ABSTRACT

This paper analyses the role of the sampling protocol for travel time estimation with low frequency probe vehicle data by likelihood-based methods, such as maximum likelihood or Bayesian estimation. In the literature, there are reported cases where the vehicle positions are sampled at either certain time intervals, say, every minute, or at certain distance intervals, say, every 500 meters. We show that whether sampling is distance-based or time-based determines the proper formulation of the likelihood function. Furthermore, an incorrect likelihood formulation (for example, treating sampling as distance-based when it is time-based in actuality) often leads to biased parameter estimates. For the special case in which the path is partitioned into segments with constant, independent travel speeds we derive explicit formulas for the likelihood function for each of the two sampling principles. We also study the consistency and bias of the estimators in numerical experiments.

Keywords: travel time, estimation, sampling, GPS, probe

1. INTRODUCTION

In recent years, GPS devices already installed for other purposes in vehicle fleets (e.g., personal cars, taxis, trucks, etc.) and smartphones have emerged as a new type of traffic sensor. These opportunistic sensors have a great potential for provision of data for traffic monitoring and management applications. Unlike stationary sensors such as loop detectors and automatic vehicle identification (AVI) sensors, they can collect travel time data, for any part of the network where equipped vehicles move. Unlike designated probe cars, they can continuously collect data for any time and day that equipped vehicles are active. However, a severe obstacle for the widespread adoption of these data is that more advanced and sophisticated methods are needed to process the data and generate useful information, compared to traditional sensors (Leduc, 2008). In particular, low sampling frequency creates difficulties in inferring the true path of the vehicle between two position reports, which may involve a considerable number of network segments (Rahmani and Koutsopoulos, 2012; Miwa et al., 2012). It also becomes difficult to identify the fraction of the travel time spent on each individual segment, and different local methods have been developed for this task (Hellinga et al., 2008; Miller et al., 2010; Zheng and van Zuylen, 2012).

The purpose of this paper is to analyse the role of the sampling protocol for travel time estimation with low frequency probe vehicle data by likelihood-based methods, such as maximum likelihood or Bayesian estimation. In the literature, there are reported cases where the vehicle positions are sampled at certain time intervals, say, every minute (e.g., Hunter et al., 2009; Hofleitner et al., 2012), and other cases where
they are sampled at certain distance intervals, say, every 500 meters (e.g., Westgate et al., 2011), or even a combination of both time and distance (Jenelius et al., 2012). The only information considered is the observed travel times and distances between reports; thus, we do not assume data on, for example, instantaneous speeds. We show that whether sampling is distance-based or time-based determines the probability distribution for how the observed travel time is distributed along the observed distance, which in turn determines the proper formulation of the likelihood function. Furthermore, an incorrect likelihood formulation (for example, treating sampling as distance-based when it is time-based in actuality) often leads to biased parameter estimates.

The remainder of the paper is organized as follows. A general likelihood formulation of the travel time estimation problem is presented in Section 2. In Section 3 we consider the special case in which the path is partitioned into segments with constant, independent travel speeds and we derive explicit formulas for the likelihood function for each of the two sampling principles. In Section 4 we study the consistency and bias of the estimators using the correct and the incorrect likelihood formulations in numerical experiments. Section 5 concludes the paper.

2. METHODOLOGY

2.1 General Problem

The problem considered in this paper is to estimate the distribution of travel times between any two positions along any specified path in a road network. The observations used for the estimation come from vehicles that travel along the path, reporting the time and their positions at certain intervals. Any position along a given path is uniquely parameterized by the distance \( x \) from the start point of the path. The travel time \( T(x, d) \) between two positions \( x \) and \( x + d \) is stochastic and described by a probability density function \( p_T(\tau | x, d, \theta) \), where \( \theta \) is a set of parameters characterizing the probability distribution for any \( x \) and \( d \). We will refer to this probability density function for all \( x \) and \( d \) as the travel time model. Our interest is thus in estimating the parameters \( \theta \) of the travel time model using a likelihood-based approach.

Each vehicle reports its position \( x_i \) and the corresponding time \( t_i \) at certain discrete points \( i = 0, 1, ..., n \), producing a trace \( (x_0:n, t_0:n) \). Rather than absolute positions and time stamps, it is convenient to express the sequence of observations in terms of the distances and time intervals between reports, i.e., \( d_i \equiv x_i - x_{i-1} \) and \( \tau_i \equiv t_i - t_{i-1} \) for \( i = 1, ..., n \). The first position and time stamp \( (x_0, t_0) \) are considered as given. The travel time distribution is assumed to be independent of the timing of the trip for simplicity, which means that \( t_0 \) does not influence the likelihood function \( p(d_{1:n}, \tau_{1:n} | x_0, \theta) \).

A dual concept to the travel time \( T(x, d) \) between two known positions \( x \) and \( x + d \) is the travel distance \( D(x, \tau) \) from a known position \( x \) during a known time interval \( \tau \). We refer to the probability density function of this distance for all possible start positions \( x \) and time intervals \( \tau \), denoted \( p_D(d | x, \tau, \theta) \), as the travel distance model. When fully specified, the travel time model and the travel distance model contain the same information, and are characterized by the same set of parameters \( \theta \). Given a feasible specification of one of the models, the specification of the other model can, at least in principle, be derived.

2.2 Distance-based Sampling

With a distance-based sampling protocol, the vehicle trajectories are sampled at certain distance intervals according to some pdf \( p_{sd}(d_{1:n} | \beta_{sd}) \), where \( \beta_{sd} \) is a set of parameters describing the sampling protocol. In other words, the distances between reports are independent of the travel time model and its parameters \( \theta \). The likelihood function with distance-based sampling can then be
separated as \( p(d_{1:n}, \tau_{1:n} | x_0, \theta) = p(\tau_{1:n} | x_0, d_{1:n}, \theta) \cdot p_{sd}(d_{1:n} | \beta_{sd}) \), where the second term is just a constant in this context. We can factor the likelihood further into the travel time model of each observed travel time conditional on the preceding travel times. Hence,

\[
p(d_{1:n}, \tau_{1:n} | x_0, \theta) \propto p_T(\tau_1 | x_0, d_1, \theta) \cdots p_T(\tau_n | x_0, d_{1:n}, \tau_{1:n-1}, \theta)
\]  

(1)

Thus the likelihood function is directly proportional to the travel time model for successive parts of the path. This means that maximizing the likelihood given the observations corresponds directly to maximum likelihood estimation of the parameters of the travel time model.

### 2.3 Time-based Sampling

With a time-based sampling protocol, probe vehicle trajectories are sampled at certain time intervals according to some probability density function \( p_{st}(\tau_{1:n} | \beta_{st}) \), where \( \beta_{st} \) is some set of parameters, independent of the travel time model and parameters \( \theta \). The likelihood function is

\[
p(d_{1:n}, \tau_{1:n} | x_0, \theta) = p(d_{1:n} | x_0, \tau_{1:n}, \theta) \cdot p_{st}(\tau_{1:n} | \beta_{st}) \propto p(d_{1:n} | x_0, \tau_{1:n}, \theta). \]

Like for the travel times above, we can factor this into the travel distance model of each observed distance conditional on the preceding travel distances. Hence,

\[
p(d_{1:n}, \tau_{1:n} | x_0, \theta) \propto p_d(d_1 | x_0, \tau_1, \theta) \cdots p_d(d_n | x_0, d_{1:n-1}, \tau_{1:n}, \theta)
\]  

(2)

The likelihood of the observations is thus proportional to the travel distance model for consecutive time intervals. An important observation to make at this point is that the principle for the sampling protocol determines the proper formulation of the maximum likelihood estimation problem. For a given sequence of observations \((d_{1:n}, \tau_{1:n})\), the two likelihood functions will not in general be maximized by the same parameter values \( \theta \). If the incorrect likelihood function is used, for example if the travel time model is used even though the sampling is time-based, the parameter estimates may therefore be biased.

### 3. SPECIAL CASE: CONSTANT SPEED SEGMENTS

A special case of the general framework is when the path is partitioned into segments (for example, network links or some smaller unit), along which the travel speed of each vehicle is assumed to be constant. Furthermore, the speeds are assumed to be independent between segments. To formalize the assumptions of the model, the speed on each segment \( j \) is a stochastic variable with probability density function \( p_{ts,j}(v | \theta_j) \), where \( \theta_j \) is a set of parameters. The reciprocal of the travel speed is the travel time rate. The travel time rate on each segment \( j \) is a stochastic variable \( U_j \) with pdf \( p_{tr,j}(u | \theta_j) \). The pdf of the travel speed is related to that of the travel time rate as \( p_{ts,j}(v | \theta_j) = 1/v^2 p_{tr,j}(1/v | \theta_j) \).

We consider a single observation and assume without loss of generality that start position \( x \) is on segment 1 and end position \( x + d \) on segment \( m \geq 1 \). Given \( x \) and \( x + d \) and the length \( L_j \) of each segment, the distance traversed on each segment, denoted \( l_j \), is known. For \( m = 1 \) we have \( l_1 = d \); for \( m \geq 2 \), we have \( l_1 = L_1 - x, l_j = L_j \) for \( j = 2, ..., m - 1 \), and \( l_m = x + d - \sum_{i=1}^{m-1} L_i \). In contrast, the time spent on each segment \( j \), denoted \( \tau_j \), is not observed, expect in the case \( m = 1 \) where \( \tau_1 = \tau \). The sum of the times spent on each segment must equal the observed travel time, i.e., \( \sum_{j=1}^{m} \tau_j = \tau \).

#### 3.1 Distance-based Sampling

With distance-based sampling, the travel time \( T(x, d) \) between the two given positions \( x \) and \( x + d \) is a linear combination of the travel time rates on the covered segments, \( T(x, d) = \sum_{j=1}^{m} l_j U_j \). For any \( x \) and \( x + d \) on the same segment, i.e., for \( m = 1 \), we have \( T(x, d) = d U_1 \) and the travel time model:

\[
p_T(\tau | x, d, \theta) = \frac{1}{d} p_{tr,1}(\frac{\tau}{d} | \theta_1) \quad m = 1.
\]  

(3)
In the typical, low frequency cases that we consider, \( x \) and \( x + d \) are on different segments, i.e., \( m \geq 2 \). Only the total travel time \( \tau \) across all covered segments is observed, while the travel time rates \( U_j \) on the individual segments remain latent variables. To derive the travel time distribution between \( x \) and \( x + d \), the individual segment travel time models must therefore be convoluted. Let \( f_j = \tau_j/\tau \) denote the unobserved fraction of the total travel time \( \tau \) spent on segment \( j \), with \( f_m = 1 - \sum_{j=1}^{m-1} f_j \). Given fraction \( f_j \) the travel time rate on segment \( j \) is thus \( \tau f_j/l_j \), and the travel time model is:

\[
p_T(\tau \mid x, d, \theta) = \int_0^1 \int_0^{1-f_1} \cdots \int_0^{1-\sum_{j=1}^{m-2} f_j} \prod_{j=1}^m \frac{1}{l_j} p_{tr,j} \left( \frac{\tau}{l_j} f_j \mid \theta_j \right) df_1 df_2 \cdots df_{m-1} \quad m \geq 2.
\]  

(4)

### 3.2 Time-based Sampling

With time-based sampling, the travel distance \( D(x, \tau) \) from a given position \( x \) during a given time interval \( \tau \), given that the final position \( x + d \) on segment \( m \), is a linear combination of the travel speeds on the traversed segments, \( D(x, \tau) = \sum_{j=1}^m \frac{\tau}{U_j} \). Here, both the travel speeds \( 1/U_j \) and the times \( \tau_j \) (for \( m \geq 2 \)) are unobserved. However, the distance \( l_j \) traversed on each segment is known given the total travel distance \( \tau \). Hence, we actually observe every term in the sum, i.e., the stochastic variables \( D_j(x, \tau) = \tau/U_j \). For \( m = 1 \), we have \( D(x, \tau) = \tau/U_1 \) and the travel distance model

\[
p_D(d \mid x, \tau, \theta) = \frac{1}{\tau U_1} \left( \frac{d}{\tau} \right) \frac{\tau}{d^2} p_{tr,1} \left( \frac{\tau}{d} \mid \theta_1 \right) \quad m = 1
\]  

(5)

For \( m \geq 2 \), the travel distance model is given by the joint pdf of the segment travel distances \( D_j(x, \tau) \), \( j = 1, \ldots, m \). Since the times spent on each segment is unobserved, they must be integrated out similar to in equation (4). Expressed in terms of travel speeds and travel time rates, respectively, the travel distance model is then

\[
p_D(d \mid x, \tau, \theta) \equiv \int_0^1 \int_0^{1-f_1} \cdots \int_0^{1-\sum_{j=1}^{m-2} f_j} \prod_{j=1}^m \frac{1}{l_j} p_{tr,j} \left( \frac{\tau}{l_j} f_j \mid \theta_j \right) df_1 df_2 \cdots df_{m-1} \quad m \geq 2
\]  

(6)

We can draw several important conclusions from the analysis above. First, whether sampling is distance-based or time-based does not affect likelihood-based estimation if observations cover only a single segment. This is intuitive since the speed by assumption is constant along the segment, so that it does not matter when and where the vehicle is sampled. Second, if observations cover multiple segments, the sampling protocol influences likelihood-based estimation, in contrast to the single segment case. Comparing equations (4) and (6), we see that the feasible travel time allocations \( (f_1, \ldots, f_m) \) are weighted uniformly for distance-based sampling, while they are weighted proportionally to \( \prod_{j=1}^m f_j \) for time-based sampling. In the latter case, the weights are concentrated towards the center of the simplex \( \Delta = \{ (f_1, \ldots, f_m) \mid \sum_{j=1}^m f_j = 1 \} \) of feasible allocations. The fact that the allocations must be weighted in this way implies that time-based sampling otherwise overrepresents unequal allocations in which a large portion of the travel time is spent on a small number of segments. This impacts the likelihood of the observations through the product of the segment travel time rate probability density functions for each allocation. Thus, the likelihood function in general attains its maximum at a different set of parameters \( \theta \). The sign and magnitude of this bias depends on the characteristics of the segment pdfs.

### 4. Consistency and Bias of Estimators

We now study numerically the impact of distance-based versus time-based sampling protocols for travel time estimation bias and consistency for the constant speed segment model in Section 3. We consider \( m = 2 \) segments and \( K \) independently drawn observations \( \{d^k, \tau^k\}_k \), \( k = 1, \ldots, K \), each starting at a given position \( x^k \) on segment 1 and ending at position \( x^k + d^k \) on segment 2. Given the known transition point \( L_1 \) between the two segments, the distances traveled on the two segments are
$l^k_1 = L_1 - x^k$ and $l^k_2 = x^k + d^k - L_1$, respectively. Given the generated data, our aim is to estimate the parameters $\theta = (\theta_1, \theta_2)$ of the travel time rate probability density functions of the two segments using maximum likelihood estimation. From Sections 3.1 and 3.2 we know the likelihood function is:
\begin{equation}
L(\theta) \propto \prod_{k=1}^{K} \int_0^1 f(1-f)p_{tr,1}(\frac{t^k}{l^k_1} \mid \theta_1) \cdot p_{tr,2}(\frac{t^k}{l^k_2}(1-f) \mid \theta_2) df,
\end{equation}
(7)
\begin{equation}
L(\theta) \propto \prod_{k=1}^{K} \int_0^1 f(1-f)p_{tr,1}(\frac{t^k}{l^k_1} \mid \theta_1) \cdot p_{tr,2}(\frac{t^k}{l^k_2}(1-f) \mid \theta_2) df,
\end{equation}
(8)
in the cases of distance-based and time-based sampling, respectively. In each case we compare the parameter estimates obtained from the two likelihood formulations (7) and (8) with each other and with the known, true parameter values used to generate the observations, in order to investigate the errors introduced by misspecifying the likelihood function.

### 4.1 Travel Time Rate Distributions and Data Generation

To make sure that the conclusions from the experiments do not hinge too much on the properties of a particular distribution, we consider three different distributions for the segment travel time rates: the (1) lognormal, (2) gamma, and (3) inverse gamma distributions. All three distributions have been found in the literature to be suitable representations of empirical speed or travel time distributions. To make comparisons as easy as possible, we parameterize each distribution in terms of its mean $\mu$ and variance $\sigma^2$. The three distributions are very similar when the mean is high and the variance is low, while the differences are bigger with lower mean and higher variance. For all parameter combinations, the gamma distribution has the lowest skewedness and the inverse gamma distribution has the highest skewedness (for the inverse gamma distribution the skewedness is infinite for the combination $\mu = 2, \sigma^2 = 6$).

To limit the number of parameter configurations, the mean parameters $\mu_j, j = 1, 2$, of the two segment travel time rate distributions are given two levels; a baseline value $\mu_B = 2$ and a high value $\mu_H = 10$. The variance parameters $\sigma^2_j$ can take a baseline value $\sigma^2_B = 1.5$ or a high value $\sigma^2_H = 6$. These are chosen to be representative of the scale and shape of real travel time rate distributions while demonstrating some interesting features in the estimation results. For each segment, there are thus four different parameter combinations, which we may denote compactly as BB ($\mu_j = 2, \sigma^2_j = 1.5$), HB ($\mu_j = 10, \sigma^2_j = 1.5$), BH ($\mu_j = 2, \sigma^2_j = 6$), and HH ($\mu_j = 10, \sigma^2_j = 6$). A full parameter configuration is specified by a combination of parameter levels for both segments. For example, HHBB ($\mu_1 = 10, \sigma^2_1 = 6, \mu_2 = 2$ and $\sigma^2_2 = 1.5$).

For each observation $k = 1, \ldots, K$ the travel time rate $u^k_j$ of the vehicle on each of the two segments is generated by a random draw from an associated distribution $p_{u,j}(u \mid \mu_j, \sigma^2_j)$ with mean and variance parameters according to a specific configuration. When sampling by distance, the travel distance spent on segment 1, $l^k_1$, is drawn from a uniform(0,1) distribution, and the travel distance on the second segment is $l^k_2 = 1 - l^k_1$. The observed travel time $\tau^k$ is finally calculated as $\tau^k = l^k_1 u^k_1 + (1 - l^k_1) u^k_2$. When sampling by time, the unobserved time spent on segment 1, $\tau^k_1$, is drawn from a uniform(0,1) distribution, and the observed travel distances on the two segments are then calculated as $l^k_1 = \tau^k / u^k_1$ and $l^k_2 = (1 - \tau^k_1) / u^k_2$.

### 4.2 Results

The parameters of the segment travel time rate pdfs were estimated by numerical integration and maximization of the likelihood functions (7) and (8). Maximum likelihood estimation ensures that the estimators are consistent if the correct likelihood formulation is used, that is, if the travel time model is used with distance-based sampling and the travel distance model is used with time-based sampling.
Figure 1 shows parameter estimates and standard errors for one of the segments in the symmetric parameter configuration BBBB ($\mu_1 = \mu_2 = 2, \sigma_1^2 = \sigma_2^2 = 1.5$). To the left the observations are sampled by distance, which means that the travel time model provides the correct likelihood formulation. As can be seen, the parameter estimates from the travel time (TT) model converge to the true parameter values. The estimates from the travel distance (TD) model also converge, but not to the true parameter values. The estimator is therefore not consistent, but systematically overestimates both the mean and the variance. To the right in Figure 4 the observations are sampled by time, which means that the travel distance model provides the correct likelihood formulation. In accordance with theory, the parameter estimates from the travel distance (TD) model converge to the true parameter values. The estimates from the travel travel (TT) model, however, converge to values that underestimate both the mean and the variance. In other words, the estimator is not consistent.

We next analyze the sign and magnitude of the estimation bias under different parameter configurations and distributions; more detailed results are presented in Jenelius and Koutsopoulos (2012). We focus here on the case with time-based sampling; for distance-based sampling, the signs of the biases are the opposite throughout. As expected, there is no evidence of bias in the parameter estimates using the correct travel distance model formulation. With the travel time model, meanwhile, there is statistically significant bias for all distributions and parameter configurations. To begin with, the signs of the biases in each parameter configuration are almost always the same for all three distributions. The magnitudes vary, however: in general the bias for the mean parameters appears to be least severe for the gamma distribution and most severe for the inverse gamma distribution, while the order is reversed for the variance parameters.

Configuration BBBB is symmetric between the segments with relatively skewed travel time rate distributions. Here the travel distance model underestimates both the mean (7%-11% depending on the distribution) and the variance (8%-32%) of the distributions; see also Figure 1. Configuration HBHB is also symmetric but with almost non-skewed distributions; here the bias is much more moderate. Thus, the magnitude of the bias appears to increase with the skewedness of the travel time rate distributions. Configurations HBBB, HHBB and BHHB are non-symmetric with one segment having a higher mean than the other, while the variance is the same, higher and lower, respectively. For these three configurations, both parameters of the high-mean segment are overestimated (mean 0-22%, variance 7%-250%) while the parameters of the low-mean segment are underestimated (mean 10%-40%, variance 17%-51%). The bias is least severe for the HBBB configuration and most severe for the BHHB configuration. Configuration BHBB, finally, is non-symmetric with one segment having a higher variance than the other, while the means are the same. Here the mean and the variance of the high-variance segment are underestimated (mean 21%-50%, variance 58%-66% depending on

![Distance-based sampling](image1.png)

![Time-based sampling](image2.png)
the distribution), while the parameters of the low-variance segment are overestimated for the lognormal and gamma distributions (mean 4%-25%, variance 15%-74%) and somewhat underestimated (mean 1%, variance 6%) for the inverse gamma distribution. This shows that the characteristics of the travel time rate distributions can influence the sign and magnitude of the estimation bias in non-trivial ways.

Figure 2 illustrates how the biased parameter estimates impact on the travel time rate distributions in the case of the lognormal distribution and the BHHB parameter configuration ($\mu_1 = 1.473$, $\sigma_1^2 = 3.472$, $\mu_2 = 11.095$, $\sigma_2^2 = 3.030$). The histograms show the unobserved empirical distributions. The estimated distributions using the travel distance (TD) model match the empirical distributions closely, while the estimated distributions using the incorrect travel time (TT) model deviate significantly.

Figure 2: Empirical relative frequencies and estimated probability density functions of segment travel time rates. Time-based sampling, lognormal distributions and parameter configuration BHHB ($\mu_1 = 2$, $\sigma_1^2 = 6$, $\mu_2 = 10$, $\sigma_2^2 = 1.5$). Sample size $K = 20,000$.

5. CONCLUSION

The paper analysed the role of the sampling protocol for travel time estimation from low frequency probe data. We presented a general likelihood formulation of the problem and highlighted the fundamental difference between distance-based and time-based sampling protocols. Estimates of the travel time distribution parameters may be biased if the sampling protocol is not properly considered in the estimation. The paper next considered the important special case in which the path is partitioned into segments with constant, independent travel speeds. The bias of the estimators using the incorrect likelihood formulations was investigated for several different travel time distribution forms (lognormal, gamma and inverse gamma) and parameter configurations. When sampling is treated as distance-based while being time-based in reality, the results indicate that the bias in the mean and variance parameters is small if the travel time rate distributions are similar among the segments and roughly symmetric. However, the bias increases towards underestimation as the distributions become more skewed to the right. When one segment has a higher true mean than the other, the incorrect model overestimates the mean and variance of the high-mean segment and underestimates the parameters of the low-mean segment. The results suggest that the principles of the sampling protocol used by the probe vehicles must receive more attention than has been the case until now (with more awareness of the magnitude and sign of the error due to misspecification). The statistical model should incorporate the sampling protocol. Further work is needed to develop estimators for the mixed protocols based on both distance and time protocols that are sometimes used in practice.
REFERENCES


