NISS

Frequency of Probe Vehicle Reports and Variance of Arterial Link Travel Time Estimates

Ashish Sen, Piyushimita Thakuriah,
Xioquon Zhu, and Alan F. Karr

Technical Report Number 54
October, 1996

National Institute of Statistical Sciences
19 T. W. Alexander Drive
PO Box 14006
Research Triangle Park, NC 27709-4006
www.niss.org

Frequency of Probe Vehicle Reports and Variance of Arterial Link Travel Time Estimates*

Ashish Sen, Piyushimita (Vonu) Thakuriah, Xioquon Zhu and Alan Karr

Abstract

Because of dependence among travel time observations on signalized arterials, the estimation of variances of key statistics is not straightforward. The variance of the estimate of expected (mean) travel times obtained from n probe vehicles for the same link over a fixed time period may be shown to be of the form $a + b/n$ where a and b are link-specific parameters. Using data from a set of arterials, it is shown that a is positive for well traveled signalized links. This implies that the variance of the estimate does not go to zero with increasing n. Consequences of this fact for probe-based system are explored. While the results presented are for a specific set of links, we argue that because of the nature of the underlying travel time process, the broad conclusions would hold for most links with signal control.

1 Introduction

One of the primary functions of an Advanced Traveler Information System (ATIS) is route guidance. Route guidance may be provided to participating vehicles in real-time where the vehicle receives information based on current and predicted travel time conditions. Route guidance may also be provided on an autonomous basis, where the vehicle has access to historical information of travel times. In both cases, the statistical quality of the travel time estimates are a crucial determinant of the quality of route information that the driver finally receives from the ATIS (Thakuriah and Sen, 1995).

In a probe-based ATIS, the information that is collected on network conditions may be on travel times prevailing on links of the network or on the route that a monitoring vehicle used in order to travel from an origin to a destination. In order to supply vehicles requesting route guidance with the necessary information, a short-term travel time prediction problem is involved — that is, a prediction of the travel time on a link at the clock time that the vehicle receiving guidance is expected to traverse that link. This, in turn, involves making suitable travel time estimates for the relevant time periods.

There has been, thus far, virtually no large-scale [area-wide] travel time data collection efforts on signalized arterials in dense urban networks and consequently, little empirical analysis of these processes. The nature of travel time processes on signalized arterials are affected by a large number of factors, which include network volumes, traffic control factors such as signal timings and progressions factors and so on. Temporal and spatial dependence in these processes are critical determinants of the quality of the estimates that one would obtain from estimation and forecasting procedures.

Two such dependencies are important to consider in the travel time estimation procedure of any route guidance system where link travel times are the primary data generated: (i) the level of dependence

*This work was partially supported by the National Science Foundation's Division of Mathematical Sciences Grant Numbers 9313013 and 9208758. The authors would also like to thank the ADVANCE project.

between travel time realizations on the same link and (ii) the level of dependence between realizations on consecutive links that form routes. Regarding the second kind of dependence, volumes on contiguous links are correlated because, within a discrete time slice, many of the vehicles on them are in fact the same. While ignoring this correlation would be convenient in obtaining an estimate of route travel time, it would also be inappropriate. Factors such as progression and platoon-formation would exacerbate this dependence. And if link volumes and hence travel times on links along a route were to be independent, route travel times would merely be the sum of independent link travel times. The Law of Large Numbers would then reduce the variance in travel times for long routes to the point where there would be very little need for guidance, except under situations of non-recurrent congestion!

While we have commented briefly on the second issue of dependence, our present analysis is motivated by the first issue — that travel time observations on a link are not independent. Ignoring dependence among travel time observations will lead to incorrect estimates of the variances of travel time estimates, which are necessary to compute in order to have an idea of precision of estimates. Since the computation of the necessary sample size results directly from the desired precision of the estimate, it is critical that the variance be correctly computed. In this context, we may quote Wilson and Halferty (1929): '... that reliance on such formula as $\sigma/\sqrt{n}$ is not scientifically satisfactory in practice, even for estimating unreliability of means'.

The key conclusions that this paper reaches from an analysis of the relationships among travel times on the same link are:

- that high levels of deployment may not offer substantial amounts of improvement in the quality of guidance because the marginal improvement in the precision of link travel time estimates drop off after only a few observations per discrete time interval are obtained and
- that the variances of the estimates under high levels of congestion never go to zero and one must take explicit cognizance of this fact in constructing route guidance strategies.

In our analysis, we have used travel time data on signalized and unsignalized arterials collected by suitably equipped vehicles from ADVANCE, a large-scale ATIS project in suburban Chicago (Boyce, *et al.*, 1994). In the ADVANCE project, the transmission of information in both directions — vehicle to central computer [called the Traffic Information Center or TIC] and TIC to vehicle is by radio frequency [RF]. Probes collect and transmit several kinds of data, the key one being *link travel time,* which is the time the vehicle took to traverse each link it traveled over (other variables are described in Section 2). For each link and for each (5-minute) time period, mean link travel times were estimated based on data gathered by different probe vehicles and these estimates act as building blocks for most of the route guidance supplied. If a link has detectors, then data from detectors and from probe vehicles are used to compute link travel time estimates. For links without detectors [most links are not detectorized], estimates of mean travel times from probe data only, are supplied to vehicles as 'current travel times.' Forecasts of link travel times made 5, 10 and 15 minutes into the future are also computed. Five-minute estimates are used as inputs in the formulæ used for constructing these forecasts. The on-board computer in each equipped vehicle computes routes based on current or forecasted travel times, as appropriate.

Therefore, the key statistic in this paper is that of the estimate of mean link travel time. The question that naturally arises, then, is: given that there is dependence in the data, how many probes traversals do we need on a link per unit time in order to get reasonably precise estimates of the mean? This issue is of some importance for probe-based ATIS systems. Unless the market penetration of such systems are moderately high, the number of probe traversals per unit time on a link would usually be low and the estimate of the mean would be poor in the sense it would have fairly high variance. Also certain links might not get covered by probes at all. On the other hand, if this information is poor and consequently, the resultant route guidance not good, market penetration would be small, as the information may be of little value to users of the system.

The estimation of the variance of the estimated mean link travel time is a fairly simple matter if observations are independent. On the other hand, if the observations are not independent, then one needs to estimate covariances. The lack of independence could occur due to several reasons, of which, the following are especially noteworthy: (i) similarity of entry times from upstream links; (for example, a vehicle following another would tend to have similar link travel time as the leader vehicle; vehicles close together in a platoon would have similar travel times) (ii) similarity of cycle phase encountered (under otherwise similar conditions, two vehicles arriving at a traffic signal ten seconds after the onset of the red phase will have similar link travel times). [Clearly similarity of travel times occur in other ways too; two such examples are similarity of turning-movement executed to enter the link from an upstream link or of the turning movement executed in order to depart from the link]. Thus, not only would one conjecture that link travel times are correlated but that covariances are functions of headways and are variable.

In this paper, we explore the effects of such correlations on the estimates of mean travel times by explicitly taking such dependence into account. We use the ADVANCE data, described in Section 2 for this purpose. In Sections 3 and 4, we explore the implications of this dependence on the variability of estimates and on the number of observations required to make precise estimates. We present the estimates and the necessary analysis in Section 5. We present some implications of the analysis and our conclusions in the context of probe-based ATIS in Sections 6 and 7.

2 The Data

The data used for the entire analysis were collected as part of the evaluation of the ADVANCE project during the summer of 1995 and are from a suburban area of Chicago. The link travel time data were transmitted in real-time from ADVANCE probe vehicles that drove down two pre-specified routes that consisted of a total of 12 links. All links in the route (except 2) are signalized by means of demand-actuated strategies. The purpose of driving these vehicles over a small number of links was to simulate fairly high levels [1–2 percent] of deployment of such vehicles, using very few equipped vehicles, by concentrating them on these links. The probes were released at the beginning of the route at clock times that allowed the formation of randomized headways. Paid drivers were used for this effort.

The major variables on which data were collected are (i) link ID (a link is identified by turning-movements at exits so that each one-directional road segment may be common to three ADVANCE links) (ii) link travel time (in seconds) (iii) congested time (in seconds) or the amount of time spent on a link during which the vehicle traveled at or below 2 meters per second (iv) congested distance (in meters) or the distance on the link covered at speeds less than 10 meters per second.

We have used only link travel times in this analysis, although the other variables have offered much insights into the nature of travel time processes in signalized arterials. We would like to point out from one such secondary analysis that the data indicate that travel times actually *decrease* on the average on some links analyzed, as volumes increase during peak periods, because the progression effects are excellent during the evening peak period in the direction of heavy traffic in this area. This happens because although speeds drop during the peak period with heavy traffic, stopped delay due to traffic signals (as indicated by the congested time variable) also decreases substantially during peak periods and the effect of decrease in stopped delay is stronger than the effects of speed reduction.

We have detailed data on signalization (recorded as events over time) for a few of these links and for a limited number of days. However, we have not made use of these data in the present analysis, with the understanding that most actual or even concept large-scale ATIS do not obtain data on this, albeit important, variable.

While these data were collected from 1:00 p.m. to 8:00 p.m. on Mondays through Thursdays for 10 weeks, we have used observations from a total of 14 days, between June 6, 1995 and July 11, 1995, with several days around July 4 removed, in the analysis presented in this paper. The data were screened for incidents and unusual observations based on detailed logs that drivers kept on pre-printed forms.

In the present analysis, means of travel times were estimated for pre-determined 5-minute intervals on each day. This leads to an estimate based on an unequal number of observations per time period. Our analysis depends on having an unequal number of observations in each interval. We present a representative breakdown of the number of observations within each five-minute interval in Table 1, for a selection of time intervals of a representative day.

3 Variance of Estimated Mean Link Travel Times

The variance of the estimated mean $\bar{x} = n^{-1}\sum_{i=1}^{n} x_i$ of probe reports, $x_1, x_2, \ldots x_n$, for the same link over some interval is

$$\begin{aligned} \text{var}\,[\bar{x}] &= \text{E}[\bar{x} - \text{E}(\bar{x})]^2 = \text{E}[n^{-1}\sum_{i=1}^{n}(x_i - \text{E}[x_i])]^2 \\ &= n^{-2}\,\text{E}[\sum_{i=1}^{n}(x_i - \text{E}[x_i])^2 + \sum_{\substack{i,j \\ i\neq j}}(x_i - \text{E}[x_i])(x_j - \text{E}[x_j])] \\ &= n^{-2}\sum_{i=1}^{n}\ \text{var}\,[x_i] + \sum_{\substack{i,j \\ i\neq j}}\ \text{Cov}\,[x_i, x_j] \end{aligned} \tag{1}$$

where Cov $[\cdot]$ stands for covariance. Let $\eta = n^{-1}\sum_{i=1}^{n}$ var $[x_i]$ be the average variance of x_i's and let

$$\nu = [n(n-1)]^{-1}\sum_{\substack{i,j \\ i\neq j}}\ \text{Cov}\,[x_i, x_j]$$

be the average covariance of all pairs (x_i, x_j). Then, the expression in equation (1) may be written as

$$\begin{aligned} \text{var}\,[\bar{x}] &= n^{-2}[n\eta + n(n-1)\nu] = \eta n^{-1} + [1 - n^{-1}]\nu \\ &= \nu + n^{-1}[\eta - \nu]. \end{aligned} \tag{2}$$

Of course, if the covariances were zeros, then $\nu = 0$, and we would get var $[\bar{x}] = n^{-1}\eta$, and if all variances of x_i were equal to σ^2, then we would get the familiar var $[\bar{x}] = n^{-1}\sigma^2$. In this case as $n \to \infty$, var $[\bar{x}] \to 0$, as is very well known. On the other hand, if the correlation between every pair of x_i's were one and all variances were equal, then for all i and j, Cov $[x_i, x_j] =$ var $[x_i]$ and $\nu = \eta$. Then var $[\bar{x}] = \nu$ and then var $[\bar{x}] = \nu$ stays the same no matter what n is. This is to be expected, because then all the observations x_i would be exactly the same. In practice we would expect something in between, where $\eta > \nu > 0$. Even then as $n \to \infty$, var $[\bar{x}] \to 0$. While our discussion has been in terms of the mean, a similar situation would hold for most reasonable estimators.

The implication of these facts is that if $\nu > 0$, then no matter how many link travel times of distinct vehicles we measure over a short time interval, the variance of the mean would be large. Clearly, how large is a critical question and can be answered by examining values of ν's. In fact, even if we computed the mean of all vehicles during a given time interval, the variance of the mean would remain above the value of ν for that link.

This might appear counter-intuitive, since when a sample becomes the same as the population one might expect the variance of the sample mean to go to zero. This does not happen here because each x_i is assumed to be one out of a continuum of [and hence an infinite number of] random variables. This is appropriate for forecasting applications, because it is not nearly as important for us to know the exact mean travel

Interval	Clock	Date	Link	Link	Link
No.	Time		1	9	11
1	13:00:00–13:04:59	June 12	0	0	0
2	13:05:00–13:09:59	June 12	4	0	0
3	13:10:00–13:14:59	June 12	0	0	0
4	13:15:00–13:19:59	June 12	1	4	3
5	13:20:00–13:24:59	June 12	5	1	0
6	13:25:00–13:29:59	June 12	1	0	0
7	13:30:00–13:34:59	June 12	1	5	6
8	13:35:00–13:39:59	June 12	3	0	0
9	13:40:00–13:44:59	June 12	4	3	0
10	13:45:00–13:49:59	June 12	0	4	5
11	13:50:00–13:54:59	June 12	1	1	3
12	13:55:00–13:59:59	June 12	3	3	0
13	14:00:00–14:04:59	June 12	1	3	5
14	14:05:00–14:09:59	June 12	3	2	3
15	14:10:00–14:14:59	June 12	3	4	3
16	14:15:00–14:19:59	June 12	2	1	2
17	14:20:00–14:24:59	June 12	3	2	2
18	14:25:00–14:29:59	June 12	1	0	1
19	14:30:00–14:34:59	June 12	1	0	0
20	14:35:00–14:39:59	June 12	0	1	1
21	14:40:00–14:44:59	June 12	1	2	3
22	14:45:00–14:49:59	June 12	2	3	2
23	14:50:00–14:54:59	June 12	2	5	2
24	14:55:00–14:59:59	June 12	3	0	1
25	15:00:00–15:04:59	June 12	3	3	3
26	15:05:00–15:09:59	June 12	3	2	2
27	15:10:00–15:14:59	June 12	1	3	2
28	15:15:00–15:19:59	June 12	3	4	4
29	15:20:00–15:24:59	June 12	4	0	1
30	15:25:00–15:29:59	June 12	1	3	0
31	15:30:00–15:34:59	June 12	2	2	4
32	15:35:00–15:39:59	June 12	3	0	1

Table 1: Probe Traversals by 5-minute intervals on a representative day of data-collection.

time for a given set of vehicles during some interval in the past, as it is to know what travel times will be in the future, if similar conditions persist and, for this purpose, sets of travel times need to be considered as samples from an infinite population.

4 Test of Hypothesis of No Correlation

If the x_i's were uncorrelated, the average covariance ν would be zero. Therefore, all we need to do is test if $\nu = 0$. One method of testing the hypothesis $H : \nu = 0$ against the alternative $A : \nu \neq 0$ is afforded by equation (2) itself. If we have reasonable estimates of $\text{var}[\bar{x}]$, we could regress this against the corresponding values of n^{-1}, where n is the number of observations used to compute $\bar{x}$. The intercept term would then be an estimate of ν and could be used to test H against A.

One estimate of $\text{var}[\bar{x}]$ is $[\bar{x} - \widehat{\text{E}[\bar{x}]}]^2$, where $\widehat{\text{E}[\bar{x}]}$ is an estimate of $\text{E}[\bar{x}]$. In order to estimate $\widehat{\text{E}[\bar{x}]}$, we used the model

$$\text{E}[x_{d,t}] = \gamma + \alpha_d + \beta_t, \tag{3}$$

where α_d and β_t are respectively day effects and time-of-day effects. The model given in equation (3) is implied by

$$\text{E}[x_{d,t,i}] = \gamma + \alpha_d + \beta_t \tag{4}$$

where $x_{d,t,i}$ is the ith observation during day d and time-period t. The model in (4) can be estimated by least squares, after coding the independent variables corresponding to α_d's and β_t's as indicator [or dummy] variables [one indicator variable for each time interval t and one for each day d] with the restrictions $\sum_d \alpha_d = 0$ and $\sum_t \beta_t = 0$, in order to avoid multicollinearity. The residuals $e_{d,t,i}$ obtained from the model in (4) are

$$e_{d,t,i} = x_{d,t,i} - \hat{\gamma} - \hat{\alpha}_d - \hat{\beta}_t \tag{5}$$

where $\hat{\alpha}_d$ denotes an estimate of the parameter α_d. For every day and time period, the mean over all i of these residuals is, therefore,

$$\bar{e}_{d,t} = \bar{x}_{d,t} - \hat{\gamma} - \hat{\alpha}_d - \hat{\beta}_t. \tag{6}$$

Thus, if the model in (3) holds, a reasonable estimate of $[\bar{x} - \text{E}[\bar{x}]]$ is $\bar{e}_{d,t}$, and $[\bar{e}_{d,t}]^2$ estimates $\text{var}[\bar{x}]$. Therefore, the model estimated is

$$[\bar{e}_{d,t}]^2 = \nu + \gamma_1[1/n_{d,t}] + \epsilon_{d,t} \tag{7}$$

where the $\eta = \nu + \gamma_1$ and $\epsilon_{d,t}$ is the error term relating to the tth time interval on day d. These parameters may be estimated by some regression procedure. Thus a test of the hypothesis H against A could be conducted by using least squares to carry out the regression and then testing $\nu = 0$ in the usual way using the t-statistic.

However, before conducting the estimation, we need to make sure that the underlying assumptions of least squares are at least approximately met. Figures 1 and 2 show plots of the dependent variable $y_{d,t} = [\bar{e}_{d,t}]^2$ against the independent variable n^{-1} for links 1 and 11 during the peak period. It would appear that we have a wedge or funnel-shaped pattern of points indicating the presence of heteroscedasticity or unequal variance violating the assumption that $\text{E}[\epsilon^2_{d,t}] = \sigma^2$.

One solution to this problem is to weight the regression. In order to find appropriate weights, ignoring subscripts for the moment, we write the dependent variable as $y = u^2$ where $u = \bar{e}$, with $\bar{e}$ as in (7). For any differentiable function $f(u)$, the variance $\text{var}[f(u)]$ of $f(u)$, may be approximately written as $\text{var}[f(u)] \approx [f'(z)]^2 \text{ var}[u]$, where a prime denotes a derivative and z is the mean of u. Since $f(u) = u^2$ here, and an estimate of $\text{var}[u]$ is y, we get $\text{var}[f(u)] \approx 4\,\text{E}[y]^2 \propto \text{E}[y]^2$. Thus a proper weight would appear to be the reciprocal of the square of an estimate of $E[y]$. In our examination of diagnostics from the different regressions weighted as above, no heteroscedasticity was noticeable.

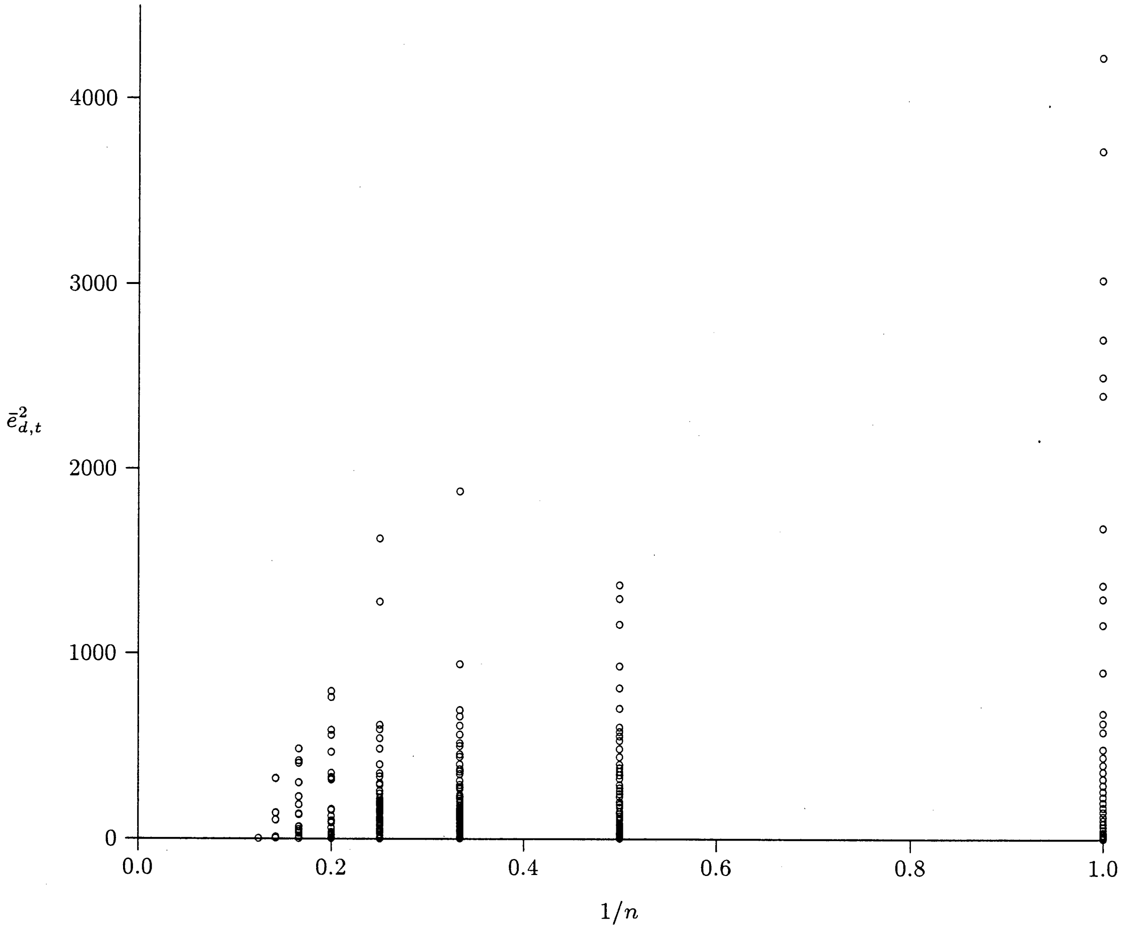

Figure 1: Plot of $[\bar{e}_{d,t}]^2$ against $1/n$ for Link 1 during peak

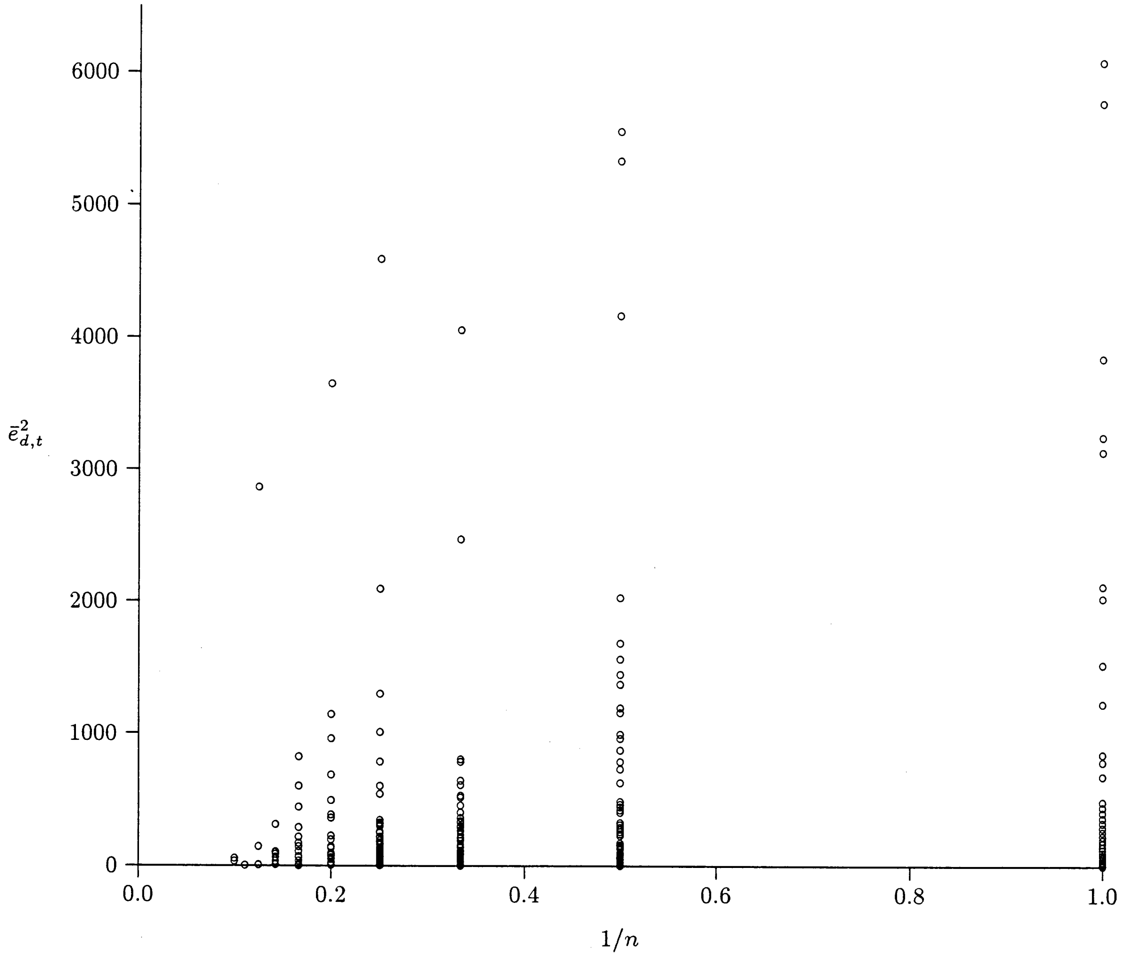

Figure 2: Plot of $[\bar{e}_{d,t}]^2$ against $1/n$ for Link 11 during peak

link		no. of							
id	mean	observations	$\hat{\nu}$	s.e.	t	$\widehat{\eta - \nu}$	s.e.	t	$\hat{\eta}$
1	69.73	1233	102.77	26.04	3.95	246.88	73.17	3.37	349.65
2	59.29	1242	93.36	40.06	2.33	518.30	119.67	4.33	611.66
31	64.90	886	228.16	71.25	3.20	553.98	171.28	3.23	782.14
32	39.14	308	113.69	44.56	2.55	55.95	79.93	0.70	169.64
4	114.47	320	440.99	162.43	2.72	298.12	236.11	1.26	739.11
5	38.20	327	36.55	61.15	0.60	44.97	87.53	0.51	81.70
6	53.63	321	100.13	58.03	1.73	124.56	121.29	1.03	224.69
7	196.3	309	966.11	248.69	3.89	24.26	332.95	0.07	990.37
8	85.36	297	30.50	94.02	0.32	804.41	209.61	3.89	834.91
9	167.98	355	1082.86	228.85	4.73	172.97	325.73	0.53	1255.83
10	81.38	1253	102.98	50.18	2.05	479.67	91.35	5.26	582.66
11	47.55	1275	122.25	45.96	2.67	279.56	82.75	3.38	401.91
12	88.83	1260	250.41	46.20	5.42	249.33	82.94	3.01	499.74

Table 2: Estimates from the model in (7) for Peak Period

An alternative to weighted regression in this case would be logarithmic transformation, since by the same formula as above, $\text{var}\,[\log[y]] = 2\ \text{var}\,[\log[u]] \approx [z^{-1}]^2\ \text{var}\,[u] = \text{E}[y]^{-1}\,\text{E}[y] = 1$. However, we felt it was simpler to weight the regression.

Notice that under the hypothesis of no correlation, the dependent variable values are correlated only slightly — the only correlation would be that due to the parameters γ, α_d and β_t being common to different $\bar{e}_{d,t}$'s. The model given in (2), is based on a mathematical identity. Therefore, there would appear to be no significant violations to the underlying assumptions of least squares *under the hypothesis*, save for the presence of outliers. The presence of outliers would only tend to increase rather than decrease standard errors. If there is any slight heteroscedasticity left, that too, by violating the minimum variance property of least squares estimates, would only tend to raise standard errors. Thus the test we conducted [using t-values to examine the size of the intercept] is an appropriate test, albeit perhaps lacking somewhat in power, that is, we would err on the side of accepting rather than rejecting H. Notice further that given the large sample sizes, it is appropriate to assume that the estimate of the intercept is approximately normally distributed [Sen and Srivastava, (1990), Ch. 5], although the dependent variable values might not be — and are, in fact, highly skewed.

5 Estimation of Parameters

Column 6 in Tables 2 and 3 show the t-values for ν for various links for the peak and non-peak periods respectively. The two tables also give $\hat{\nu}$, $\widehat{\eta - \nu}$, and their standard errors and $\hat{\eta}$ along with the total sample size and $\bar{x}$, or the mean travel times for each link that was analyzed. A description of the data has been given in Section 3 and the time interval over which means were computed was 5 minutes. It is easily seen that all estimates of ν are positive and are significant at the 5 percent level or better, with the exception of links 5, 6 and 8 in the peak period and 2, 5 and 6 in the off-peak period.

The situation in Link 5 is easily explainable; it is a very lightly traveled link with no traffic control. On this link, a vehicle rarely affects the travel time of another vehicle. Link 6 is also very lightly traveled but has a stop sign. Link 8 is more difficult to explain, particularly because the lack of significance occurs only for the peak. However, a partial explanation is afforded by the fact that all the probe vehicles that were the source of the data, entered the link via a right turn — some making the turn on red. However, because traffic during the peak period on the major westbound route is heavy, most vehicles would execute the turn on green, and would, therefore, often encounter sparse traffic until they reached the end of the link. This is particularly the case during the peak when progression was excellent.

With the exception of the links and time periods mentioned above, the t-values were always large indicating that the hypothesis of no correlation and of $\nu = 0$ may be rejected for them.

The least squares procedure we used above to test for lack of correlations is also a reasonable method for estimating η and ν. There are both advantages and disadvantages to such an approach. A key advantage is that least squares estimates are mean-like [for example, if we estimated a measure of location or central tendency of $x_1, x_2, \ldots, x_n$ by least squares, the estimate would be the mean $\bar{x}$]. This has benefits when we are dealing with means elsewhere in the analysis and also has theoretical advantages. Also, it is consistent with the idea of average [that is, mean] variances and covariances. It is particularly for this last reason that we used least squares estimates of ν and $\eta - \nu$ in this work.

Figures 1 and 2 show that in spite of the fact that observations under incident conditions have been removed, such plots can still be very messy. This is partly because each dependent variable value is an estimate of the variance of $\bar{x}$ based on a single observation on $\bar{x}$. Thus the distribution of the dependent variable would be akin to a scaled version of a [non-central] chi-square distribution which is, of course, a highly skewed distribution. The messy appearance of the data might suggest a more robust regression. There are many such. The one we like consists of partitioning the data points in plots like the ones in Figures 1 and 2 by values of n and then computing medians of the dependent variable values for each such partition. Then a line can be fitted by eye to points representing these medians. An alternative is to minimize the sum of absolute values of errors, which is sometimes also called L^1 regression, and M-estimation (Andrews, 1974; see also Montgomery and Peck, 1982, Chapter 9).

One advantage of treating average covariances and variances as parameters in a regression model is that we do not have to directly measure them. Direct measurement would be complicated because, owing to traffic signals, the mean link travel time at any instant would be very difficult to obtain without using additional information, such as signal timing.

Since, as mentioned earlier, signal control themselves could be major contributors to correlation, it might have been desirable to include a term reflecting them in the model given by (3). However, as mentioned in Section 2, since information on signal control are unlikely to be readily available in ATIS of the near future, we decided not to use such a model at this time. It is well known that if a variable is left out of a regression model, parameter estimates can become biased. Thus, leaving out a variable reflecting traffic

link id	mean	no. of observations	$\hat{\nu}$	s.e.	t	$\widehat{\eta - \nu}$	s.e.	t	$\hat{\eta}$
1	69.86	1746	56.98	20.98	2.72	287.83	63.74	4.52	344.81
2	47.40	1728	31.23	28.56	1.09	615.80	88.62	6.95	647.03
31	60.43	627	205.93	64.70	3.18	412.19	113.45	3.63	618.12
32	31.56	1028	39.61	19.15	2.07	140.42	53.38	2.63	180.03
4	93.08	1048	346.53	74.06	4.68	258.09	126.69	2.04	604.62
5	36.29	1063	20.81	47.22	0.44	56.24	94.46	0.60	77.05
6	44.78	1059	11.15	6.24	1.79	57.62	17.18	3.35	68.77
7	96.22	1068	516.91	91.08	5.68	494.69	177.27	2.79	1011.60
8	58.59	961	77.73	29.42	2.64	163.45	48.51	3.37	241.17
9	65.72	1042	277.22	58.40	4.75	282.80	114.66	2.47	561.00
10	58.86	1680	203.63	53.27	3.82	623.81	89.47	6.97	827.44
11	50.98	1712	148.54	43.27	3.43	546.22	75.20	7.26	694.76
12	75.18	1588	159.53	31.35	5.09	170.94	55.32	3.09	330.47

Table 3: Estimates from the model in (7) for Non-peak Period

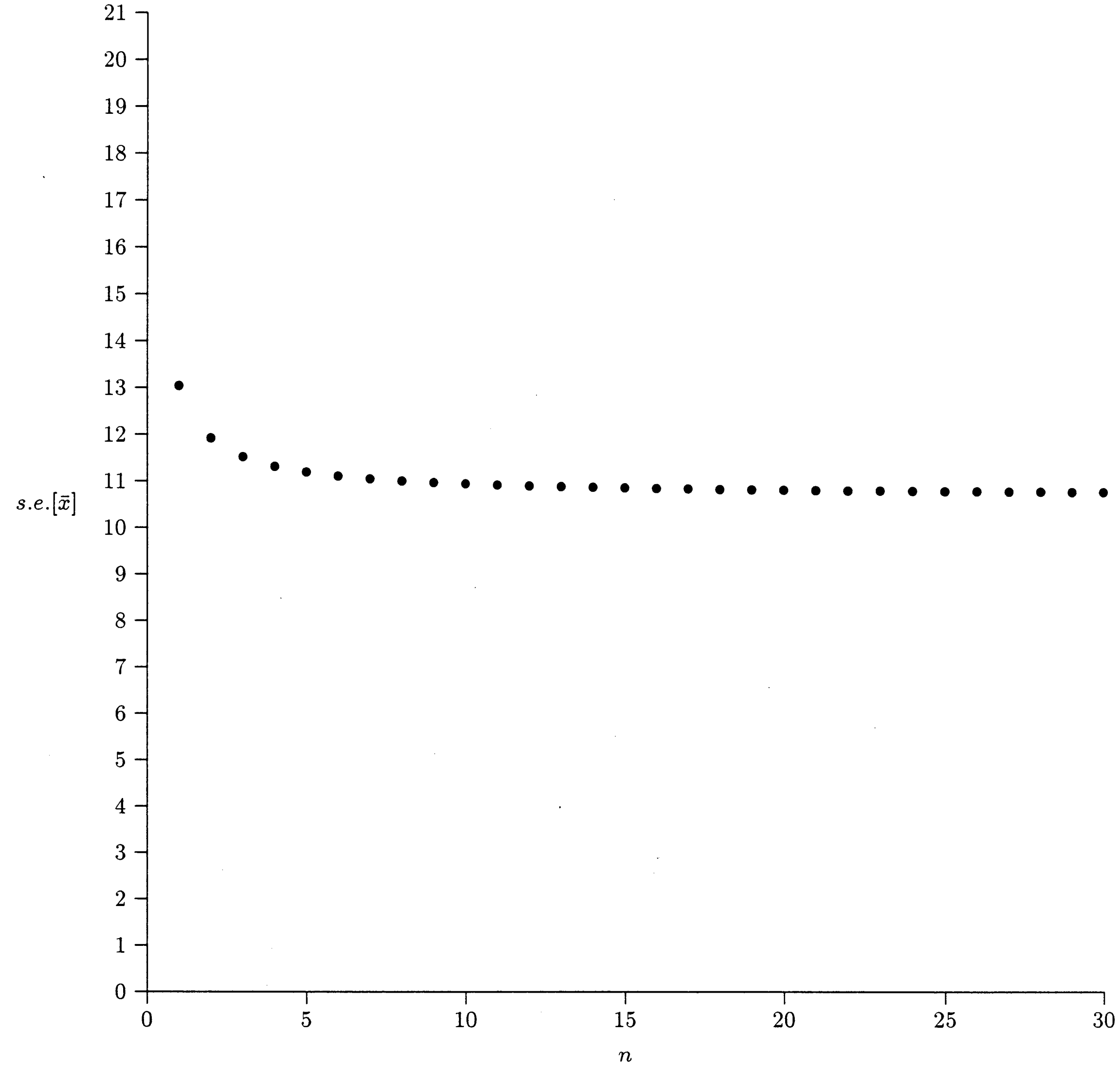

Figure 3: Relationship between standard error $s.e.[\bar{x}]$ and frequency n of probes on Link 32 during peak period estimated from model $[\bar{e}_{d,t}]^2 = \nu + \gamma_1[1/n_{d,t}] + \epsilon_{d,t}$.

signal timings can bias the estimates of the parameters in the model in (4), particularly the ones on time of day effects. An examination of the formulæ involved [Sen and Srivastava, 1990, p. 235 *et seq.*] reveals that, if such a bias exists, it is sometimes positive and sometimes negative following the periodicity of the signals. The effects on estimates of ν and η we conjecture would be minimal.

For the purpose of this paper, the main point is that the intercept be a positive number and this follows from our test in Section 4 and also from the derivation of the model given in (2), coupled with the fact that travel time observations for pairs of vehicles separated by small headways are correlated.

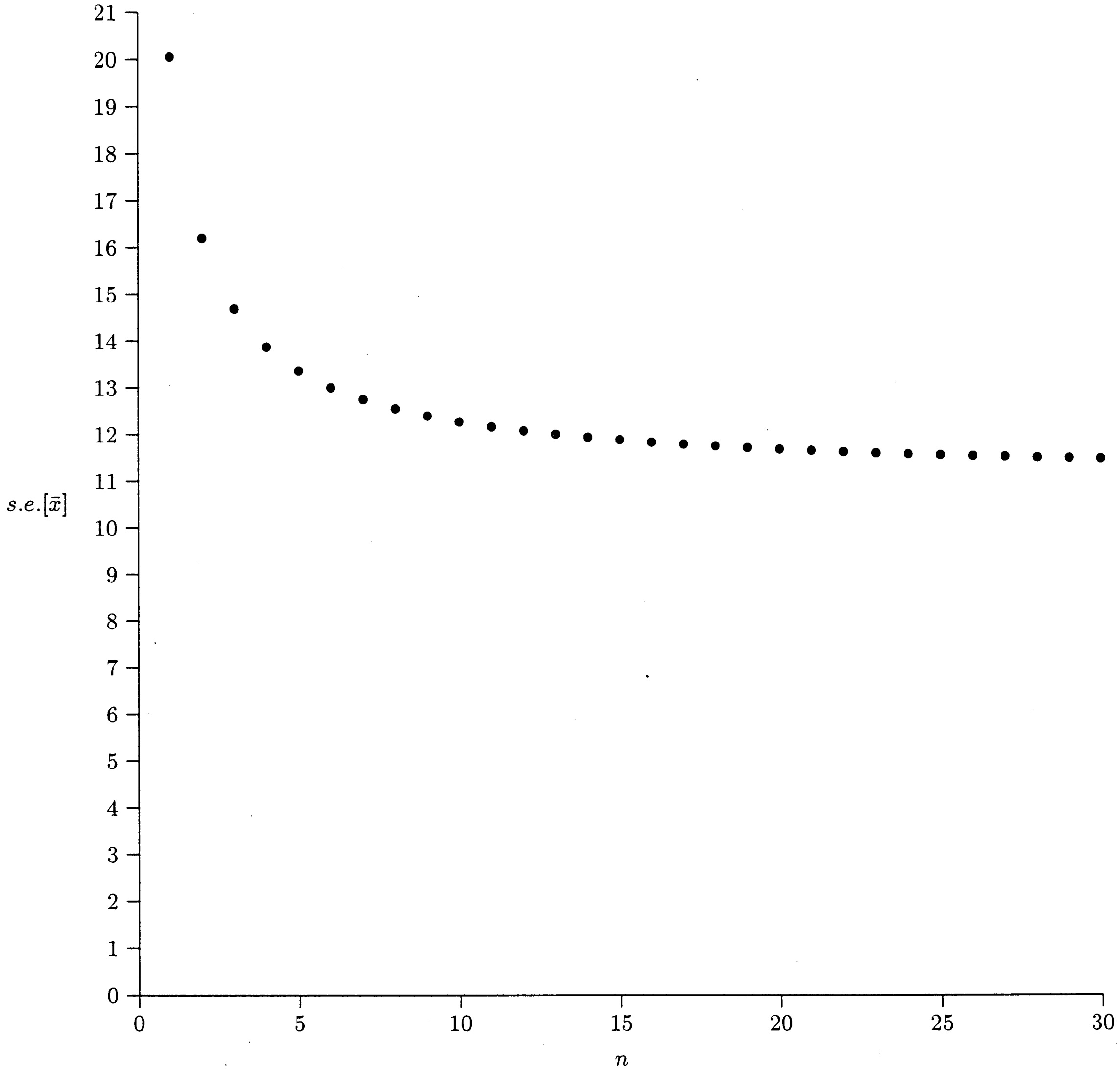

Figure 4: Relationship between standard error $s.e.[\bar{x}]$ and frequency n of probes on Link 11 during peak period estimated from model $[\bar{e}_{d,t}]^2 = \nu + \gamma_1[1/n_{d,t}] + \epsilon_{d,t}$.

6 Implications of Dependence

Although, we have examined only a small number of links, we would expect the model in (2), with $\nu > 0$, to be true for signalized links with dense traffic, in general. In this section, we draw some implications of this result for probe-based ATIS. Figures 3 and 4 show the relationship between the standard error of mean link travel times *computed using the model given in (2) and parameter values from Table 2*, for links 32 and 11, against n, the number of probes during a five-minute interval. While figures comparable to Figures 3 and 4 would vary from link to link, we would expect these figures to illustrate curves of generally the same shape provided the links were well traveled and signalized.

Notice from Figure 3, the plot for link 32 during the peak period, that the points flatten out rapidly with increasing n. For $n \to \infty$, the standard errors would approach $\hat{\nu} = 10.66$. Therefore, if we are willing to accept a roughly 10 percent increase in standard errors, we could get by with two observations and if we accept a 20 per cent increase in standard error, even a single observation could suffice.

The rate of flattening out depends on the relative sizes of $\hat{\nu}$ and $\hat{\eta} - \hat{\nu}$ and we chose Link 32 to illustrate a case where $\hat{\eta}$ was somewhat close to $\hat{\nu}$. Consider now the case of link 11 for the peak period [Figure 4]. Here the ratio of $\widehat{\eta - \nu}$ to $\hat{\nu}$ is considerably larger than for Link 32. For link 11, then, a 10 percent increase in standard errors is achieved with about 11 observations, and a 20 percent increase with 5 observations.

Usually highly used links have high correlations between travel times of vehicles. Therefore, under situations of high congestion the sample sizes required would be smaller. Under recurrent congestion, these are the situations where ATIS would have much impact, since under low congestion, travel times would tend to be closer to historic averages (suitably conditioned on day-of-week and time-of-day considerations). And under high congestion levels, because volumes would be higher, even low deployment rates would usually achieve reasonable probe frequencies. For example, suppose we have moderately high congestion level represented by 1500 vehicles per hour for a two lane arterial or 125 vehicles for a five minute period. Then a 1 percent deployment level would get us 1 vehicle per five minutes on the average and a 5 per cent deployment would get us 6 vehicles.

Two rather important conclusions emerge from this :

- The variance of the estimated mean link travel time remains quite far from zero no matter how many probes use the link.
- After a certain number of probes per unit time, additional probes do not decrease the variance of this estimate very much.

High levels of deployment would of course be necessary to cover and monitor a wider area of the network. However, the second conclusion above suggests that very high levels of market penetration by probe-based ATIS may be unnecessary in order to improve estimates of link travel times. Methods must therefore be devised which take the correct variance of link travel time estimates explicitly into account in any sample size computation for probe-based systems. Our analyses show that estimating this variance under conditions of dependence (as exhibited by probe-based travel time observations) is far from straight forward.

7 Conclusion

In order to obtain estimates of link travel times that are of reasonable quality, the quality of data from the network must also be of reasonably good. In order to estimate travel times of some desired precision, we would have to have some idea of the number of travel time realizations needed to obtain that level of precision. Therefore, the variance of the estimate plays a fundamental role. The dependence among realizations is critical to model in order to obtain a correct estimate of the variance. This paper addresses this problem and illustrates the implications of this analysis with an empirical analysis.

Clearly one consequence of the model in equation (2) is that estimated mean travel times over a short time period has a standard error which will never be smaller than a positive number. This lower bound is the average of covariances among travel times. This is not only true for data from probes but owing to the generality of the model in (2) itself, for any method of measuring travel times of vehicles [for example, video surveillance technology].

Moreover, we strongly suspect that this phenomenon holds for any reasonable estimate [not just the mean] of expected travel times. Lack of independence would affect other statistical procedures as well. Most linear estimators would be affected, including least squares estimates. In Section 5, we use (2) to test whether travel times x_i's are uncorrelated for each of a set of links for which data were available. The results show that except for extremely uncongested links, this hypothesis can usually be rejected.

Since variances of the estimates of expected travel time never go to zero, it is unlikely that deterministic models would ever be too useful in working with estimates obtained from probe-reported travel times. Even if we were to obtain travel times from every vehicle on a link, we would need carefully constructed probabilistic models in order to make good use of such data.

Another consequence of the model in (2) is that illustrated by Figures 3 and 4. A small number of probes within a five minute interval yields a standard error which is not substantially improved by making the number of probes much larger. Thus, high levels of probe deployment is not necessary in order to have a link travel time estimate of reasonably good quality, as long as all important links are covered by at least a few probes (see also Hicks, *et al.*, 1992). Since market penetration by any ATIS will at best proceed slowly, we see this as a very important fact *in favor* of probe-based systems.

References

1. Andrews, D. F. (1974). A Robust Method for Multiple Linear Regression. *Technometrics*, **16**, pp. 523-531.

2. Boyce, D. E., A. M. Kirson and J. L. Schofer (1994). ADVANCE — The Illinois Navigation and Route Guidance Demonstration Program. In *Advanced Technology for Road Transport*, Ian Catling (ed.), Artech House, Boston.

3. Thakuriah, P. and A. Sen (1995). An Investigation into the Quality of Information given by an Advanced Traveler Information System. *In Review.*

4. Hicks, J. E., D. E. Boyce and A. Sen (1992). Static Network Equilibrium Models and Analyses for the Design of Dynamic Route Guidance Systems. A Technical Report in support of the design phase of the ADVANCE project, Urban Transportation Center, University of Illinois at Chicago, Chicago, Illinois.

5. Montgomery, D. C. and E. A. Peck (1982). *Introduction to Linear Regression Analysis.* Wiley.

6. Sen, A. and M. Srivastava (1990). *Regression Analysis: Theory, Methods and Applications.* New York: Springer-Verlag.

7. Wilson, E. B. and M. M. Hilferty (1929). Note on C. S. Peirce's Experimental Discussion of the Law of Errors. In Proceedings of the National Academy of Sciences, **15**(2), pp. 120-125.