A Nonparametric Bayesian Estimation of Freeway Capacity Distribution from Censored Observations

Eren Erman Ozguven, M.Sc.
Graduate Research Assistant,
Department of Civil and Environmental Engineering,
Rutgers, The State University of New Jersey,
623 Bowser Road, Piscataway, NJ 08854 USA,
Tel: (732) 445-0597 / 119
Fax: (732) 445-0577
e-mail: ozguvene@rci.rutgers.edu

Kaan Ozbay, Ph.D. (Corresponding Author)
Associate Professor,
Department of Civil and Environmental Engineering,
Rutgers, The State University of New Jersey,
623 Bowser Road, Piscataway, NJ 08854 USA,
Tel: (732) 445-0579 / 127
Fax: (732) 445-0577
e-mail: kaan@rci.rutgers.edu

Word count: 5991 + 3 Figures + 3 Tables = 7491
Abstract: 248

Resubmission Date: November 15, 2007

Paper Submitted for Presentation and Publication at the Transportation Research Board’s 87th Annual Meeting, Washington, D.C., 2008
ABSTRACT

The usefulness and effectiveness of survival analysis in transportation and traffic engineering studies with incomplete data have been previously studied in (1) where the Kaplan-Meier estimate is proposed for determining the traffic capacity distribution. However, the well-known estimators like Kaplan-Meier and Nelson-Aalen have several disadvantages that make it difficult to obtain the traffic capacity distribution. First of all, both estimators are not defined for all values of traffic flows possible. That is, the maximum flow followed by a breakdown defines the final point of the estimated distribution curve. Therefore, parametric fitting tools have to be applied to obtain the remaining portion of the curve. Moreover, the discontinuity and non-smoothness of the Kaplan-Meier and Nelson-Aalen estimates make it difficult to ensure the robustness of the estimation. In this paper, firstly Kaplan-Meier and Nelson-Aalen nonparametric estimators are used to obtain the traffic capacity function of four freeway sections. Then, a Bayesian nonparametric estimator, which is shown to be a Bayesian extension of Kaplan-Meier estimator, has been introduced for estimating the capacity distribution. This estimator assumes a Dirichlet process prior for the survival function under the minimization of a squared-error loss function. The results indicate that the curves obtained by using the Bayesian estimation method are smoother than the other estimator curves. This smoothness also ensures the continuity in the vicinity of censored observations. Furthermore, the Bayesian estimates can be obtained for any traffic flow regardless of the availability of data only for certain ranges of observations (including censored data).
INTRODUCTION

Several researchers have recently recognized the necessity of characterizing highway capacity as a stochastic phenomenon. This is because it is clear that the traffic flow is a stochastic process and thus it is necessary to add a random component to deterministic parameters. This can be achieved by the introduction of stochasticity in the estimation of the parameters of a traffic model mainly to obtain a more realistic representation of observed data which will in turn improve understanding of real-life situations.

Lorenz and Elefteriadou proposed a new definition for the traffic capacity by including probabilistic concepts (2), (3). Persaud et al. (4) also studied the probabilistic breakdown phenomenon in freeway traffic. The results convincingly confirmed that breakdown is stochastic in nature. Chung et al. (5), and Hall and Agyemang-Duah (6) worked on the freeway capacity drop and the definition of the capacity whereas Cassidy and Bertini (7), and Smilowitz et al. (8) studied some important traffic features with respect to the bottleneck flows.

An analytical but deterministic method has been proposed by Cassidy and May (9) for estimating the traffic capacity and level of service of major freeway weaving sections. Wang et al. (10), and Lertworawnich and Elefteriadou (11) also worked on the capacity of weaving sections. Another approach that employs neural network techniques for estimating traffic capacity of weaving segments has been proposed by Awad (12). Yeon et al. (13), and Kuhne et al. (14) have also worked on the development of a stochastic theory of freeway traffic. Kuhne and Mahnke (15) have derived a formulation for obtaining the cumulative distribution for traffic breakdowns defining the breakdown as a car cluster formation process. They have also used Weibull curves as fitting curves to match a broad variety of cumulative distribution functions (16).

Banks (17), (18), (19), (20) conducted several important studies to develop a new approach to bottleneck capacity analysis. Although these works provide very useful insights, they do not attempt to determine a probability distribution for the capacity. Polus and Pollatschek (21) tried to explore the stochastic nature of freeway capacity by fitting speed-flow diagrams based on empirical data. Laval (22) came up with approximate formulas for highway capacity using the presence of slow vehicles.

However, none of these research studies focus on the use of statistically robust methodologies to estimate the distribution of the freeway capacity. Recently, Brilon et al. (1) used a practical estimation method based on an analogy to the statistics of lifetime analysis to obtain the distribution of freeway capacity. They used Kaplan-Meier estimator to obtain the capacity distribution function, and tried to estimate the parameters of this distribution with various plausible functions like Weibull, Normal and Gamma distributions, where Weibull function appeared to be the best fit. The lifetime method had been proposed by van Toorenburg (23) and discussed by Minderhoud et al. (24). Regler (25), then, applied the capacity analysis in (1) to other freeway sections. Ozbay and Ozguven (26), improving the method given in (1), studied the effects of spatial and temporal differences of the freeway section on the capacity distribution. On the other hand, Bernard and Axhausen (27) worked on designing a load concept that treats flow as a normal distributed random variable.
Based on this growing body of literature related to this topic, we can conclude that to obtain a robust stochastic model of the freeway capacity, it is important to address several issues:

- A theoretically sound model that captures the stochastic nature of freeway capacity.
- As data related problems such as insufficient and unreliable observations are frequently encountered in this context (1), the appropriate consideration of censored data has great importance in order to obtain a reliable distribution of the freeway capacity.
- The robust estimation of the model parameters incorporating the spatial and temporal effects of the data.

In this paper, we propose to address these important issues as follows:

- An analogy to lifetime data statistics will be adopted, and the survival analysis will be used to capture the stochastic nature of the traffic capacity, improving the modeling methodology first proposed in (1).
- In addition to the conventional nonparametric estimators, a Bayesian estimation approach, which is useful while obtaining the capacity distribution in the presence of insufficient and unreliable data, will be employed for the first time.
- Different freeway sites will be studied to determine the effects of the spatial and temporal variations on the estimated parameters.

In summary, a survival model for estimating the probabilistic distribution function of highway capacity will be developed using both conventional and Bayesian estimators mainly to address the issue of censored data.

SURVIVAL ANALOGY

As indicated by previous studies (1-27), a constant value for traffic capacity cannot be used to accurately represent the real life situations. Actually, the capacity of a freeway should be regarded as a random variable which has its own distribution as clearly stated in (1) and (15). Therefore, the problem is how to determine the capacity distribution of the traffic volumes. In this paper, this problem is addressed by using survival analogy which was first proposed by van Toorenburg (23) and discussed by Minderhoud et al. (24). The conventional nonparametric estimation approach that is used to compare with the Bayesian estimation approach proposed in this paper is first presented by Brilon et al. (1). They presented a lifetime analysis application for determining freeway capacities without any assumption for the estimation of the capacity distribution function. Their capacity distribution function will reach a probability value of 1 only if the maximum observed volume value is followed by a breakdown in the freeway. This simple yet very important fact requires additional attention to the problem of analyzing the data with right censored observations. As in the lifetime analysis, the breakdown point is unknown for the censored data points as there is a free flowing traffic at that volume. On the other hand, this does not suggest that traffic will never breakdown at that point. It only suggests that it might not breakdown during the limited observation period. This problem is also encountered in the study of Brilon et al. (1). They could not obtain a complete distribution function, despite the large size of a one-year sample, since
the highest flow values were not followed by a breakdown. Hence, it can be suggested that the distribution function is incomplete if only the available data points are used. However, more robust and realistic parameter estimates can be obtained by using advanced techniques developed for the data with censored information (see (28) for detailed information) such as:

- Kaplan-Meier Estimator,
- Nelson-Aalen Estimator, and
- Bayesian Estimator with Dirichlet process as a prior.

These techniques will be briefly described in the next sections.

**Kaplan-Meier Estimation (Product Limit Method)**

Basically, the survival function describes the probability that a variate $T$ takes on a value greater than a number $t$ in the lifetime analysis. That is, we can define the Survival Function $S(t)$, as the probability a unit survives beyond time $t$:

$$ S(t) = P(T > t) $$

The standard and efficient estimator of the survival function is Kaplan-Meier estimator (Product Limit Estimator) proposed by Kaplan and Meier (29). The Kaplan-Meier method is so widely used and so well known, that in research papers survival curves are more often called Kaplan-Meier curves. The Kaplan-Meier estimator can be used to calculate the survival function as follows:

$$ \hat{S}_K(t) = \prod_{j:t_j<t} \frac{n_j-d_j}{n_j} $$

where

- $\hat{S}_K(t)$: estimated Kaplan-Meier survival function,
- $n_j$: number of individuals who are at risk at $t_j$,
- $d_j$: number of deaths at time $t_j$.

The estimator is simply a step function with jumps at the observed failure times. The size of the jumps depends on the number of events observed and censored observations.

The so-called lifetime (unreliability) distribution is simply

$$ F(t) = 1 - S(t) = P(T \leq t) $$

To perform the lifetime analogy, the traffic breakdown should be considered as a failure event, like the failure of a bulb. If $C$ and $q$ represent capacity and traffic volume, respectively, then the capacity distribution function $F(q)$ can be written as:

$$ F(q) = 1 - S(q) = P(C \leq q) $$

---

TRB 2008 Annual Meeting CD-ROM
Paper revised from original submittal.
Ozguven E. E., Ozbay K.

Using the Kaplan-Meier approach, the traffic capacity distribution function can be obtained using the lifetime analogy between traffic capacity and the lifetime. It will be referred as the “capacity function” from now on.

\[
P_K(q) = 1 - \hat{S}_K(q) = 1 - \prod_{i:q_i<q} \frac{k_i-d_i}{k_i}
\] (5)

where

- \(P_K(q)\): estimated Kaplan-Meier capacity distribution function,
- \(\hat{S}_K(q)\): estimated Kaplan-Meier survival function,
- \(q\): traffic volume (vehicles per 5 minutes),
- \(q_i\): traffic volume in interval \(i\) (vehicles per 5 minutes),
- \(k_i\): number of intervals that do not cause breakdown just before \(q_i\),
- \(d_i\): number of breakdowns at a volume of \(q_i\).

**Nelson-Aalen Estimation**

The cumulative hazard function, which is defined as

\[
H(t) = -\ln [S(t)]
\] (6)

can also be used for estimating the survival function.

To estimate the cumulative hazard rate, Nelson-Aalen estimator can be used. This estimator was first suggested by Nelson in 1972 (30) in a reliability context and re-derived by Aalen in 1978 (31) using modern counting process techniques. In the context of traffic capacity problem, the Nelson-Aalen estimator is formulated as follows:

\[
\hat{H}_N(q) = \sum_{i:q_i<q} \frac{d_i}{k_i}
\] (7)

where

- \(\hat{H}_N(q)\): estimated Nelson-Aalen cumulative hazard function,
- \(k_i\): number of intervals that do not cause breakdown just before \(q_i\),
- \(d_i\): number of breakdowns at a volume of \(q_i\).

Based on this estimator for the cumulative hazard function, an alternate estimator for the survival function can be obtained as,

\[
\hat{S}_N(q) = \exp \left[ -\hat{H}_N(q) \right]
\] (8)

Then, the capacity distribution estimator can be obtained as

\[
P_N(q) = 1 - \hat{S}_N(q) = 1 - \exp \left[ -\left( \sum_{i:q_i<q} \frac{d_i}{k_i} \right) \right]
\] (9)

where

- \(P_N(q)\): estimated Nelson-Aalen capacity distribution function,
- \(\hat{S}_N(q)\): estimated Nelson-Aalen survival function,
Ozguven E. E., Ozbay K.

$q$: traffic volume (vehicles per 5 minutes),
$q_i$: traffic volume in interval $i$ (vehicles per 5 minutes),
$k_i$: number of intervals that do not cause breakdown just before $q_i$,
$d_i$: number of breakdowns at a volume of $q_i$.

Bayesian Estimation

The pioneering study on the Bayesian estimation of a survival function with right censored data was conducted by Susarla and Ryzin (32), (33), (34), Phadia and Ferguson (35), (36). The idea is to obtain the nonparametric estimation of the survival function and therefore the cumulative lifetime distribution function when the observations are censored from the right using a Bayesian approach. In the presence of censored data, Bayesian analysis appears to be a very useful tool. (see (37) and (38) for details). The difference of Bayesian statistics from Classical statistics is that it incorporates the prior knowledge along with a given set of current observations so as to make statistical inferences. The prior information could come from operational or observational data, from previous experiments or from engineering knowledge.

In this paper, a nonparametric Bayesian estimation of the traffic capacity analysis will be given under the squared-error loss notion with Dirichlet process priors. With this information, the posterior distribution of the cumulative capacity distribution function will be derived. The resulting Bayesian estimator under Dirichlet process priors has been shown to be very useful and efficient for the survival analysis described by Susarla and Ryzin in (32) especially right censored observations are present. This prior process is perhaps the most important and widely used prior in nonparametric Bayesian inference.

To obtain an estimate of the survival function, a loss function has to be specified as a base decision rule. This squared error loss function is given as

$$L(\hat{S}, S) = \int_0^\infty \left( \hat{S}(q) - S(q) \right)^2 \, dw(q)$$

(10)

where $w$ is a nonnegative weight function. This loss function simply presents the weighted integrated difference between the true value of the survival function and the estimated one.

The Dirichlet distribution, introduced by Ferguson (39) in his seminal paper, is selected as the prior that will lead to closed form estimates for the survival function using the loss function given in Equation (10). That is, we assume that the survival function is sampled from a Dirichlet process with a parameter function $\alpha$, which is a kind of probability measure. This prior is chosen because it is a conjugate prior for the survival function. A conjugate prior is one which, when combined with the likelihood and normalized, produces a posterior distribution which is of the same type as the prior. This makes the computation of the posterior particularly simple.

The Bayesian estimator of the survival function with the Dirichlet prior is given as
Ozguven E. E., Ozbay K.

\[ \hat{S}_B(q) = \left( \frac{\alpha(q, \infty) + n_{i+1}}{\alpha(0, \infty) + n_{\text{total}}} \right) \prod_{k=1}^{i} \frac{\alpha(q_k, \infty) + n_{i+1} + c_i}{\alpha(q_k, \infty) + n_{i+1}} \]  \hspace{1cm} (11)

where
\[ \hat{S}_B(q): \text{estimated Bayesian survival function}, \]
\[ n_{i+1}: \text{number of observations in the interval } i + 1, \]
\[ n_{\text{total}}: \text{total number of observations}, \]
\[ c_i: \text{number of censored observations in the interval } i. \]

This estimator has been shown to be a Bayesian extension of the Kaplan-Meier nonparametric estimator in (32). To assign a prior distribution to the survival function, the parameter function is taken to be of the form \( \alpha(q, \infty) = \beta S_0(q) \) where \( S_0(q) = \alpha(0, \infty) \) is the prior guess at the survival function and \( \beta \) is a kind of measure that indicates how much weight to put on the prior guess (prior mass). With the Dirichlet prior, we take
\[ S_0(q) = \alpha(0, \infty) = e^{-\theta q} \]  \hspace{1cm} (12)

where \( \theta \) can be heuristically calculated using the Kaplan-Meier data of breakdown intervals. Therefore, we can write
\[ \alpha(q, \infty) = \beta S_0(q) = \beta \alpha(0, \infty) = \beta e^{-\theta q} \]  \hspace{1cm} (13)

Now, all the variables in Equation (11) have been introduced for the estimation procedure. Of course, the careful selection of parameters has utmost importance for an efficient Bayesian estimation. This issue will be discussed in the model application section of this paper.

Finally, the estimated capacity function via Bayesian analysis, \( \hat{F}_B(q) \), can be calculated by
\[ \hat{F}_B(q) = 1 - \hat{S}_B(q) = 1 - \left[ \left( \frac{\alpha(q, \infty) + n_{i+1}}{\alpha(0, \infty) + n_{\text{total}}} \right) \prod_{k=1}^{i} \frac{\alpha(q_k, \infty) + n_{i+1} + c_i}{\alpha(q_k, \infty) + n_{i+1}} \right] \]  \hspace{1cm} (14)

As stated in Suzarla and Ryzin (32), “Bayesian estimator is a continuous function between the distinct breakdown points, has jumps at these breakdown points and it is continuous between the breakdowns”. Note that the function can be calculated for all the flows including the censored and unobserved flows. For large samples, the Bayesian estimator naturally reduces to the Kaplan-Meier estimator, so that the prior information plays no role in estimating the capacity distribution function. For small samples, on the other hand, the prior function will dominate, and the estimator will be close to the prior guess adopted for the capacity function.

To sum up, the use of Bayesian estimation method for obtaining the traffic capacity is useful when the following conditions are met:
- There is substantial amount of censored data in the traffic flow observations.
- There is not sufficient data to be able to study the behavior of the traffic capacity accurately and efficiently.
- Traffic observations are not 100% reliable.
Ozguven E. E., Ozbay K.

In the context of the probabilistic modeling of the capacity problem mentioned in previous studies, these conditions are in fact all present. Therefore, there is definitely a need for using a Bayesian estimator.

DATA AND BASIC ASSUMPTIONS

Two freeway sites used in this study are sections of I-880 Highway, California, and Gardiner Expressway, Toronto/Canada. Data for the first section were obtained from a California PATH research project (40). Two sections have been studied for this highway, with 4 and 5 lanes, respectively. The second data set was obtained from Gardiner Expressway which has been used in previous capacity related studies (7), (10). There are two sections studied for this expressway, with 3 and 4 lanes.

Short observation intervals are shown to be useful for the analysis of highway traffic capacity. Otherwise, the causal relationship between traffic volume and breakdown can be too weak. For instance, 1-hour traffic counts are meaningless and not sufficient to represent the aforementioned relationship. The observation period, therefore, is taken to be 5 minute for the speed and volume values, as also suggested in (41).

The concept of determining how to define the breakdown is very crucial in terms of the accuracy and efficiency of the proposed methodology. Therefore, the flow values are classified, using a 5-minute interval denoted as \( i \), and shown in TABLE 1. The only difference between the definitions given by Brilon et al. (1) is that another interval type, namely D3, is added.

<table>
<thead>
<tr>
<th>Interval</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>Traffic is fluent in time interval ( i ), but the observed volume causes a breakdown, i.e. the average speed drops below the threshold speed in the next time interval ( i + 1 ).</td>
</tr>
<tr>
<td>C</td>
<td>Traffic is fluent in interval ( i ) and in the following interval ( i + 1 ). This interval ( i ) contains a censored value. Its information is that the actual capacity in interval ( i ) is greater than the observed volume ( q_i ).</td>
</tr>
<tr>
<td>D1</td>
<td>Traffic is congested in interval ( i ), i.e. the average speed is below the threshold value. As this interval ( i ) provides no information about the capacity, it is disregarded.</td>
</tr>
<tr>
<td>D2</td>
<td>Traffic is fluent in interval ( i ), but the observed volume causes a breakdown. However, unlike classification B, traffic is congested at a downstream cross section during interval ( i ) or ( i - 1 ). In this case, the breakdown at the observation point is supposed to be due to a tailback from downstream. As this interval ( i ) does not contain any information for the capacity assessment at the observation point, it is disregarded.</td>
</tr>
<tr>
<td>D3</td>
<td>Traffic data may include low traffic flow values having a wide range of corresponding speed values. As they have nothing to do with the breakdown, they are disregarded within a certain threshold for the data.</td>
</tr>
</tbody>
</table>

There are several approaches available for the definition of freeway “breakdown” (see (1) for details). One approach is to determine the breakdown value by using the speed difference between consecutive intervals. Another approach is including the
average space mean speed to detect the breakdown point. This is a logical approach as the traffic breakdown is usually followed by a significant amount of reduction in speed. The latter approach will be used throughout this paper. Firstly, the idea is to consider an interval $i$ as a breakdown interval whenever the speed falls below a threshold value in the next interval. A threshold speed of 70 km/h is selected for both freeways studied in this paper. This threshold selection is consistent with the thresholds selected in Brilon et al. (1) and Kuhne and Mahnke (15), 70 km/h and 75 km/h, respectively. Of course, this threshold may be different for other types of roadways or conditions. Now, it is appropriate to briefly explain the concept of breakdown intervals.

First of all, every interval that is followed by a speed lower than the threshold value cannot be considered as a breakdown interval (B). That is, if each congested volume is considered as a B-value, we would be considering all the breakdown flows again and again until the end of the analysis after the failure of a specific road segment. Hence, the first interval that causes breakdown is considered as a B-value, and the lower-speed intervals following that B-interval are disregarded (D1-values).

Before using the method proposed in this paper, one should be aware of some basic concepts like the censored data. C-values represent the right-censored data in the analysis. That is, C-values are accompanied by speed values higher than the threshold speed and they do not cause breakdown.

Finally, D2-values are similar to B-values; however they are due to a breakdown in a downstream cross-section. That is, the reason that the speed goes below the threshold value is not the interval itself, but rather downstream congestion. D3-values, on the other hand, include low traffic flow values that do not contribute to the study. Therefore, both interval values are disregarded.

MODEL APPLICATION

The different estimating techniques briefly described in Section 2 are applied to the data obtained from I-880 Highway in California, and Gardiner Expressway in Canada. To apply the Bayesian estimation, the parameter $\alpha$ of the Dirichlet prior and $\beta$ in Equation (13) should be carefully selected.

In the model application, the $\alpha$ parameters are calculated based on the following heuristic argument given in (32) for each freeway section separately. Note that if one were to consider the Bayesian estimator without any data, the estimator would simply be $\hat{S}_B(q) = e^{-\theta q}$. The idea is to force this estimator to satisfy the following equation:

$$e^{-\theta M} = quantile(P_K, 0.5)$$

(15)

where $M:quantile(q_K, 0.5)$, and $P_K$ is the estimated Kaplan-Meier breakdown probability vector. Here, $M$ represents the 0.5 quartile of the Kaplan-Meier estimated flow vector. Using this heuristic argument, and by the double use of the observations, $\theta$ is obtained for each section. This approach is found to be useful in (32) as other values of $\theta$ tend to pull the Bayesian estimator away from the Product-Limit estimator.

For the $\beta$ values, the idea is to define the prior mass on the sample size for the data. The larger $\beta$ becomes, the smoother the Bayesian estimator will be in comparison to the Kaplan-Meier estimator. This indicates that the jumps at the breakdown points will be smaller. The extreme case is to let $\beta \to \infty$ in which case the Bayesian estimator reduces
to $e^{-\theta q}$ with the calculated $\theta$ values in Equation (14). Several values of $\beta$ are plotted to see the difference, roughly speaking, 0.15, 0.25, 0.5 and 1 are taken as the prior mass on the observations. So, 0.15, 0.25, half and equal the sample size are used in the calculations. The values for the sections studied in this paper are given in TABLE 2.

<table>
<thead>
<tr>
<th>Freeway</th>
<th>Section</th>
<th>$\alpha$</th>
<th>$\beta = (0.15, 0.25, 0.5, 1) \times \text{Sample Size}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>I-880 Highway</td>
<td>Section 1 (4 Lanes)</td>
<td>0.0010</td>
<td>(127, 212, 423, 846)</td>
</tr>
<tr>
<td></td>
<td>Section 2 (5 Lanes)</td>
<td>0.0006</td>
<td>(74, 124, 247, 494)</td>
</tr>
<tr>
<td>Gardiner Expressway</td>
<td>Section 1 (3 Lanes)</td>
<td>0.0023</td>
<td>(397, 663, 1325, 2649)</td>
</tr>
<tr>
<td></td>
<td>Section 2 (4 Lanes)</td>
<td>0.0009</td>
<td>(372, 621, 1241, 2482)</td>
</tr>
</tbody>
</table>

The choice of $\beta$ can be done by using the guidelines in (32). In fact, value of $\beta$ solely depends on the sample size of the data obtained for the sections. The variation of $\beta$ values shown in TABLE 2 is due to the difference of the number of observations for each section especially after applying Brilon’s filter (see TABLE 1) to each section’s data set. First, various values of $\beta$ are tried and plotted. A comparison of the curves with different $\beta$ parameters for the section 1 of I-880 Highway is given in FIGURE 1. Out of these values, 0.15 appears to give a plot close to the Kaplan-Meier estimator. As seen from FIGURE 1, other values of $\beta$ tend to pull the Bayes estimator away from the Kaplan-Meier estimator (see (32) for detailed information). Therefore, the figures are plotted for all sections with that value. FIGURE 2 and FIGURE 3 represent the estimated capacity distribution curves for the sections of I-880 Highway and Gardiner Expressway, respectively.
FIGURE 2 Estimation Comparison of I-880 Highway (Figure 2a) for Section 1 (Figure 2b) for Section 2
FIGURE 3 Estimation Comparison of Gardiner Expressway (Figure 2a) for Section 1 (Figure 2b) for Section 2
DISCUSSIONS

As expected, Bayesian estimator is observed to be smoother than the Kaplan-Meier estimator in the sense that the jumps at censored observations are not as large for the Bayesian estimator. Moreover, the unobserved (censored) observations can be represented in the Bayesian estimating curve. That is, the maximum value of the capacity distribution function for the Kaplan-Meier and Nelson-Aalen estimates will only reach 1 if the maximum observed volume \( q \) belongs to the B-breakdown interval (followed by a breakdown). This is the case for I-880 Highway section. On the other hand, since the censored observations prevent one to find the upper tail of the distributions, the curve is not complete for other sections. The Bayesian estimator solves this problem and gives a complete curve for the estimated capacity distribution. For instance, a comparative idea is given in TABLE 3 to observe the fact that the Bayesian estimator exists for all the flow values and can be plotted till the breakdown probability of 1. This situation is encountered by Brilon et al. (1), where they stated that despite the large sample they used, no complete distribution function could be estimated since the highest flow values observed were not followed by breakdown. Therefore, the capacity distribution function terminated at a value lower than 1 at its upper end in their case. Similar arguments are also presented by Kuhne and Mahnke (15). They found out that the distribution dies out after the maximum flow with a breakdown pointing the fact that all possible breakdowns already happened below that flow value. Therefore, the lower and upper tails of the capacity distribution cannot be estimated by non-Bayesian estimation techniques.

### TABLE 3 Bayesian, Kaplan-Meier and Nelson-Aalen Estimates for Section 1 of Gardiner Expressway

<table>
<thead>
<tr>
<th>Flow ((q)) in the interval</th>
<th>Bayesian Estimate (1 - \alpha(q, \infty) = 1 - \beta e^{-\theta q})</th>
<th>Kaplan-Meier Estimate</th>
<th>Nelson-Aalen Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>([0,3580))</td>
<td>(1 - \left(\frac{\beta e^{-\theta q} + 2649}{\beta + 2649}\right))</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>([3580,3900))</td>
<td>(1 - \left[\left(\frac{\beta e^{-\theta q} + 2647}{\beta + 2649}\right) \times 1.28\right])</td>
<td>0.0005</td>
<td>0.0005</td>
</tr>
<tr>
<td>([3900,3980))</td>
<td>(1 - \left[\left(\frac{\beta e^{-\theta q} + 2642}{\beta + 2649}\right) \times 1.49\right])</td>
<td>0.0016</td>
<td>0.0016</td>
</tr>
<tr>
<td>([5880,5980))</td>
<td>(1 - \left[\left(\frac{\beta e^{-\theta q} + 9}{\beta + 2649}\right) \times 31.74\right])</td>
<td>0.2441</td>
<td>0.2423</td>
</tr>
<tr>
<td>([5980,6060))</td>
<td>(1 - \left[\left(\frac{\beta e^{-\theta q} + 8}{\beta + 2649}\right) \times 42.99\right])</td>
<td>0.2711</td>
<td>0.2689</td>
</tr>
<tr>
<td>([6060,\infty))</td>
<td>(1 - \left[\left(\frac{\beta e^{-\theta q}}{\beta + 2649}\right) \times 56.37\right])</td>
<td>Undefined</td>
<td>Undefined</td>
</tr>
</tbody>
</table>
Another important point is that the actual breakdown points are sufficient to calculate the Kaplan-Meier and Nelson-Aalen estimates of the capacity distribution. That is, one does not need to know the actual censored observations. In fact, the number of censored observations between the two breakdown points (uncensored observations) is enough to obtain the Kaplan-Meier or Nelson-Aalen curves. On the other hand, all the data (both censored and uncensored observations) is used to obtain the Bayesian estimating curve. The fact that Bayesian estimate uses all the data makes it preferable to the other estimates. In this sense, the Bayesian estimator serves as a function of the sufficient statistic by fully utilizing the data.

For the Kaplan-Meier and Nelson-Aalen estimators, the breakdown probability is zero up to the flow with first breakdown, which may not be realistic for every case. For instance, consider Section 1 of Gardiner Expressway (see FIGURE 3). With a flow value of 6000 veh/h, the Kaplan-Meier and Nelson-Aalen curves have a probability of breakdown value of 0.33 roughly. On the other hand, the breakdown probability is 0.04 for 5500 veh/h. This means that the traffic is almost free flowing (no chance of breakdown) at that flow value that is merely 500 veh/h less than 6000 veh/h. This unrealistic situation can be addressed by the Bayesian estimation. Bayesian estimator calculates the breakdown probabilities for all traffic flow values possible including the relatively lower flows for which breakdowns are not actually observed, as seen in TABLE 3 for flow values between 0 and 3580 veh/h. It is important to note that breakdowns can also occur at low flow levels. Therefore, Bayesian estimation becomes more effective while defining the capacity profile of the freeway section.

Furthermore, both the conventional and Bayesian estimate are discontinuous at uncensored observations. However, Bayesian approach smoothes the estimator at censored observations by achieving continuity at the censored observations. Of course, although the Bayesian curves are continuous functions and smoother than the other estimating curves, they are not as smooth as the ones obtained by a parametric fit. This smoothness depends on the value of $\beta$, and the percentage of censored observations in the data, which is very high in the case study presented in the paper.

**CONCLUSIONS AND FUTURE RESEARCH**

Using the deterministic traffic capacity values is an incomplete representation of the stochastic real life conditions. Therefore, there is definitely a need to try to estimate the robust probabilistic capacity distributions. The major point of interest in this research was to study the distribution of traffic capacity in the presence of censored observations, and compare the conventional estimation methods with the Bayesian estimation techniques. The evaluations and calculations are based on five-minute interval observations from I-880 Highway in California, USA and Gardiner Expressway in Toronto, Canada.

The significant point in this study is that it is impossible to obtain the actual capacity distribution as the flow observations (from the data) have right-censored points. However, several nonparametric estimation methods that can handle data with censored observations are available in the literature. Kaplan-Meier and Nelson-Aalen methods are the most well-known estimation techniques to obtain the survival distribution which leads us to the cumulative distribution of capacity. In the presence of sufficient amount of data, these techniques work efficiently. However, with a small sample that possibly contains a
substantial amount of censored observations, they may be inefficient. First of all, these estimators are not defined for all values of traffic flow values possible from zero to infinity. The last observation defines the final point of the estimated distribution curve (see (1) and (15) for details). Since a complete curve cannot be obtained, parametric distribution fitting tools have to be applied. Moreover, the discontinuity and non-smoothness of the Kaplan-Meier estimator curve makes it difficult to rely on the estimation results with a high level of confidence.

Hence, a Bayesian estimator used in lifetime analysis which assumes a Dirichlet process prior for the survival function under the minimization of a squared-error loss function has been introduced. There are four main advantages of Bayesian Estimator over others:

- All the data points (both censored and uncensored observations) are used to obtain the curves estimated by the Bayesian approach. Hence, the Bayesian estimator serves as a function of the sufficient statistic by fully utilizing the data.
- The results indicate that the curves obtained by using Bayesian estimation are smoother than the ones obtained by Kaplan-Meier and Nelson-Aalen estimators. This smoothness ensures the continuity of the estimated functions everywhere in the functional space including around the censored observations.
- Furthermore, unlike Kaplan-Meier estimation, the Bayesian estimators can be obtained for any value of traffic flow regardless of the available data points. For example, up-to a flow value, the conventional estimation techniques always give zero breakdown probability. This case is observed and mentioned for the distribution functions obtained by Brilon et al (1), and Kuhne and Mahnke (15). This problem is also solved by using the Bayesian estimator that generates a smoother curve, which in turn gives the breakdown probabilities, even for low traffic flow values.
- Finally, only if the maximum observed traffic flow value is followed by a breakdown, the maximum value of the capacity distribution function will reach 1 for the Kaplan-Meier and Nelson-Aalen estimates. As stated by Kuhne and Mahnke (15) and Brilon et al. (1), the distribution dies out at a point where the maximum flow causing the breakdown is obtained. Therefore, given the limitations of the data set, a complete distribution cannot be obtained when non-Bayesian estimators are used. However, Bayesian estimation approach gives a complete curve including the upper tail portion of the estimated capacity distribution where there are no observations available for that portion.

As a future work, another distribution from the exponential family, namely Beta distribution will be selected as the prior distribution for the Bayesian estimation procedure, and compared with the Dirichlet prior. The optimal way of finding $\theta$ and $\beta$ depending on the data (especially, sample size) is also an interesting area for future research. Finally, additional data obtained from different sites will be used to test the location specific differences among the estimated distributions. This information can then be used to estimate hierarchical models.
ACKNOWLEDGMENTS

We would like to thank Prof. Bruce Hellinga for providing us the traffic data from Gardiner Expressway. Special thanks are due Ozlem Yanmaz-Tuzel for useful discussions and/or assistance.

REFERENCES

Ozguven E. E., Ozbay K.