Paper No. 02-2129

Reconsideration of Sample Size Requirements for Field Traffic Data Collection Using GPS Devices

By

Shuo Li, P.E.
Research Engineer
Division of Research, Indiana Department of Transportation
1205 Montgomery Street, West Lafayette, IN 47906
sli@indot.state.in.us
(765) 4631-521

Karen Zhu, Ph.D.
Senior System Analyst
Division of Research, Indiana Department of Transportation
1205 Montgomery Street, West Lafayette, IN 47906
kzhu@indot.state.in.us
(765) 4631-521

van Gelder B.H.W, Ph.D.
Associate Professor
Geomatics Engineering, 1284 Civil Engineering Building
Purude University, West Lafayette, IN 47907
vngelder@ecn.purdue.edu
(765) 494-2165

John Nagle
Safety/Congestion Management Engineer
Program Development Division, Indiana Department of Transportation
100 North Senate Avenue, Indianapolis, In 46204
jnagle@indot.state.in.us
(317) 232-5464

Carl Tuttle
Operation Support Engineer
Operation Support Division, Indiana Department of Transportation
100 North Senate Avenue, Indianapolis, In 46204
ctuttle@indot.state.in.us
(317) 233-4726

Length of Paper: 6370 words

October 22, 2001

ABSTRACT

In recent years, the use of GPS technologies has expanded to perform traffic data collection for transportation studies such as work zone studies. To generate reliable results from the data acquired using GPS devices, it is necessary to investigate the factors such as sample size requirements that may have an impact on a specific study and to establish a consistent method for data collection. Based on the discussions and the real life GPS data presented in this paper, it is confirmed that the Institute of Transportation Engineers' Manual of Transportation Engineering Studies usually underestimates the sample sizes for travel-time and delay studies. However, the hybrid method developed by Quiroga et al. overestimates the sample sizes. A modified equation is presented to estimate the minimum sample sizes for field data collection using GPS devices. Travel speed may be more stable and can be easily measured for travel-time and delay studies. Stopped delay varies considerably at intersections and the sample sizes depend to a large extent on the permitted errors. Work zone layout and construction activities will create variations in vehicle flow within the work zone. To estimate the sample size requirements, it is advisable to use the standard deviation to measure the data dispersion and a minimum of three initial test runs is required. For GPS devices with sufficient accuracy, it usually requires five to ten samples for travel-time and delay studies and work zone studies. For stopped delay studies, it may require a large sample of up to 30 test runs.

INTRODUCTION

In recent years, applications using the global positioning system (GPS) have been introduced into transportation studies and the GPS devices have evolved to provide the basic traffic data needs for various transportation studies such as work zone studies (1), congestion management studies (2) and car following analyses (3). With differential capability, a GPS device can provide its users with real time position data accurate to within sub-meter under good conditions. Based on these accurate GPS spatial and temporal data, transportation engineers can easily calculate the basic traffic data such as travel time, travel speed and travel distance, and then determine travel delay and vehicle trajectories. Therefore, GPS data collection is one of the most important tasks for transportation engineers to perform transportation studies using GPS devices. To produce reliable traffic information using GPS devices, it is of significance to investigate the factors such as sample size requirements that may have an impact on a certain study and to establish a consistent method for data collection.

Over the past decades, the Institute of Transportation Engineers (ITE) has provided technical guidance for transportation engineers to conduct field studies (4, 5). As an illustration, the Manual of Transportation Engineering Studies presents a method to determine the sample size requirements for travel-time and delay studies. It has been reported, however, that the ITE method contains systematic numerical errors and may result in lower than expected sample sizes (2). To the authors' knowledge, inconsistencies may also arise if GPS devices are used to collect field traffic data by following the ITE method. This is because no matter what field study it is, the basic data collected using GPS devices will not change and the fundamentals for estimating the sample size requirements should remain consistent. This paper addresses the sample size issues based on the statistical fundamentals and presents a modified equation for estimating sample size requirements for traffic data collection. Field studies were conducted using a GPS device with differential capability switched on and the data were examined so as to verify the results of theoretical analysis.

CURRENT METHODS FOR ESTIMATING SAMPLE SIZE REQUIREMENTS

To provide guidance for transportation engineers to carry out field transportation studies such as travel-time and delay studies and intersection studies, the ITE manual provides various methods for estimating the minimum sample sizes (5). For example, the minimum sample sizes for travel-time and delay studies along a roadway segment are determined using the following equation

$$R = \frac{\sum A}{N-1} \tag{1}$$

where R = average range in running speed, ΣA = sum of calculated speed difference, and N = number of the test runs that are available before data collection starts. It is required by the manual to perform 2 to 4 initial test runs so as to calculate the average range in running speed.

After the average range, R is determined, the minimum sample size with a specific level of confidence can be found according to a permitted error. As the average range increases and the permitted error decreases, the required sample size grows. It should be pointed out that while the ITE manual requires use of the running speed to compute the average range, it might be more reasonable to use the travel speed rather than the running speed in Eq. 1. This is because the real world data demonstrate that in many cases, the running speed fluctuates more significantly than the travel speed, especially when the study route is over one-mile long. In addition, it is much easier to measure the travel speed rather than the running speed during the field studies and transportation engineers need not to choose a predesignated speed to define running time.

Quiroga et al. have also examined the above method and concluded that the ITE method contains systematic numerical errors and may seriously underestimate the required sample sizes (2). To correct the ITE method, Quiroga et al. provided a so-called hybrid method for estimating the required sample sizes as follows:

$$n = \left[\frac{t_\alpha \overline{R}}{d\varepsilon} \right]^2 \tag{2}$$

where α = significance level, t_α = t-value from two-tailed t distribution with n-1 degrees of freedom for a confidence level of 1-α, and ε = user-selected allowable error in the estimate of the mean speed, d = ratio of $|\overline{R}|$ to σ (standard deviation of the population), and $\overline{R}$ = sample range based on available data, which is computed using the equation below

$$\overline{R} = \max_{i=1}^{m} v_i - \min_{i=1}^{m} v_i \tag{3}$$

where m = sample size available, and v_i = ith speed observation of the initial study.

POTENTIAL ERRORS IN DETERMINATION OF MINIMUM SAMPLE SIZES

Determination of the sample size requirements is one of the typical statistical problems. To estimate the population mean using the sample mean, the sample size should satisfy the following relationship (6)

$$N = \left(\frac{\overline{X} \times CV \times t_{\alpha/2}}{\overline{X} - \mu} \right)^2 \tag{4}$$

where N = required sample size, $\overline{X}$ = sample mean (estimator), μ = population mean (true value), CV = coefficient of variation in the sample, α = significance level, and $t_{\alpha/2}$ = t-value of the two-tailed t distribution with a confidence level of 1-α.

Letting $\sigma = \overline{X} \times CV$ and $\varepsilon = \overline{X} - \mu$, the above equation can be reduced into the following equation

$$N = \left(\frac{\sigma t_{\alpha/2}}{\varepsilon} \right)^2 \quad (5)$$

where σ = population standard deviation, and ε = permitted error. To the authors' knowledge, Eq. 5 serves as the foundation for estimating sample sizes. As an illustration, Eq. 2 becomes Eq. 5 when d is replaced by $|\overline{R}| / \sigma$.

It is illustrated that in Eq. 5, determination of the sample sizes involves estimating the value of $t_{\alpha/2}$ and the standard deviation of the population, σ. Consequently, two types of errors may be involved in the resulting sample sizes. First, estimating the $t_{\alpha/2}$ value requires knowledge of the number of the degrees of freedom that equals (N-1) and is unknown before data collection. It is a common practice to use the Z-value of the standard normal distribution, $Z_{\alpha/2}$, to approximate $t_{\alpha/2}$. Figure 1 shows the discrepancies between t-value and Z-value for 90% and 95% confidences, respectively. It is observed that the Z-value is independent of sample size and nevertheless, the t-value varies with the sample size. A specific t-value is always greater than the corresponding Z-value. The largest discrepancy occurs when the number of the degrees of freedom, (N-1) equals one, i.e., the sample size, N equals two. As the sample size grows to three, the discrepancy drops significantly. It is also observed that after the sample size is approximately greater than five, the discrepancy becomes negligible. These observations can be extended to conclude that two samples are always not sufficient for travel-time and delay studies and the sample sizes estimated by the ITE manual may be small. To estimate the required sample sizes, a minimum of three initial test runs should be performed so as to reduce the potential errors in estimation of the t-value.

The second type of the potential errors may arise associated with estimating the population standard deviation, σ. The standard deviation is a measure of dispersion or variability. Likewise, the range is also used to measure the dispersion. In statistics, the range is defined below

$$R = \text{Max}(X) - \text{Min}(X) \quad (6)$$

where R = range, i.e., the difference between the largest and smallest values, and X = set of data.

The range is the simplest measure of the dispersion. Once the largest and smallest values are picked, the range can be readily computed. Since the range depends only on two values and ignores the other values, it may not provide a full picture of the dispersion and may overestimate the dispersion because of the presence of an unusually extreme value. As a result, the range is considered to be the least satisfactory of all measures of dispersion. Unfortunately, the hybrid method developed by Quiroga et al. uses the range and may result in an estimate of the population standard deviation with the least satisfaction.

MODIFIED EQUATIONS FOR ESTIMATING MINIMUM SAMPLE SIZES

A large sample is always anticipated to provide more information about the population. As the sample size increases, the data will become more representative of the real world conditions, and analysts will be more confident about the results. On the other hand, the cost

for data collection will increase in terms of time and manpower. Therefore, determination of the minimum sample size is also a trade-off between the required accuracy and the potential cost. In addition, to obtain a reasonable sample size, transportation engineers should also choose an appropriate permitted error and a significance level. Selection of the permitted error depends on the study purpose and the capability of the equipment. With respect to significance levels, transportation engineers should consider the consequence of committing the two types of errors in hypothesis test. For many real life problems, however, the consequence of favoring alternative hypothesis (the so-called Type II error) is more serious and it is advisable to pick a large value of α. The widely used α value is 0.05.

As discussed previously, the use of t-value causes considerable problems mathematically since the t-value depends on the sample size and iteration must be performed to search the possible sample size. As a common practice, the Z-value is employed in place of the t-value. While this treatment results in errors and usually generates lower values than the t-value does, these errors are predictable as shown in Figure 2. In Figure 2, the required sample sizes are computed using both t-value and Z-value for a selected level of confidence, 90%, 95% and 99%, respectively. It is likely that the errors are independent of the sample size and vary only with the selected confidence level. As the confidence increases, the error grows slightly. Based on the numerical results, it is recommended that the minimum sample sizes should be determined using the modified equation below

$$N = \left(\frac{\sigma Z_{\alpha/2}}{\varepsilon} \right)^2 + \varepsilon_N \tag{7}$$

where ε_N = sample size adjustment and the other variables are as defined earlier. The ε_N and $Z_{\alpha/2}$ values are given in Table 1.

CASE STUDIES

Travel-Time and Delay Studies

Travel-time and delay studies were conducted on SR26 in downtown Lafayette, Indiana. The route is about 3.2 km long with eleven signals. A GPS device with differential capability switched on was installed on a car to collect spatial and temporal data. Dispersions of travel time and speed were computed as the range (Eq. 6), the average range (Eq. 1) and the sample standard deviation. Figure 3 shows the travel speeds and running speeds measured in each test run. It is demonstrated that running speeds varies more significantly than travel speeds. This agrees with the finding obtained earlier, which states that the travel speed may be a more stable measure than the running speed. Therefore, there is no guarantee that use of the running speed can generate better results.

Figure 4 shows the calculated measures of the dispersions of travel time and speed. It is observed that the range is the largest value of all measures and experienced significant rise in both travel time and speed because of an unusual value. Notice that this significant rise is probably independent of the number of available data and depends to a large extent on the presence of an unusual value. When only two samples are used, the average range equals the

range. The average range and sample standard deviation vary with a similar trend and become very stable as the number of initial test runs increases. There exist discrepancies between the average range and the standard deviation. These observations can be used to confirm the findings presented in the preceding sections, i.e. use of the range may generate results with the least satisfaction and the hybrid method developed by Quiroga et al. will generate greater than expected sample sizes. In any case, the minimum sample size should be greater than two. It is also recommended that the sample standard deviation should be used to measure the data dispersion while estimating the minimum sample sizes.

Figure 5 presents the minimum sample sizes computed using the ITE method, the hybrid method and the modified equation, respectively. It is demonstrated that the hybrid method produces the largest sample sizes and the ITE method the smallest sample sizes for travel-time and delay studies. The minimum sample size obtained using the hybrid method solely depends on the extreme values and is not stable as the test runs increase. While, the hybrid method employs an iteration algorithm to calculate the sample size requirement with respect to the actual t-value instead of Z-value, such an effort is reduced since the hybrid method uses the range to estimate the dispersion, and consequently, the sample sizes may be overestimated. The minimum sample size produced by the ITE method approaches two as the number of initial test runs grows. As discussed earlier, the largest discrepancy occurs between t-value and Z-value when the sample size equals two. Therefore, the ITE method may underestimate the minimum sample sizes. While the minimum sample size produced by the modified equation may be close to that produced by the hybrid method, it drops and becomes predictable as the number of initial test runs increases. It is likely that five to ten samples may generate reliable results for travel-time and delay studies.

The length of the test route will also affect the sample size requirements. It is illustrated that in Figure 6 (a), the dispersion of the travel speed decreases and the travel speed becomes more stable as the route length increases. A short study route may experience great data dispersion and usually requires a large sample. As illustrated in Figure 6 (b), the minimum sample sizes decrease with an increase in the length of the study route. When the length of the study route is less than one mile, the minimum sample sizes grow considerably. This confirms the recommendation by the ITE manual that requires a minimum one-mile (1.6 kilometers) long study route. Also, this may be used to explain why the minimum sample sizes obtained by the hybrid method (2) are very large for the three 0.2-mile (0.32 kilometers) segments.

Stopped Delay Studies

Currently, most stopped delay studies are performed manually or using electronic counting boards. Use of the GPS devices for stopped delay studies depends on the rate of GPS data output. Although some GPS devices can be enhanced to output data at a fast rate of 10 Hz (10 times per second), an output rate of 1 Hz (once per second) should be fast enough to satisfy the requirement of accuracy. With a GPS device at an output data rate of 1 Hz, transportation engineers can record the vehicle position every second and then establish an accurate vehicle trajectory and measure the stopped delay by adding the time when the test car is standing still.

This study measured the stopped delay at the intersection of SR26 with Creasey Lane in Downtown Lafayette. The stopped delay varied considerably, ranging from 0 to 69 seconds during the peak hours. Figure 7(a) shows the mean value and dispersion of the measured stopped delay data. The bold solid line indicates the variation of the mean value with the sample size. As the sample size grows, the average stopped delay becomes stable. Again, the range gives the greatest estimate of the dispersion. While both the range and average range fluctuate significantly, the sample standard deviation varies smoothly, especially after the sample size exceeds five. Notice that the range also appears to be a constant when the sample size is greater than five samples. However, it may rise significantly if another extremely large value is experienced.

The permitted errors also affect the sample size requirements remarkably, as demonstrated in Figure 7(b). Selection of the permitted errors for stopped delay studies using GPS devices is still under investigation and will be discussed elsewhere. For a permitted error of 5 seconds per vehicle, the minimum sample sizes are not stable and may be very large. It appears that a 5-second permitted error is very strict and will incur high study costs. For example, it usually takes about five minutes to complete one test run, depending on the traffic and road conditions. For a 15-minute period, only three to five test runs can be completed. If the permitted error of five seconds is used, it may take up to ten to thirty days to complete the data collection under similar traffic conditions. When a permitted error of 10 seconds per vehicle is used, the sample sizes drop significantly and become more predictable. If the permitted error drops to 15 seconds per vehicle, the minimum sample size is reduced to less than 30. However, such a permitted error may not satisfy the requirements of accuracy for level-of-service analysis of controlled intersections.

Work Zone Studies

GPS technology is a promising tool for work zone studies because GPS data can be used to establish an accurate vehicle trajectory. With GPS devices, it is possible for transportation engineers to investigate the acceleration and deceleration features of the vehicle driving through a work zone, evaluate the geometric layout of the work zone, and identify potential safety problems. This study measured GPS spatial and temporal data in three work zones. The first work zone is a 6-kilometers long, partial closure work zone on I74. The second work zone is a crossover work zone with a length of about 10 kilometer on I69. The third work zone is an 11-kilometers work zone on I65. This work zone is a mixed work zone with a partial closure layout at some locations and a crossover layout at other locations.

Figure 8(a) shows three vehicle-trajectories recorded using GPS devices within the three work zones, respectively. It is demonstrated that the test car traveled very steadily within the work zone on I69. Within the work zone on I65, the test car slowed down at the location of about 7200 meters due to the vehicles merging into the work zone from the on-ramp. Turbulence also occurred to the test car on I74 because of the effects of construction activities, especially construction vehicles. Turbulence within a work zone will result in large variations in vehicle movement, leading to a large sample requirement. Figure 8(b) shows the minimum sample sizes obtained using the modified equation. For a work zone with minor effect of construction activities, four or five test runs may produce reliable information. If turbulence occurs very often, for example, when the traffic volumes are high on the on- and

off-ramps within the work zone or when construction vehicles use or encroach upon the traffic lanes frequently, it may require a large sample of about ten test runs.

CONCLUSIONS

This paper examines the current methods for estimating sample size requirements presented in the ITE manual and discusses the issues raised in some studies associated with the use of the ITE method. Also, this paper provides a modified equation for estimating the sample size requirements for GPS traffic data collection. Based on the discussion in the preceding sections, several conclusions have been drawn as follows:

The errors associated with determining the sample size requirements arise mainly from estimating the t-value and the data dispersion. Use of the Z-value from the standard normal distribution rather than the t-value from the t distribution will underestimate the sample size requirements by two, three and four samples for a confidence level of 90%, 95% and 99%, respectively. To avoid significant errors, three initial test runs should be at least performed to estimate the minimum sample sizes. The range may overestimate the data dispersion because of the possible presence of extreme values and result in greater than expected sample sizes. It is advisable to use the sample standard deviation, rather than the range or the average range, to measure the data dispersion.

For travel-time delay and delay studies, it is demonstrated by the field GPS data that the ITE method underestimates the sample size requirements and the hybrid method overestimates the sample size requirements. This is probably because the ITE method uses the Z-value to replace the t-value and the hybrid method uses the range that is the least satisfactory of the dispersion measures. Travel speed may be more stable than running speed and can be easily measured. The length of study route will also affect the dispersion of traffic data. As the route length increases, travel speed and travel time become more stable and the required sample size decreases. Usually, five to ten samples can yield reliable results for travel-time and delay studies.

A GPS device with an output data rate of 1 Hz is suitable for stopped delay studies. The range may generate an extremely large sample because of the presence of an unusual value. It is recommended that five initial test runs should be performed so as to determine the required sample sizes. The permitted errors also affect the sample size requirements remarkably. It appears that a permitted error of 10 seconds per vehicle may produce a reasonable sample size taking into account the accuracy requirement and cost incurred. Because the stopped delay varies considerably, a large sample is often required for stopped delay studies.

The sample size requirements for work zone studies depend on the effect of construction activities, traffic volumes on off-and on-ramps, work zone layout and the length of work zone. If the merging and diverging traffic volumes are high and the construction vehicles use traffic lanes within a work zone, traffic flow can be easily disturbed and travel speed may experience considerable variations. As a result, a large sample may be required for work zone studies. In general, five to ten test runs may provide sufficient data for work zone studies.

Acknowledgement

This study was conducted as a part of the Joint Transportation Research Program (JTRP) research project SPR-2392. The authors acknowledge the support of the Federal Highway Administration and the Indiana Department of Transportation. Sincere thanks are extended to the Study Advisor Committee members, Kirk Mangold, Steve Smith, Dennis Lee and Andy Fitzgerald for their expertise and technical guidance in the process of performing this study.

Disclaimer

The contents of this paper reflect the views of the authors, who are responsible for the facts and the accuracy of the data presented herein. The contents do not necessarily reflect the official views or policies of the Indiana Department of Transportation or the Federal Highway Administration. This paper does not constitute a standard, specification or regulation.

References

1. Yi Jiang and Shuo Li. Measuring and Analyzing Vehicle Position and Speed Data at Work Zones Using Global Positioning System. Accepted for Publication in the *Journal of Institute of Transportation Engineers*, 2001.
2. Quiroga, C. A., and Darcy, B. *Development of CMS Monitoring Procedures*. Louisiana Transportation Research Center, Louisiana, 1998.
3. Wolshon, B., and Y. Hatipkarasulu. Results of Car Following Analyses Using Global Positioning System. *Journal of Transportation Engineering*, ASCE, Vol. 126, No. 4, July/August 2000, pp. 324-331.
4. Institute of Transportation Engineers. *Manual of Traffic Engineering Studies*, 4th Edition, Washington DC, 1976.
5. Institute of Transportation Engineers. *Manual of Transportation Engineering Studies*, Washington DC, 2000.
6. Shuo Li. *Reliability Theories and Algorithms in Road Engineering (in Chinese)*. Shaanxi Science and Technology Publishing Corporation, China, 1990.

List of tables and figures

TABLE 1 Values of ε_N and $Z_{\alpha/2}$ for Determining Sample Sizes

Level of Confidence	ε_N	$Z_{\alpha/2}$
90%	2	1.64
95%	3	1.96
99%	4	2.58

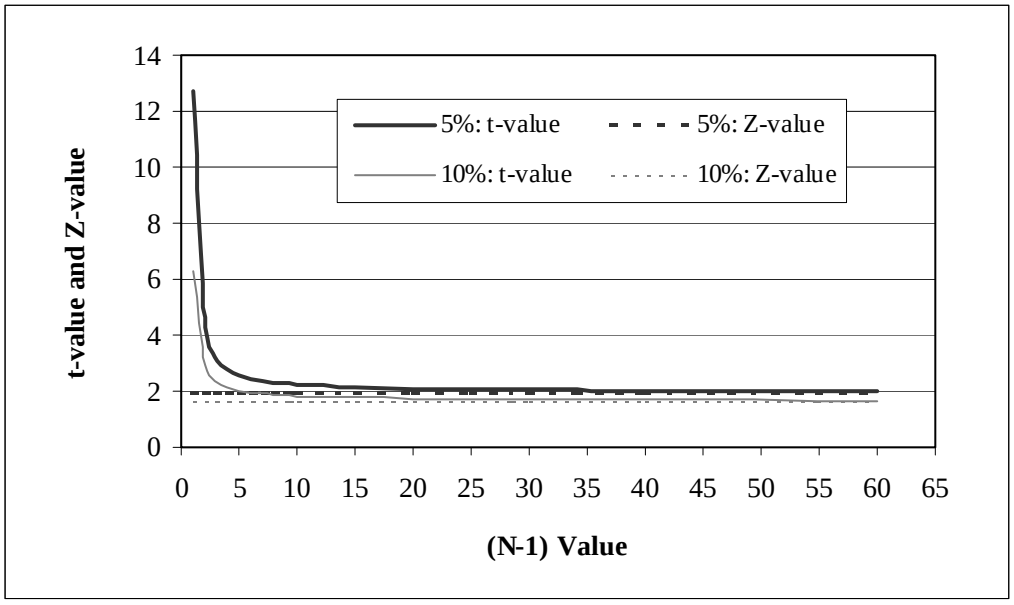


FIGURE 1 Variations of t-value and Z-value with sample sizes

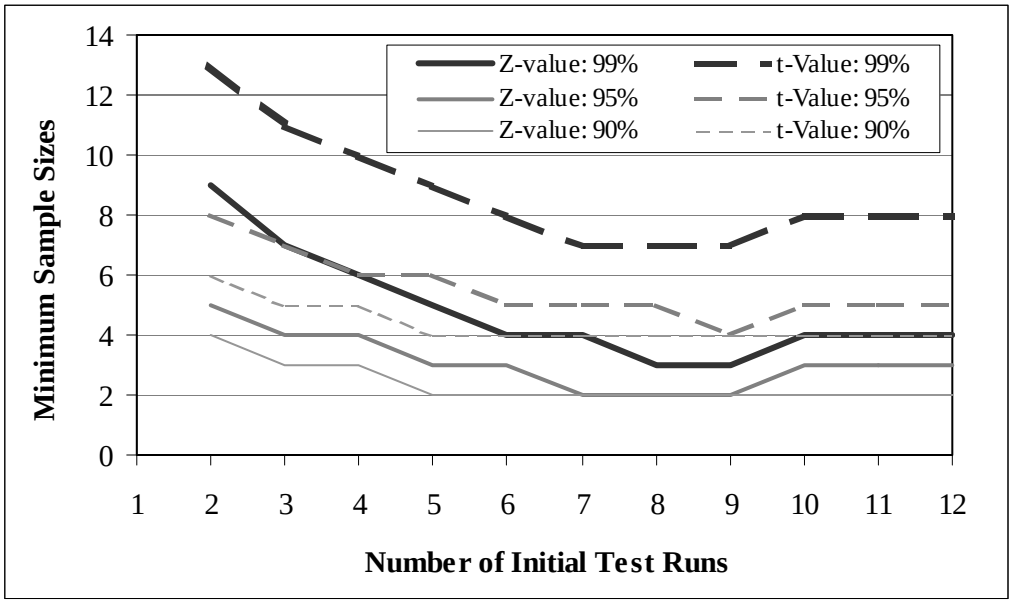


FIGURE 2 Sample sizes computed using t-value and Z-value

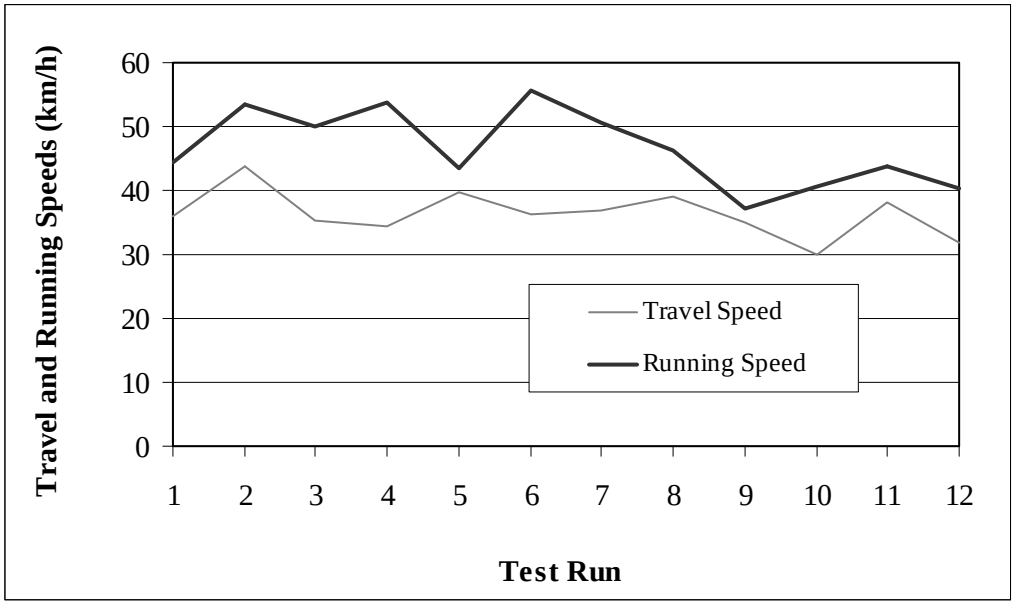
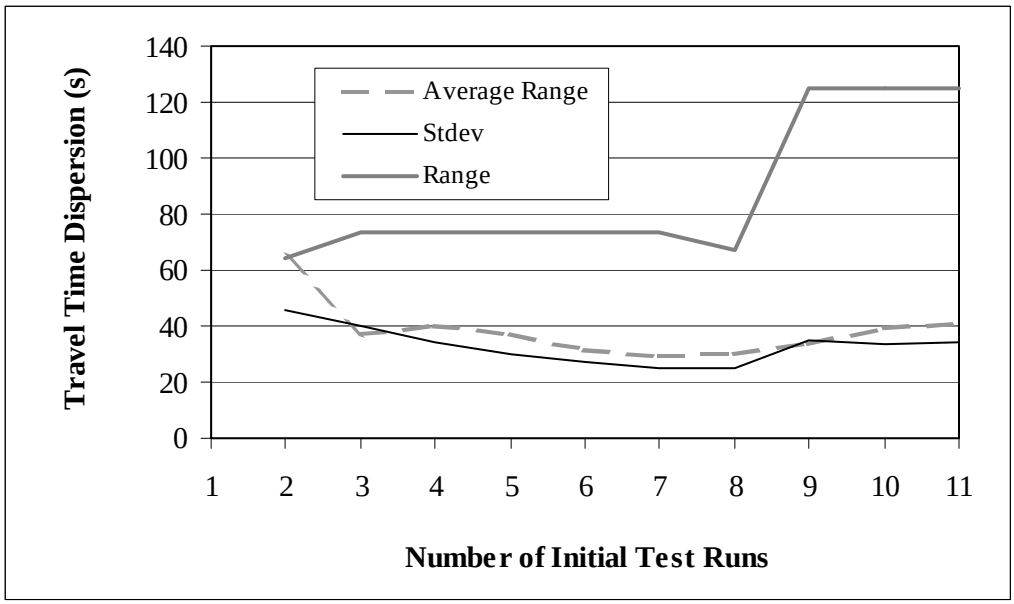
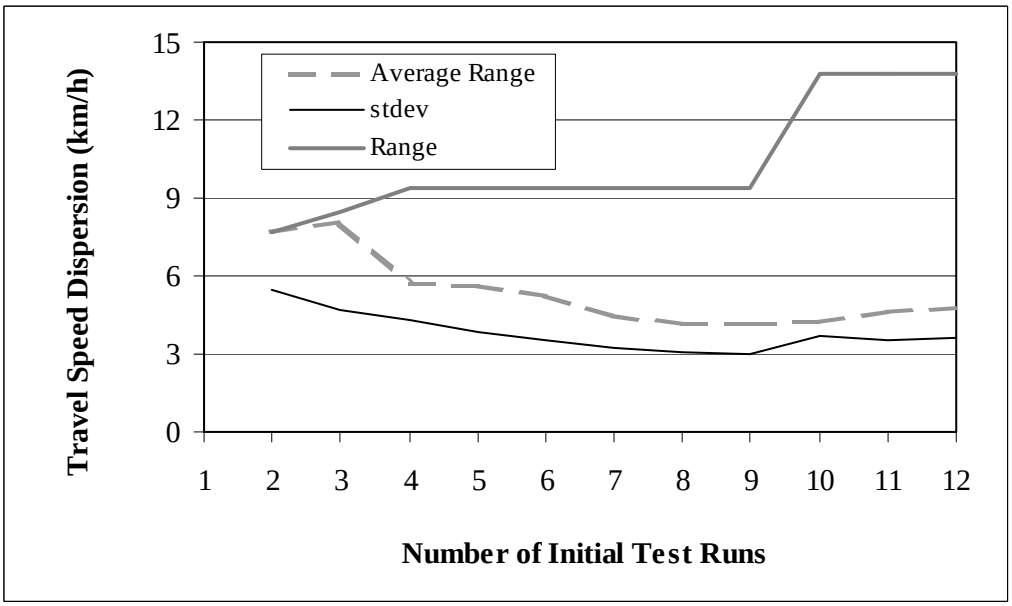


FIGURE 3 Variations of travel and running speeds



(4.a) Measures of travel time dispersion



(4.b) Measures of travel speed dispersion

FIGURE 4 Dispersions of travel time and speed

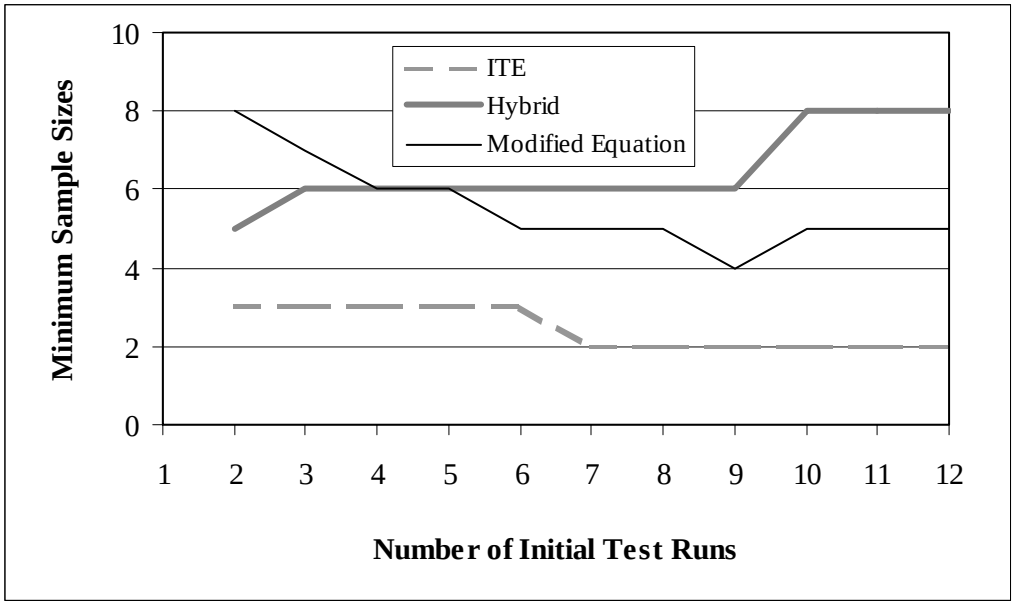
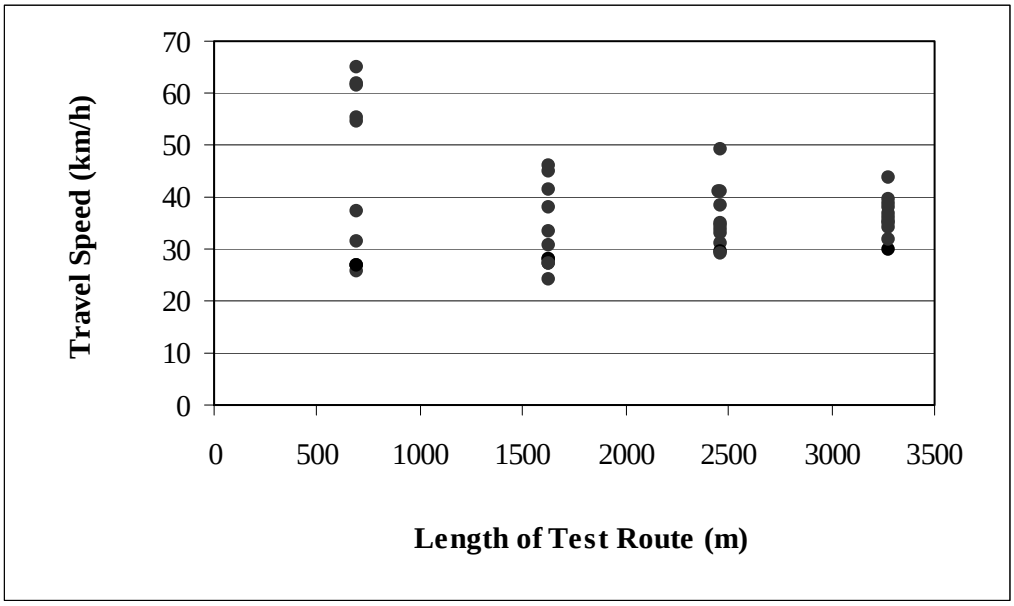
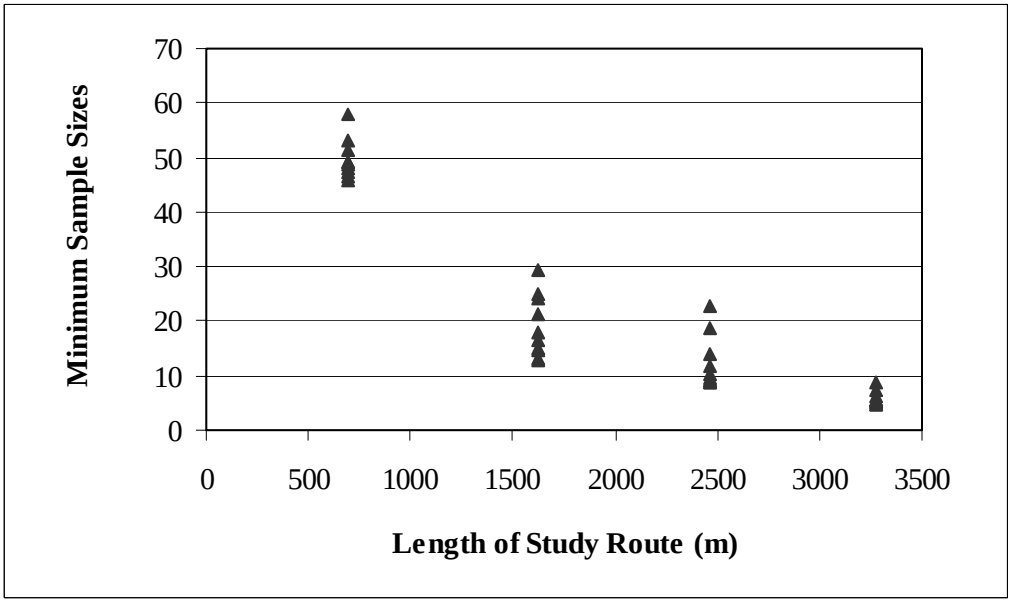


FIGURE 5 Sample sizes computed using various methods

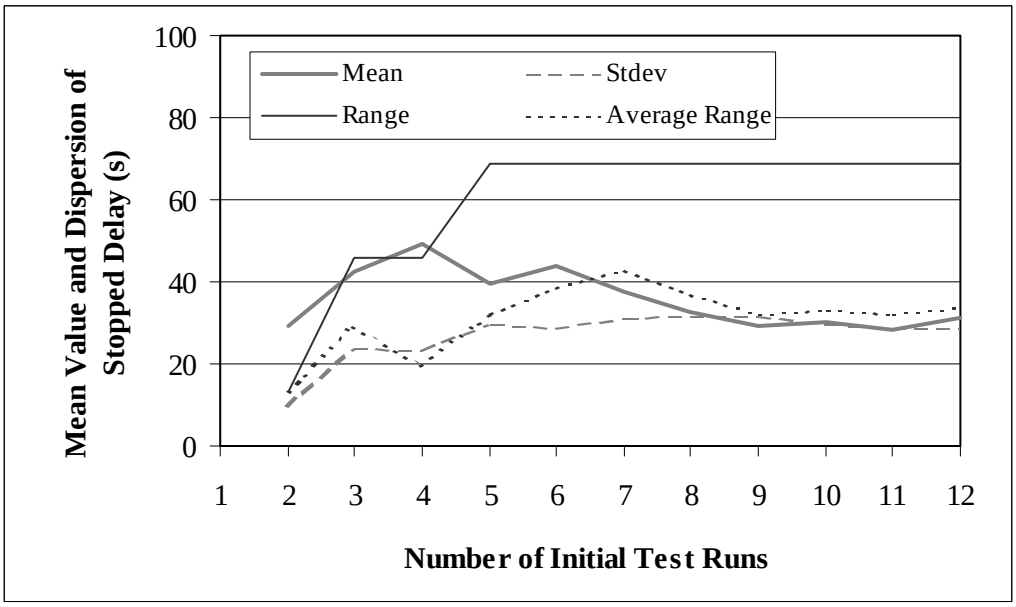


(6.a) Variations of travel speed with route length

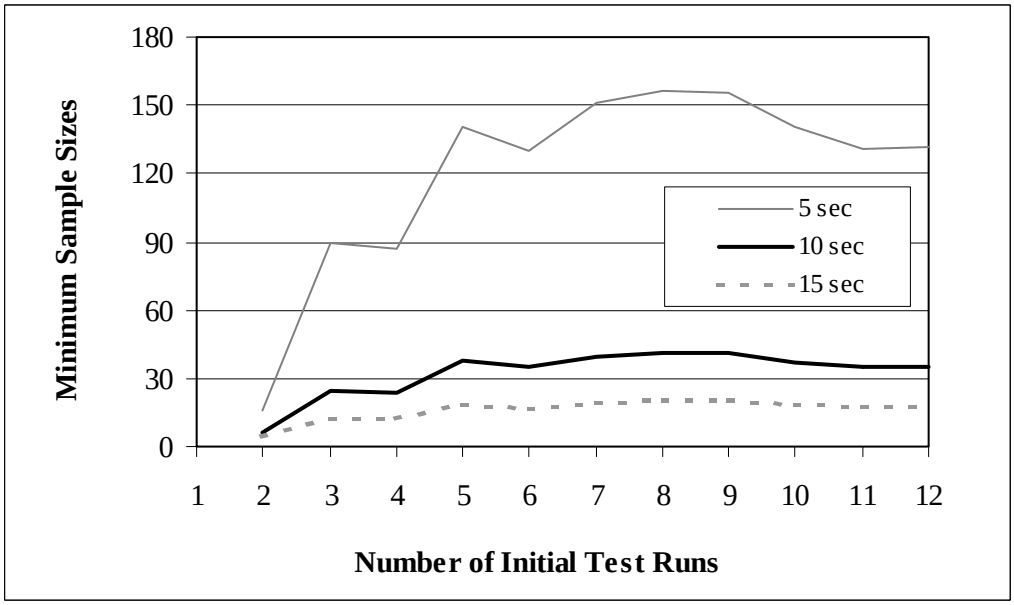


(6.b) Variations of minimum sizes with route lengths

FIGURE 6 Effect of the length of study route on sample sizes

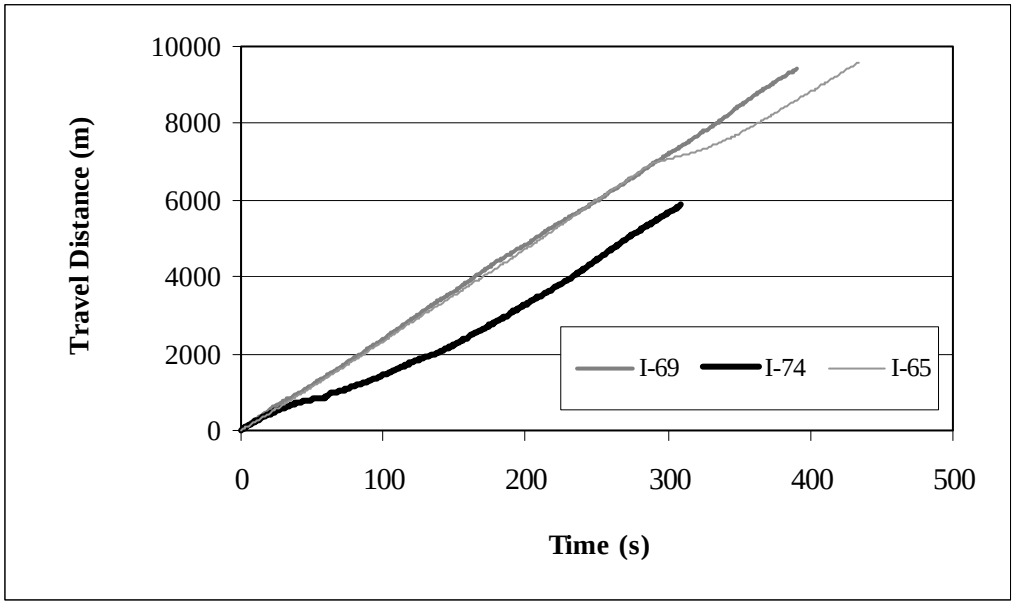


(7.a) Average and dispersion of stopped delay

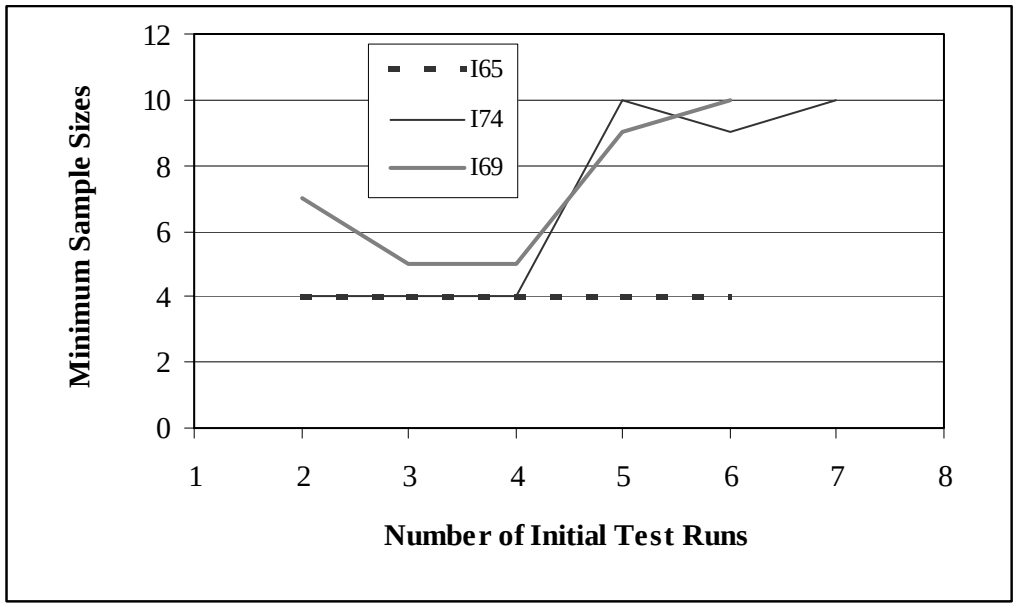


(7.b) Variations of minimum sample sizes with permitted errors

FIGURE 7 Measured stopped delay and minimum sample sizes



(8.a) Vehicle trajectories



(8.b) Minimum sample sizes on I74

FIGURE 8 Trajectories and minimum sample sizes within work zones