Hierarchical Capacity Analysis of Freeways via Nonparametric Bayesian Estimation with Censored Data

Eren Erman Ozguven, M.Sc. (Corresponding Author)
Graduate Assistant,
Department of Civil and Environmental Engineering,
Rutgers University
623 Bowser Road, Piscataway, NJ 08854 USA,
Tel: (732)445-4012
Fax: (732)445-0577
e-mail: ozguvene@rci.rutgers.edu

Kaan Ozbay, Ph.D.
Professor,
Department of Civil and Environmental Engineering,
Rutgers University
623 Bowser Road, Piscataway, NJ 08854 USA,
Tel: (732)445-2792
Fax: (732)445-0577
e-mail: kaan@rci.rutgers.edu

Word count: 5498 + 3 Figures + 5 Tables = 7498
Abstract: 195
Re-submission Date: November 15, 2010

Paper Submitted for Presentation and Publication at the Transportation Research Board’s 90th Annual Meeting, Washington, D.C., 2011
Traffic capacity of a freeway differs depending on its distinct sections with different spatial characteristics such as the number and width of lanes, existence and type of shoulders and/or medians, traffic characteristics (such as the number of breakdowns defined using the sudden changes in the speed and density values that occur during the flow phase transition), and population characteristics (rural and urban areas). To account for these spatial differences, this paper investigates the hierarchical estimation of the traffic capacity distribution on a highway using a nonparametric Bayesian approach assuming two prior distributions, namely Dirichlet and Gamma process priors under the minimization of a squared-error loss function. This approach addresses the difficult problem of the censored observations while treating the model parameters as random variables represented by a probability distribution. The methodology is applied on the highway sections with different spatial characteristics. An application of the method for on and off-ramps of a highway is also presented. Finally, the results are discussed hierarchically with presenting a methodology to simulate censored and breakdown observations and to analyze them statistically using a bootstrap approach in order to obtain the capacity distributions for sections without sufficient data.
INTRODUCTION

Traffic capacity has utmost importance in every phase of the planning, maintenance and operations for various sections of a roadway. The overestimation of the capacity leads to delays and traffic incidents due to the congestion and excessive volumes whereas the underestimation can lead to high costs and unnecessarily large ramp/mainline delays. Thus, there is a need to study the spatial characteristics of the probabilistic distribution of traffic capacity utilizing the traffic state information including free flow and breakdown observations. Although they have not explicitly considered a concept of capacity distribution, several researchers have recognized the necessity of characterizing roadway capacity as a stochastic phenomenon.

Lorenz and Elefteriadou defined the traffic capacity by including probabilistic concepts (2), (3). Kim and Elefteriadou (4) performed a simulation analysis to estimate the capacity of 2-lane/2-way highways focusing on the difference from the deterministic capacity values. Persaud et al. (5) studied the probabilistic breakdown phenomenon in freeway traffic and confirmed that breakdown is stochastic in nature. Chung et al. (6), and Hall and Agyemang-Duah (7) worked on the freeway capacity drop and the definition of the capacity whereas Cassidy and Bertini (8), and Smilowitz et al. (9) studied some important traffic features with respect to the bottleneck flows. A deterministic method was proposed by Cassidy and May (10) for estimating the traffic capacity of major freeway weaving sections. Wang et al. (11), and Lertworawich and Elefteriadou (12) also worked on the capacity of weaving sections. Another approach employing neural network techniques for estimating traffic capacity of weaving segments has been proposed by Awad (13). Yeon et al. (14), and Kuhne et al. (15) worked on the development of a stochastic theory of freeway traffic. Kuhne and Mahnke (16) derived a formulation for obtaining the cumulative distribution for traffic breakdowns defining the breakdown as a car cluster formation process. Kuhne (17) used Weibull curves to match a broad variety of cumulative distribution functions. Uchida and Muniero (18) proposed a traffic capacity and covariance estimation methodology using a macroscopic traffic model based on a generalized car following model. Jun et al. (19) used a simultaneous-spline regression model for the joint estimation of traffic variables. Similarly, for the estimation of the key traffic stream parameters, a dual regime model was proposed by Yao et al. (20) whereas Banks (21), (22), (23), (24) conducted several important studies to develop a new approach to bottleneck capacity analysis. Polus and Pollatschek (25) explored the stochastic nature of freeway capacity by fitting speed-flow diagrams. Laval (26) came up with approximate formulas for highway capacity using the presence of slow vehicles.

However, none of these studies focus on the use of statistically robust methodologies to estimate the probabilistic distribution of the freeway capacity spatially. Brilon et al. (27), (28) used a practical estimation method based on an analogy to the statistics of lifetime analysis to obtain the distribution of freeway capacity. They used the Kaplan-Meier estimator to obtain the capacity distribution function, and estimated the parameters of this distribution with various functions, where Weibull function appeared to be the best fit. This lifetime method had been first proposed by van Toorenburg (29) and discussed by Minderhoud et al. (30), Regler (31), applied the capacity analysis in (28) to other freeway sections. Recently, Geistefeldt and Brilon (I) and Geistefeldt (32)
Ozguven E. E., Ozbay K.

made a comparative assessment between the direct estimation of breakdown probabilities and estimation of capacity distribution functions via Kaplan-Meier method. They concluded that Kaplan-Meier method performs better than the direct breakdown probability estimation techniques. Ozbay and Ozguven (33), studied the effects of spatial/temporal differences of freeway sections on the capacity distribution using a non-informative Bayesian prior distribution called Jeffrey's prior. Ozguven and Ozbay (34), improving the method given in (27), introduced a nonparametric Bayesian estimation under an informative Dirichlet process prior. for the first time, to robustly estimate the traffic capacity model parameters in the presence of insufficient and unreliable data.

In this article, we make the following contributions:

- Improving the Bayesian estimation performed by Ozguven and Ozbay (34), we propose a hierarchical nonparametric Bayesian approach to estimate the traffic capacity distribution where the prior knowledge about the traffic parameters specific to a freeway section is represented by a probabilistic function.
- We use two different probability distributions, namely Dirichlet and Gamma processes as the priors to compare the traffic capacity distributions for different sections of the same highway, and to test the location specific differences among the estimated distributions.
- We apply this estimation methodology to the ramp sections using the occupancy data which can be used to enhance freeway control strategies such as ramp metering.
- With the proposed model, a freeway can be studied hierarchically with respect to the following factors:
  - Number of lanes,
  - Population affects (rural and urban areas),
  - Traffic features (number of breakdowns, free flows and flow phase transition).

**HIERARCHICAL BAYESIAN ESTIMATION WITH CENSORED DATA**

The use of Bayesian estimation to obtain 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 insufficient data to study the behavior of the traffic capacity accurately and efficiently.
- Traffic observations are not 100% reliable.

In the context of the probabilistic traffic capacity modeling mentioned in previous studies, these conditions are present. Therefore, there is definitely a need for using hierarchical Bayesian estimators.

We first consider the following model. Let $q = (q_1, q_2, ..., q_n)$ be the traffic flow data and $V = (V_1, V_2, ..., V_n)$ be the speed data of $n$ observations obtained from the sensors on a freeway section through a specific time period. Data is censored from the right by a threshold speed value of $Y$ where $q$ consists of breakdown and censored flows. Censored flows are the ones where the speed values are higher than the threshold speed so that they do not cause breakdown. Therefore, the data is in the form of:
$$q_j, j = 1, ..., n$$

$$\delta_j = \begin{cases} 1 & \text{if } V_j \leq Y \\ 0 & \text{if } V_j > Y \end{cases} \quad (1)$$

If $\delta_j = 1$, there is a traffic breakdown at volume $q_j$, and, if $\delta_j = 0$, traffic is fluent above the threshold speed without any breakdown (censored). Therefore, the important point is to make use of the traffic observations, including the censored volumes, to calculate the probability distribution of the capacity. With this information, we obtain the capacity distribution function $F(q)$ as:

$$F(q) = 1 - S(q) = P(C \leq q) \quad (2)$$

Here, $C$ represents the section capacity, and $S(q)$, the survival function. The Kaplan-Meier estimation approach to find this capacity distribution was presented in (27). The difference of Bayesian statistics from this approach is that it incorporates the prior knowledge along with a given set of current observations to make statistical inferences. The prior information could come from operational/observational data, from previous experiments or engineering knowledge. The pioneering studies on the Bayesian estimation with right-censored data were conducted by Susarla and Ryzin (35), and Ferguson and Phadia (36). The idea is to obtain the nonparametric estimation of the capacity distribution function using a Bayesian approach with right-censored observations. As a base decision rule to obtain an estimate of the traffic capacity function, a loss function is specified as

$$L(\tilde{F}, F) = \int_{0}^{\infty} (F(q) - \tilde{F}(q))^2 dw(q) \quad (3)$$

where $w$ is a nonnegative weight function. This loss function simply presents the weighted integrated difference between the actual and estimated values of the capacity function.

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

The standard estimator of the survival function is the Kaplan-Meier estimator (Product-Limit Estimator) proposed by Kaplan and Meier (37), and applied into traffic engineering by Brilon et al. (27) to obtain the distribution of freeway capacity. Using this approach, traffic capacity distribution function, $\tilde{F}_K(q)$ is obtained as:

$$\tilde{F}_K(q) = 1 - \tilde{S}_K(q) = 1 - \prod_{j: q_j \geq q} \frac{n_j - \rho_j}{n_j} \quad (4)$$

where

- $\tilde{F}_K(q)$: estimated Kaplan-Meier capacity distribution function,
- $\tilde{S}_K(q)$: estimated Kaplan-Meier survival function,
- $n_j$: number of intervals with a flow rate $\geq q_j$,
- $q_j$: traffic volume in interval $j$,
- $\rho_j$: number of breakdowns at a volume of $q_j$. 

TRB 2011 Annual Meeting
Paper revised from original submittal.
This product is calculated over all time intervals \( j \) with a traffic volume \( q_j < q \) that were followed by a traffic breakdown. With a comparative assessment, Geistefeldt and Brilon (1) concluded that Kaplan-Meier method performs better than the direct breakdown probability estimation techniques.

**Hierarchical Bayesian Approach**

In this paper, a nonparametric hierarchical Bayesian estimation of the traffic capacity analysis is given under the squared-error loss notion with stochastic process priors. We consider a nonparametric approach in which \( F_0(q) \), the prior guess on the estimated capacity function, is drawn from a stochastic process \( f(\gamma, H) \) based on the observations \( (\delta, q) \). Here, \( \gamma > 0 \) and \( H \) are the baseline probability measures for \( f(\gamma, H) \). With this information, the posterior distribution of the capacity distribution function, \( F(q) \) is derived. Therefore, the hierarchical stochastic process becomes (see Teh et al. (38) for details):

\[
\begin{align*}
F_0(\gamma) &\mid \gamma, H \sim f(\gamma, H) \\
F(q) &\mid \beta, F_0(q) \sim f(\beta, F_0(q))
\end{align*}
\]

(5)

where \( \beta \) is a parameter of the prior distribution \( F_0(q) \). The selection of the prior guess has utmost importance. This prior guess based on \( f(\gamma, H) \) can be obtained from previous empirical work, data, or researcher’s subjective beliefs. In the proceeding sections, we try two different priors, namely Dirichlet and Gamma process priors, to obtain the posterior distribution of the traffic capacity function.

**Dirichlet process prior**

Dirichlet distribution is the first prior distribution selected to lead to closed-form estimates for the survival function using the loss function given in Equation (3). This prior is chosen because it is a conjugate prior (the one which produces a posterior distribution of the same type as the prior) for the survival function. This makes the computation of the posterior particularly simple.

When we use a Dirichlet process (DP) in Equation (5), the general hierarchical model becomes as follows:

\[
\begin{align*}
F_0(q) &\mid \alpha(0, \infty) \sim DP(\alpha(0, \infty)) \\
F_D(q) &\mid \beta, F_0(q) \sim DP(\beta, F_0(q))
\end{align*}
\]

(6)

To assign a prior distribution to the capacity function, the parameter function is taken to be of the form \( \alpha(q, \infty) = \beta(1 - F_0(q)) \) where \( F_0(q) = 1 - \alpha(0, \infty) \) is the prior guess at the traffic capacity function and \( \beta \) is a kind of measure that indicates how much weight to put on the prior guess. Here, we take the prior guess at the survival function \( S_0(q) = 1 - F_0(q) \) with a Dirichlet prior given as \( \alpha(0, \infty) = e^{-\theta q} \) and our problem becomes:

\[
\begin{align*}
F_0(q) &= 1 - \alpha(0, \infty) = 1 - e^{-\theta q} \\
F_D(q) &\mid \beta, \alpha(q, \infty) \sim DP(\beta, \alpha(q, \infty))
\end{align*}
\]

(7)
As observed, there is a precise relationship between the prior and posterior distributions; therefore, we obtain the posterior distribution as:

$$\alpha(q, \infty) = \beta(1 - F_q(q)) = \beta \alpha(0, \infty) = \beta e^{-\theta q}$$ (8)

where $\theta$ is heuristically calculated using the Kaplan-Meier data of breakdown intervals.

To find the Bayesian estimator with Dirichlet prior, the conditional $p^{th}$ moment of the breakdown probability $F_D(q) = 1 - S_D(q)$ given $(\delta, q)$ is calculated as follows, where $\alpha(q, \infty)$ is distributed as a Dirichlet process on $\mathbb{R}^+ = (0, \infty)$:

$$E[(1 - F_D(q))^p | (\delta, q)] = \prod_{s=0}^{p-1} \left\{ \frac{\alpha(q, \infty) + s + n_q}{\alpha(0, \infty) + s + n} \right\} \mathcal{G}(j)$$ (9)

which is adapted from (35) where

$$\mathcal{G}(j) = \begin{cases} 
\text{while } q_j \text{ censored:} & \prod_j \left\{ \frac{\alpha(q_j, \infty) + n_j}{\alpha(q_j, \infty) + n_j - \lambda_j} \right\} \\
\text{while } q_j \text{ uncensored:} & 1
\end{cases}$$

$n_j$: number of intervals with a flow rate greater than or equal to $q_j$,

$n$: total number of observations,

$n_q$: total number of observations with a flow rate greater than $q$,

$\lambda_j$: number of censored observations at a volume of $q_j$.

$q_j$: traffic volume in interval $j$,

$p$, $s$: nonnegative integers.

Using Equation (9), the Bayesian estimator of the survival function with the Dirichlet prior is calculated taking $p=1$. That is, we are focusing on the conditional mean of $1 - F_D(q)$ given $(\delta, q)$. Then, our Bayesian estimator becomes:

$$\hat{F}_D(q) = 1 - \hat{S}_D(q) = 1 - \left( \frac{\alpha(q, \infty) + n_q}{\alpha(0, \infty) + n} \right) \mathcal{G}(j)$$ (10)

where

$$\mathcal{G}(j) = \begin{cases} 
\text{while } q_j \text{ censored:} & \prod_j \left\{ \frac{\alpha(q_j, \infty) + n_j}{\alpha(q_j, \infty) + n_j - \lambda_j} \right\} \\
\text{while } q_j \text{ uncensored:} & 1
\end{cases}$$

$\hat{S}_b(q)$: estimated Bayesian survival function.
To compare the Dirichlet results with another distribution, we choose the Gamma prior, namely $F_G(q)$. With the Gamma prior having a continuous shape parameter $\gamma(q)$ and reciprocal scale parameter $\tau$, the posterior expectation is calculated for the traffic capacity as follows (see (36) for the original formulation):

$$E[(1-F_G(q))|\delta, q] = \left(\frac{n_q + \tau}{n_q + \tau + 1}\right)^{\gamma(q)} \prod_{j=1}^{n-1} \left(\frac{(n_{j-1} + \tau)(n_j + \tau + 1)}{(n_{j-1} + \tau + 1)(n_j + \tau)}\right)^{\gamma(q_j)} \left(\frac{\phi_G(n_j + \lambda_j + \tau + 1, \delta_j)}{\phi_G(n_j + \lambda_j + \tau, \delta_j)}\right)$$

(11)

where

- $n_j$: number of intervals with a flow rate $\geq q_j$,
- $n$: total number of observations,
- $n_q$: total number of observations with a flow rate greater than $q$,
- $\delta_j = \begin{cases} 1 & \text{if } V_j \leq Y \\ 0 & \text{if } V_j > Y \end{cases}$
- $\lambda_j$: number of censored observations at a volume of $q_j$.
- $q_j$: traffic volume in interval $j$,
- $\tau$: scale parameter of the Gamma prior,
- $\gamma(q)$: shape parameter of the Gamma prior,
- $\phi_G(\eta, \omega) = \sum_{i=0}^{\omega-1} \binom{\omega-1}{i} (-1)^i \log\left(\frac{\eta+i+1}{\eta+i}\right)$.

Here, the prior Gamma guess at the capacity function is selected as:

$$F_0(q) = 1 - S_0(q) = 1 - \left(\frac{\tau}{\tau+1}\right)^{\gamma(q)}$$

(12)

Then, for all $q$, we have $\gamma(q)$ for fixed $\tau$ as:

$$\gamma(q) = \frac{\ln(1-F_0(q))}{\ln\left(\frac{\tau}{\tau+1}\right)}$$

(13)

Finally, the Bayesian capacity function estimator with the Gamma prior is calculated as in the Dirichlet prior case, and our Bayesian estimator becomes:

$$\tilde{F}_G(q) = 1 - \tilde{S}_G(q) = 1 - \left(\frac{n_q + \tau}{n_q + \tau + 1}\right)^{\gamma(q)} \prod_{j=1}^{n-1} \left(\frac{(n_{j-1} + \tau)(n_j + \tau + 1)}{(n_{j-1} + \tau + 1)(n_j + \tau)}\right)^{\gamma(q_j)} \left(\frac{\phi_G(n_j + \lambda_j + \tau + 1, \delta_j)}{\phi_G(n_j + \lambda_j + \tau, \delta_j)}\right)$$

(14)

where $\gamma(q)$ is given as in Equation (13) and $\phi_G(\cdot)$ as in Equation (11).
The Bayesian estimator, using the Dirichlet or Gamma process priors, estimates all possible probability values between the breakdown and censored observations (achieving continuity) and therefore smoothes the posterior distribution curve at all discontinuities. Suzarla and Ryzin (39) state that "the Bayesian estimator is a better admissible estimator smoothing the resulting posterior on the discrete survival data using the expectation and variances of the cumulative distribution given in Equation (9) under the loss function used and under the weak convergence topology". Survival models are based on data that measure time to some event such as death, transition or failure. In our case, this time period is replaced by the traffic volume and the discrete failure event becomes the discrete traffic breakdown event. Therefore, assuming the breakdown as a failure event, the statistical methods developed for the survival data analysis can be successfully used to estimate the traffic capacity, which is the analog of the lifetime in this context.

COMPARATIVE MODEL APPLICATION

Basic Assumptions

The concept of determining how to define the breakdown is very crucial in terms of the accuracy and efficiency of the proposed methodology. There are several approaches available to define the freeway breakdown. One approach is to determine the breakdown value by using the speed difference between consecutive intervals whereas another approach is including the average space mean speed to detect the breakdown point. Recently, Geistefeldt and Brilon (1) used a more sophisticated set of criteria to determine the breakdown value. In this study, the flow values are classified, using a 5-minute interval denoted as $i$, and shown in Table 1 where $Y$ is a threshold speed, and $\varepsilon$ is a small speed value selected for comparison to determine the breakdown. This approach is logical as the traffic breakdown is usually followed by a significant amount of reduction in speed, however every interval followed by a speed lower than the threshold value cannot be considered as a breakdown interval (B). If each congested volume was 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 intervals that causes breakdown with satisfying the conditions in Table 1 is considered as a B-value.

The traffic congestion usually leads to a breakdown, which is usually determined by a sudden decrease in the speed level accompanied with a simultaneous steep increase of the traffic density level that occurs during the phase transition from free-flow to breakdown flow. When the traffic breakdown ends, the speed and density both return to the normal level of the uncongested conditions. Therefore, while considering the land usage for the proposed spatial analysis in this study, the phase transition is also considered during the determination of the breakdown points used in Table 1.
TABLE 1 Breakdown Interval Definitions (Based on (1))

<table>
<thead>
<tr>
<th>Interval</th>
<th>Definition</th>
</tr>
</thead>
</table>
| **B**    | Traffic is fluent in time interval $i$, but the observed volume causes a breakdown, satisfying the following conditions:  
- $v_{i-1} > Y$,  
- $v_i > Y$,  
- $v_{i+1} < Y$,  
- $v_{i+2} < Y$,  
- $\frac{1}{2}(v_{i-1} + v_i) - \frac{1}{2}(v_{i+1} + v_{i+2}) > \varepsilon$. |
| **C**    | Traffic is fluent in the intervals $i$ and $i+1$. This interval $i$ contains a censored value. Its information is that the actual capacity in interval $i$ is greater than the observed volume $q_i$. 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. |
| **D1**   | 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. These are mainly those intervals representing congested flow conditions, which do not contain any information about the traffic capacity. |
| **D2**   | 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. 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 $i-1$, 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. |
| **D3**   | 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. |

**Data**

Data obtained from different sections of the I-5 Highway, California, USA from the PEMS database (40) is used to test the location specific differences among the estimated distributions. The information about the sections studied are given in Table 2.
### TABLE 2 I-5 Highway Section Information

<table>
<thead>
<tr>
<th>Vehicle Detection System (VDS) Number</th>
<th>Type</th>
<th>Number of Lanes</th>
<th>Location</th>
<th>Maximum Observed Volume per lane (PEMS^{(40)})</th>
</tr>
</thead>
<tbody>
<tr>
<td>1015010</td>
<td>Mainline</td>
<td>2</td>
<td>Rural (Stanislaus County, CA)</td>
<td>1524 veh/h</td>
</tr>
<tr>
<td>1108613</td>
<td>Mainline</td>
<td>4</td>
<td>Urban (Washington St, San Diego, CA)</td>
<td>1914 veh/h</td>
</tr>
<tr>
<td>1205157</td>
<td>Mainline</td>
<td>6</td>
<td>Urban (Orange County, CA)</td>
<td>1732 veh/h</td>
</tr>
<tr>
<td>1214241</td>
<td>On Ramp</td>
<td>2</td>
<td>Urban (Orange County, CA)</td>
<td>-</td>
</tr>
<tr>
<td>1211850</td>
<td>Off Ramp</td>
<td>1</td>
<td>Urban (Orange County, CA)</td>
<td>-</td>
</tr>
</tbody>
</table>

The important point of the study while using the data is to obtain plausible cumulative distribution functions for different sections of the freeway (which results in obtaining data from several single loop detectors on different lanes of the freeway sections) so that this can be used to make it possible for a traffic engineer/planner to choose the proper pairs of the breakdown probability and traffic capacity for the whole freeway. After analyzing one-year (2009) traffic data obtained for these highway sections, threshold speeds for the mainline sections are selected as 70 m/h, whereas it was taken as 40 m/h for the ramp sections, and $\epsilon$ is selected 5 m/h for all sections.

We have also checked the validity of the breakdown speed values from consecutive detectors. For instance, as observed from Table 2, Vehicle Detection System (VDS) No. 1108613 is one of the detection systems used in the analysis. However, one of the consecutive systems namely, VDS-1108603 is also tested using the model to determine if the breakdowns and the corresponding traffic flow per lane values are reasonably similar. This is simply a way of validating the data.

#### Example Calculation

Firstly, we work on an example to illustrate the concept of the Bayesian estimation. Kaplan-Meier estimator is indeed a limit of the Bayesian estimator given in Equation (6) when $\alpha(R^+) \rightarrow 0$. This reduction and better performance of Bayesian estimation using censored data is clearly seen on an example with the following data (8 observations obtained from VDS-1015010):

- **Breakdown flows**: 40, 70, 90, 120 veh/5 min
- **Censored flows**: 50, 60, 100, 130 veh/5 min

Given the data and selecting the prior guess of the survival function as $S_0(q) = 1 - F_0(q) = e^{-\epsilon q}$, the Bayesian estimators and Kaplan-Meier estimates are calculated in Table 3. Based on the prior guess of the survival function, and taking $\tau = 1$, we obtain $\gamma(q) = 1.443\theta q$ from Equation (13).
### Table 3: Bayesian and Kaplan-Meier Estimates of $\tilde{S}(q)$ for the Example

<table>
<thead>
<tr>
<th>Flow (veh/5 min) in the interval</th>
<th>Bayesian Estimate with Dirichlet prior ($\tilde{S}_D(q)$)</th>
<th>Bayesian Estimate with Gamma prior ($\tilde{S}_G(q)$)</th>
<th>Kaplan-Meier Estimate ($\tilde{S}_K(q)$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>[0,40)</td>
<td>$\beta e^{-0q} + \frac{8}{\beta + 8}$</td>
<td>$\left( \frac{9}{10} \right)^{1.44\theta q}$</td>
<td>1.0</td>
</tr>
<tr>
<td>[40,50)</td>
<td>$\beta e^{-0q} + \frac{7}{\beta + 8}$</td>
<td>$\left( \frac{8}{9} \right)^{1.44\theta q} \left( \frac{81}{80} \right)^{57.72\theta} \frac{\ln(\frac{10}{9})}{\ln(\frac{9}{8})}$</td>
<td>7/8</td>
</tr>
<tr>
<td>[50,60)</td>
<td>$\beta e^{-0q} + \frac{6}{\beta + 8} \ast \beta e^{-500q} + \frac{7}{\beta e^{-500q} + 6}$</td>
<td>$\left( \frac{7}{8} \right)^{1.44\theta q} \left( \frac{81}{80} \right)^{57.72\theta} \frac{\ln(\frac{10}{9})}{\ln(\frac{9}{8})} \ast \left( \frac{64}{63} \right)^{72.15\theta}$</td>
<td>7/8</td>
</tr>
<tr>
<td>[60,70)</td>
<td>$\beta e^{-0q} + \frac{3}{\beta + 8} \ast \beta e^{-500q} + \frac{7}{\beta e^{-500q} + 6} \ast \beta e^{-600q} + \frac{6}{\beta e^{-600q} + 5}$</td>
<td>$\left( \frac{6}{7} \right)^{1.44\theta q} \ast \left( \frac{49}{48} \right)^{86.58\theta}$</td>
<td>7/8</td>
</tr>
<tr>
<td>[70,90)</td>
<td>$\beta e^{-0q} + \frac{3}{\beta + 8} \ast \beta e^{-500q} + \frac{7}{\beta e^{-500q} + 6} \ast \beta e^{-600q} + \frac{6}{\beta e^{-600q} + 5}$</td>
<td>$\left( \frac{5}{6} \right)^{1.44\theta q} \ast \left( \frac{36}{35} \right)^{101\theta} \frac{\ln(\frac{7}{6})}{\ln(\frac{6}{5})}$</td>
<td>7/10</td>
</tr>
<tr>
<td>[90,100)</td>
<td>$\beta e^{-0q} + \frac{3}{\beta + 8} \ast \beta e^{-500q} + \frac{7}{\beta e^{-500q} + 6} \ast \beta e^{-600q} + \frac{6}{\beta e^{-600q} + 5}$</td>
<td>$\left( \frac{4}{5} \right)^{1.44\theta q} \ast \left( \frac{25}{24} \right)^{129.87\theta} \frac{\ln(\frac{6}{5})}{\ln(\frac{5}{4})}$</td>
<td>21/40</td>
</tr>
<tr>
<td>[100,120)</td>
<td>$\beta e^{-0q} + \frac{2}{\beta + 8} \ast \beta e^{-500q} + \frac{7}{\beta e^{-500q} + 6} \ast \beta e^{-600q} + \frac{6}{\beta e^{-600q} + 5}$</td>
<td>$\left( \frac{3}{4} \right)^{1.44\theta q} \ast \left( \frac{16}{15} \right)^{144.30\theta}$</td>
<td>21/40</td>
</tr>
<tr>
<td>[120,130)</td>
<td>$\beta e^{-0q} + \frac{2}{\beta + 8} \ast \beta e^{-500q} + \frac{7}{\beta e^{-500q} + 6} \ast \beta e^{-600q} + \frac{6}{\beta e^{-600q} + 5}$</td>
<td>$\left( \frac{2}{3} \right)^{1.44\theta q} \ast \left( \frac{9}{8} \right)^{173.16\theta} \frac{\ln(\frac{4}{3})}{\ln(\frac{3}{2})}$</td>
<td>21/80</td>
</tr>
<tr>
<td>[130, $\infty$)</td>
<td>$\beta e^{-0q} + \frac{7}{\beta + 8} \ast \beta e^{-500q} + \frac{6}{\beta e^{-500q} + 6} \ast \beta e^{-600q} + \frac{5}{\beta e^{-600q} + 5}$</td>
<td>$\ast \beta e^{-1000q} + \frac{3}{\beta e^{-1000q} + 2}$</td>
<td>Not defined</td>
</tr>
</tbody>
</table>

Note: $\theta$ and $q$ are parameters specific to the example.
As $\alpha(R^+) = \beta e^{-\theta q} \to 0$, the ratios coming from the censored data naturally vanish as the values common to both products cancel out reducing the Bayesian estimator with Dirichlet prior to the Kaplan-Meier. However, the Gamma prior estimate does not converge to the Kaplan-Meier. It converges to an estimate which gives more weight to the right tails than the Kaplan-Meier estimate. This takes the Bayesian estimator with Gamma prior away from the other two estimators.

**Comparative Spatial Results**

Distributions for the Dirichlet and Gamma priors, namely $\alpha(0, \infty)$ and $\left(\frac{\tau}{\tau+1}\right)^{\gamma(q)}$ are calculated based on the following heuristic argument given in (35) for each freeway section separately. The idea is to force the estimators to satisfy the following equation:

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

(15)

where $M : \text{quantile}(P_K, 0.5)$ for the estimated Kaplan-Meier breakdown probability vector $P_K$. Therefore, $M$ represents the 0.5 quartile of the Kaplan-Meier estimated flow vector. Using this method, and by doubly using the observations, $\theta$ is obtained for each section. This approach is found to be useful in (35) as other values of $\theta$ tend to pull the Bayesian estimator away from the Product-Limit estimator.

The selected parameters for the Gamma and Dirichlet distributions are given in Table 4. The parameters $\beta$ for Dirichlet prior and $\tau$ for Gamma prior define the prior mass on the traffic volume data. For instance, when the data sample size is $n$, the prior mass on $(0, \infty)$ becomes $\frac{\beta}{n}$. The extreme case is to let $\beta \to \infty$ in which case the Bayesian estimator reduces to $e^{-\theta q}$. $\beta$ values are optimally obtained for each section iteratively tending not to pull the Bayesian estimator from the Kaplan-Meier curve. On the other hand, the intensity parameter $\tau$ is selected as 1 according to the argument given in (36) and by choosing that value, we correspond reasonably well to the prior sample size for the Gamma process given in this study. With this information, $\gamma(q)$ is calculated with Equation (13).
TABLE 4 Distribution Parameters for the Highway Sections Studied

<table>
<thead>
<tr>
<th>Section</th>
<th>$\theta$</th>
<th>$\alpha(q)$</th>
<th>$\beta$</th>
<th>$\gamma(q)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>VDS-1205157 (6 Lanes)</td>
<td>0.1071</td>
<td>$e^{-0.1071q}$</td>
<td>0.0050</td>
<td>$\frac{\ln(e^{-0.1071q})}{\ln(\frac{\tau}{\tau+1})}$</td>
</tr>
<tr>
<td>VDS-1108613 (4 Lanes)</td>
<td>0.0498</td>
<td>$e^{-0.0498q}$</td>
<td>0.0040</td>
<td>$\frac{\ln(e^{-0.0498q})}{\ln(\frac{\tau}{\tau+1})}$</td>
</tr>
<tr>
<td>VDS-1015010 (2 Lanes)</td>
<td>0.0126</td>
<td>$e^{-0.0126q}$</td>
<td>0.0045</td>
<td>$\frac{\ln(e^{-0.0126q})}{\ln(\frac{\tau}{\tau+1})}$</td>
</tr>
<tr>
<td>VDS-1211850 (1 Lane)</td>
<td>0.0344</td>
<td>$e^{-0.0344q}$</td>
<td>0.0035</td>
<td>$\frac{\ln(e^{-0.0344q})}{\ln(\frac{\tau}{\tau+1})}$</td>
</tr>
<tr>
<td>VDS-1214241 (2 Lanes)</td>
<td>0.0536</td>
<td>$e^{-0.0536q}$</td>
<td>0.0048</td>
<td>$\frac{\ln(e^{-0.0536q})}{\ln(\frac{\tau}{\tau+1})}$</td>
</tr>
</tbody>
</table>

Figure 1 represents the estimated capacity distribution curves for the 2-lane, 4-lane and 6-lane sections of I-5 Highway and the spatial comparative plot for the estimators obtained using these parameters.
For this section, Kaplan-Meier curve ends at the probability of 0.56 where the flow (2126 veh/h) is the last flow followed by a breakdown.

There is a clear difference between the Gamma prior and other estimators.

For this section, Kaplan-Meier curve ends at the probability of 0.33 where the flow (1959 veh/h) is the last flow followed by a breakdown.

Kaplan-Meier estimator have a probability of zero.

Kaplan-Meier estimator have a probability of zero.
FIGURE 1 Capacity estimators of I-5 highway (FIGURE 1a) for section 1, (FIGURE 1b) for section 2, (FIGURE 1c) for section 3, and (FIGURE 1d) capacity comparison with Bayesian estimation for all sections.
The maximum value of the Kaplan-Meier curve only reaches 1 if the maximum observed volume \( q \) is followed by a breakdown. This is encountered in (27), where despite the used large sample, no complete distribution function could be obtained since the highest flow values observed were not followed by breakdown. Bayesian estimators solve this problem and give more complete curves for the estimated capacity distributions. To sum up, the lower and upper tails of the capacity distribution cannot be estimated by non-Bayesian estimation techniques.

Moreover, Gamma prior does not work as well as the Dirichlet prior. The Dirichlet prior estimate is closer to the Kaplan-Meier estimate than the Gamma prior estimate. This is because the estimator with Dirichlet prior assigns more mass on the breakdown flows and less mass between these flows and at the tails than Gamma prior estimate. Unlike for the Dirichlet process where the relative change due to a breakdown in the estimator occurs only at the observation itself, for the Gamma prior, some of the information obtained from a breakdown is used to change the relative weight of the points of the estimator to the left of it. Therefore, based on the observations obtained from empirical data, Dirichlet distribution is chosen as a reasonable prior for our model that represents the distribution of traffic capacity.

For the Kaplan-Meier estimator, the breakdown probability is 0 up-to the flow with first breakdown, which is not realistic for every case. For instance, consider the section with 4-lanes (VDS-1108613) of I-5 Highway. With a flow value of 1959 veh/h, the Kaplan-Meier curve has a breakdown probability of 0.33 roughly whereas the breakdown probability is 0 for 1515 veh/h. This means that the traffic is almost free flowing (no chance of breakdown) at that volume which is merely 444 veh/h less than 1959 veh/h. This unrealistic situation is addressed by the Bayesian estimator calculating the breakdown probabilities for all traffic flow values possible including the relatively lower flows for which breakdowns are not actually observed. Therefore, Bayesian estimation becomes more effective while defining the capacity profile of a freeway section.

Bayesian estimators are smoother than the Kaplan-Meier estimator in the sense that the jumps at censored observations are not as large for the Bayesian estimators. Actually, both the conventional and Bayesian estimate are discontinuous at uncensored observations. However, Bayesian approach smoothes the estimator at censored observations by achieving continuity. This smoothness depends on the value of \( \beta \), and the percentage of censored observations which differs from section to section. For example, for the rural section (VDS-1015010), the breakdowns are very limited, and at the same time, the censored observations for higher traffic volumes are not present. This makes it even difficult for the Bayesian estimation to obtain the upper part of the curve. Therefore, possible trajectories (trend lines) for the sections without enough data are plotted based on the shape of the section with VDS-1205157 which has the satisfactory data. A procedure to obtain a trend line mathematically is given in the proceeding sections.

**Application to Ramps**

PEMS ramp data (40) do not include the density and volume observations. It only has the number of vehicles passing through the detector, and their occupancies.
Therefore, we apply the following formula to calculate the speed values from the single-loop detector data (41):

\[
V = \left( \frac{Q}{O} \right) \left( \frac{\hat{L}_v + L_D}{\mu} \right)
\]

(16)

where

\( Q \): number of vehicles per 5 minutes,
\( O \): occupancy, percentage of the time the loop is occupied by vehicles during the 5 minutes interval,
\( \hat{L}_v \): average vehicle length,
\( L_D \): length of detector,
\( \mu \): conversion constant.

Choosing the average vehicle length as 16 ft, and the length of detectors obtained from the PEMS data, we obtain the speed observations for the ramp sections where \( Y \) is selected as 40 m/h, and \( \varepsilon \) as 5 m/h. Despite the large data size, it is difficult to come up with a reasonable amount of traffic observations (both breakdowns and censored observations) for the off-ramp section using Table 1. The breakdown values are very limited, and the largest flow value is followed by a breakdown. Hence, Kaplan-Meier and Bayesian estimator curves are almost the same. For VDS-1214241 (on-ramp with 2-lanes), we have better data to obtain the capacity distribution. This is possibly because the section with 2-lanes is more suitable to apply a breakdown analysis suggested in this paper. Actually, the section can be viewed as a short mainline freeway section. Moreover, the comparison figure indicates that the capacity of the urban on-ramp is significantly higher than the urban off-ramp section.
For this section, Kaplan-Meier curve reaches the probability of 1 where the flow (172 veh/h) is the last flow followed by a breakdown.

Despite the size of one-year data, it was not possible to obtain a reliable amount of breakdown observations. With the limited data, Kaplan-Meier and Bayesian Estimators are almost the same.

Capacity Range: (150-250 veh/h/lane)

For this section, Kaplan-Meier curve ends at the probability of 0.46 where the flow (189 veh/h) is the last flow followed by a breakdown.
FIGURE 2 Capacity estimators of two ramp sections of I-5 highway (FIGURE 2a) for section 1, (FIGURE 2b) for section 2, and (FIGURE 2c) capacity comparison with Bayesian estimation for all sections

Ramp metering strategies that heavily depend on the accurate estimation of the section capacities can be significantly improved using this methodology. For example, for the VDS-1214241 section, rather than having just one probability value for the capacity, a capacity range from 150 to 250 veh/h/lane can be used as (Figure 2):

- Capacity interval: 150–185 → Breakdown probability: 0.25
- Capacity interval: 185–210 → Breakdown probability: 0.50
- Capacity interval: 210–250 → Breakdown probability: 0.90

These intervals indicate that the breakdown probability can occur for any traffic flow value between the boundaries of these intervals. For instance, knowing that VDS-1214241 section has a 0.50 probability of breakdown for 185 to 210 veh/h/lane, it is possible to adjust the traffic signaling system at the end of the ramp accounting for the higher flows.

Discussion of the Hierarchical Modeling Approach

The choice of the traffic capacity depends on the engineer’s judgment/purposes at the specific period s/he is working in. In general, the main idea is to choose a strategy that ensures the maximum flow with minimum breakdown probability. Moreover, the breakdown probability initially selected at the design stage may change due to operational characteristics, external conditions, or even a new construction. Using the
presented methodology, the operating agency has the opportunity to revise the breakdown probability.

Spatially, the steepness of the curve is related to the quality and quantity of the data, which changes the parameters of the posterior distributions. For instance, $\theta$ parameter directly depends on the number of breakdown and censored observations. As a rule of thumb, the larger it is, a better data set (better representative information) we have for the roadway section. The parameter $\beta$ for Dirichlet prior, on the other hand, defines the prior mass on the traffic volume data. When the sample size is $n$, the prior mass on $(0, \infty)$ becomes $\frac{\beta}{n}$. The larger it is, the smoother the Bayesian estimator becomes when compared to the Kaplan-Meier estimator.

To calculate the Kaplan-Meier estimator, we do not have to know the actual censored observations, but the number of censored observations between two uncensored traffic volumes is sufficient. The Bayesian estimators, however, are calculated with using all the data including both uncensored and censored observations. The fact that Bayesian estimates use all the data makes it preferable to the other estimates. In this sense, the Bayesian estimators serve as a function of the sufficient statistic by fully utilizing the data.

The spatial results presented indicate the need for a hierarchical sampling approach to obtain the freeway capacity curves incorporating the stochastic nature of the traffic volume and the section characteristics. For instance, for the rural section of I-5 highway (VDS-1015010), it is difficult to obtain a full curve because data obtained do not include reasonable amount of breakdown and even censored observations (Figure 1). On the other hand, for the urban section of I-5 highway (VDS-1205157), we have a better capacity distribution estimate where substantial amount of data including breakdowns are present. Although these two sections differ in terms of driver population characteristics and number of lanes, the lanes were actually built with similar design principles, and therefore the actual capacity distribution curves of the roadways should be similar in shape. Therefore, given the sufficient amount of data, it can be possible to obtain a capacity estimator curve for VDS-1015010 by changing the parameters of the capacity estimating distribution. With this idea, data obtained for the urban section (with sufficient amount of breakdown and censored observations) is used as the base to simulate the prior distribution of the hierarchical model presented. This is also a way of assessing the robustness of the estimation results by analyzing the subsets of the traffic data samples. Geistefeldt and Brilon (1) chose half of their data set as the subset sample, and obtained the estimator curves calculating the parameters of a Weibull distribution. Similarly, we focus on the data obtained from VDS-1205171, and we take a sample of 3-months from the data set to obtain reliable parameters for the Dirichlet estimator.

To recall, the data should be in the form of:

$$q_j, j = 1, \ldots, n$$

$$\delta_j = \begin{cases} 1 & \text{if } V_j \leq Y \\ 0 & \text{if } V_j > Y \end{cases}$$

If $\delta_j = 1$, there is a traffic breakdown at volume $q_j$, and, if $\delta_j = 0$, traffic is fluent above the threshold speed without any breakdown (censored). From Equation (10), the Bayesian
estimate of the capacity curve for any breakdown $q_b$ value with the Dirichlet prior, 

\[
\hat{F}_b(q) = 1 - \left( \frac{\alpha(q, \infty) + n_q}{\alpha(0, \infty) + n} \right) \tag{17}
\]

whereas the estimate for the censored observations ($\hat{F}_c(q)$) for any censored value $q_c$ is as follows:

\[
\hat{F}_c(q) = 1 - \frac{\alpha(q, \infty) + n_q}{\alpha(0, \infty) + n_c} \prod_{j, q \text{ censored}} \frac{\alpha(q_j, \infty) + n_j}{\alpha(q_j, \infty) + n_j - \lambda_j} \tag{18}
\]

We obtain these breakdown and censored observations, namely $q_b$ and $q_c$, where $\delta = 0$ or 1 as $q = q_b$ or $q = q_c$, respectively. Then, we estimate the parameter values $\beta, \theta$ to obtain $\hat{F}_c(q)$ and $\hat{F}_b(q)$ based on the observations $(\delta, q)$ of the VDS-120517 section where $\alpha(q, \infty)$ is the empirical distribution function assigning the prior mass on each observed value of traffic volume in $(0, \infty)$ as $\frac{\beta}{n}$. To perform that, we employ the well-known bootstrap method (Efron (42)) and assign a measure of accuracy to the estimated parameters. The bootstrap methodology used for this model can be described as follows:

- We randomly obtain a 3-months data sample of size $l$ from the VDS-120517. In this data set, $\hat{F}_b(q)$ puts mass only at those observations $q = q_b$ where $\delta_j = 1$, and $\hat{F}_c(q)$ puts mass only at those observations $q = q_c$ where $\delta_j = 0$.
- This gives a bootstrap empirical distribution function $\alpha^*(q, \infty)$, the empirical sample distribution $q_j, j = 1,...,l$, and the corresponding bootstrap value $\hat{F}^* = F(q^*)$.
- These steps are repeated independently for a large number of iterations, say $N$, obtaining bootstrap values $\hat{F}^{*1}, \hat{F}^{*2},...,\hat{F}^{*N}$.
- The value of the measure of accuracy $\hat{\sigma}_{BOOT} = \sigma(\hat{F})$ is approximately calculated using the sample standard deviation of the $\hat{F}^*$ values:

\[
\hat{\sigma}_{BOOT} = \sqrt{\frac{\sum_{j=1}^{N} (\hat{F}^{*j})^2 - \left( \frac{\sum_{j=1}^{N} (\hat{F}^{*j})^2}{N} \right)^2}{N-1}} \tag{19}
\]
With this approach, we obtain the bootstrap distribution percentiles \( \hat{F}^* \), and standard deviation estimates based on those percentiles. We select \( N \) as 1000, and use the Monte Carlo simulation procedure to obtain values in Table 5. The results indicate that the sampling estimator performs sufficient enough in terms of representing the original Bayesian capacity distribution using the VDS-1205157 data. However, it is important to note that the reduced data samples still do contain a considerable amount of at least 95 breakdown observations. That is actually why this section is selected as an input for the analysis. For the data samples including only a few traffic breakdowns, it will be harder to maintain a reasonable sampling bootstrap distribution.

**TABLE 5 Bootstrap Distribution Results**

<table>
<thead>
<tr>
<th>( q )</th>
<th>( q_{278} )</th>
<th>( q_{548} )</th>
<th>( q_{818} )</th>
<th>( q_{1097} )</th>
<th>( q_{1371} )</th>
<th>( q_{1643} )</th>
<th>( q_{1918} )</th>
<th>( q_{2250} )</th>
<th>( q_{2465} )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>260</td>
<td>514</td>
<td>769</td>
<td>1042</td>
<td>1297</td>
<td>1550</td>
<td>1803</td>
<td>2125</td>
<td>2340</td>
</tr>
<tr>
<td>( \hat{F}(q) )</td>
<td>0.053</td>
<td>0.056</td>
<td>0.059</td>
<td>0.062</td>
<td>0.071</td>
<td>0.097</td>
<td>0.218</td>
<td>0.449</td>
<td>0.995</td>
</tr>
<tr>
<td>( \hat{\sigma}_{\text{BOOT}} )</td>
<td>0.088</td>
<td>0.070</td>
<td>0.065</td>
<td>0.058</td>
<td>0.051</td>
<td>0.046</td>
<td>0.042</td>
<td>0.037</td>
<td>0.031</td>
</tr>
</tbody>
</table>

Percentiles of the bootstrap distribution of \( \hat{F}^*(q) \)

<table>
<thead>
<tr>
<th>Percentile</th>
<th>( q_{278} )</th>
<th>( q_{548} )</th>
<th>( q_{818} )</th>
<th>( q_{1097} )</th>
<th>( q_{1371} )</th>
<th>( q_{1643} )</th>
<th>( q_{1918} )</th>
<th>( q_{2250} )</th>
<th>( q_{2465} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>10%</td>
<td>0.048</td>
<td>0.052</td>
<td>0.056</td>
<td>0.058</td>
<td>0.068</td>
<td>0.082</td>
<td>0.153</td>
<td>0.367</td>
<td>0.876</td>
</tr>
<tr>
<td>25%</td>
<td>0.049</td>
<td>0.054</td>
<td>0.058</td>
<td>0.060</td>
<td>0.070</td>
<td>0.089</td>
<td>0.193</td>
<td>0.402</td>
<td>0.914</td>
</tr>
<tr>
<td>50%</td>
<td>0.052</td>
<td>0.056</td>
<td>0.060</td>
<td>0.062</td>
<td>0.072</td>
<td>0.095</td>
<td>0.224</td>
<td>0.443</td>
<td>0.995</td>
</tr>
<tr>
<td>75%</td>
<td>0.055</td>
<td>0.057</td>
<td>0.062</td>
<td>0.065</td>
<td>0.075</td>
<td>0.102</td>
<td>0.301</td>
<td>0.496</td>
<td>0.996</td>
</tr>
<tr>
<td>90%</td>
<td>0.057</td>
<td>0.061</td>
<td>0.064</td>
<td>0.068</td>
<td>0.077</td>
<td>0.107</td>
<td>0.367</td>
<td>0.537</td>
<td>0.999</td>
</tr>
</tbody>
</table>

Now, we change the parameters \( \beta, \theta \) of the sampling distribution to obtain a reasonable capacity distribution curve for VDS-1015010. The selected values for the parameters \( \beta, \theta \) are given in Table 3 for VDS-1015010 roughly representing the sectional and traffic characteristics of the section. As observed in Table 5, the simulated distribution has breakdown probability values similar to those of the input data set. The issue is to which extent the sampled capacity distribution can estimate the capacities for VDS-1015010. Figure 3 shows the actual Bayesian estimator and Kaplan-Meier curve obtained for VDS-1205157, the actual Kaplan-Meier curve for VDS-1015010, and its estimated bootstrap curve obtained via sampling from the input data of VDS-1205157. The result shows that the new sampled curve gives a reasonably good fit for VDS-1015010 curve. This can be interpreted as if we had sufficient data for that section to create a capacity distribution curve.
This paper proposes a hierarchical estimation methodology to determine roadway traffic capacity distribution using nonparametric Bayesian estimation techniques based on two different priors, namely Dirichlet and Gamma process priors. The distribution of traffic capacity in the presence of censored observations is studied incorporating the spatial features of the roadway sections. Then, Kaplan-Meier estimator is compared with the Bayesian estimators. The analysis is based on five-minute interval observations from I-5 Highway in California, USA. The results indicate that, especially for urban roadways and ramp sections where breakdowns are most likely to occur, the use of a stochastic approach is necessary and crucial.

It is mostly impossible to obtain the actual capacity distribution as the flow observations have right-censored data points. However, several non-parametric estimation methods that can handle data with censored observations are available in the literature. Kaplan-Meier method is the most well-known estimation technique to obtain the capacity distribution (Geistefeldt and Brilon (1)). In the presence of sufficient amount of data, these techniques can work efficiently. The comparative results presented in this paper suggest that, for a section with sufficient amount of breakdown values, and without too many censored volume values, it is possible to use the Kaplan-Meier estimator. However, in cases where there are substantial amount of censored traffic volumes in the data, it is necessary to use a hierarchical Bayesian estimation approach to obtain a better capacity estimation curve. As shown before, Kaplan-Meier estimator is indeed a limit of the Bayesian estimator with the Dirichlet prior. Thus, we can state that the estimators

![FIGURE 3 Capacity estimator for VDS-1015010 via Sampling](image-url)
complement each other, and the traffic engineer should be careful to decide which one to
use for the specific case on hand. Moreover, the Kaplan-Meier estimator is not defined
for all possible values of traffic flow values ranging from zero to infinity. Since a
complete curve cannot be obtained, parametric distribution fitting techniques have to be
applied on the Kaplan-Meier estimator. On the other hand, the proposed non-parametric
Bayesian estimators serve as a function of the sufficient statistic by fully utilizing the
data.

Moreover, it is found that Gamma prior performs worse than Dirichlet prior. This
is basically because Dirichlet prior estimator assigns more weight on the breakdown
flows and less weight between these flows and at the tails than the one with Gamma
prior. Unlike for the Dirichlet process where the relative change due to a breakdown in
the estimator occurs only at the observation itself, for the Gamma prior, some of the
information obtained from a breakdown is used to change the relative weight of the
points of the estimator to the left of it. Therefore, based on the observations obtained
from empirical data, Dirichlet distribution is chosen as a reasonable prior for our model
that estimates the capacity distribution.

The effects of the distribution parameters on the shape of estimated capacity
curves have shown to have the utmost importance. For instance, \( \theta \) parameter directly
depends on the number of breakdown and censored observations. As a rule of thumb, the
larger it is, a better data set (better representative information) we have for the given
section. The parameter \( \beta \) for the Dirichlet prior, on the other hand, defines the prior
weight on the traffic volume data. When the sample size of the observations is \( n \), the
prior mass on \((0, \infty)\) becomes \(\frac{\beta}{n}\). The larger it is, the smoother the Bayesian estimator
becomes compared to the Kaplan-Meier estimator.

Finally, the capacity intervals for the breakdown probabilities of ramp sections
make it possible for a traffic engineer/planner to choose the proper pairs of the
breakdown probability and traffic capacity. Freeway control strategies such as ramp
metering that heavily depend on the accurate estimation of the section capacities can be
improved by using this methodology. The temporal changes in the freeway capacity can
also be studied regarding the weather conditions, seasonal demand, and daylight/darkness
conditions.

ACKNOWLEDGMENTS

We would like to acknowledge California DOT for providing us with the PEMS
data.

REFERENCES

1. Geistefeldt, J., and Brilon, W., “A Comparative Assessment of Stochastic Capacity
   Estimation Methods”, Transportation and Traffic Theory 2009, Edited by W.H.K.
   Capacity and Breakdown”, Proceedings of the 4th International Symposium on
   Highway Capacity, pp. 84-95, TRB-Circular-E-C018, Transportation Research


