Accurate estimation of travel times from single-loop detectors*

Karl F. Petty[a], Peter Bickel[b], Jiming Jiang[c], Michael Ostland[b], John Rice[b], Ya'acov Ritov[d], Frederic Schoenberg[b]

[a] *Department of Electrical Engineering and Computer Science, University of California, Berkeley, Berkeley, California*

[b] *Department of Statistics, University of California, Berkeley, Berkeley, California*

[c] *Department of Statistics, Case Western Reserve University, Cleveland, Ohio*

[d] *Department of Statistics, The Hebrew University of Jerusalem, Jerusalem, Israel*

Abstract

As advanced traveler information systems become increasingly prevalent the importance of accurately estimating link travel times grows. Unfortunately, the predominant source of highway traffic information comes from single-trap loop detectors which do not directly measure vehicle speed. The conventional method of estimating speed, and hence travel time, from the single-trap data is to make a common vehicle length assumption and to use a resulting identity relating density, flow, and speed. Hall and Persaud (Hall and Persaud, 1989) and Pushkar, Hall, and Acha-Daza (Pushkar et al., 1994) show that these speed estimates are flawed. In this paper we present a methodology to estimate link travel times directly from the single-trap loop detector flow and occupancy data without heavy reliance on the flawed speed calculations. Our methods arise naturally from an intuitive stochastic model of traffic flow. We demonstrate by example on data collected on I-880 data (Skabardonis et al., 1994) that when the loop detector data has a fine resolution (about one second), the single-trap estimates of travel time can accurately track the true travel time through many degrees of congestion. Probe vehicle data and double-trap travel time estimates corroborate the accuracy of our methods in our examples.

1 Introduction

Accurate estimation of freeway travel times based on loop detector data is an important goal in transportation. In addition to providing a real-time measurement of congestion, automated travel time estimates provide a useful measure of throughput and could possibly be employed to detect incidents. Also, accurate prediction, a related problem, could provide valuable information for scheduling in the shipping industry. Since most existing freeways are equipped only with single-trap loop detectors, it is highly desirable to find estimation techniques using

*Funding for this research was provided in part by NSF-DMS 9313013 to NISS.

only the flow and occupancy data that these these detectors provide. In this paper we develop methods for using these inexpensive, existing detectors to obtain accurate estimates of freeway travel times. We demonstrate our methods on data from I-880 (Skabardonis et al., 1994).

Our methods are based on a simple stochastic model in which vehicles that arrive at an upstream point during a given interval of time have a common probability distribution of travel times to a downstream point. In estimating this travel time distribution, we make use of an approximate relationship between flow, occupancy, and speed. We discuss the relationship of our method to previous work based on cross-correlations (Dailey, 1993). We also suggest extensions to situations involving multiple entrances and exits and to transition periods between congestion and free flowing traffic when we would not expect our model to hold even approximately.

The I-880 data on which we demonstrate our methods is an excellent test set for several reasons. First, it contains a wide range of road and traffic conditions. Second, in addition to the flow and occupancy measurements, double-trap speed measurements are available for use as a fairly accurate standard of comparison. Finally, probe vehicle data provides a loop-independent source of travel time validation.

The organization of the paper is as follows: In Section 2 we give a brief introduction to previous work in this area. In Section 3 we build a stochastic model of freeway flow from which our estimation techniques arise naturally and discuss variations and extensions. In Section 4 we demonstrate the methods on data from the I-880 data set (Skabardonis et al., 1994). We explore the limitations of our method and compare it to related work (Dailey, 1993). The estimates resulting from our methodology compare very favorably to travel time estimates from both the double-trap speed measurements and probe vehicles.

2 Previous Work

Because of the importance of accurate travel time measurements for traffic management operations, there have been a number of studies that have attempted to determine link travel times on both freeways and arterial streets. An excellent overview of methods to determine travel time on arterial streets is given by Sisiopiku and Rouphail (Sisiopiku and Rouphail, 1994b). One common technique involves using regression analysis to relate the flow and occupancy reported by single-trap loop detectors to the travel time on the link. The details of these different approaches are given in Sisiopiku and Rouphail (Sisiopiku and Rouphail, 1994b) and the references therein, and hence we do not review them here. Another approach is based on a real time simulation of the freeway, either microscopic (Sisiopiku and Rouphail, 1994a) or macroscopic (Sanwal et al., 1996). The measurements from the existing loop detectors are used as input to the model and the link speed, and hence travel time, are taken as output. Since this approach requires a model of traffic behavior at either the microscopic or macroscopic level it is very complex, and necessarily model dependent.

On freeways, a common method of estimating travel time is to estimate the speed from the loop detectors and to convert. Since most systems have single-trap loop detectors, which only measure flow and occupancy, the speed is calculated using the relationship:

$$\text{speed} = \frac{\text{flow}}{\text{occupancy} * g} \tag{1}$$

where $1/g$ is the average effective car length—the sum of the car length and the width of the loop detector—in the correct units. The factor of g simply converts occupancy to density. Hall and Persaud (Hall and Persaud, 1989) and Pushkar, Hall, and Acha-Daza (Pushkar et al., 1994) investigated this relationship and discovered many problems. They showed that the accuracy of (1) is a function of many factors including location and weather. They also presented results suggesting that it is prone to a systematic bias with respect to occupancy. Single-trap loop detector speed estimates have consequently been shunned by the transportation research community.

Dailey (Dailey, 1993) uses cross-correlation of the flow at the upstream and downstream detectors to estimate travel time between two single-trap loop detectors placed 0.5 miles apart. In Section 3.2 we discuss the relationship of this approach to ours. However, as we will see in Section 4.3, while cross-correlation can lead to accurate estimates, without modification it is unable to track the travel time even during moderate levels of congestion. Dailey aggregates flow to five second intervals prior to cross-correlation. In Section 4.2 we explore in detail the effects of aggregation on the accuracy of the travel time estimate. Our findings show that too much aggregation can result in a loss of information, and in our examples we show it is desirable to have a small level of aggregation.

3 Methodology

The methodology for our procedures for estimating travel times between single loop detectors is based on a stochastic model, which, although over-simplified, suggests procedures whose effectiveness will be demonstrated in subsequent sections.

We consider point processes of arrivals during a time interval $[T_B, T_F]$, initially in a single lane. For simplicity we assume that the arrivals are exchangeable (thus not distinguishing say between cars and trucks, for example) and let $X(t)$ be the cumulative number of arrivals upstream and $Y(t)$ the cumulative number of arrivals downstream. Denoting the arrival times upstream by σ_i and the travel times by τ_j, where we use i to label vehicles independently of arrival order and the summation is over all possible σ_i, not just those in $[T_B, T_F]$, we have

$$dX(t) = \sum_i \delta(t - \sigma_i)dt \tag{2}$$

$$dY(t) = \sum_j \delta(t - \sigma_j - \tau_j)dt \tag{3}$$

where $\delta(\cdot)$ is Dirac's delta function. Our model postulates that conditional on σ_i (and all other upstream events), τ_i has a distribution independent of i, σ_i (and all upstream events) for $T_B \leq \sigma_i \leq T_F$. That is, the τ_i are also exchangeable given this information (but not necessarily mutually independent). This assumes a certain homogeneity during the interval $[T_B, T_F]$, precluding, for example, a change of regime during this time. Let $f(\cdot)$ denote the marginal density of τ_i under these assumptions, $p(\cdot)$ the conditional density of σ_i given $T_B \leq \sigma_i \leq T_F$ and $q(\cdot)$ the density of $\sigma_i + \tau_i$, the arrival time at the downstream detector given $T_B \leq \sigma_i \leq T_F$. If $T_B = -\infty$ we then have that, conditional on $X(\cdot)$, the expected process of downstream arrivals satisfies

$$E[dY(t)|X] = \int_{-\infty}^{t} \sum_j \delta(t-\sigma_j-\tau_j) f(\tau_j) d\tau_j \; dt \tag{4}$$

$$= \sum_j f(t-\sigma_j) \; dt \tag{5}$$

$$= \left(\int_{-\infty}^{t} f(t-v) dX(v)\right) dt \tag{6}$$

where $E(\cdot)$ is the expectation operator.

However, $T_B = -\infty$ is unrealizable. On the other hand, if cars arrive before T_B, (6) does not hold without further requirements. The most natural condition is that $0 < a \leq \tau_i \leq b$ which implies that $f = 0$ outside $[a, b]$. We will refer to the interval of possible travel times $[a, b]$ as the fit window. If d is the distance between the upstream and downstream detectors this condition corresponds to assuming that cars only travel at speeds between $\frac{d}{b}$ and $\frac{d}{a}$. In this case for $T_B + b \leq t \leq T_F + a$, (6) holds since $-\infty$ can then be replaced by T_B.

We can rewrite the expectation of (6) in terms of p, f, q as

$$q(t) = \int_{T_B}^{t} f(t-v) p(v) dv \tag{7}$$

for $T_B + b \leq t \leq T_F + a$.

In practice the processes are aggregated in discrete time units of length Δ; we will assume that T_B, T_F, a and b are multiples of Δ. If we approximate f, p, q by discrete mass functions f_s, p_s, q_s, where q_s is the mass in the interval $[s\Delta, (s+1)\Delta)$, for example, we are led to a discrete approximation of (7), for $(T_B + b)/\Delta \leq t \leq (T_F + a)/\Delta - 1$

$$q_t = \sum_{v=T_B/\Delta}^{t/\Delta - 1} f_{t-v} p_v. \tag{8}$$

This in turn leads to a natural estimation scheme in which we replace q_t and p_v by the counts y_t, x_v of arrivals at the downstream and upstream detectors in consecutive intervals and apply a natural measure of discrepancy, least squares, to fit (8). That is we estimate f_s, $a/\Delta \leq s \leq b/\Delta - 1$, by minimizing

$$\sum_{t=(T_B+b)/\Delta}^{(T_F+a)/\Delta - 1} \left(y_t - \sum_{s=a/\Delta}^{b/\Delta - 1} x_{t-s} f_s \right)^2 \tag{9}$$

over $\{\mathbf{f} : f_s \geq 0, \Delta \sum f_s = 1\}$. This is the basis of our scheme.

We can arrive at (9) from (6) by arguing conditionally on $X(\cdot)$. This has the advantage of allowing the possibility of "constant" p as in the purely stationary case, but makes considering aggregated processes conceptually more awkward.

We note three significant modifications:

1. Other forms of smoothing are possible as well. For example, the density f could be approximated by a density depending on a finite number of parameters.

2. The least squares criterion (9) is computationally by far faster, but it may be more effective to use an information measure between the observed and expected downstream flows such as
$$\sum_{t=(T_B+b)/\Delta}^{(T_F+a)/\Delta-1} y_t \log \sum_{s=a/\Delta}^{b/\Delta-1} x_{t-s} f_s \tag{10}$$
Alternatively this can be looked at as a nonparametric maximum likelihood estimate conditional on the $X(\cdot)$ process.

3. We can allow the possibility of cars enterning and/or exiting from the lanes between detectors by adding probabilities π for entering and γ for exiting leading to (in the discretized version)
$$q_t = \pi + (1-\gamma) \sum_{v=T_B/\Delta}^{t/\Delta-1} f_{t-v} p_v. \tag{11}$$

However, we do not pursue these modifications further here.

Having determined estimates $\hat{f}_s$, of the f_s we now have a choice as to which measure of the travel time distribution we use as a summary. The most natural is the mean of the estimated travel time distribution. However, a somewhat more stable choice of parameter (in view of the difficulties of selecting a and b which we discuss below) is the mode, i.e. $\max^{-1} f_s$, where we again plug in our estimate. If f_s is symmetric unimodal the two measures, of course, agree.

The essential difficulty we face is the choice of the parameters a and b and to a lesser extent the "stationary" periods, $[T_B, T_F]$. To an even lesser extent we face, as in all deconvolution problems, a choice of Δ in (9). If Δ is very small the number of parameters f_s being estimated is large and this can lead to well known instabilities. On the other hand, taking Δ too large may result in bias if the unaggregated travel time distribution is not constant on consecutive Δ units of time. In fact we find that 1 second aggregation seems to lead to more satisfactory results than coarser aggregations for our data.

The choice of a and b:

If the model (7) holds for $a = a_0$ and $b = b_0 > a_0$ there appears to be no penalty for using $a < a_0$ and/or $b > b_0$ in (9). This is however illusory since both increase the number of parameters to be estimated and $b > b_0$ decreases the amount of data usable in (9).

We have found empirically that an adaptive choice of a, b is required for generally satisfactory results in both congested and uncongested periods. The width $b - a$ of what we call the fit window can be taken fairly small and fixed; the range in speed of drivers during homogeneous regimes is not great. But the center of the window, $\frac{a+b}{2}$, has to move. We have found that using (1) with a reasonable value of g works well.

The choice of T_B, T_F:

In a homogeneous period choosing $|T_B - T_F|$ as large as possible gives us maximum precision. However, the larger $|T_B - T_F|$ the more we run the risk of (7) failing due to a transition in regime (e.g. from free flow to congested traffic) at some $T_C \in (T_B, T_F)$. In practice moderate values of $|T_B - T_F|$ give travel time estimates which track well throughout the day.

We mention only briefly that it may be possible to detect the transitions described in the previous paragraph. Suppose for instance that a transition from free flow to congestion occurs at T_C where $T_B < T_C < T_F$. Then the density of τ_i given $\sigma_i \leq T_C$ is f_L and given $\sigma_i > T_C$ is f_H. Then (7) changes to

$$q(t) = \int_{T_B}^{t \wedge T_C} f_L(t-v)p(v)dv + \int_{T_C}^{t \vee T_C} f_H(t-v)p(v)dv \tag{12}$$

where f_L is concentrated on $[a_L, b_L]$, f_H on $[a_H, b_H]$ and $a_L < a_H$, $b_L < b_H$. Model (12) can be fit in exactly the same way as (7) is fit by (9). However, with $a = a_L$ and $b = b_H$ we see that $1 + b_L - a_H$ additional f's plus the parameter T_C need to be fit. We do not pursue this here but note that if (7) is fit rather than (12), then having a large fit window permits us to observe that the distribution fit by (9) is bimodal thus signaling lack of fit. However, its mode and mean may be poor guides to the "actual" travel times.

3.1 Extending Methodology to Multiple Lanes

The formulation thus far is for a single lane of traffic but can be extended to model multiple lanes with exits and entrances. For the most part we will not pursue such a formulation here except to note a useful special case. Consider, for example, a four lane freeway with lane 1 being the high occupancy vehicle (HOV) lane and lane 4 the exit/entrance lane. We might assume that the flows in lanes 2 and 3 are similar, ignore lane changes into and out of them, and model the travel times in these two lanes as having the same probability mass function, f. The flows in these two lanes could be used jointly to estimate f. Let the upstream and downstream flows be denoted by x_t^i and y_t^i, $i = 2, 3$. The travel time distribution, f, is then estimated by minimizing

$$\sum_{i=2}^{3} \sum_{t=(T_B+b)/\Delta}^{(T_F+a)/\Delta - 1} \left(y_t^i - \sum_{s=a\Delta}^{b/\Delta - 1} f_s x_{t-s}^i \right)^2 \tag{13}$$

analogously to (9). Using the data from the two lanes in this fashion will hopefully result in a better estimate of f than that resulting from either lane alone. (Note that this method of combining data from the two lanes is not the same as adding the two flows.)

3.2 Relation to Cross Correlation

We next relate this formulation to a commonly used method for finding delays between stationary time series—cross correlation. Suppose that $U(t)$ and $V(t)$ are stationary time series with $U(t) = V(t-\tau) + Z(t)$, where the stationary noise series Z is independent of V. Let the covariance function of $V(t)$ be denoted by $\rho_{VV}(u)$ and let the cross covariance function of U and V be $\rho_{UV}(u) = \text{Cov}(U(t+u), V(t))$. Then

$$\rho_{UV}(u) = \rho_{VV}(u - \tau) \tag{14}$$

and since $\rho_{VV}(u)$ is maximal at $u = 0$, the cross covariance function is maximal at τ, the lag between the two series. This relationship has been used (Dailey, 1993) to estimate travel times between loop detectors.

What if the lag between the series is not fixed, but is random, there being variation in travel time from vehicle to vehicle? We will consider the same model for travel times as that used earlier, following the development of Brillinger (Brillinger, 1974). Let X and Y be the point processes of arrivals defined above and assume additionally that they are *stationary*. $X(I)$ is the total number of arrivals in the interval I, etc. For stationary point processes, the analogous objects to the mean and covariance function of ordinary time series are the first moment measure and the covariance measure (Brillinger, 1972). The first moment measure is $E[dX(t)] = \lambda dt$, where λ is the mean flow rate, and this is also the first moment measure of Y. The second moment measure of X is $\mu_{XX}(u)du\ dt = E[dX(t+u)dX(t)]$, the second moment measure of Y is similarly defined and the second cross moment measure is $\mu_{YX}(u)du = E[dX(t)dY(t+u)]$. We will make use of the covariance densities, $\rho_{XX}(u) = \mu_{XX}(u) - \lambda^2$, with $\rho_{YY}(u)$ and $\rho_{YX}(u)$ being similarly defined.

Having defined this notation, the starting point of the analysis, using the same assumption on travel times as that above, is (6). We have, assuming $T_B = -\infty$, and since $f(s) = 0,\ s < 0$,

$$\begin{aligned} \mu_{YX}(u)du\ dt &= E_X\left[E\left(dX(t)dY(t+u)|X\right)\right] && (15)\\ &= E_X\left[\int_{-\infty}^{\infty} f(t+u-v)dX(v)dX(t)\ dv\right] && (16)\\ &= \left(\int_{-\infty}^{\infty} f(s)\mu_{XX}(u-s)ds\right) du\ dt. && (17)\end{aligned}$$

Converting to cross covariances we have

$$\rho_{YX}(u) = \int_{-\infty}^{\infty} f(s)\rho_{XX}(u-s)\ ds. \tag{18}$$

The cross covariance function is the convolution of the travel time density function with the autocovariance function of X. The degenerate case of constant travel time τ, $f(s) = \delta(s-\tau)$, gives $\rho_{YX}(u) = \rho_{XX}(u-\tau)$, as in (14).

From (18), we can derive an expression for the mean travel time:

$$\begin{aligned} \int_{-\infty}^{\infty} u\rho_{YX}(u)\ du &= \int_{-\infty}^{\infty} f(s)\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} u\rho_{XX}(u-s)\ du\ ds && (19)\\ &= \int_{-\infty}^{\infty} f(s)\int_{-\infty}^{\infty}(r+s)\rho_{XX}(r)\ dr\ ds && (20)\\ &= \left(\int_{-\infty}^{\infty} sf(s)\ ds\right)\left(\int_{-\infty}^{\infty}\rho_{XX}(r)\ dr\right) && (21)\end{aligned}$$

where we have used the fact that $\rho_{XX}(r)$ is an even function. From this expression, we see that the mean travel time can be expressed in terms of the first moment of the cross covariance function and the integrated autocovariance function.

The development is unsatisfactory since it assumes both $T_B = -\infty$, $T_F = \infty$ and stationarity of X and Y on all of (T_B, T_F). Suppose however, in analogy to the development leading to (7), we only ask that

$$E[dX(t)dX(t+u)] = \mu_{XX}(u)dudt \tag{22}$$

for $\{(u,t) : T_B - t \leq u \leq T_F - t, T_B \leq t \leq T_F\}$ where $\mu_{XX}(u)$ is necessarily defined for $|u| \leq T_F - T_B$ and

$$E[dX(t)dY(t+v)] = \mu_{YX}(v)dvdt \tag{23}$$

for $T_B \leq t \leq T_F$, $a \leq v \leq b$. Then, we are led to,

$$\begin{aligned}
\mu_{YX}(v)dtdv &= E_X\left(E(dX(t)dY(t+v)|X)\right) && (24)\\
&= E(dX(t)\int_a^b f(s)dX(t+v-s)) && (25)\\
&= \left(\int_a^b \mu_{XX}(v-s)f(s)ds\right)dvdt && (26)
\end{aligned}$$

valid for $a \leq v \leq b$ if $b - a \leq T_F - T_B$. We can no longer obtain a relation between the means of μ_{YX}, μ_{XX}, and f since the first two functions are not defined on all of R. However, suppose f is symmetric unimodal and $\mu_{XX}(v)$ is (necessarily also symmetric) unimodal for $|v| \leq b - a$. Then the modes of f and μ_{YX} coincide on (a,b) by Wintner's theorem (Dharmadhikari and Joag-dev, 1988). If we estimate $\mu_{YX}(v)$ as usual from aggregated data by $(T_F - T_B - a)^{-1}\Delta\sum_{t=(T_B+a)/\Delta}^{T_F/\Delta} x_t y_{t+v}$ and maximize for $a \leq v \leq b$ we see that we have arrived essentially at Dailey's method since the cross correlation and second cross moment functions realize their maxima at the same point. Under the symmetry and unimodality conditions we again have a consistent estimate of $\max^{-1} f_s = \sum_{s=a}^{b} s f_s$. This correspondence is borne out by our analysis of the I-880 data.

Finally, a simple relation between cross-correlation and our method can be seen by noticing that the normal equations for least squares based on (9) are

$$\sum_{t=(T_B+b)/\Delta}^{(T_F+a)/\Delta-1} y_t x_{t-v} = \sum_{s=a/\Delta}^{b/\Delta-1}\left(\sum_{t=(T_B+b)/\Delta}^{(T_F+a)/\Delta-1} x_{t-s}x_{t-v}\right) f_s \tag{27}$$

for $a \leq v \leq b$. Now, under the assumption of stationarity and appropriate ergodicity conditions the left hand side of (27), after normalization, is a reasonable estimate of $\mu_{YX}(v)$, and the coefficients of f_s on the right hand side of (27) are similarly reasonable estimates of $\mu_{XX}(v-s)$. Thus, our approach may be thought of as solving an empirical version of (26).

4 Results

In this section we present some results of the application of our methodology to real freeway data. The data comes from a seven mile section of highway I-880 in Hayward, California (Skabardonis et al., 1994). This section of freeway was instrumented with type 170 loop controllers spaced approximately 1/3 of a mile apart. Each loop controller monitored eight to ten mainline, double-trap loop detectors. In addition to flow and occupancy, double-trap loop detectors allow one to make fairly reliable measurements of vehicle speed by observing the time it takes for a vehicle to pass over both detectors. This speed data is used to form a travel time estimate that serves as a standard of comparison for our methods. As a side point we should note that it is possible for the double-trap speed estimates to be incorrect as well. In order to calculate speed the loop controller tries to match up pulses on the downstream trap with the corresponding pulses on the upstream trap. Since the loops are separated by less than 20 feet the loop controller usually has no problem matching up the correct pairs of pulses. But if a heavy truck goes over the detectors then every axle might trigger a pulse. In this case the loop controller can inadvertently match up two pulses incorrectly and hence calculate an incorrect speed that is much higher than the true speed.

Although the loop detector signals are sampled by the 170 controller at a rate of 1/60'th of a second, we aggregated this to $\Delta = 1$ second resolution for the purpose of this study. In Section 4.2 we investigate the effects of using a larger value of Δ. Loop data was collected during the peak commute times of 5:00 - 10:00 a.m. and 2:00 - 7:00 p.m. during 24 weekdays in the spring of 1993. A very small amount of missing flow data was replaced with zeros, and clock off-sets were estimated by hand since the loop controller clocks were not properly synchronized.

In addition to the loop detector data, there were four probe vehicles that were driving up and down the freeway during the same time period. The probe vehicles, which maintained a headway of approximately 7 minutes, were equipped with computers that accurately recorded the car's movement, and hence travel time, down the freeway. The probe vehicle travel times are an important piece of information because they give us a loop-independent measurement of the travel time down the freeway.

4.1 Effect of Fit Window

The first results that we present illustrate the importance of choosing the fit window, $[a, b]$. For this experiment we focus on two loop detectors separated by 2200 feet and we concentrate on only the third lane from the center. At free-flow speeds of 60 miles per hour the travel time is 25 seconds. We use (9) to estimate the travel time distribution, and we use a data window of $|T_B - T_F| = 300$ seconds and $\Delta = 1$ second. However, we explore two different fit windows, a static one and an adaptive one, and demonstrate that the adaptive window does significantly better. The first window is based on the assumption that the vehicles will always be traveling between 15 and 70 mph, and hence, for these two detectors $[a, b] = [21, 100]$. The second window is an adaptive window that is centered at an estimate of the travel time based on the loop detector data. In this scheme, the speed is first estimated from the upstream single-trap loop detector using the relationship in (1), with $1/g = 22.6$ feet. The travel time is then calculated for this pair of loop detectors assuming that the speed is constant for the entire link. Finally the window is centered at this calculated travel time and given a width of 20 seconds (i.e. $b - a = 20$). Note that the center of this window is adjusted based on the single-trap loop data at every time period and is therefore adaptive. Figures 1 and 2 show the travel times estimated every 2 minutes from 5:00 a.m. until 10:00 a.m. for these two loop detectors. During the early morning, 5:00 a.m. to 6:30 a.m. when the traffic is light, the travel time is roughly between 25 and 30 seconds. Around 6:30 a.m. there was a incident that caused congestion to form upstream and hence caused an increase in the travel time. Figure 1 gives the estimate of the travel time for the static fit window. As one can see the estimate of this travel time is quite noisy even during the free flowing traffic conditions of the early morning. But when we restrict the size of our fit window to only 20 parameters and center it on the car-length travel time estimate we see a significant improvement as shown in Figure 2. Not only are we now able to accurately track the travel time in free flowing conditions, but

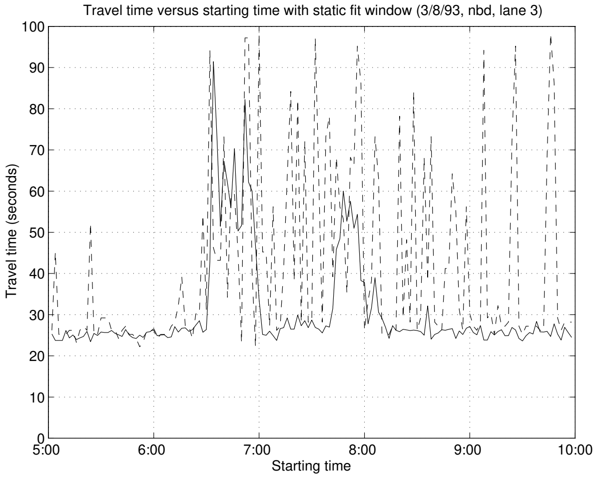

Figure 1: Travel time for static fit window based on a rational speed assumption. Solid line shows double-trap estimates.

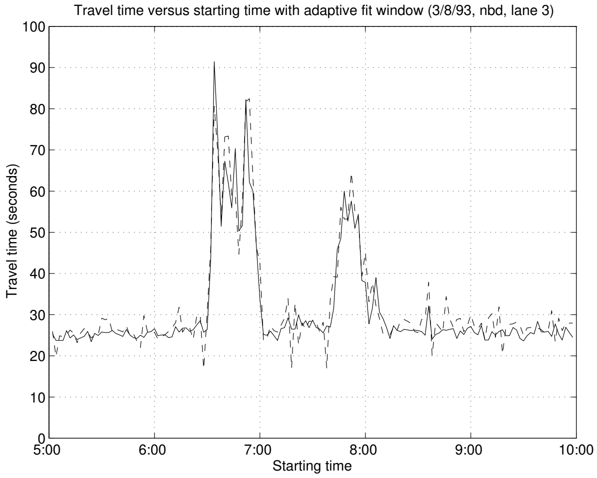

Figure 2: Travel time for adaptive fit window based on a constant car length assumption. Solid line shows double-trap estimates.

we are also able to give reasonable estimates during the periods of congestion.

To examine this further, we give plots of the estimate of the travel time distribution, $\hat{f}_s$, for a particular time point. We choose the time point of 9:08 a.m. because the static, wide fit window in Figure 1 fails to find the correct mode but the adaptive, narrow fit window in Figure 2 succeeds. Figure 3 is a plot of $\hat{f}_s$ for the static fit window $[a, b] = [22, 100]$. The

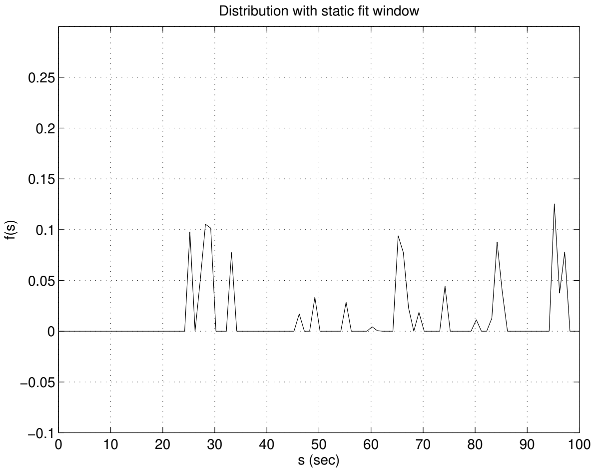

Figure 3: Distribution of $\hat{f}_s$ for static fit window based on rational speed.

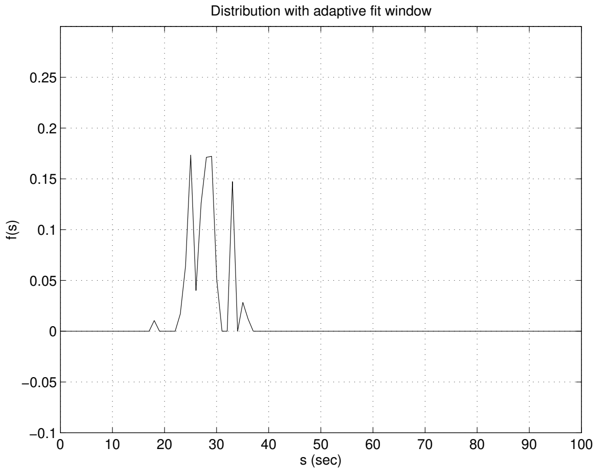

Figure 4: Distribution of $\hat{f}_s$ for adaptive window based on constant car length.

true travel time is taken as the mode of the distribution and in this plot the mode is $s = 96$ seconds, which corresponds to a speed of 15.6 mph. In this figure one can see that the mass centered around 28 and 29 seconds is much more likely to be the true travel time. Indeed, when we restrict the fit window to be $[a, b] = [17, 37]$, as in Figure 4, we find that the mode is now taken as $s = 24$ seconds—which, according to the double-trap loop detectors, is the correct travel time.

Note that the center of the fit window is simply the estimate of travel time based on the single-trap speed calculation. Although it is known that this speed calculation is highly inaccurate and prone to various biases (Hall and Persaud, 1989), for our purpose it is quite sufficient because it gives us a rough estimate of the true travel time. Since we then center our fit window on this estimate and only fit over 20 parameters, this first guess allows us to essentially filter out any extraneous peaks that might appear in $\hat{f}_s$ as we saw in Figure 3.

4.2 Effect of Aggregation

In this section we describe some of our results with respect to data aggregation. As discussed, if we assume the travel time distribution f is constant on consecutive Δ second intervals and T_B, T_F, a, b are multiples of Δ, then (9) can be used to estimate the coarsened travel time distribution f. When these assumptions hold, the Δ-fold reduction in parameters reduces the variance in $\hat{f}_s$. However, the decrease in variance may come at a great cost in terms of bias.

Since by assumption the mode of f corresponds to a Δ interval of travel times on which f is constant, we need a convention for choosing a single summary travel time. A natural candidate is the midpoint of the modal class, $(\max^{-1} f_s - 0.5)\Delta$, where we once again plug in $\hat{f}_s$ for our travel time estimate. For level Δ of aggregation, denote our estimate of the travel time at the i^{th} time point as $\hat{\tau}_i^{\Delta}$. Also, let $\tilde{\tau}_i$ stand for the travel time estimate at the i^{th} time point based on the double-trap speed measurement. To compare results at different levels of aggregation we consider the average L_1 distance between our travel time estimates and the double-trap estimates as a function of aggregation:

$$\frac{1}{N}\sum_{i=1}^{N} |\hat{\tau}_i^{\Delta} - \tilde{\tau}_i| \tag{28}$$

where N is the number of times at which travel time estimates are made.

For example, we estimated f at $N = 77$ equally spaced 3 minute intervals from 6 a.m. to nearly 10 a.m. between a certain northbound pair of detectors using $\Delta = 1, \ldots, 12$ seconds. In each case the fit window at the unaggregated level was chosen by using (1) to find an initial center, c. Then a was taken as the largest Δ multiple less than or equal to $c-10$, and b as the smallest Δ multiple greater than or equal to $c+10$. This insured that the fit window $[a, b]$ was always a Δ multiple, yet continued to cover the middle of the common car length based travel time estimate. $T_F - T_B = 600$ seconds was used as well. Figure 5 shows the plot of (28) versus Δ. Indeed the dips at $\Delta = 3$ seconds and $\Delta = 11$ seconds show that aggregation can improve the estimates in this L_1 sense. However, the nearby peaks demonstrate the other edge of the sword. Figures 7 and 8 show the travel time estimates side by side for $\Delta = 10$ seconds and $\Delta = 11$ seconds, respectively. Although the levels of aggregation differ by only a second, the results are dramatically different. $\Delta = 11$ seconds does well because the middle of the 4^{th} aggregated travel time interval is 38.5 seconds, which is very close to the free flow travel time. Conversely, $\Delta = 10$ seconds does poorly because the corresponding mid-interval misses the mark for most of the morning.

Of the 77 time points in our example, 50 were during periods of low density. One might therefore ask if the L_1 error analysis is being driven mostly by the performance of the estimators during low occupancy. Further, one could argue for judging the magnitude of

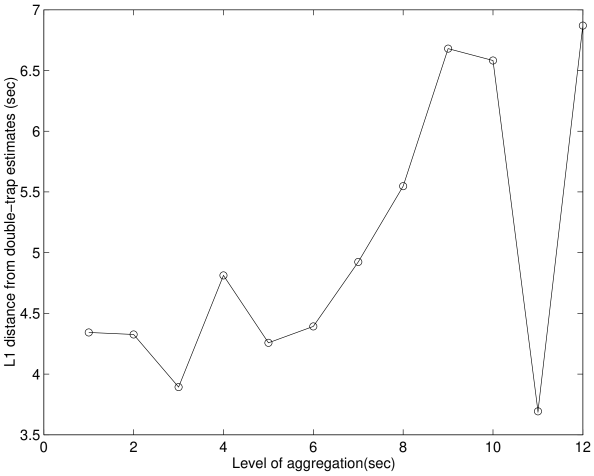

Figure 5: L_1 distance between estimates based on (9) and double-trap speed based estimates as a function of Δ.

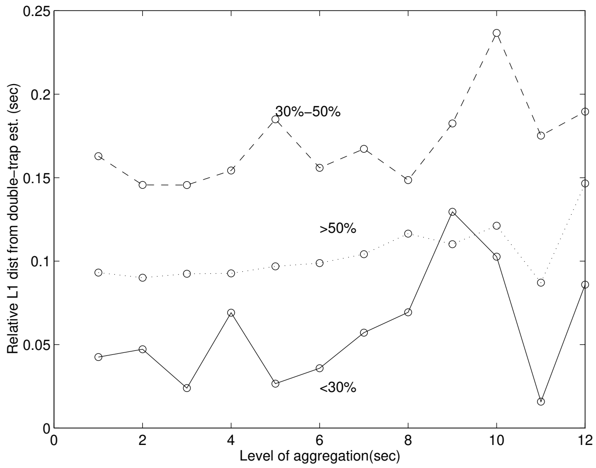

Figure 6: Relative L_1 distance between estimates based on (9) and double-trap speed based estimates as a function of Δ grouped by density level.

errors relative to the (double-trap) travel times, rather than absolutely. Figure 6 addresses these points by showing average relative L_1 distances as a function of Δ, with the averages taken over estimates corresponding to different levels of density. By relative L_1 error, we mean that the summand in (28) is replaced with $|\hat{\tau}_i^{\Delta} - \tilde{\tau}_i|/\tilde{\tau}_i$. As expected, the results during low occupancy ($< 30\%$) are very similar to Figure 5. However, two new striking features appear. First, periods of medium density ($30\% - 50\%$) are more difficult to estimate than high density ($> 50\%$) in the relative sense. The nine periods of medium density take place roughly from 7:15 until 7:30 and again from about 8:20 until 8:30. The $\tilde{\tau}_i$ are lower during these transition periods than during high density (7:30 to 8:20), but the absolute errors are comparable. Hence, the relative errors are in the 15% - 20% range. This agrees with our stochastic model that associates transition with bi-modality of the travel time distribution. The second striking feature is that during periods of high density the relative error is nearly constant as a function of aggregation. This occurs since during high occupancy the estimates gravitate towards the center of the fit window, which is basically the same for all Δ. This point is discussed further in section 4.5. It is important to keep in mind that the L_1 distances are distances between two sets of estimates, not estimates and the "truth". Although the $\tilde{\tau}_i$ are believed to be fairly accurate, they may also suffer some degree of breakdown during transition and/or high occupancy.

This example demonstrates our general finding that when the "true" travel time minus $\Delta/2$ (or more generally minus whatever point from the Δ interval of equally likely travel times that is returned as a summary) is nearly a multiple of Δ for large portions of the estimation period, then aggregation can be beneficial. But when this is not true the ensuing bias can outweigh the variance reduction benefits. This bias is most noticeable during low occupancy when fairly constant free-flow travel times are present for much of the day. Since one never knows the true travel times in advance, this strongly suggests leaving the data at a

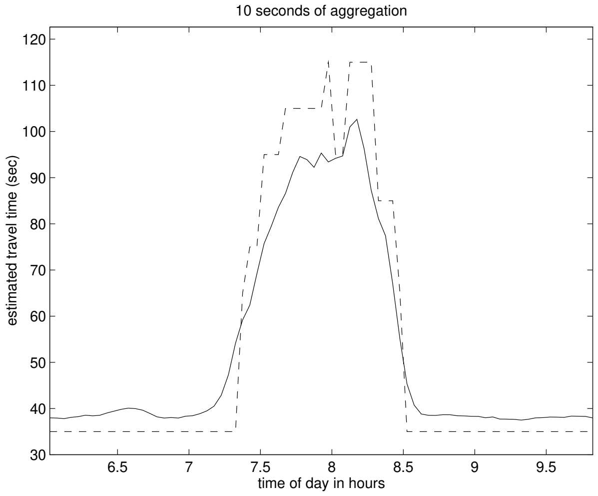

Figure 7: Travel time estimates using car length based fit window and $\Delta = 10$ seconds. Solid line shows double-trap estimates.

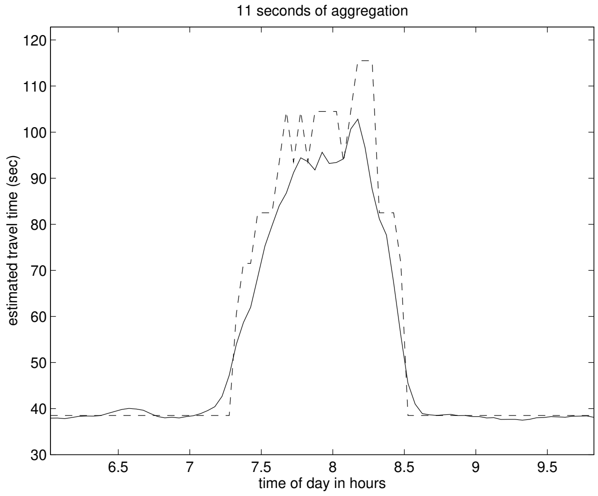

Figure 8: Travel time estimates using car length based fit window and $\Delta = 11$ seconds. Solid line shows double-trap estimates.

small enough level of aggregation that the bias is tolerable, yet not so small that the variance gets too large. We find $\Delta = 1$ second to be a good level for the data at hand.

4.3 Cross-Correlation Estimates

In Section 3.2 we noted that the regression method and the cross-correlation method are related. In this section we empirically investigate this by using a cross correlation method to estimate travel times. As was noted in Section 3.2, we do this by estimating the cross covariance function, $\rho_{XY}(u)$, as first indicated in equation (18). This is essentially Dailey's method (Dailey, 1993, equations 20 and 21). In order to properly compare this method with our regression method presented earlier we use the same pair of loop detectors and the same time period as was used in Section 4.1. We also use the same amount of data, $T_F - T_B = 300$ seconds.

Figure 9 is an estimate of the travel time using the cross correlation method with an aggregation of $\Delta = 5$ seconds. This is essentially Dailey's method with the slight modification of having a static window $[a, b] = [22, 100]$ (without this window Dailey's method is even more noisy). As in the regression method, the choice of $[a, b]$ is very important. In the regression method the interpretation of $[a, b]$ is that the support of the distribution of travel times, f_s, only exists on $[a, b]$. Therefore we computed the estimate $\hat{f}_s$ only over the interval $[a, b]$ and assumed that it was zero everywhere else and referred to $[a, b]$ as the fit window. In the cross correlation method the interpretation is slightly different because we compute the cross correlation over the entire range of $[T_B + a, T_F]$. Then we *search* for the mode of the distribution over the range $[a, b]$. Hence, the meaning of the range $[a, b]$ has changed from being a fit window to being a search window. The importance of the search window can be seen in Figure 10 where the search window is now the same adaptive window that was used in Section 4.1 with $b - a = 20$ seconds.

Note how in both of these figures the estimated travel time is only a multiple of

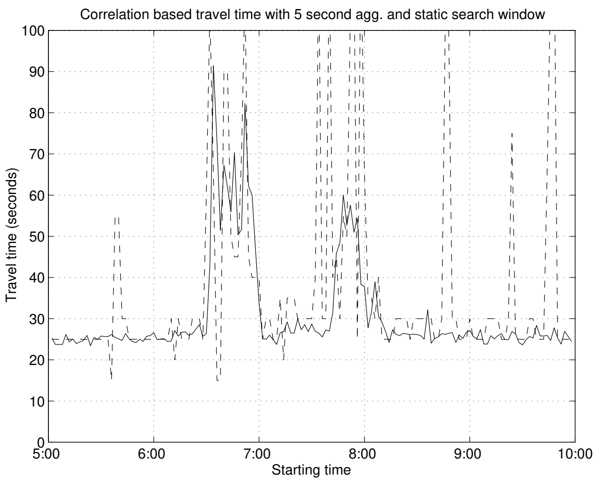

Figure 9: Travel time for single link for cross-correlation method and static fit window, $\Delta = 5$ seconds. Solid line shows double-trap estimates.

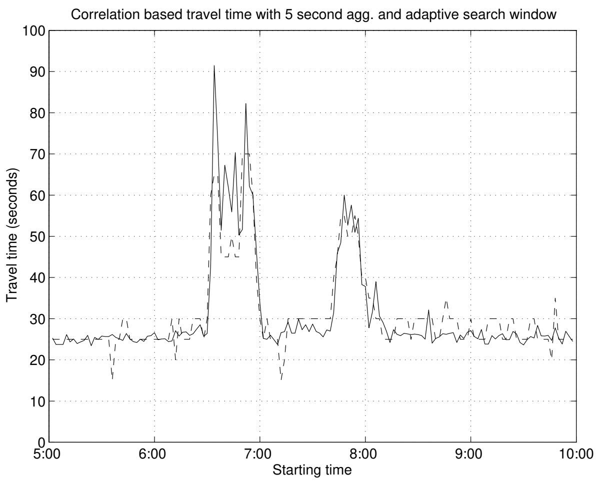

Figure 10: Travel time for multiple links for cross-correlation method and adaptive fit window, $\Delta = 5$ seconds. Solid line shows double-trap estimates.

$\Delta = 5$ seconds. As was discussed in Section 4.2, the reason that the cross correlation estimate in Figures 9 and 10 is doing so well during the early morning is because the true travel time is 25 seconds, which is a multiple of Δ (we chose a multiple of Δ instead of the center of the aggregation window). In the late morning, the true travel time is 26 or 27 seconds and hence the estimate jumps back and forth between 25 and 30 seconds as can be seen in Figure 10. If we decrease the level of aggregation to $\Delta = 1$ second we can overcome this chattering effect. Figure 11 is the travel time estimate with $\Delta = 1$ second and a static fit window $[a, b] = [22, 100]$. This should be compared to Figure 9. Figure 12 is the travel time estimate with $\Delta = 1$ second with an adaptive fit window and should be compared to Figure 10. It is clear from these figures that it is essential to properly center and restrict the search window, $[a, b]$, in order to have satisfactory results. Finally, one should compare Figure 12 to Figure 2 and note that they are quite similar. These two figures represent a direct comparison between the regression and the cross correlation methods with the same fit/search window, $b - a = 20$ seconds, and the same level of aggregation, $\Delta = 1$ second.

Figure 13 gives the estimate of the cross covariance, $\rho_{XY}(v)$, at three consecutive time slices of 5:22, 5:24, and 5:26. The correct travel time for all three time points is around 25 seconds. The mass centered at 25 seconds in the cross covariance functions can easily be visually tracked through the three time slices. In the first and third time slice the global peak of the cross covariance function agrees with the results from the double-trap speed measurement. But in the second estimate, the global peak of the cross covariance function occurs at $\tilde{v} = 52$ seconds, which is incorrect. Although the proper peak in the cross covariance estimate at $\tilde{v} = 25$ seconds is not a global maximum, one can easily see that it is still a local one. Therefore the restriction of the search window to a neighborhood of this maximum will usually allow us to locate it.

Although we do not pursue it here, it is interesting to note that the spurious global maxima that overshadow our "true" maxima are usually very skinny. Whereas the "true"

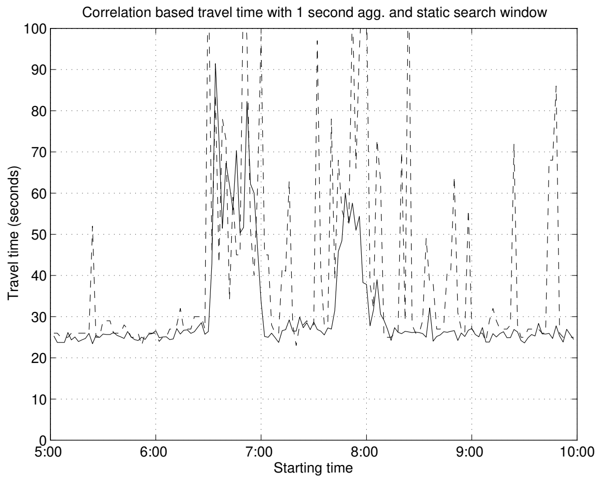

Figure 11: Travel time for single link for cross correlation method and static search window, $\Delta = 1$ second. Solid line shows double-trap estimates.

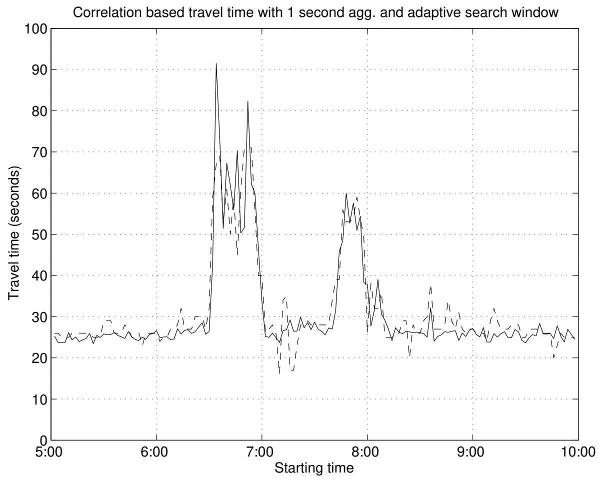

Figure 12: Travel time for multiple links for cross correlation method and adaptive search window, $\Delta = 1$ second. Solid line shows double-trap estimates.

peaks that we are looking for usually have a lot of mass associated with them. This can be seen in Figure 13 as well as in Figure 3. This would lead one to believe that filtering the cross covariance function with a simple low pass filter before searching for the global maximum would enhance our results by flattening any spurious, skinny peaks. This can be seen to be equivalent to smoothing both the upstream and downstream counts, x_t, y_t, prior to performing the cross correlation.

4.4 Multiple Freeway Links

An important question to investigate is whether this method will work for an extended section of the freeway that is covered by many single-trap loop detectors. The northbound direction of I-880 has 17 working single-trap loop detectors spread over 5.8 miles. The distance between them ranges from 1000 feet to 3400 feet. To estimate the travel time down this entire section of freeway we estimated the travel time on each link (where a link is from loop detector to loop detector) and summed them up. The day that we picked to demonstrate this method, the afternoon rush hour of 3/5/93, had an unusually high amount of congestion. There were multiple stalled vehicles on the side of the road that needed assistance—all of which were located close to a major off-ramp. This caused a two hour period of stop-and-go traffic that stretched for approximately 2 miles. The results of our estimate for all 17 loops is given in Figure 14. The dotted line is our estimate of the travel time and the solid line is the double-trap loop detector measurement of travel time. Note that out estimates did not use corrections for entrances and exits such as we indicated could be made via equation (13).

It is clear that the double-trap travel time estimates are noisy for high travel time estimates. Since the period from 4:00 p.m. until 5:30 p.m. corresponds to very slow moving traffic these travel time estimates are very sensitive to small fluctuations in the speed. Indeed, a fluctuation from 5 to 10 mph would double the travel time. Since we are estimating the

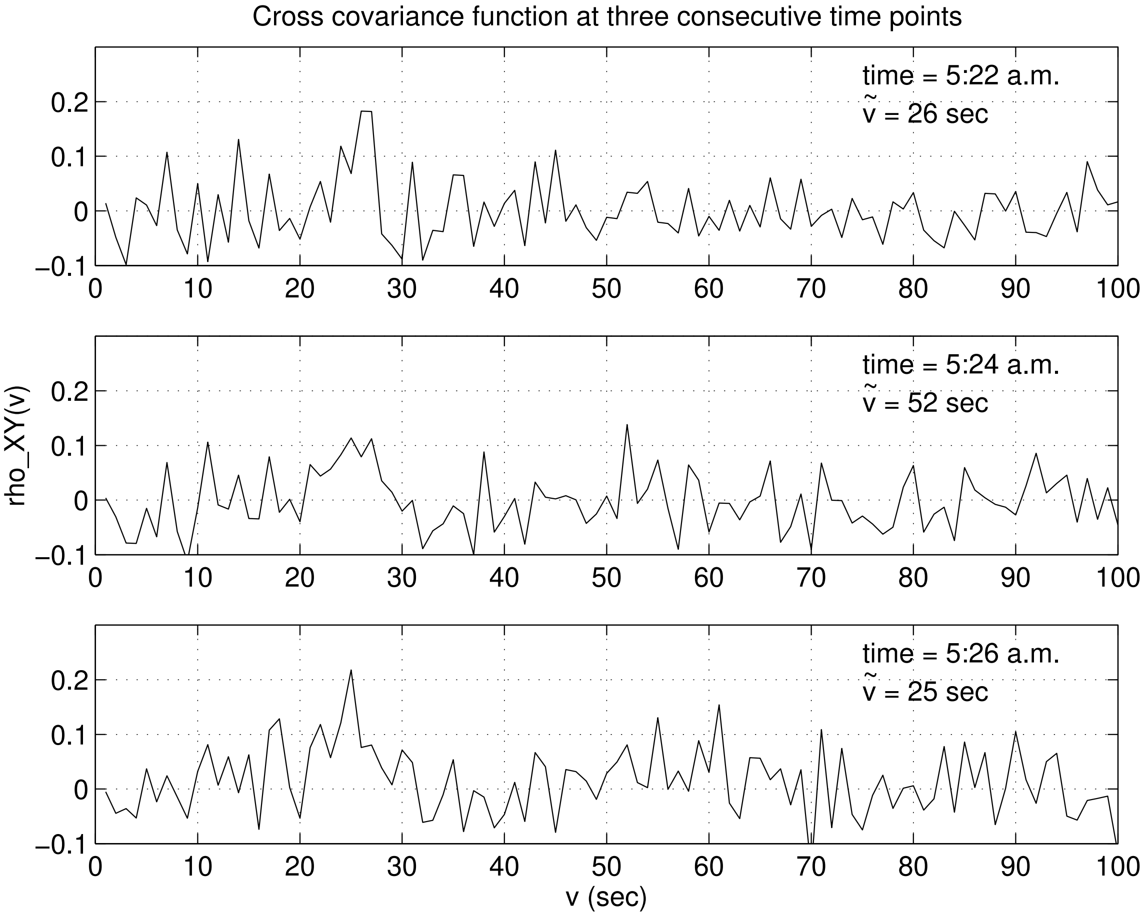

Figure 13: Three consecutive slices of the cross covariance, $\rho_{XY}(v)$. The middle estimate at 5:24 a.m. has a global maximum at $v = 52$ seconds.

double loop travel time by assuming a constant speed on the freeway link, it is clear that when the traffic flow and speed are not homogeneous then this estimate will be incorrect. Therefore, in this regime of high congestion, a much more reliable source of travel time information is the probe vehicles. The travel times reported by the probe vehicles for the same section of freeway are also placed on Figure 14 as asterisks. During the highly congested period the probe vehicle travel times consistently match up with our regression-based estimates.

It should be noted that the probe vehicle travel times reported in Figure 14 could possibly have some bias. For this calculation both the double and single loop travel time estimates are for the third lane of the freeway. The probe vehicles on the other hand, were not restricted to drive in only the third lane and it is possible that they switched between lanes 2, 3 and 4 (they couldn't drive in lane 1, the HOV lane). The correspondence between the probe vehicles and our regression estimate is quite remarkable, especially considering the distances involved.

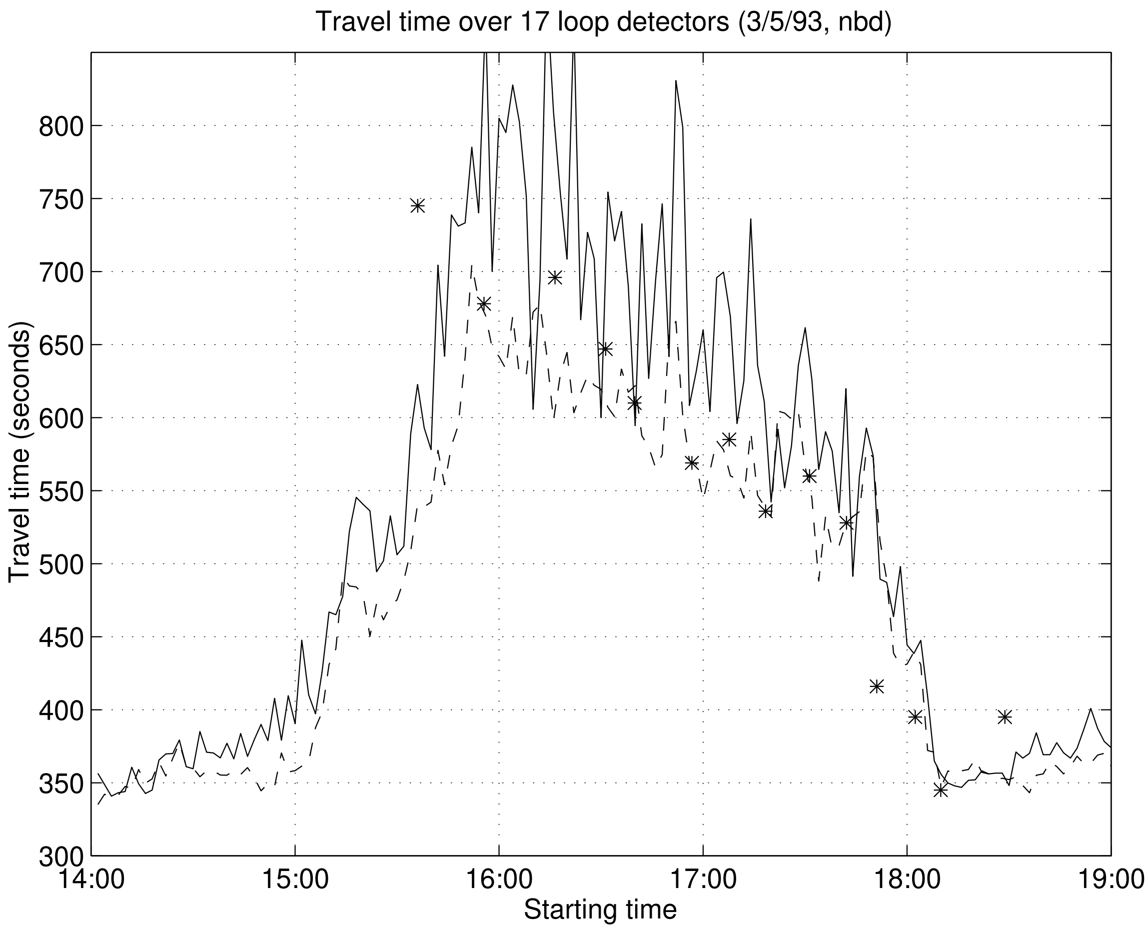

Figure 14: Travel time down entire length of freeway. Solid line is the double-trap estimate.

4.5 Reliance of Methods on Fit Window

We have mentioned that the fit window, $[a, b]$, is crucial to the success of both the regression and cross correlation methods. In both methods the fit window is centered on the estimate of travel time based on the speed obtained from equation (1). At this point the most intriguing question to ask is to what extent the regression and cross correlation methods rely on the accuracy of the fit window. If the value used for g is inaccurate what effect does this have on the two methods? This really gets to the issue of the whether these method are working or not. For example, it's not quite clear from Figure 4 that our estimate of the mode of $\hat{f}_s$ isn't really the center of the fit window in the first place. Hence, would the center of the fit window be a good estimate for the travel time?

In order to investigate this we perform our regression method down the entire length of the freeway and then compare this to the travel time estimated from the center of the fit window. In equation (1) $1/g$ = average effective vehicle length in feet, provided that the other quantities have correct units. A typical value for the average effective vehicle length is around 22 feet. This value is obviously a function of the types of vehicles in each lane. Hence the value will be smaller in the left-most lanes, where small passenger cars typically travel, and larger in the right-most lanes where trucks tend to drive. Since the density of trucks varies from freeway to freeway, the value of g would need to be calibrated to reflect the local conditions. Despite these deficiencies in using equation (1) we will try to determine if the center of the fit window provides a good estimate for travel time and what effect the accuracy of the parameter g has on our method.

We do this computation with two different values of g. The first is with $1/g = 24$ feet and the second is with $1/g = 20.5$ feet. The width of the fit window is $b - a = 20$ seconds in both cases. The day that we chose for this experiment is 3/8/93. There was moderate congestion on this day caused by various breakdowns along the right-hand side of the road.

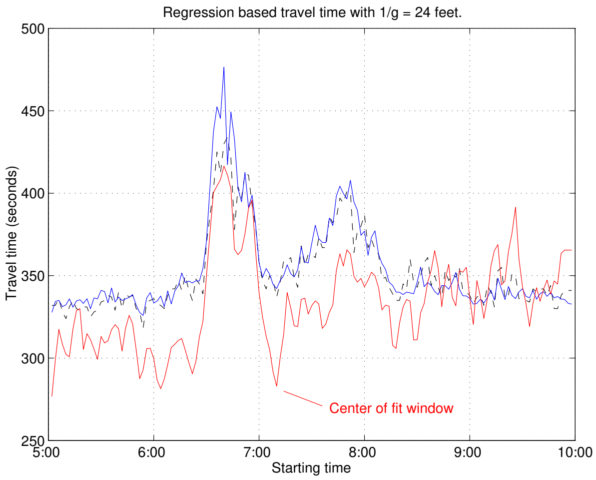

Figure 15: Travel time for adaptive fit window based on $1/g = 24$ ft. Solid line shows double-trap estimates, dashed line shows regression estimate, and light line shows center of fit window.

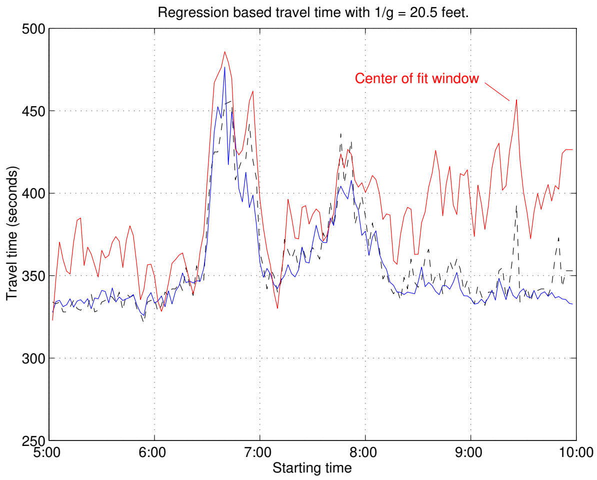

Figure 16: Travel time for adaptive window based on $1/g = 20.5$ ft. Solid line shows double-trap estimates, dashed line shows regression estimate, and light line shows center of fit window.

Figures 15 and 16 show the results of our two experiments. In each figure the true travel time is represented by the dark solid line and the regression based estimate, which is the dashed line, is right on top of it. The third line in each figure is the center of the fit window $[a, b]$. In Figure 15 the center of the fit window is almost always below the true travel time. Yet the regression based method can still accurately track the double loop based travel time. We have analogous results in Figure 16 where the center of the fit window is consistently above the true travel time.

The important thing to note is that the regression based estimate of the travel time is basically the same in both figures. The L_1 distance between the regression based travel time and the double loop based travel time is $L_1 = 6.8$ seconds in Figure 15 and $L_1 = 8.8$ seconds in Figure 16, whereas the L_1 distance between the center of the fit window and the loop based travel time is $L_1 = 27$ seconds in Figure 15 and $L_1 = 36$ seconds in Figure 16.

Hence, the reliance of our regression method on the actual value of car length, $1/g$ is indeed small. This suggests that as long as the value of g is relatively close to the correct value then our method can pick out the true peak in the travel time distribution. On the other hand, travel time estimates based on equation (1) are certainly not robust to slight variations in g.

5 Conclusion

This paper has presented a method for using high-resolution flow and occupancy data from single-trap loop detectors to obtain accurate estimates of travel time. The model upon which our method is based assumes that a homogeneous (with respect to traffic) time interval, $[T_B, T_F]$, exists, and that within this interval the travel times are exchangeable given the upstream arrival times. This leads to a simple estimation procedure which is given by equations (8) and (9). One would like to keep $|T_B - T_F|$ as large as possible to increase precision. However, as $|T_B - T_F|$ gets larger the estimation procedure loses its ability to track the travel time through different regimes. Therefore a balance must be made between the opposing goals.

We noted that using an adaptive fit window $[a, b]$ to specify a reduced support for the travel time distribution is essential for satisfactory results in both the regression and cross correlation methods. The adaptive window that we have chosen is centered on the commonly used relation for speed from single-trap loop detectors given in equation (1). Although the methods depend on equation (1) and hence on the value of g, it was shown in Section 4.5 that the dependence is not too great. The methods can work when the value of g is incorrect and give significantly better estimates than using equation (1) alone.

As noted in Section 3.1 our regression method can easily be extended to incorporate multiple lanes of the freeway. Properly modeling entrances and exits is the subject of future work. Section 4.2 notes that it is desirable to aggregate the data if the true travel time down the link is a multiple of the level of aggregation. Otherwise, any aggregation can lead to a bias possibly as large as the level of aggregation itself. So, while the computational gains from aggregation are significant, the resulting loss of accuracy is almost always unavoidable.

We have demonstrated that this methodology can estimate the travel time down an extended section of the freeway even during extremely congested conditions. The accuracy of our estimates is confirmed by the probe vehicle data. The implications of having a method to accurately estimate travel times on freeways using single-trap loop detectors are significant. From the freeway management perspective this will mean cost savings. In many freeways the single-trap loop detectors are the only source of measurement information about the traffic conditions. If it is possible to obtain a fine enough resolution from these existing detectors then our methodology will allow reliable estimates of travel times to be made. This is a substantially cheaper alternative than installing new measuring devices on the freeway (i.e. new loop detectors, video cameras). From the ATIS application end it will mean a reliable source of travel time information. Most traffic management and control algorithms and all routing algorithms are based on knowing the link travel time so that they can assign a cost to each link. Hence the need for accurate travel times is very important.

Travel time is also a good indication of congestion. In this vein, the possibility of using this methodology for detecting a rapid change of regime that accompanies an incident is certainly intriguing. Whether an incident can be detected by the distribution $\hat{f}_s$ becoming bimodal is the subject of future work.

Acknowledgements—The authors gratefully acknowledge the support of NISS and the California PATH Program of the University of California, in cooperation with the State of California Business, Transportation, and Housing Agency, Department of Transportation. The authors also wish to thank Pravin Varaiya for his helpful comments and valuable suggestions throughout the course of the project.

References

Brillinger, D. R. (1972). The spectral analysis of stationary interval functions. In *Proc. Sixth Berkeley Symp. Math. Statist. Prob.*, pages 483–513. University of California Press.

Brillinger, D. R. (1974). Cross-spectral analysis of processes with stationary increments including the stationary g/g/infinity queue. *Annals of Probability*, 2:815–827.

Dailey, D. J. (1993). Travel-time estimation using cross-correlation techniques. *Transportation Research, Part B*, 27B(2):97–107.

Dharmadhikari, S. and Joag-dev, K. (1988). *Unimodality, Convexity and Applications*. Academic Press.

Hall, F. and Persaud, B. N. (1989). Evaluation of speed estimates made with single-detector data from freeway traffic management systems. *Transportation Research Record*, 1232:9–16.

Pushkar, A., Hall, F. L., and Acha-Daza, J. A. (1994). Estimation of speeds from single-loop freeway flow and occupancy data using cusp catastrophe theory model. *Transportation Research Record*, (1457):149–157.

Sanwal, K., Petty, K., Walrand, J., and Fawaz, Y. (1996). An extended macroscopic model for traffic flow. *Transportation Research, Part B (Methodological)*, 30B(1):1–9.

Sisiopiku, V. P. and Rouphail, N. M. (1994a). Analysis of correlation between arterial travel time and detector data from simulation and field studies. *Transportation Research Record*, (1457):166–173.

Sisiopiku, V. P. and Rouphail, N. M. (1994b). Toward the use of detector output for arterial link travel time estimation: A literature review. *Transportation Research Record*, (1457):158–165.

Skabardonis, A., Noeimi, H., Petty, K., Rydzewski, D., Varaiya, P. P., and Al-Deek, H. (1994). Freeway service patrols evaluation. Technical Report UCB-ITS-PRR-95-5, Institute of Transportation Studies, University of California, Berkeley.