

# Local MAD Method for Probe Vehicle Data Processing

Xuegang (Jeff) Ban <sup>1\*</sup>, Yuwei Li <sup>2</sup>, Alex Skabardonis <sup>3</sup>

<sup>1</sup>:CCIT, Institute of Transportation Studies, University of California, Berkeley; 2105 Bancroft Way, Suite 300, Berkeley, CA 94720; Phone: 510-642-5112; Fax: 510-642-0970; Email: [xban@calccit.org](mailto:xban@calccit.org)

<sup>2</sup>: Yuwei Li, PATH of ITS, University of California, Berkeley; Email: [yuwei@path.berkeley.edu](mailto:yuwei@path.berkeley.edu)

<sup>3</sup>: Alex Skabardonis, PATH of ITS, University of California, Berkeley;

Email: [skabardonis@ce.berkeley.edu](mailto:skabardonis@ce.berkeley.edu)

\*: Corresponding Author

## Abstract

This paper presents a local median absolute deviation (MAD) method to process travel times from raw probe vehicles data. The method is applied locally to each time window (band) with a fixed duration. To determine a proper band width, a sensitivity analysis is conducted to study how the band width impacts the performance of the method. The analysis reveals that if data samples are fairly large, a band width of 15-30 minutes can be chosen in order to capture the trend of travel times and the statistical rigor of the local MAD method. The method and the analysis presented in this paper are expected to provide some insights to practitioners who plan to use probe vehicle data for traffic analyses.

## 1. Introduction

Probe vehicles can provide a rich set of traffic information for various traffic management and traveler information applications. Among all the data that can be obtained from probe vehicles, travel time information is one of the most important. This is because, firstly, travel time is one crucial measure to assess the performance of traffic conditions. For example, recent research on travel time reliability aims to quantify travel time variations across different days (Chen et al., 2003; Liu et al., 2006). On the other hand, as more and more technologies are applied in the field nowadays, travel time has become the most critical traveler information based on which drivers can make informed decisions. One such example is that Federal Highway Administration (FHWA) has recently recommended that “no new Changeable Message Signs (CMS) should be installed in a major metropolitan area or along a heavily traveled route unless the operating agency and the jurisdiction have the capability to display travel time messages.”

Although travel time can be estimated from other types of sources, e.g., loop detector data, probe vehicles can, however, provide the so-called “ground-truth” travel times, i.e., travel times that are actually experienced by individual vehicles. This type of travel times turns out to be very valuable for the above traffic management and traveler information applications.

Travel times generated by probe vehicles, however, may contain significant amount of outliers that must be filtered. The filtering may be further complicated by the inherent time-dependent trend of travel times within a given day. In this paper, we apply a local Median Absolute Deviation (MAD) method to remove outliers in the raw *FasTrak* data, a kind of probe vehicle data produced by electronic toll tag readers in the San Francisco Bay Area. Due to the obvious

time-dependent trend of travel times, the MAD method is applied locally to data points that are within a given time window. We show that for properly selected time window, the local MAD method is very effective to remove outliers. We further conduct the sensitivity analysis of how the length of the time-window impacts the performance of the method, based on which a time window with proper length can be selected. It turns that a time window with the length about 15 – 30 minutes is appropriate to remove outliers. Further, the performance of the method is not sensitive to lengths of time windows within this range.

This paper is organized as follows. The local MAD method is presented in Section 2, together with a brief description of the *FasTrak* data. Section 3 provides the sensitivity analysis of performances of the MAD method with respect to varied lengths of the time window. This is followed by conclusion remarks and future study directions in Section 4.

## 2. Local MAD Method

### 2.1 Description of Probe Vehicle Data

The probe vehicle data was obtained from *FasTrak* in the San Francisco Bay Area. *FasTrak* is used statewide in the State of California, US to automatically collect road and bridge tolls. *FasTrak* readers are currently installed at each toll booth, as well as along the road side in a spacing of 5 to 10 miles. The data contains individual vehicle travel times between two readers and thus is rich information to obtain accurate and reliable travel times. Figure 1 below depicts a route along US-101 near the Golden Gate Bridge (north of the City of San Francisco). The starting and ending points of the route are the locations of two *FasTrak* readers, as shown in the figure. This route is about 2.5 miles (4 km) with normal travel time 150 seconds at 60 mph (96 km/hour). The route travel time can be obtained directly from the raw *FasTrak* data (i.e., passage times at the starting and ending *FasTrak* readers). The raw *FasTrak* data, however, needs to be processed to remove outliers.

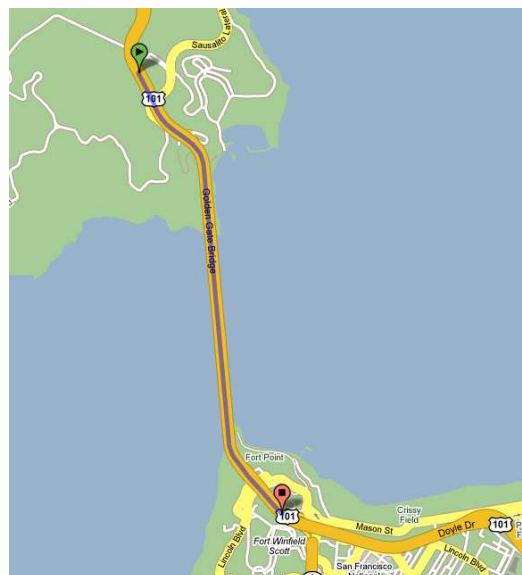


Figure 1 The Test Route (source: [www.map.google.com](http://www.map.google.com))

## 2.2. Local MAD Method

Figure 2 depicts the raw *FasTrak* travel time data for the route in Figure 1. Note that Figure 1 only shows travel times within -50 – 500 seconds, while those outside this range were ignored. The time-dependent pattern of the travel time is obvious in Figure 2, but the raw data contains a significant amount of outliers. Outliers include those vehicles that took excessively long time to travel the route, possibly because they left and re-entered the freeway at some point. Vehicles that used the HOV lane are also treated as outliers, as we are interested in predicting travel time for those in the general lanes.

To remove outliers, we applied the Median Absolute Deviation (MAD) method. MAD is a statistical measure for capturing the variation of a given set of data points. Assume  $x_i, i = 1, \dots, N$  is the set of data points. Then MAD can be defined as

$$MAD = median(|x_i - \tilde{x}|). \quad (1)$$

Here  $\tilde{x}$  is the median value of  $x_i, i = 1, \dots, N$ . In practice, MAD in (1) is less impacted by outliers in the data set since extreme points have less influence on the calculation of the median than they do on the mean.

To detect whether  $x_i$  is an outlier, a  $z$ -score needs to be computed for each data point:

$$z_i = \frac{|x_i - \tilde{x}|}{MAD} \quad (2)$$

Then if  $z_i \geq \bar{z}$  for a given threshold  $\bar{z}$ ,  $x_i$  is an outlier. Here we use  $\bar{z} = 4.5$ .

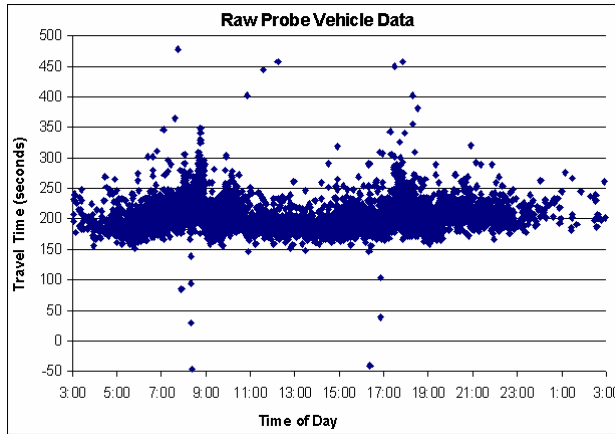


Figure 2 Raw *FasTrak* Data

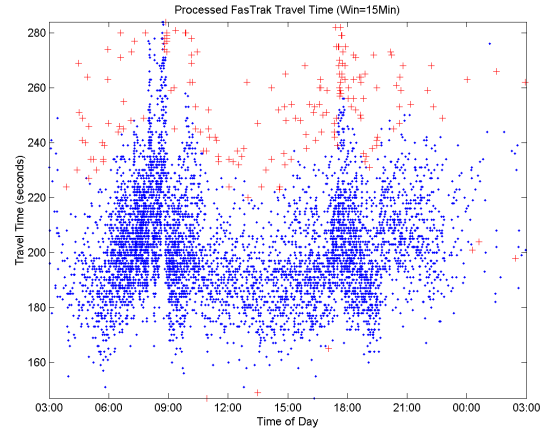


Figure 3 Processed *FasTrak* Data

Since the raw *FasTrak* data in Figure 2 has a clear time-dependent pattern, we divided the entire time window to time “bands” and applied the MAD method locally to each band – the so-called local MAD method. The width of the band, denoted as  $h$ , can be specified. Figure 3 shows the

processed *FasTrak* data by setting band width to 15 minutes. The red “plus” signs in Figure 3 represent outliers, while the blue dots denote “good” data points. One can see clearly that the local MAD method is effective to remove outliers. Also note that even after being processed, the true travel time at a given time instant is random (i.e., spans over a range) rather than deterministic (i.e., a single value). The latter has been assumed by most previous travel time related studies (Lindveld et al., 2000).

The width of the band is expected to have some impacts on the performance of the local MAD method. This will be discussed in more detail in the next section.

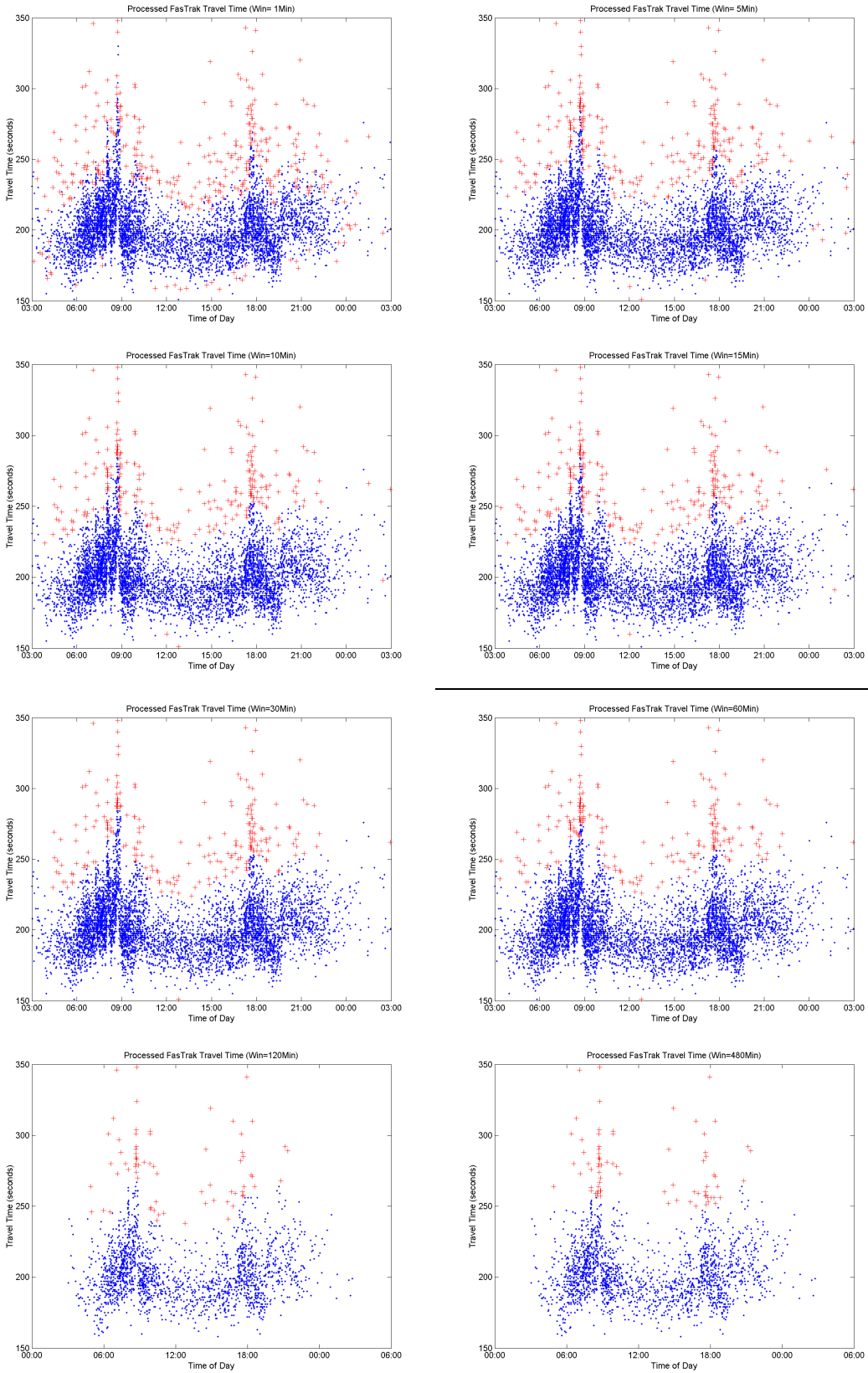
### 3. Sensitivity Analysis of Band Width

To study how different band widths impact the performances of the local MAD method, we set  $h=1, 5, 10, 15, 30, 60, 120, 480$  minutes and run the local MAD method once for each  $h$ . Figure 4 depicts the processed travel times. In these plots, similarly as that in Figure 3, blue dots represent the “good” data, whereas red “plus” signs denote outliers identified by the local MAD method.

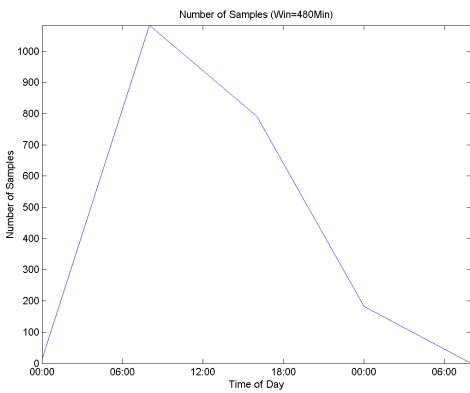
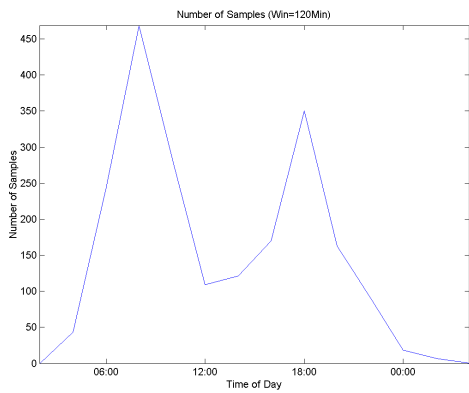
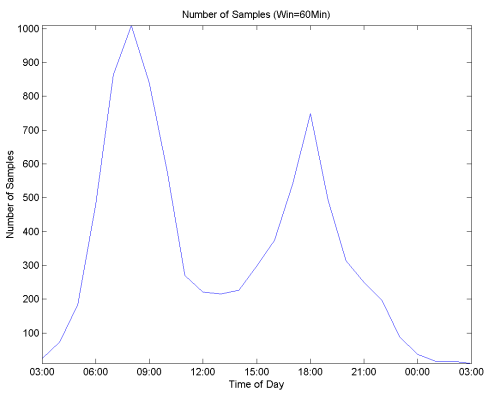
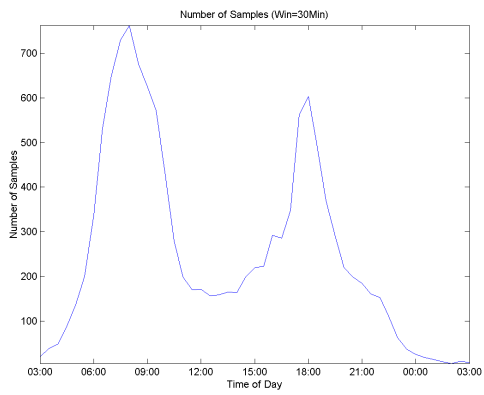
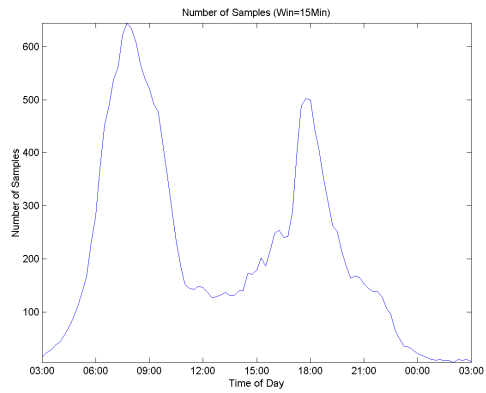
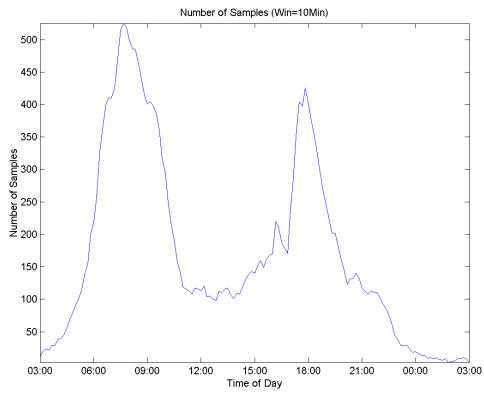
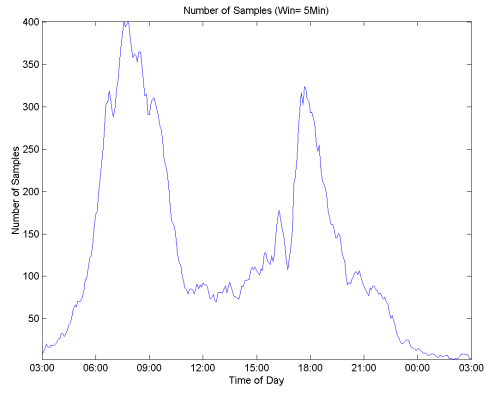
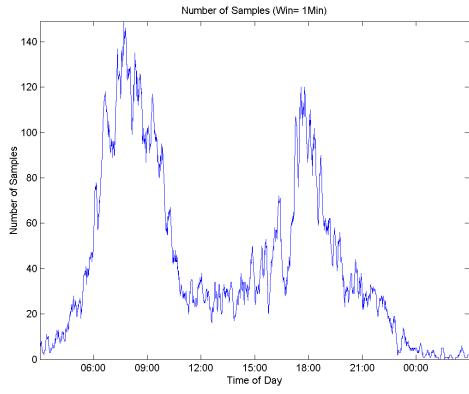
These plots clearly show that the local MAD method is not very sensitive to the widths of the bands when  $h \leq 60$  minutes. One reason for this may be that, as shown in Figure 5, the number of data points in each band is sufficiently large, even when the band width is small. For example, for  $h=1$  minute, the number of data points for most times of the day is larger than 30. For peak hours, the numbers goes up to more than 100. Because there are enough data points in each band, the local MAD method is statistically meaningful and thus the method is not very sensitive to the band width for small  $h$ 's. As  $h$  becomes too large (e.g.,  $h=120$  and 480 minutes), however, more data points during the peak periods are identified as outliers. Therefore, to retain the time-dependent trend of travel times within a day,  $h$  may not be greater than 60 minutes.

Note that the above data analysis is based on a single day with sufficient *FasTrak* samples (more than 7,000 per day) for the studied route in Figure 1. For other days, the number of *FasTrak* samples may be much smaller. To test how band width impacts the local MAD method in these days, we run the method for one of such days with the same band width settings. Figure 6 and Figure 7, depict, respectively, the processed travel times and the number of data points in each band. First, there are less than 2,500 data points in this particular day for the studied route. We can easily observe that when band width is too small (1 – 10 minutes), the local MAD method does not perform very well since there are still some unidentified outliers, as marked in circles in the plots in Figure 6. The reason for this is evident from Figure 7, which shows that for smaller band widths, the number of data points in each band is small (less than 5 for most times of the day). As the band width becomes larger (i.e., bigger than 15), most bands will have more than 25 data points. In other words, for very small bands, the local MAD method is applied to limited number of data points which may not be statistically meaningful.

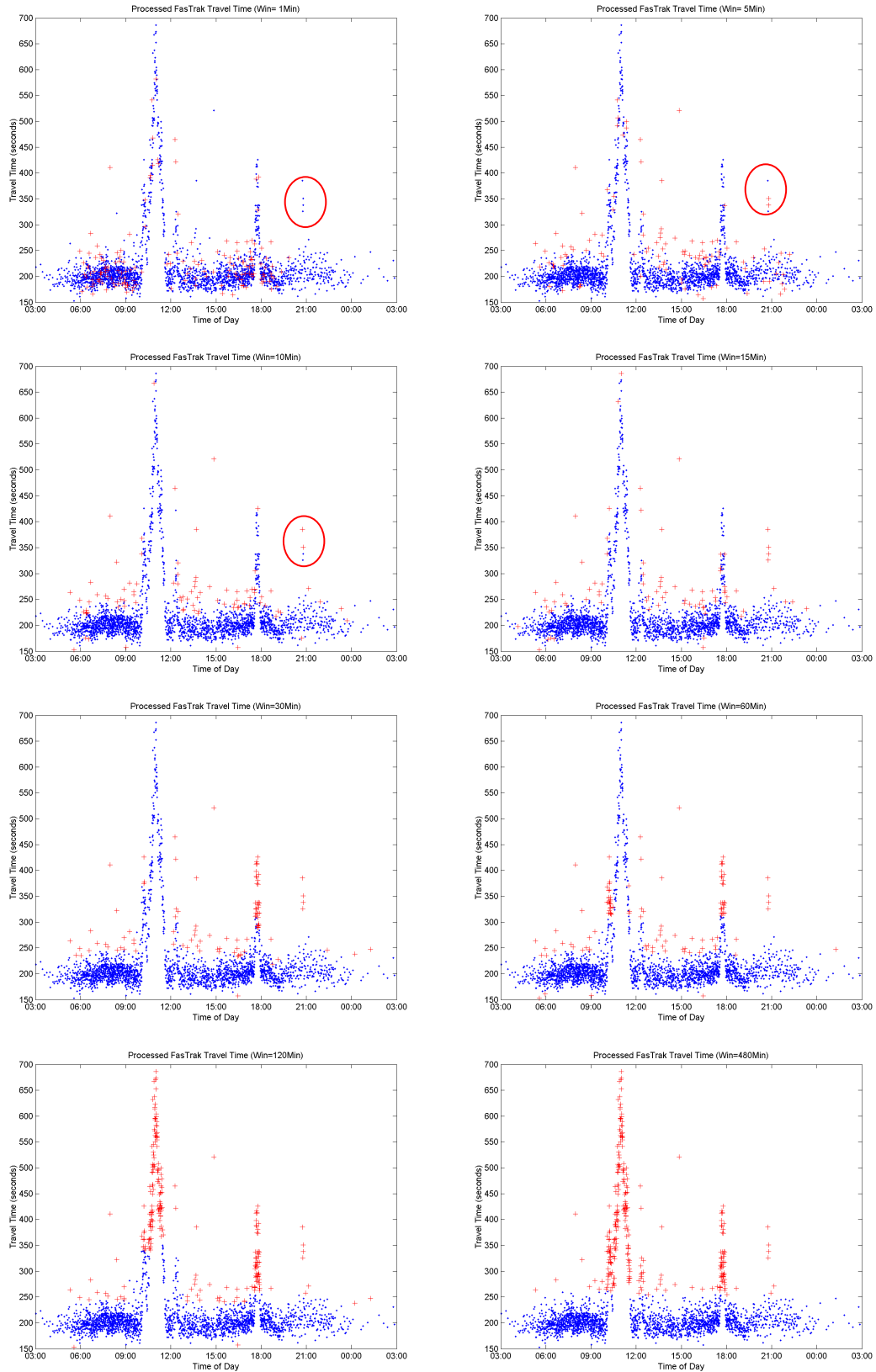
Figure 6 also depicts that as  $h$  goes to too large (i.e., 120 or 480 minutes), most data points during peak hours tend to be identified by outliers and thus removed, which is obviously not appropriate given the clear time-dependent trend of the data. This illustrate that the band width should be less than 60 minutes in order for the local MAD method to yield meaningful results.



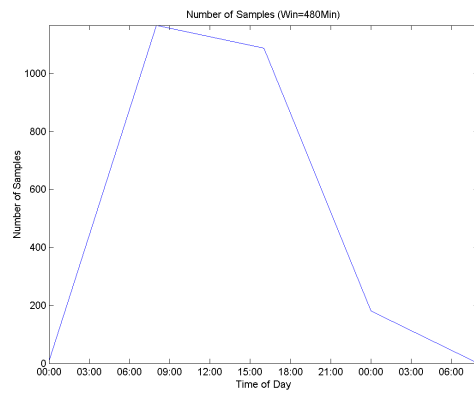
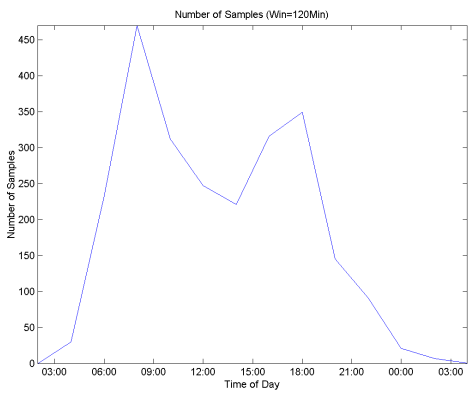
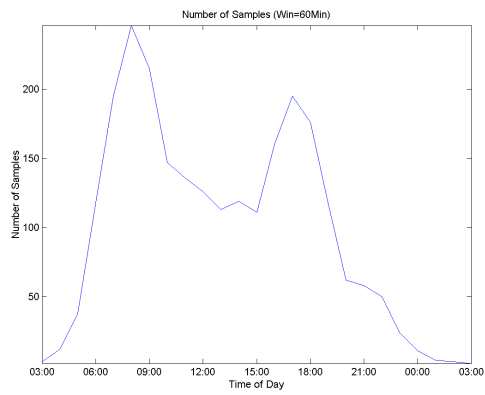
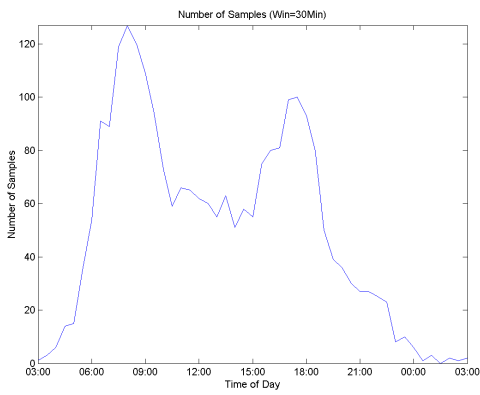
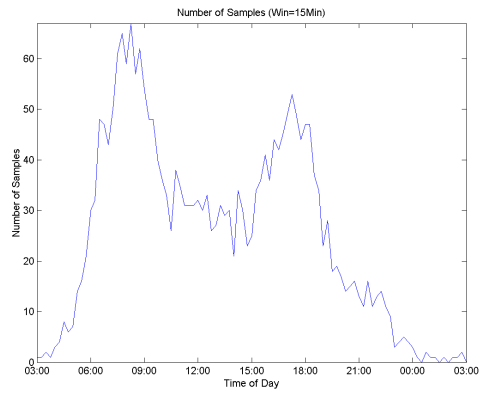
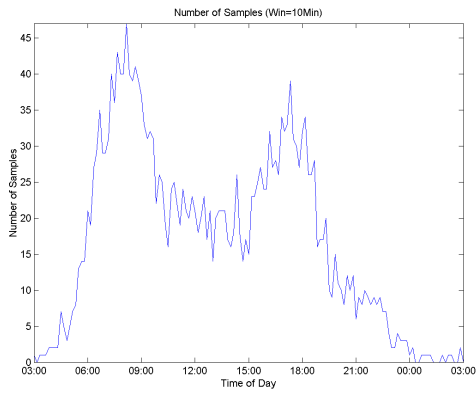
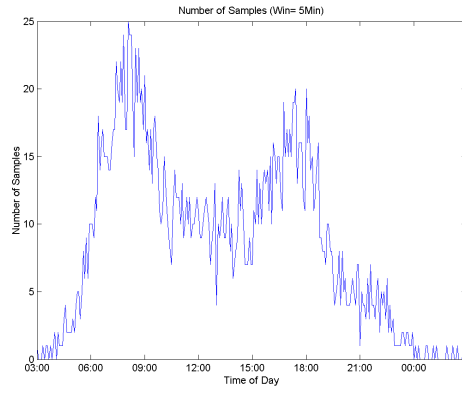
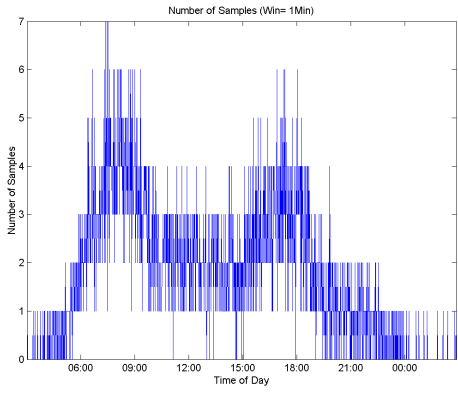
**Figure 4 Processed Travel Times Using Different Bands**



**Figure 5 Number of Data Points in Each Band**



**Figure 6 Processed Travel Times – Another Day**



**Figure 7 Number of Data Points in Each Band – Another Day**

From these figures, one may conclude that depending on the number of *FasTrak* samples within a given day, the band width should be chosen as 15 minutes or larger. On the other hand, the band width can not be set too large in order to capture the obvious time-dependent trend (especially during peak hours) of the within-day travel times. Therefore, we recommend  $h = 15 - 30$  minutes based on our analysis on the *FasTrak* data in this paper. We should point out that this recommendation is made on the basis that the number of *FasTrak* samples is no less than 2,000 for a given day. If the number of samples is much smaller than this value, more sophisticated approaches are needed to select a proper band width.

The above analysis also demonstrates that for  $h = 15 - 30$  minutes, the performance of the local MAD method is not very sensitive to the actual value of  $h$ . This is desirable when processing probe vehicles data using the local MAD method.

#### 4. Conclusion Remarks

In this paper, we presented a local MAD method to process raw *FasTrak* data to generate the ground truth travel times. The method is applied locally to each time band to capture the time-dependent trend of with-day travel times. In order to select a proper value for the band width, a sensitivity analysis was conducted in this paper to study how the band width influences the performance of the proposed method. The results showed that for data with more than 2,000 data samples per day, a band width of 15 – 30 minutes should be selected, which can capture the time-dependent trend, as well as make the method statistically significant. The analysis and recommendation in this paper, hence, may provide some insights for practitioners who are using probe vehicle data.

As mentioned in Section 3, if the number of samples is much smaller than 2,000, a more sophisticated method may be needed to choose the proper band width. In particular, from both Figure 4 and 6, the distribution of data samples within a day is not even: there are always more samples during peak hours than non-peak hours. Therefore, a variable band width based on different time-of-day seems more appropriate to capture this uneven distribution of data samples. Research in this line has shown some promising results (Ban et al., 2007) and will be further pursued in future studies.

#### References

1. Ban, X., Li, Y., Skabardonis, A., and Mugulici, JD. (2007) Performance evaluation of travel time methods for real time traffic applications. To be presented at the 11<sup>th</sup> World Congress on Transportation Research, Berkeley, California.
2. Chen, C., Skabardonis, A., and Varaiya, P. (2003) Travel time reliability as a measure of service. *Journal of Transportation Research Board* 1855, 74 - 79.
3. Lindveld, C.D.R., Thijs, R., Bovy, P.H.L, and Van der Zijpp, N.J. (2000) Evaluation of online travel time estimators and predictors. *Journal of Transportation Research Board* 1719, 45-53.
4. Liu, X., He, X., and Recker, W. (2007) Estimation of the time-dependency of values of travel time and its reliability from loop detector data. *Transportation Research* 41B, 448-461.