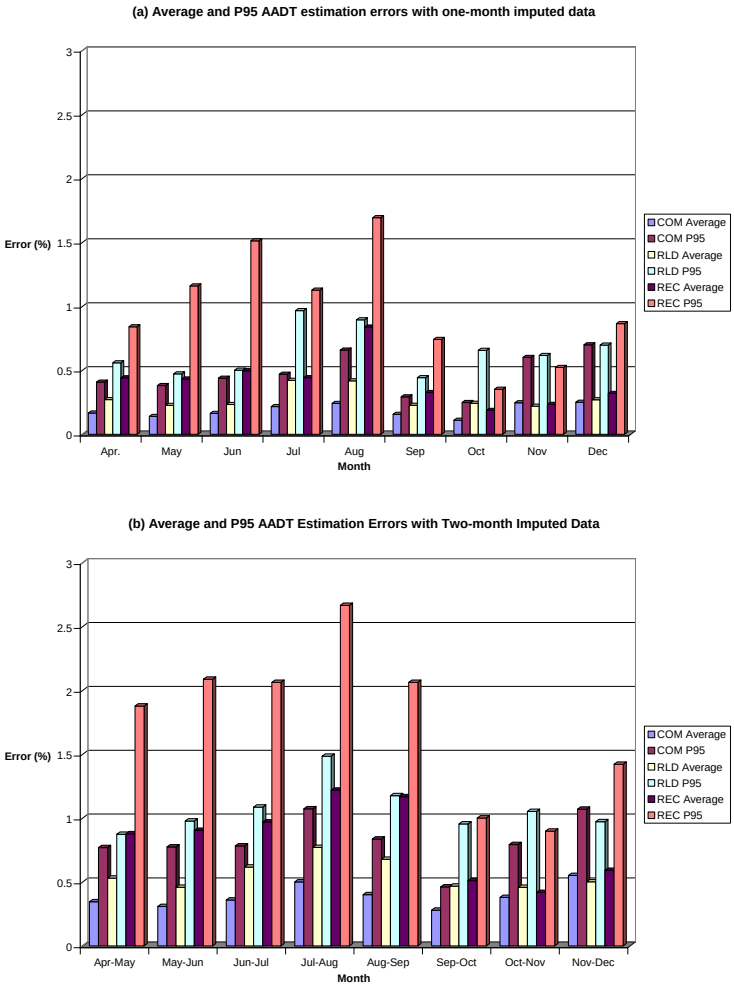

Figure 3 AADT estimation errors based on (a) one-month and (b) two-month imputed data using K-NN method

CONCLUDING REMARKS

This study investigates the sensitivity of AADT estimation accuracy based on partial and imputed PTC data. Following conclusions are drawn from the study results:

1. AADT could be significantly over- or under-estimated based on partial PTC data. For example, the average error can be up to 20% for the recreational roads. In case of one-month or two-month missing data over the summer months, AADT will be underestimated, whereas for missing data occurring during the other seasons, AADT will be overestimated.
2. AADT estimations for different road pattern groups show different sensitivity to missing data. It is found that the commuter group shows least sensitivity, whereas the recreational group shows the greatest and the rural long-distance group shows medium.
3. Simple imputation methods like K-NN can result in highly accurate AADT estimates across all road pattern groups. Most of the average errors for missing durations of one and two months are less than 0.5% and even the 95^{th} percentile errors are less than 2%.
4. Permanent traffic monitoring programs historically maintain large traffic databases, which provide enormous possibilities of using the collected information to address many new data needs. Study results show that much information can be used to accurately estimate AADTs in case of one and two complete months missing PTC data. According to *AASHTO Guidelines*, such PTCs are not eligible to be used for estimation of AADT. The study results indicate that such a standard may be too stringent and need to be reviewed.

This paper studies the AADT estimation accuracy based on the partial and imputed data with a missing duration of up to one or two months. Future research should explore this issue for larger missing durations (e.g., 4 or 6 months). Further, the impact of using partial and imputed data to estimate group expansion factors should also be evaluated in a future study.

ACKNOWLEDGEMENTS

The authors are grateful towards the Natural Science and Engineering Research Council (NSERC), Canada for their financial support, and the Alberta Transportation for the data used in this study.

REFERENCES

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

Davis, G.A. and Nihan, N.L. Nonparametric Regression and Short-Term Freeway Traffic Forecasting, *Journal of Transportation Engineering*, Vol. 117, No.2, pp 178-187, 1991.

Federal Highway Administration (FHWA). *Traffic Monitoring Guide.* Federal Highway Administration, U.S. Department of Transportation, Washington, D.C., 1985.

Federal Highway Administration (FHWA). *Traffic Monitoring Guide.* Federal Highway Administration, U.S. Department of Transportation, Washington, D.C., 2001.

Hardle, W. *Applied nonparametric regression*, Cambridge University Press, 1990.

New Mexico State Highway and Transportation Department (NMSHTD). *1990 Survey of Traffic Monitoring Practices among State Transportation Agencies of the United States*. Report No. FHWA-HRP-NM-90-05. Santa Fe, New Mexico, 1990.

Oswald, K.R., Scherer, W.T. and Smith, B.L. Traffic Flow Forecasting Using Approximate Nearest Neighbour Nonparametric Regression, *Final Report of ITS Center Project, Traffic Forecasting: Non-parametric Regressions*, Center for Transportation Studies, University of Virginia, 2000.

Smith B.L. and Demetsky M.J. "Traffic Flow Forecasting: Comparison of Model Approaches", *Journal of Transportation Engineering*, Vol. 123 No. 4, 261-266, 1997.

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

Zhong, M. and Sharma, S.C. “Examining the Imputation Accuracy of Highway Agencies”. *The Journal of ITE on the Web*, Vol. 75, No. 6, pp. 77-81, 2005.

Zhong, M., Sharma, S.C., and Dalta, S. (2006). “Studying Counting Efficiency of Continuous Traffic Monitoring Program and A Discussion about AASHTO Minimum Data Requirement”. Presented at the *CSCE 2006 Annual Conference*, Calgary, May 2006.