Title

Empirical Analysis of Traffic Breakdown Probability Distribution with Respect to Speed and Occupancy

Authors
Chow, Andy H.F.
Lu, Xiao-Yun
Qiu, Tony Z.

Publication Date
2009-04-01

CALIFORNIA PATH PROGRAM
INSTITUTE OF TRANSPORTATION STUDIES
UNIVERSITY OF CALIFORNIA, BERKELEY

Empirical Analysis of Traffic Breakdown Probability Distribution with Respect to Speed and Occupancy

Andy H.F. Chow, Xiao-Yun Lu, Tony Z. Qiu

California PATH Working Paper
UCB-ITS-PWP-2009-5

This work was performed as part of the California PATH Program of the University of California, in cooperation with the State of California Business, Transportation, and Housing Agency, Department of Transportation, and the United States Department Transportation, Federal Highway Administration.

Report for Task Order 6224

April 2009

ISSN 1055-1417

CALIFORNIA PARTNERS FOR ADVANCED TRANSIT AND HIGHWAYS

Empirical Analysis of Traffic Breakdown Probability Distribution with Respect to Speed and Occupancy

Working Paper

Andy H.F. Chow, Xiao-Yun Lu, Tony Z. Qiu

Project Number:

FHWA-EXARP (DTFH61-07-H-00038)

and

Caltrans TO6224

9 February 2009; Revised: 7 April 2009

Key Words

Traffic breakdown, Weibull distribution, distance means speed and occupancy, Copula function, VSL, ramp metering.

Abstract

From an operation viewpoint, traffic breakdown (from free-flow) was defined as when the average speed of traffic drops below a certain threshold. It is known that traffic breakdown is a stochastic phenomenon which can happen even when the traffic flow is below the capacity. The capacity has many definitions, such as that in HCM or the average of maximum daily flow. This study investigates the probability of breakdown at certain locations of freeway. The motivation is to find a practical capacity for each freeway section for active traffic control/operation purposes, which could be different from previous viewpoints. Capacity is usually expressed in terms of flow rate. Nevertheless, it is well known that a particular value of flow rate could represent two different traffic states: uncongested and congested. Therefore, simply considering flow rates as the main factor is inadequate for operational purpose. In this study, a bivariate Weibull distribution is adopted to model the probability of breakdown as a function of both mean speed and occupancy of the incoming traffic. The methodology of constructing and calibrating the bivariate distribution is introduced. In addition, three case studies are performed to test the methodology proposed herein. The case studies are carried out by using three different datasets: PORTAL, PeMS, and BHL (Berkeley Highway Lab). PORTAL is an archived data source collected from freeways in Oregon, while the other two are collected from freeways in California PATH. The datasets measure and process flow rates, occupancies, and speeds of traffic from the loop stations on the freeways. Empirical results derived and their potential applications are discussed for developing various traffic control strategies including Variable Speed Limit (VSL) and ramp metering.

Acknowledgements

This work was supported by the Federal Highway Administration (FHWA) Exploratory Advanced Research Project (Cooperative Agreement DTFH61-07-H-00038) with Match Funding from the California Department of Transportation (Caltrans) TO6224. The contents of this report reflect the views of the authors who are responsible for the facts and the accuracy of the data presented herein. The contents do not necessarily reflect the official views or policies of FHWA and the State of California. This paper does not constitute a standard, specification, or regulation.

The authors would like to acknowledge Robert Bertini of Portland State University for his advice on the PORTAL database; Karl Petty of Berkeley Transportation Systems for his advice on the PeMS database.

Contents

List of Figures and Tables

Page

Figures

Tables

Executive Summary

This study investigates the probability of traffic breakdown from free-flow on a freeway. From traffic operation point of view, traffic breakdown occurs when average speed of traffic drops rapidly to below a certain threshold. It has been revealed in the literature that such traffic breakdown is a stochastic phenomenon which can happen even when the traffic flow is below the capacity. The capacity has many definitions, such as that in HCM or the average of maximum daily flow. In practice, the capacity of the highway section is very difficult to measure due many uncertain factors including driver behaviour, road geometry, weather condition, and visibility. The motivation of this study is to find a practical capacity for each freeway section for active traffic control/operation purposes, which could be different from previous definitions/viewpoints.

Brilon et al. (2005) proposed an empirical approach to analyze the probability of breakdown based on univariate Weibull distribution with respect to flow. It looked at the traffic data collected from freeways A1 and A3 in Cologne, Germany. Nevertheless, it has been revealed that a particular value of flow rate could represent two different traffic states: uncongested and congested. Therefore, simply considering flow rates as the main factor is inadequate for operational purpose. This study adopts a bivariate Weibull distribution to model the probability of breakdown as a function of the combination of mean speed and occupancy (equivalent to density) of the incoming traffic. The methodology of constructing and calibrating the bivariate distribution is introduced.

Three case studies have been conducted to test the proposed methodology with different datasets: PORTAL (Portland Oregon Regional Transportation Archive Listing), PeMS (Freeway Performance Measurement System), and BHL (Berkeley Highway Lab). PORTAL is an archived data source collected from freeways in Oregon, while the other two are collected from freeways in California. Contour plot of the breakdown probability with respect to speed and occupancy of approaching traffic are derived at each detector station. The contour plot is useful for designing traffic control strategies such as Variable Speed Limit (VSL) and ramp metering. An optimal traffic control problem typically can be formulated as an optimization problem with an objective to maximize total benefit of the system (e.g. minimizing the total system delay) subject to a set of constraints including traffic dynamics and constraints on control variables. This study can be used to construct a constraint as upper bounds to occupancies and speeds such that the traffic breakdown probability is less than a specified threshold, say 10%.

For real time deployment, the parameters of the Weibull distribution can be updated with real time traffic data by using the rolling horizon concept. Progress and results will be reported in the future.

Chapter 1. Introduction

This study investigates the probability of traffic breakdown from free-flow on a freeway. From traffic operation point of view, traffic breakdown occurs when average speed of traffic drops rapidly to below a certain threshold (Banks, 2006). It is widely believed that breakdown occurs when the flow rate of traffic passing through a bottleneck exceeds its capacity. The Highway Capacity Manual (HCM) defines the capacity of bottleneck as the maximum sustainable flow at which vehicles and persons reasonably can be expected to traverse it during a specified time period under given roadway, geometric, traffic, environment, and control condition (Transportation Research Board, 2000). The Highway Capacity Manual assumes that there is no influence from downstream traffic operations, such as the backing up of traffic into the analysis point. Further detail of the HCM approach can be referred to Section 1.2 in Banks (2006). Jia et al. (2000) reported that such maximum throughput occurs at the free flow speed of 60 mph, and not between 35 and 45 mph, as is often assumed. Consequently, Jia et al. (2000) suggested that congestion should be measured as the additional vehicle-hours of delay traveling below 60 mph.

Nevertheless, the definition of capacity above by the HCM has been criticized as unsatisfactory for operational purposes (Zhang and Levinson, 2004; Brilon et al., 2005). The reason is that previous definition of capacity is a deterministic value while highway traffic is a complicated stochastic process. Therefore, it makes better sense to investigate the description of traffic capacity from the other side: traffic breakdown probability. It has been revealed in recent studies that breakdown occurrences are stochastic, which can happen even when the traffic flow is below the capacity (Elefteriadou et al., 1995). Evans et al. (2001) first developed a model to estimate the probability of breakdown at ramps by using Markov chains. Lorenz and Elefteriadou (2001) performed an empirical analysis of speed and flow data collected from Highway 401 in Toronto, Canada. Based on empirical observations, Lorenz and Elefteriadou (2001) defined breakdown occurs when the average speed of all lanes drops below 90 km/h (~56 mph) for a period of at least five minutes. The speed threshold proposed by Lorenz and Elefteriadou (2001) was indeed close to the one suggested by Jia et al. (2000).

Recently, Brilon et al. (2005) proposed an empirical approach to analyze the probability of breakdown based on univariate Weibull distribution with respect to flow. It looked at the traffic data collected from freeways A1 and A3 in Cologne, Germany. The data included flow rates and speeds, which were aggregated into 5-min intervals. The size of the time interval (5-min) was selected after a series of experiments (Brilon and Zurlinden, 2003). Nevertheless, it was known that a particular value of flow rate can represent two different traffic states: uncongested and congested. Therefore, simply considering flow rates is inadequate for operational purpose.

This study proposes an empirical approach to analyze traffic breakdown based on bivariate Weibull distribution with respect to both mean speed and occupancy (equivalent to density) of the approaching traffic. A methodology of calibration based on maximum likelihood estimation has been developed. The proposed methodology is then applied to three case studies using three different datasets: PORTAL, PeMS (Choe et al, 2002), and BHL (2008).
PORTAL (Portland Oregon Regional Transportation Archive Listing) is an archived data source collected from Oregon freeways. PORTAL measures flow, occupancy, and mean speed from the loop stations. The data is collected and processed in 20-sec time intervals.

Suggested by the PORTAL research team, we select a recurrent bottleneck on Northbound I-5 near Terwilliger Avenue in Portland, Oregon. It is a large horizontal curve on I-5N just south of downtown Portland. PeMS (Freeway Performance Measurement System) is a dataset for California freeways. Flow rates and occupancies at single-loop stations are first measured every 30-sec. PeMS then processes and aggregates the data into 5-min time intervals. PeMS estimates speeds from the occupancy measurements by using a 'g-factor' approach (Jia et al., 2001). With the PeMS data, we select a 4-mile section of I-80 freeway in Berkeley, CA. The section starts from Buchanan in Albany to San Francisco Bay Bridge in Emeryville. The section is considered to be one of the busiest freeway segments in the Bay Area. The Berkeley Highway Laboratory (BHL) data are collected from a 2.7 mile section of Interstate 80 in west Berkeley and Emeryville. The facility has video cameras and dual-loop installed at 8 stations to monitor traffic for each lane. The data is updated in 1Hz but contains 10Hz information (loop up and down time instant in 100ms), with which the lane (point) speed can be directly estimated from the up/down time instant of the dual loops. The outcome of this study is expected to be used for developing active traffic control strategies including VSL and ramp metering.

This report is organized as follows: Chapter 2 starts with a review of Brilon's (2005) approach. A methodology of constructing and calibrating the bivariate Weibull distribution is then introduced. Chapter 3 details the case studies. Empirical results and their implications on developing traffic control strategies are discussed. Chapter 4 gives some concluding remarks of the study. Finally, Chapter 5 suggests some possible future research direction.

Chapter 2. Methodology

Since the analysis in this report is an extension of Brilon's (2005) work, this chapter starts with a review of Brilon's (2005) approach on probability distribution analysis of traffic breakdown with respect to flow. A methodology of constructing and calibrating bivariate Weibull distribution is then introduced, which will be used in our analysis.

2.1 A Review of Brilon's (2005) Approach

Brilon et al. (2005) analyzed the data collected from freeways A1 and A3 in Cologne, Germany. The data includes flow rates and speeds, which are aggregated into 5-min intervals. This size of the time interval (5-min) is selected after a series of experiments (Brilon and Zurlinden, 2003).

Brilon et al. (2005) classified their traffic data (flow, occupancy, speed) into the following three categories for analyzing breakdown:

- Case 1: $v(t) > v^*, v(t+1) < v^*$
  Traffic at the current time t is regarded as a realization of traffic flow causing breakdown.

- Case 2: $v(t) > v^*, v(t+1) > v^*$
  It implies the system is able to accommodate the current traffic state.

- Case 3: $v(t) < v^*$
  Traffic at the current time is in congestion. The associated traffic data is discarded as it contains no useful information for analysis.

The notation v^* is the speed threshold that defines traffic breakdown. Brilon et al. (2005) set this threshold to be 70 km/h (~45 mph), which was lower than the one suggested by Jia et al. (2000) and Lorenz and Elefteriadou (2001). Brilon et al. (2005) argued that the speed threshold of 70 km/h was found to be fairly representative for German freeways although it can be different for different situations.

After classifying the traffic data into breakdown and non-breakdown flow, Brilon et al. (2005) defined the following cumulative distribution function (CDF),

$$F_c(q) = P(c \leq q), \tag{2-1}$$

where c is the capacity of the bottleneck, q is incoming traffic volume. Brilon et al. (2005) assumed that traffic breakdown follows Weibull distribution, and the probability of breakdown depends on the flow rate of the incoming traffic q. Consequently,

$$F_c(q) = 1 - e^{-\left(\frac{q}{\beta_q}\right)^{\alpha_q}} \tag{2-2}$$

with parameters α_q and β_q, the probabilistic density function of capacity is

$$f_c(q) = \frac{\alpha_q}{\beta_q}\left(\frac{q}{\beta_q}\right)^{\alpha_q - 1} e^{-\left(\frac{q}{\beta_q}\right)^{\alpha_q}} . \tag{2-3}$$

Expressions (2-2) and (2-3) are the functional forms of cumulative Weibull distribution and its density function respectively. Indeed, Weibull distribution is widely used for analyzing system reliability due to its flexibility (Wikipedia, 2008). The distribution can mimic the behavior of other statistical distributions such as the normal (with $\alpha_q = 3.4$) and the exponential (with $\alpha_q = 1$).

For Weibull distribution, the mean of the variable q is given by

$$\bar{q} = \beta_q \Gamma\left(\frac{1}{\alpha_q} + 1\right), \tag{2-4}$$

where $\Gamma(\cdot)$ is the gamma function which is defined as

$$\Gamma(n) = \int_0^\infty e^{-x} x^{n-1} dx . \tag{2-5}$$

The standard deviation of q is given by

$$\sigma_q = \beta_q \sqrt{\Gamma\left(\frac{2}{\alpha_q} + 1\right) - \Gamma^2\left(\frac{1}{\alpha_q} + 1\right)} . \tag{2-6}$$

In the work of Brilon et al. (2005), the parameters α_q and β_q are determined by maximum likelihood estimation (MLE). It defined the likelihood function of the Weibull distribution as

$$L_q = \prod_{i=1}^{n} \left\{ \left[f_c(q_i) \right]^{\theta_i} [1 - F_c(q_i)]^{(1-\theta_i)} \right\}, \tag{2-7}$$

where q_i is the flow rate at time interval *i; n* is the total number of time intervals; $\theta_i = 1$ if traffic at time interval *i* causes the speed drop below the speed threshold *v** at time interval *i*+1, $\theta_i = 0$ otherwise. Nevertheless, Brilon et al. (2005) did not give the details of the solution procedure.

2.2 Extension to Bivariate Weibull distribution – Copula Function Approach

Brilon et al. (2005) adopted a univariate Weibull distribution in which probability of breakdown was regarded as a function of incoming flow rate.

As aforementioned, traffic flow was related to the associated density and (space) mean speed as

$$q = k\bar{v}_s \,, \tag{2-8}$$

where k is the density, and $\bar{v}_s$ is the space-mean speed. It is known empirically that a particular value of flow rate q can represent two different traffic states: uncongested and congested. Uncongested state corresponds to traffic with a high speed (free-flow speed) but a low density (below critical density); congested state represents traffic with a low speed but a high density (above critical density). As a result, simply considering flow rates may be inadequate for operational purpose. This is the motivation to consider traffic breakdown with respect to both density and mean speed.

This study extends the univariate Weibull distribution function to bivariate form. Following this, the probability of traffic breakdown is represented as a function of the combination of mean speed and occupancy[1] of incoming traffic.

Copula functions (Sklar, 1973) are used as a general way of formulating a multivariate distribution. The structure of a copula function consists of two parts: one describing the dependence structure and the other describing the marginal behaviours. Denote the bivariate cumulative distribution of traffic breakdown as $H_c(\rho, v)$, which depends on the occupancy ρ and speed v of the approaching traffic. Following Sklar (1973), a bivariate distribution can be constructed as

$$H_c(\rho, v) = C(u, w) \,. \tag{2-9}$$

The density function $h_c(\rho, v)$ of the cumulative distribution $H_c(\rho, v)$ is

$$h_c(\rho, v) = u' w' c(u, w) \,, \tag{2-10}$$

in which $u = F_c(\rho) = 1 - e^{-\left(\frac{\rho}{\beta_\rho}\right)^{\alpha_\rho}}$ and $w = G_c(v) = 1 - e^{-\left(\frac{v}{\beta_v}\right)^{\alpha_v}}$ are the marginal cumulative probabilistic functions of speed and occupancy leading to breakdowns respectively; $u' = f_c(\rho) = \frac{\alpha_\rho}{\beta_\rho}\left(\frac{\rho}{\beta_\rho}\right)^{\alpha_\rho - 1} e^{-\left(\frac{\rho}{\beta_\rho}\right)^{\alpha_\rho}}$ and $w' = g_c(v) = \frac{\alpha_v}{\beta_v}\left(\frac{v}{\beta_v}\right)^{\alpha_v - 1} e^{-\left(\frac{v}{\beta_v}\right)^{\alpha_v}}$ are the associated probability density functions.

[1] In traffic engineering, density refers to the number of vehicles per unit length of road at particular time; occupancy refers to the percentage of time that a detector at a particular location is occupied by vehicles within a given time period (say 5 mins). Density and occupancy are related in which density is practically taken as the associated occupancy divided by an average effective length of vehicles (say ~16 – 17ft) on the road. Occupancy is used in the present analysis as it is directly measured from the detectors, while density is not.

Many copula functions are available and some of the examples can be referred to Genest and Rivest (1993). After comparing some candidate, through literature review and numerical experiments, the functional form proposed by Frank (1979) is selected in the present study,

$$C(u,w) = -\frac{1}{\delta}\ln\left[1+\frac{(e^{-\delta u}-1)(e^{-\delta w}-1)}{(e^{-\delta}-1)}\right], \tag{2-11}$$

in which $\delta \neq 0$. The associate density function is derived as

$$c(u,w) = \frac{\partial^2 C(u,w)}{\partial u \partial w} = \frac{-\delta(e^{-\delta}-1)e^{-\delta(1+u+w)}}{\left[e^{-\delta}-e^{-\delta(1+u)}-e^{-\delta(1+w)}+e^{-\delta(u+w)}\right]^2}. \tag{2-12}$$

The larger the absolute value of δ, the stronger the dependence between the variables, where $\delta > 0$ implies positive dependence, $\delta < 0$ implies negative dependence, $\delta \to 0$ gives $C(u,w) = uw$ which implies independence. The bivariate function is bounded by Frèchet-Hoeffding boundaries[2] as $\delta \to \pm\infty$.

The concordance of the variables $u = F_c(\rho)$ and $w = G_c(v)$ can be determined from δ by the Kendall's tau function,

$$\begin{aligned}\tau(\delta) &= 4\int_0^1\int_0^1 C(u,w)dC(u,w) - 1 \\ &= 1-\frac{4}{\delta}+\frac{4}{\delta^2}\int_0^{\delta}\frac{x}{e^x-1}dx\end{aligned}, \tag{2-13}$$

which take values in [-1, 1] and take the value of zero if $u = F_c(\rho)$ and $w = G_c(v)$ are independent of each other[3]. Two variables are concordant if large values of one variable are associated with large values of the other variables, and vice versa. Frank's bivariate function is said to be comprehensive as they accommodate all of the possible dependence.

2.3 Calibration of the Bivariate Weibull Distribution

The parameters of the bivariate distribution function can be estimated by using the method of maximum likelihood. Those parameters include: $(\alpha_\rho, \beta_\rho, \alpha_v, \beta_v, \delta)$. For this purpose, define the likelihood function as

$$L_h = \prod_{i=1}^{n} h_c(\rho_i, v_i) = \prod_{i=1}^{n} f_c(\rho_i) g_c(v_i) c(u_i, w_i), \tag{2-14}$$

---

[2] For bivariate distributions, it means $\max(0, u+w-1) \le C(u,w) \le \min(u,w)$.

[3] The term $\frac{1}{\delta}\int_0^{\delta}\frac{x}{e^x-1}dx$ in the expression is called Debye function, which does not have a closed form expression.

which leads to

$$\ln L_h = \sum_{i=1}^{n} \ln f_c(\rho_i) + \sum_{i=1}^{n} \ln g_c(v_i) + \sum_{i=1}^{n} \ln c(u_i, w_i), \tag{2-15}$$

where n is the number of samples, $u_i = F_c(\rho_i)$ and $w_i = G_c(v_i)$.

The first two terms in (2-15) are the log-likelihood functions associated with the marginal distributions, while the last term is the log-likelihood function associated with the Copula parameter. Since the three terms are independent of each other, the marginal distributions and Copula parameter δ of the bivariate distribution function can be estimated sequentially (i.e. with a two-stage estimation).

2.3.1 Estimation of marginal parameters

Brilon et al. (2005) did not detail how they solved the maximum likelihood estimation. Following Balakrishnan and Kateri (2008), it can be shown that solving the maximum likelihood function (2-7) is equivalent to jointly solving the following set of equations:

$$\frac{1}{\alpha_q} = \frac{\sum_{i=1}^{n} (\ln q_i) q_i^{\alpha_q}}{\sum_{i=1}^{n} (q_i)^{\alpha_q}} - \frac{1}{r} \sum_{i=1}^{r} \ln q_{i:n} \tag{2-16}$$

$$\beta_q = \left\{ \frac{1}{r} \sum_{i=1}^{n} (q_i)^{\alpha_q} \right\}^{\frac{1}{\alpha_q}} \tag{2-17}$$

where r is the total number of time intervals with $\theta_i = 1$; $q_{i:n}$ represents the flow rate at a time interval i in which $\theta_i = 1$.

The derivation is referred to the Appendix A.

Furthermore, define

$$\Omega(\alpha_q, \mathbf{q}) = \frac{\sum_{i=1}^{n} (\ln q_i) q_i^{\alpha_q}}{\sum_{i=1}^{n} (q_i)^{\alpha_q}} - \frac{1}{r} \sum_{i=1}^{r} \ln q_{i:n}, \tag{2-18}$$

which can be shown to be monotone increasing in α_q (Balakrishnan and Kateri, 2008). Balakrishnan and Kateri (2008) also showed that there exists a finite upper limit for $\Omega(\alpha_q, \mathbf{q})$,

$$\lim_{\alpha_q \to \infty} \Omega(\alpha_q, \mathbf{q}) = \ln q_n - \frac{1}{r} \sum_{i=1}^{r} \ln q_{i:n} = \frac{1}{r} \sum_{i=1}^{r} \ln \frac{q_n}{q_{i:n}} > 0, \tag{2-19}$$

where q_n is the maximum number in $\mathbf{q}$.

As $\frac{1}{\alpha_q}$ is decreasing in α_q, a solution of α_q can be determined by solving

$$\Omega(\alpha_q, \mathbf{q}) - \frac{1}{\alpha_q} = 0, \tag{2-20}$$

and hence β_q can also be calculated accordingly.

Likewise, for marginal distributions in occupancy and speed, the corresponding parameters are determined as:

$$\frac{1}{\alpha_\rho} = \frac{\sum_{i=1}^{n} (\ln \rho_i) \rho_i^{\alpha_\rho}}{\sum_{i=1}^{n} (\rho_i)^{\alpha_\rho}} - \frac{1}{r} \sum_{i=1}^{r} \ln \rho_{i:n}, \tag{2-21}$$

$$\beta_\rho = \left\{ \frac{1}{r} \sum_{i=1}^{n} (\rho_i)^{\alpha_\rho} \right\}^{\frac{1}{\alpha_\rho}}, \tag{2-22}$$

$$\frac{1}{\alpha_v} = \frac{\sum_{i=1}^{n} (\ln v_i) v_i^{\alpha_v}}{\sum_{i=1}^{n} (v_i)^{\alpha_v}} - \frac{1}{r} \sum_{i=1}^{r} \ln v_{i:n}, \tag{2-23}$$

$$\beta_v = \left\{ \frac{1}{r} \sum_{i=1}^{n} (v_i)^{\alpha_v} \right\}^{\frac{1}{\alpha_v}}, \tag{2-24}$$

where ρ_i and v_i represent respectively the occupancy and the speed at time interval *i*; $\rho_{i:n}$ and $v_{i:n}$ represents respectively the occupancy and the speed at a time interval *i* in which $\theta_i = 1$.

2.3.2 Estimation of Copula parameter

The Copula parameter δ is solved such that the following log-likelihood function

$$\ln L_\delta = \sum_{i=1}^{n} \ln c(u_i, w_i). \tag{2-25}$$

is maximized. Due to the complicated form of the log-likelihood function $\ln L_\delta$ and its derivative with respect to δ, the parameter δ is determined by using a golden section search such that $\ln L_\delta$ is maximized. Theoretically speaking, the parameter δ could take any value in $(-\infty, +\infty)$. However, the searching interval is set to be $(-10000, +10000)$ when the solver is implemented.

In the proposed two-step calibration, since the marginal parameters can be determined by closed-form expressions in Step 1, determining the Copula parameter (i.e. Step 2) will be more demanding computationally.

Chapter 3. Case studies

The methodology introduced in the previous chapter is tested with three case studies with three different datasets: PORTAL, PeMS, and BHL. In this chapter, the features of each dataset and the associated test site are first introduced, the empirical findings are then discussed.

3.1 PORTAL

Portland Oregon Regional Transportation Archive Listing (see PORTAL, 2008) is an archived dataset of freeways in Oregon from PORTAL database. PORTAL measures flow, occupancy, and mean speed from the loop stations. The data is collected, aggregated, and processed into 20s time intervals. Suggested by the PORTAL research team, we select a recurrent bottleneck on Northbound I-5 near Terwilliger Avenue in Portland, Oregon. It is a large horizontal curve on I-5N just south of downtown Portland. The data analyzed in this case study is collected from 0:00 to 23:59 on 4-8 August 2008 (Mon – Fri).

We adopt the classification of Brilon et al. (2005) in Section 2.1 for analyzing breakdown flow. It is reckoned that choosing the speed threshold v^* for identifying breakdown flow is crucial in the present analysis. In a recent study, Dowling et al. (2008) suggested that this speed threshold should be determined by looking at the speed variations between sequential time intervals. The speed threshold should be located with the greatest speed changes. Following Dowling et al. (2008), we select the speed threshold v^* based on speed variations between sequential time intervals. To illustrate how it is done, Figure 1 plots the speed variations between sequential 20-sec intervals at the selected bottleneck by using PORTAL data. The scatter plot depicts the greatest variability at around 45 mph. Consequently, we set v^* to be 45 mph in this case.

After defining the speed threshold, we categorize the traffic data into breakdown and non-breakdown flow. We then construct the bivariate cumulative distribution of traffic breakdown $H_c(\rho, v)$ by using Copula approach in Section 2.2. Finally, we calibrate $H_c(\rho, v)$ with the categorized data by using the two-stage method in Section 2.3.

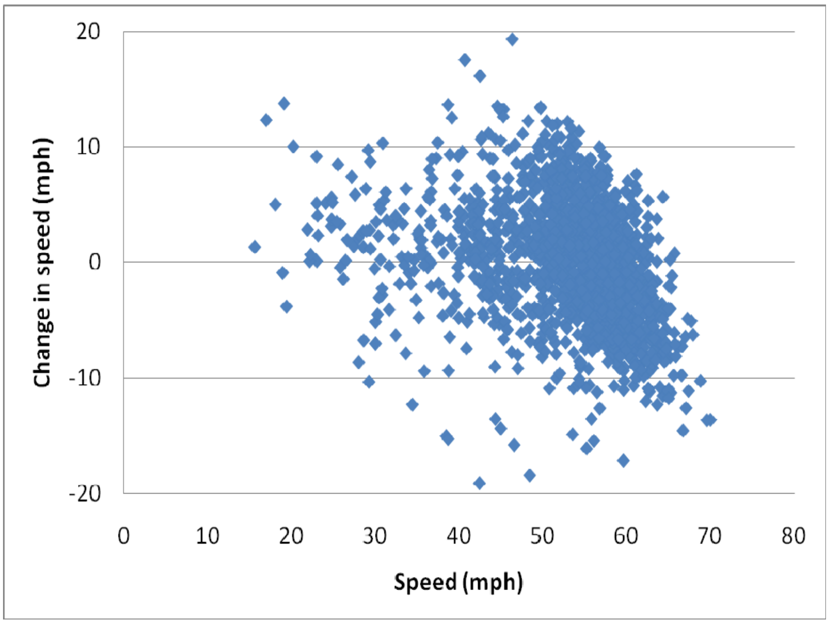


Figure 3-1: Speed variations at Terwilliger, I-5N, OR

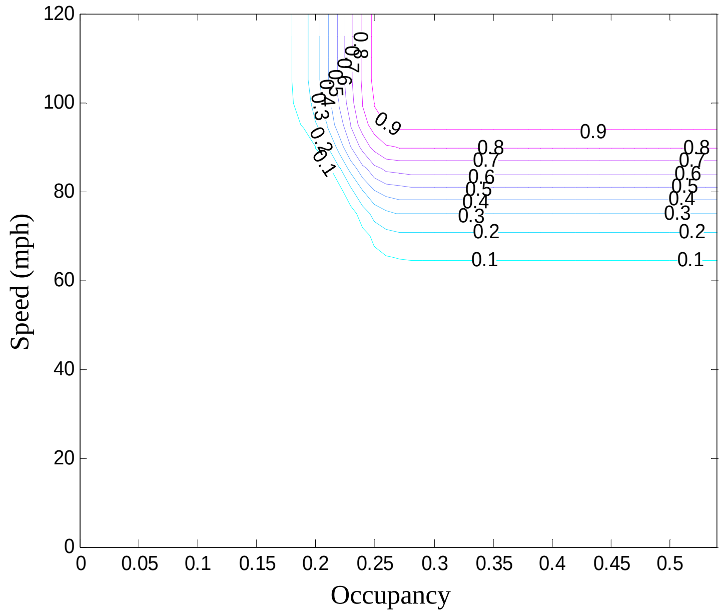


Figure 3-2: Contour of breakdown probability – PORTAL data

Figure 3-2 shows the contour of the bivariate breakdown probability, $H_c(\rho, v)$, with respect to speed and occupancy of approaching traffic. Each coordinate in the plot, which represents a combination of speed and occupancy of approaching traffic, on the same contour has the same probability of inducing breakdown. The scatter plot in Figure 3-2 provides useful information for designing traffic control strategies. For example, it can be seen that the probability of breakdown drops below 10% if the speed and the occupancy of the approaching traffic are limited to less than 60 mph and 0.15 respectively.

3.2 PeMS

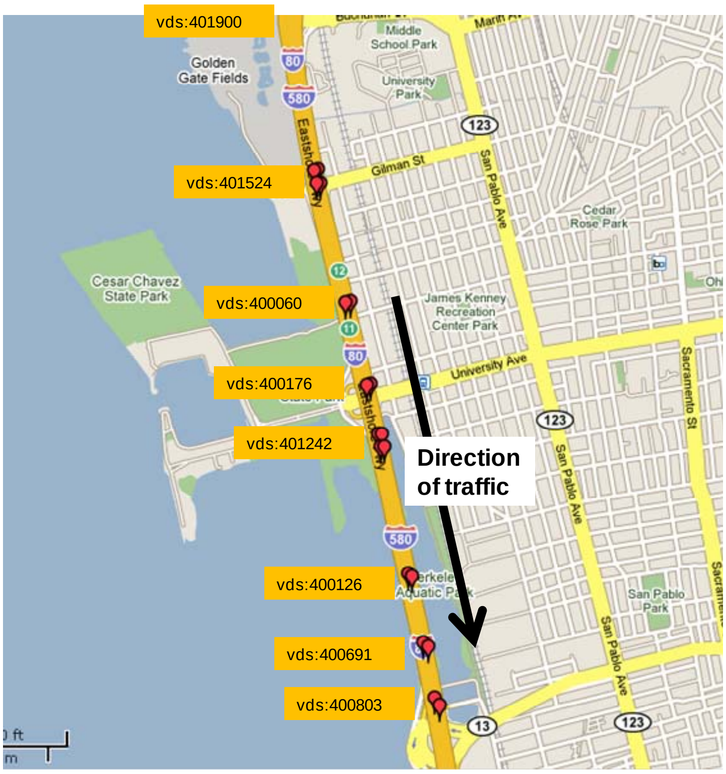
Freeway Performance Measurement System (see PeMS, 2008) is a dataset for California freeways. Flow rates and occupancies at single-loop stations are first measured every 30-sec. PeMS then processes and aggregates the data into 5-min time intervals. PeMS estimates speeds from the measurements by using a 'g-factor' approach (Jia et al., 2001), in which

$$\hat{v} = g\frac{q}{\rho}, \tag{3-1}$$

where $\hat{v}$ is the estimated speed; ρ, q are the measurements of occupancy and flow rate respectively; g is the g-factor.

The g-factor is a combination of the average length of the vehicles in the traffic stream and the tuning of the loop detector itself. Typically, a constant value for the g-factor is used which leads to inaccurate speeds because the g-factor varies by lane, time-of-day, as well as the loop sensitivity. PeMS estimates a g-factor for each loop for every 5 minutes over an average week to provide a more accurate speed estimates.

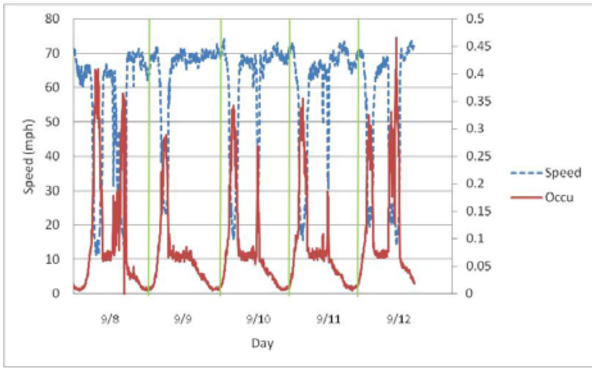
With the PeMS data, we select a 4-mile section of I-80 freeway in West Berkeley, CA, as shown in Figure 3-3. The section starts from Buchanan in Albany to San Francisco Bay Bridge in Emeryville. This section is considered to be one of the busiest freeway segments in San Francisco Bay Area.



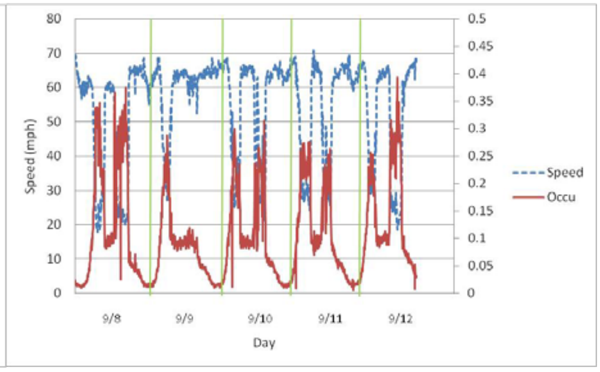
source: maps.google.com

Figure 3-3: I-80W – Buchanan to Powell, West Berkeley, CA

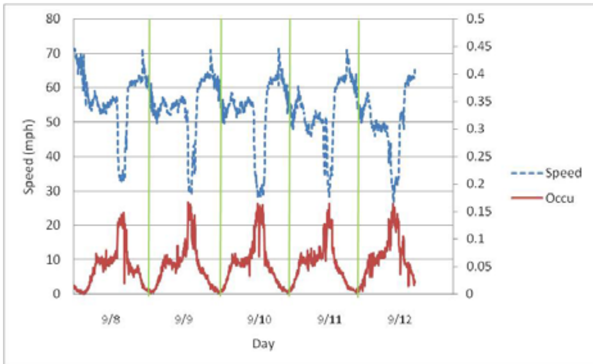
We use PeMS data collected on the weekdays (in total 21 days) in September 2008. There are 10 vehicle detector stations (VDSs) within the stretch and the profiles of speed and occupancy are plotted over a week (8 Sept 2008 – 12 Sept 2008) in Figure 3-4. The plot shows that the congestion is generally a recurrent phenomenon.



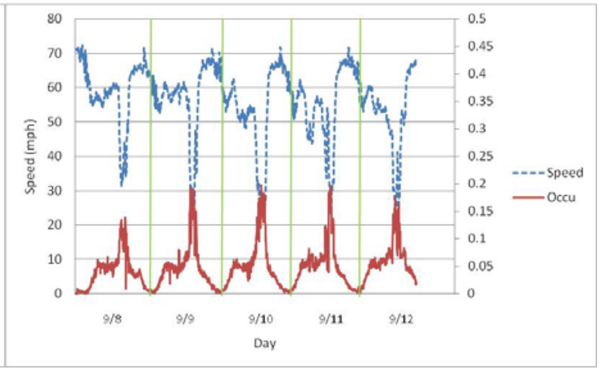
VDS 401900



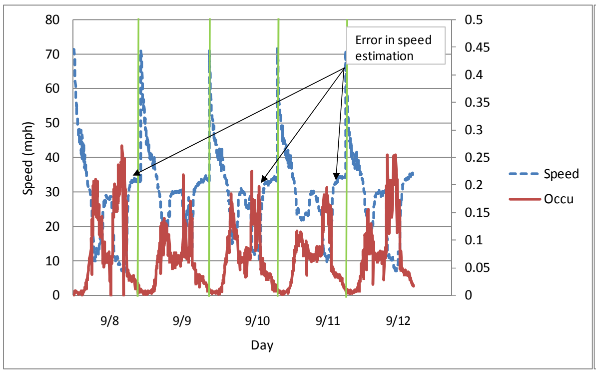
VDS 401524



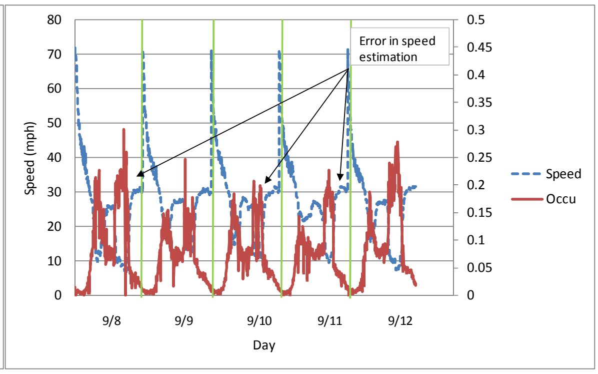
VDS: 400060



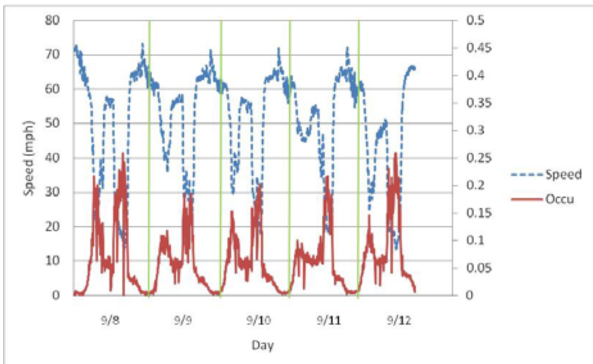
VDS 400176



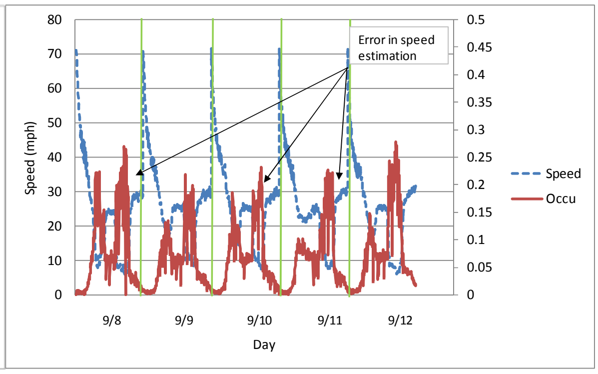
VDS: 401242



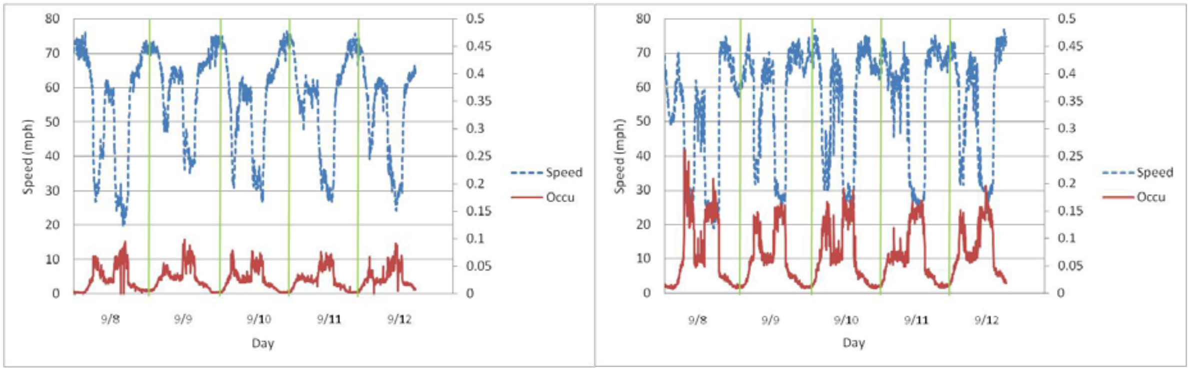
VDS 400126



VDS: 400691



VDS 400803



VDS: 401211 VDS 401698

Figure 3-4: Profiles of speed and occupancy – PeMS data

In the present analysis, we do not include data collected from VDSs 401242, 400126 and 400803 because the speed estimations at those stations do not appear to be correct. Those speed estimates are low even when the occupancy measurements are low at midnight ~21:00 and 24:00 each day, and then abruptly rise to 70 mph at the beginning of the next day. We suspect that the error is due to the 'g-factor' calculation in PeMS at those stations.

For readers' interest, we depict the geographical layout of the loop stations covered in this study in Figure 3-5. Each lane on the freeway is 12 ft wide. The 3+ HOV lane is in operation on weekdays from 0500 – 1000, and 1500 – 1900.

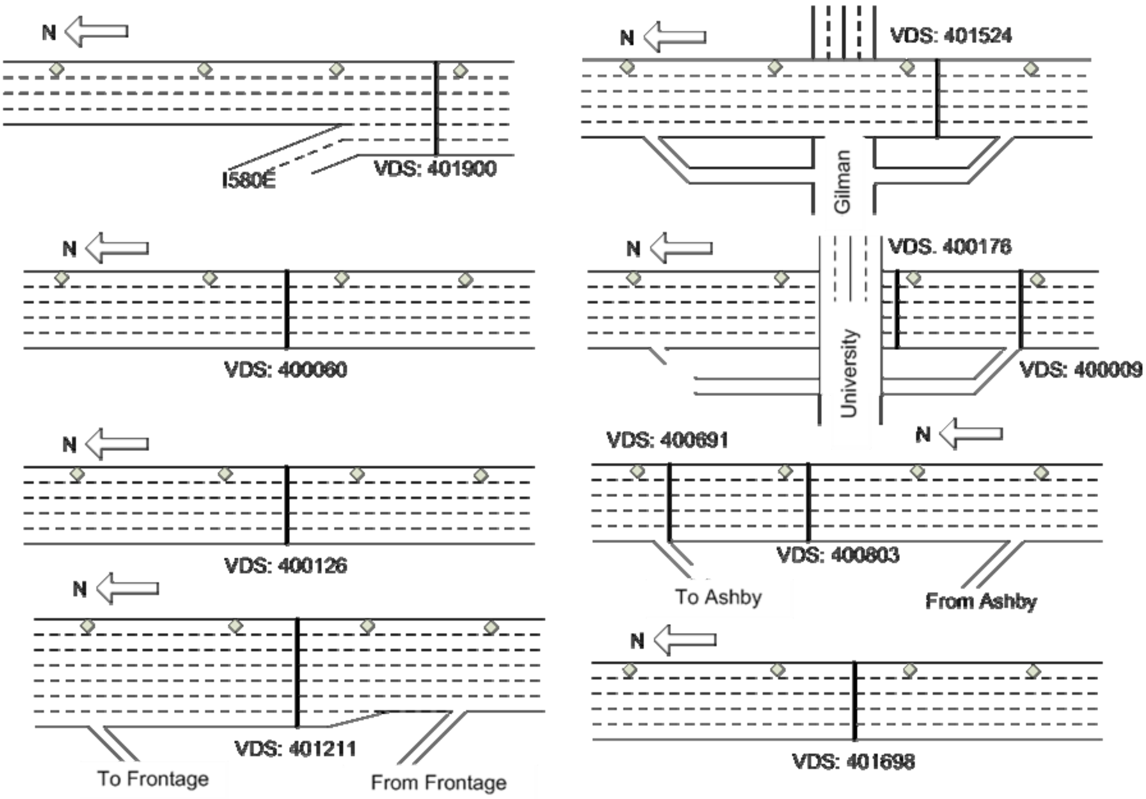
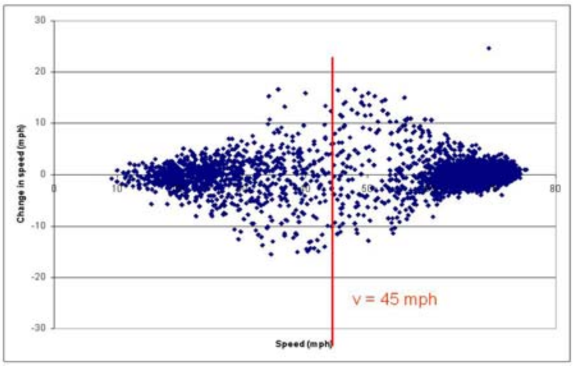
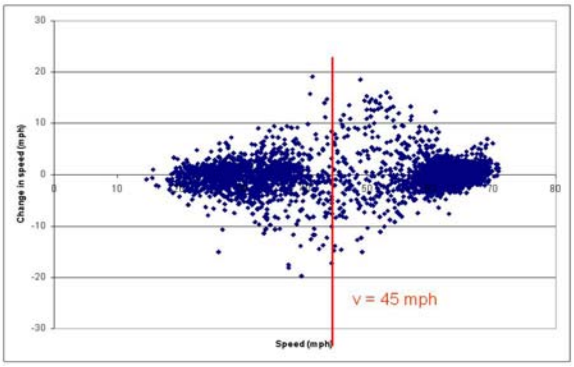


Figure 3-5: Layout of the PeMS stations

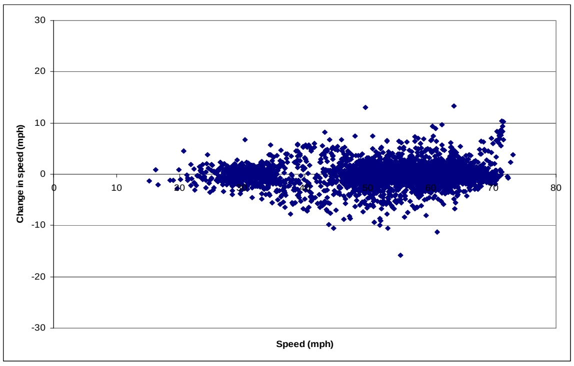
Figure 3-6 plots the speed variations between sequential 5-min intervals at the selected detector stations. The scatter plots of speed change (difference of two speeds in consecutive time steps) versus speed at stations 401900, 401524, and 401698 form a general diamond sharp with the greatest variability at around 45 mph. There are two clusters of stability in the 15-30 mph range (correspond to congested situations) and 55 – 70 mph (correspond to uncongested situations). The other four stations (400060, 400176, 400691, and 401211) generally show stability for all ranges of speeds, which suggests that the traffic is homogenized at those locations. We reckon those stations are 'non-critical' locations. Cconsequently, we set v^* to be 45 mph.



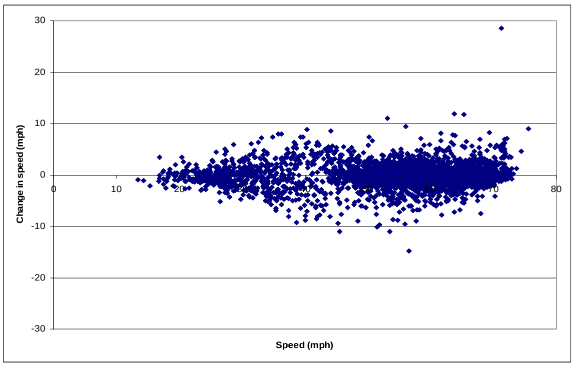
VDS: 401900



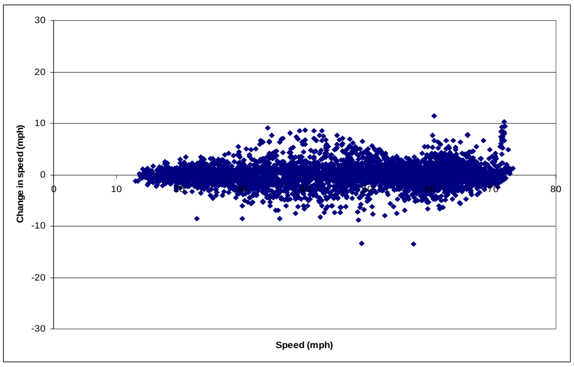
VDS 401524



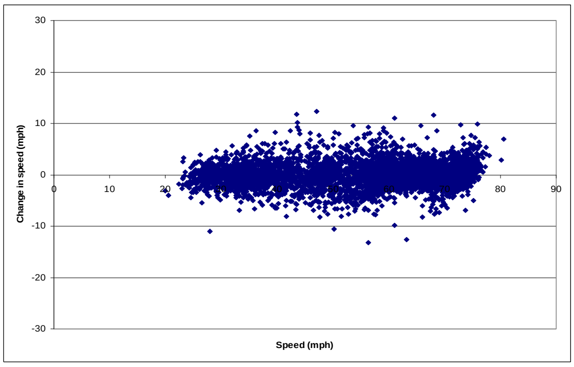
VDS: 400060



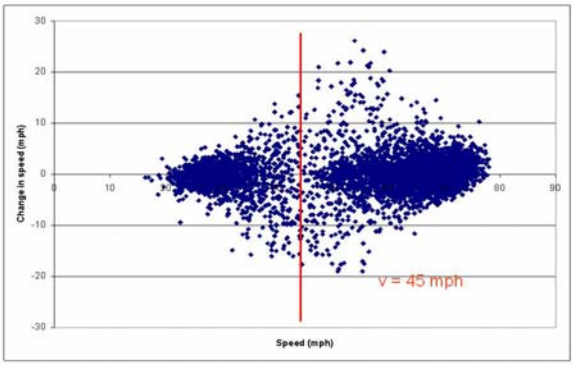
VDS 400176



VDS: 400691



VDS 401211

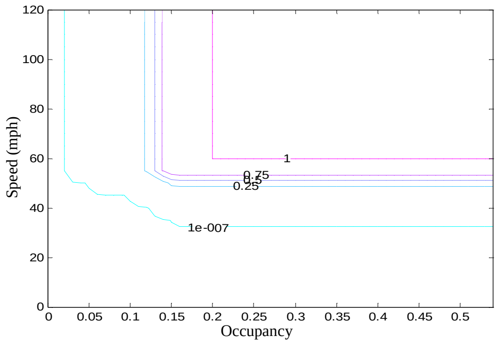


VDS: 401698

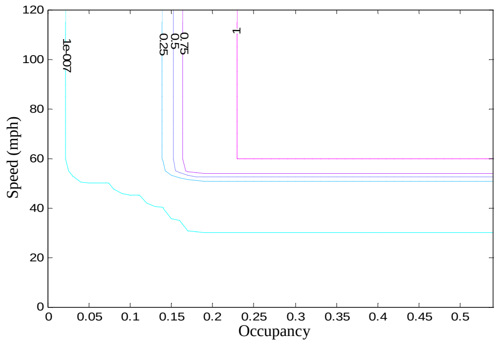
Figure 3-6: Speed variations – West Berkeley, I-80W, CA

Figure 3-7 shows the contour of the breakdown probability with respect to speed and occupancy of approaching traffic. It is interesting to note that the shapes of the contours are similar to each other along the stretch except station VDS 401211 at which the probability of breakdown is significantly lower than the other locations. It is not clear at this stage why the breakdown probability contours at VDS 401211 differ from the others significantly. Further investigation will be needed, for instance, to examine the data quality, characteristics of that site and the frequency of incidents at that location during the study period.

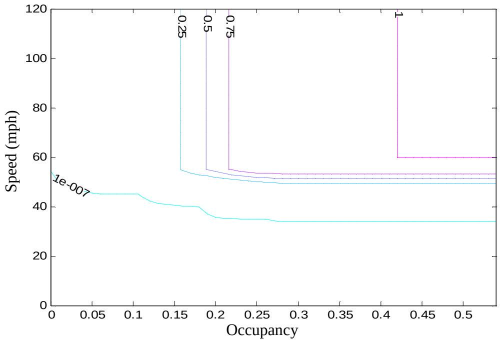
Table 3-1 summarizes the statistics based on Weibull distribution. The table gives the means and standard deviations of the variables ρ and *v*, which are the breakdown occupancy and speed respectively. The means and standard deviations are calculated by using formulae (2-4) and (2-6) respectively, in which α_q and β_q are replaced by α_v and β_v for speed, and α_ρ and β_ρ for occupancy. Table 3-1 also gives the bounds on speed and occupancy, given certain ranges of breakdown probability. In general, the results show that the probability of a breakdown occurs can be reduced to below 50% if the approaching traffic speed and occupancy are controlled to below 55 mph and 0.10 – 0.13 respectively. The highly negative Copula parameters at all stations suggest that the speed and occupancy leading to breakdowns are negatively correlated to each other. This implies that a traffic stream travelling at a high(low) speed will be associated with a low(high) occupancy, which is no surprise as it is the characteristic of traffic.



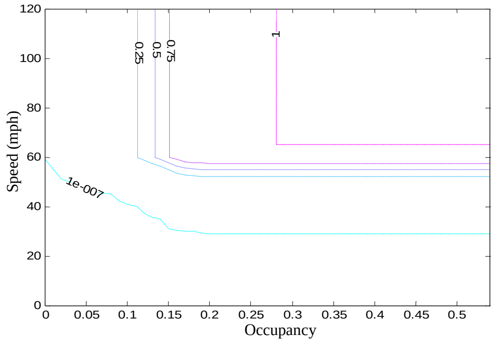
VDS: 401900



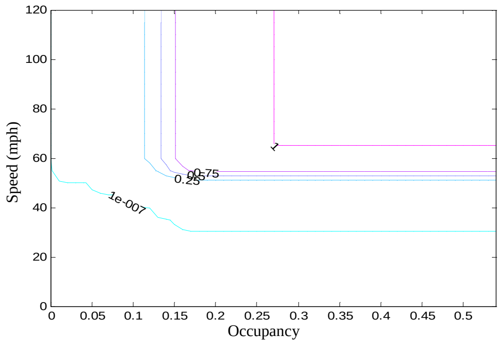
VDS 401524



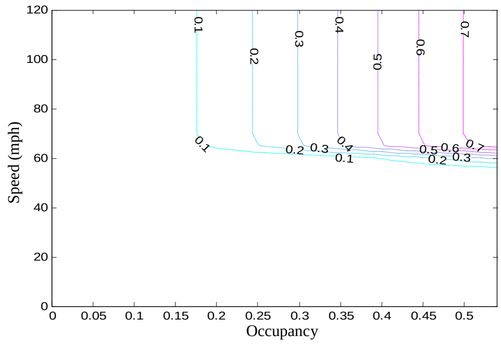
VDS: 400060



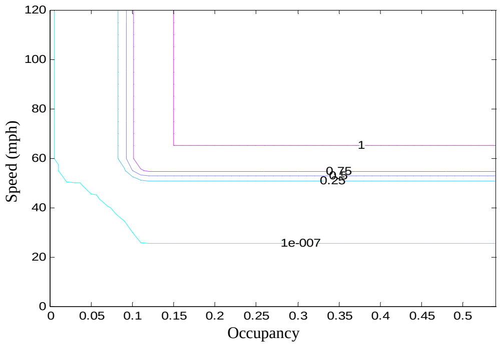
VDS 400176



VDS: 400691



VDS 401211



VDS: 401698

Figure 3-7: Contours of breakdown probability – PeMS data

Table 3-1: Summary of statistics – PeMS

	Weibull parameters									Breakdown prob. < 75%		Breakdown prob. < 50%	
VDS	δ	α_ρ	β_ρ	α_v	β_v	Mean occ	SD occ	Mean spd	SD spd	occ <	spd <	occ <	spd <
401900	**-129.75**	**10.44**	**0.14**	**45.86**	**55.98**	**0.14**	**0.016**	**55.30**	**1.52**	**0.13**	**55**	**0.13**	**55**
401524	**-310.37**	**10.33**	**0.17**	**34.11**	**58.29**	**0.16**	**0.019**	**57.36**	**2.11**	**0.15**	**55**	**0.15**	**55**
400060	-326.88	5.23	0.21	46.35	56.29	0.20	0.043	55.61	1.52	0.20	58	0.18	58
400176	-270.84	5.83	0.15	28.81	60.74	0.14	0.028	59.59	2.59	0.14	58	0.13	58
401242							--- N/A ----						
400126							--- N/A ----						
400691	-271.06	6.20	0.15	33.63	58.99	0.14	0.027	58.02	2.17	0.13	65	0.13	65
400803							--- N/A ----						
401211	-341.14	2.42	0.47	22.05	65.66	0.42	0.184	64.07	3.61	0.50	65	0.40	65
401698	**-237.40**	**9.20**	**0.11**	**26.08**	**58.99**	**0.10**	**0.013**	**57.77**	**2.77**	**0.10**	**55**	**0.10**	**55**

- Note: entries in bold indicate corresponding to the critical locations.

3.2.1 An experiment to test the persistence of the method

The contour plots in Figure 3-7 are derived from the 21-weekday data in September 2008 following the proposed statistical methodology described in Chapter 2. To test the persistence of the proposed methodology, we divide the original 21-day dataset into four subsets of weekly data in which weekends are excluded. Provided that the congestion is recurrent, the contour plots derived from these four sets of weekly data should be similar to each other and to those 'aggregated' ones in Figure 3-7, if the methodology in Chapter 2 is stable.

Stations VDS 400060 and VDS 401900 are chosen for the experiment and the results are shown in Figure 3-8 and Figure 3-9. Although there exist some variations from week to week, the patterns of the contours resemble each other. This indicates that the bivariate approach and the method for calibration in Chapter 2 is stable and persistent over time.

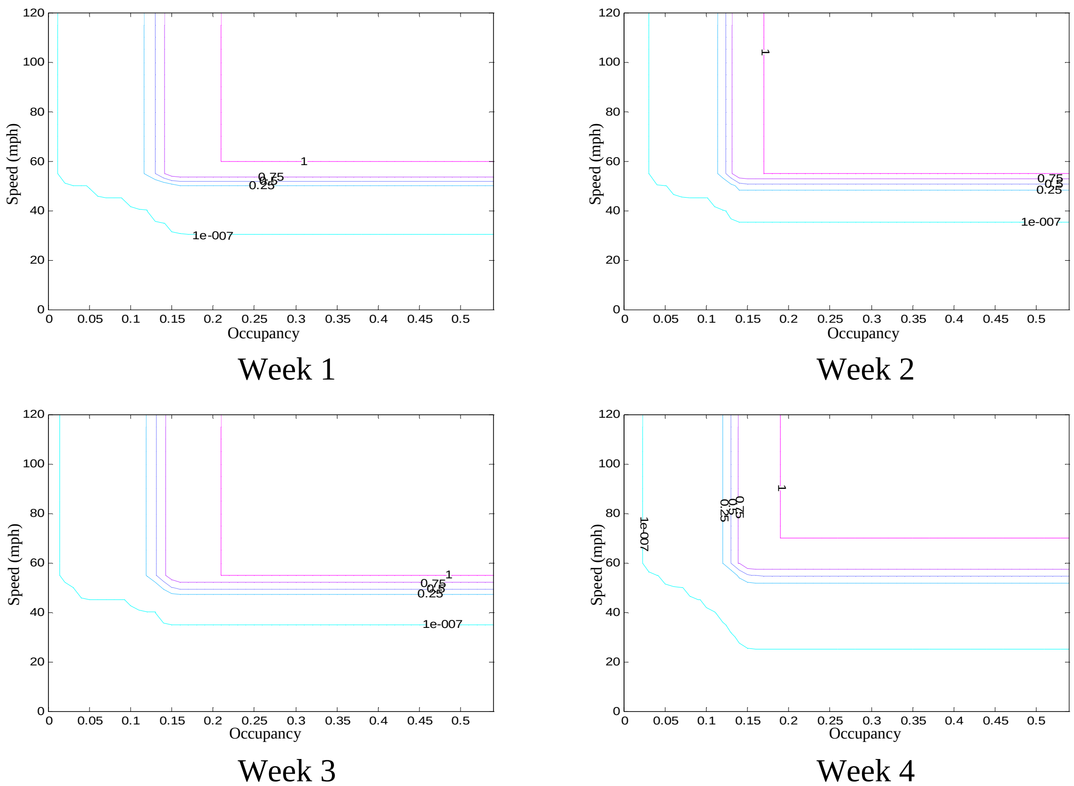


Figure 3-8: Comparison of contours over week at VDS 401900 for persistence

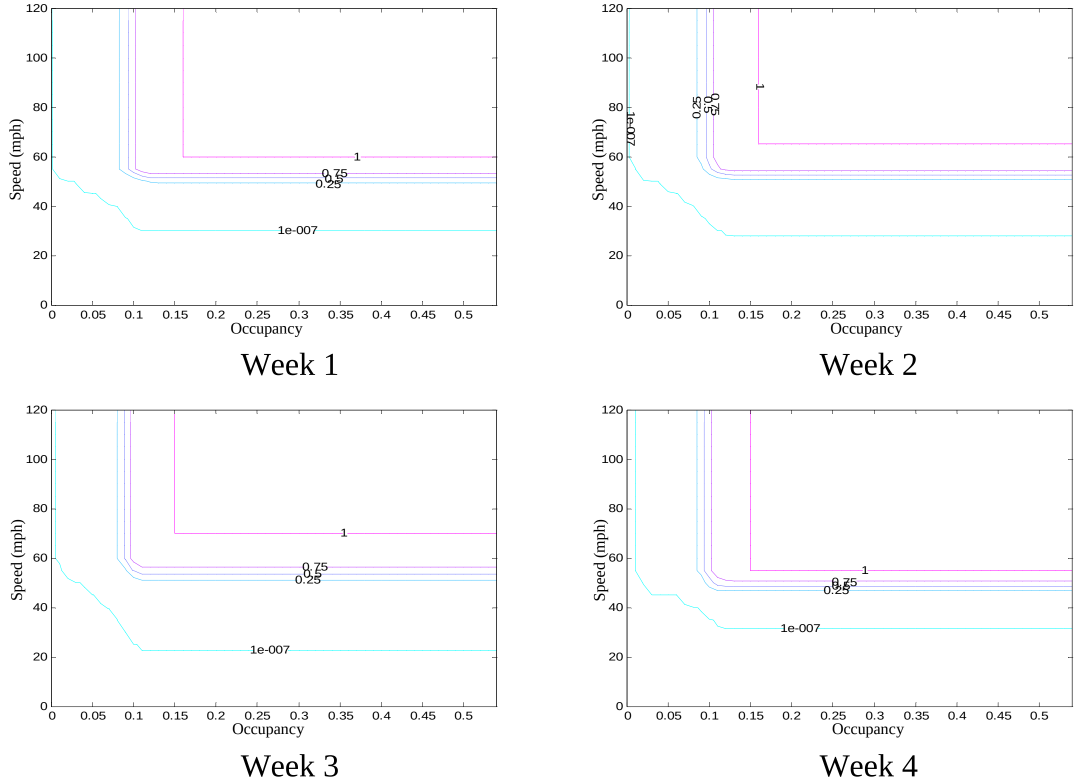


Figure 3-9: Comparison of contours over week at VDS 401698 for persistence

3.3 BHL

We adopt the BHL (Figure 3-10) data collected in the eastbound direction on 17 July 2008. It is noted that data on some of the lanes at loop stations 4 and 5 are missing and hence two stations are not included in this study. For comparison, we also download the corresponding PeMS data at the same loop stations on the same day.

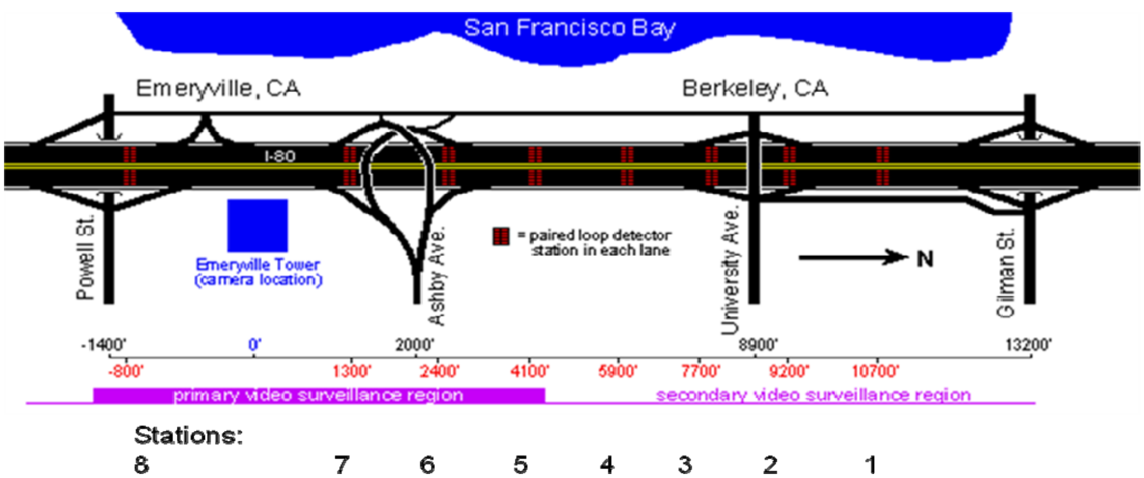
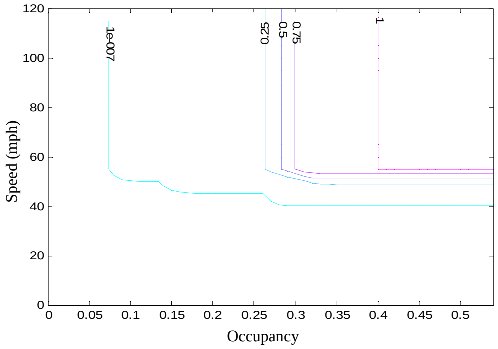
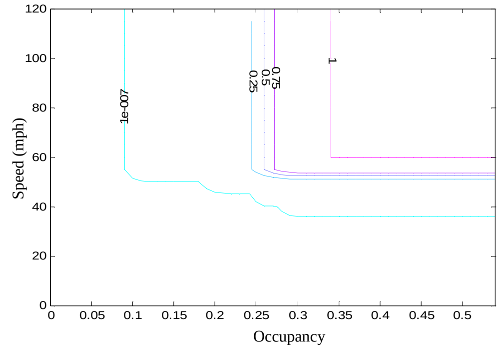


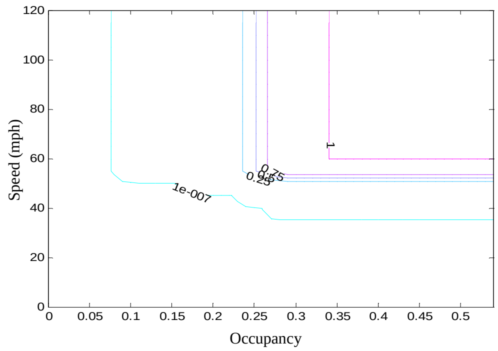
Figure 3-10: Berkeley Highway Laboratory (source: http://bhl.calccit.org:9006/bhl/)



Station 1



Station 2



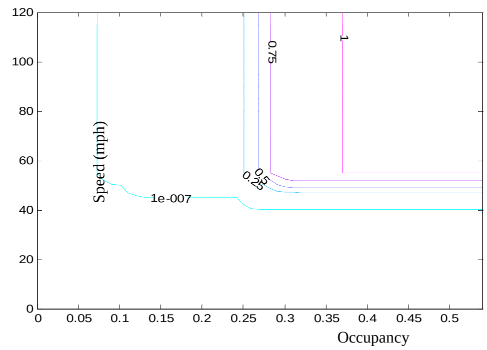
Station 3

Error in data

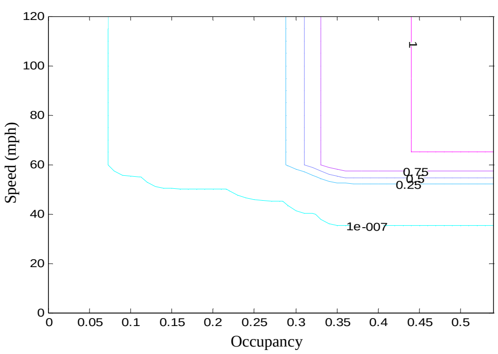
Station 4

Error in data

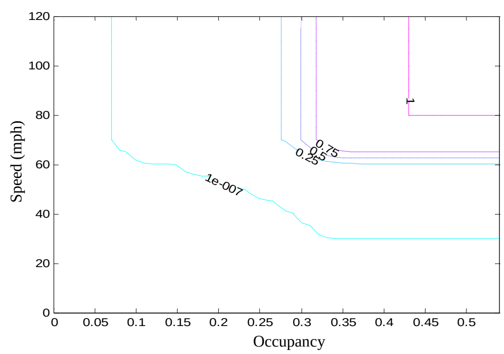
Station 5



Station 6



Station 7



Station 8

Figure 3-11: Contours of breakdown probability – BHL data

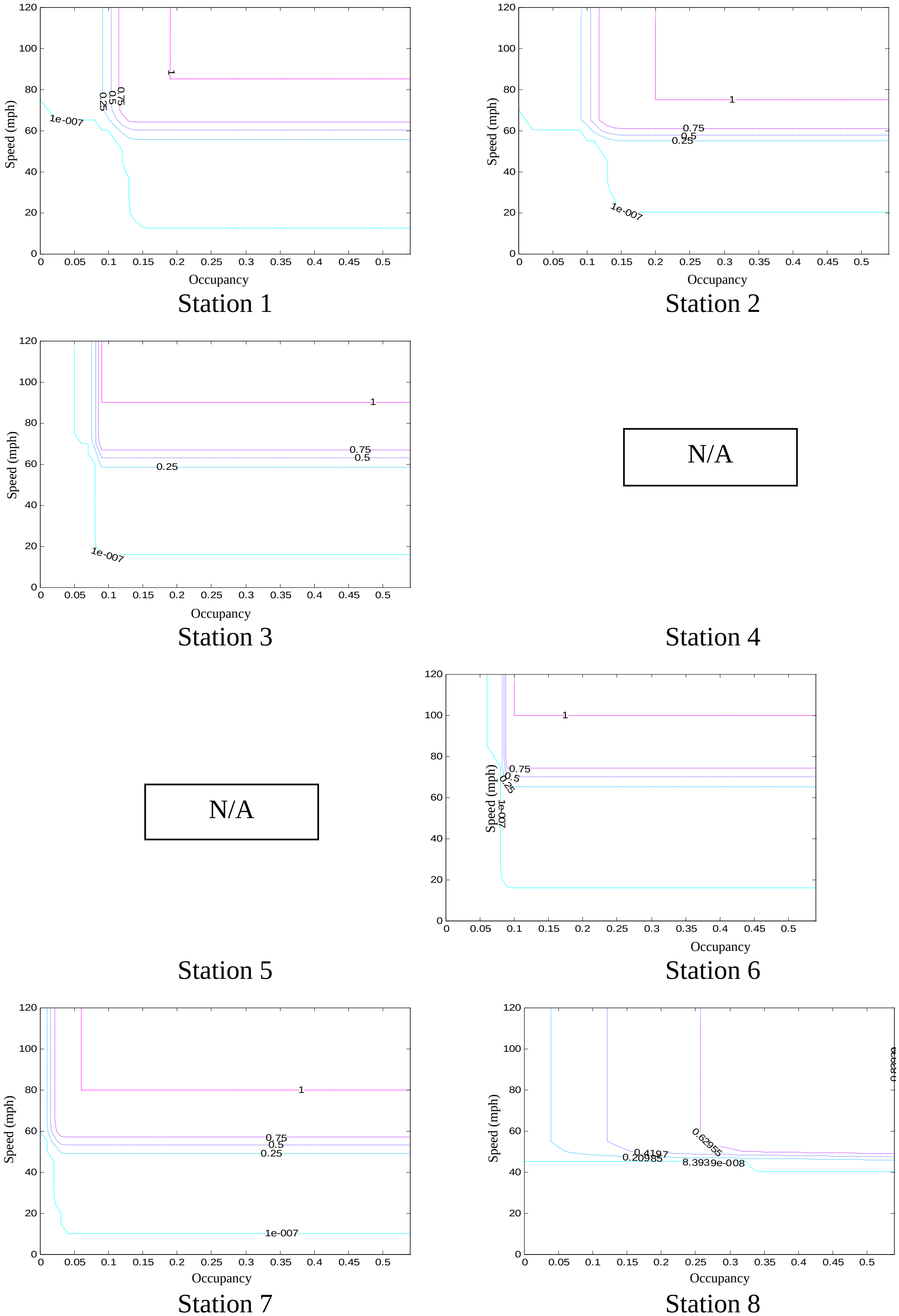


Figure 3-12 Contours of breakdown probability – PeMS data

It is noted that the BHL detector stations indeed coincide with the PeMS VDSs. The detectors are owned and managed by Caltrans. For reference, Table 3-2 lists the BHL stations alongside with the PeMS VDSs in both directions.

Table 3-2 BHL stations vs PeMS VDSs

BHL station	PeMS - I80W	PeMS - I80E
1	400060	400612
2	400176	400728
3	400009	400432
4	401242	401198
5	400126	400679
6	400691	400367
7	400803	400808
8	401211	401513

A summary of statistics with BHL data and PeMS data are shown in Table 3-3 and Table 3-4 respectively. Again, the highly negative values of the Copula parameter imply that the speeds and occupancies of traffic leading to breakdowns are negatively correlated.

BHL data are collected every 30-sec while PeMS data are updated every 5-min. With this difference, the contours obtained from BHL data indicate a significantly lower breakdown probability with respect to approaching occupancies and speeds. In fact, same observations were reported by Kerner (2004, p274) and Brilon et al. (2005), which showed that a finer data resolution results in lower estimates of breakdown probability. A possible explanation for this is that with finer resolution, the sample size (i.e. total number of time intervals) increases significantly while the number of time intervals inducing a traffic breakdown remains about the same. As a result, the proportion of breakdown intervals and hence breakdown probability drops. On the requirement of data resolution, Brilon et al. (2005) suggested that a 5-min data resolution should be enough for operational purpose. Moreover, consider that BHL data are expensive and not readily available, we suggest the use of PeMS data for practical purpose.

Finally, it is noticeable that the breakdown probability contours derived from I-80 are significantly different from those derived from I-5. It can be due to different characteristics of site and different characteristics of drivers. Effects of characteristics of site and driving behaviors will be studied in the future.

Table 3-3: Summary of Statistics – BHL data

	Weibull parameters									Breakdown prob. < 75%		Breakdown prob. < 50%	
Station	δ	α_ρ	β_ρ	α_v	β_v	Mean occ	SD occ	Mean spd	SD spd	occ <	spd <	occ <	spd <
1	-351.89	12.77	0.30	76.91	55.65	0.29	0.03	55.25	0.91	0.3	58	0.25	58
2	-351.91	16.03	0.28	59.09	57.60	0.27	0.02	57.05	1.22	0.27	55	0.25	55
3	-351.77	14.38	0.27	52.85	57.13	0.26	0.02	56.53	1.35	0.25	58	0.23	58
4							--- N/A ----						
5							--- N/A ----						
6	-351.85	13.14	0.29	81.25	55.01	0.28	0.03	54.63	0.85	0.27	50	0.25	50
7	-351.90	11.81	0.33	44.44	60.30	0.32	0.03	59.54	1.69	0.32	55	0.28	55
8	-351.85	11.65	0.32	23.72	69.02	0.31	0.03	67.46	3.54	0.32	64	0.27	64

Table 3-4: Summary of Statistics – PeMS data

	Weibull parameters									Breakdown prob. < 75%		Breakdown prob. < 50%	
Station	δ	α_ρ	β_ρ	α_v	β_v	Mean occ	SD occ	Mean spd	SD spd	occ <	spd <	occ <	spd <
1	-280.92	7.59	0.12	67.12	12.78	0.11	0.018	64.46	6.14	0.13	65	0.12	60
2	-325.49	6.94	0.12	64.42	17.95	0.11	0.019	62.54	4.30	0.13	62	0.12	58
3	-322.92	41.06	0.09	14.14	69.81	0.09	0.003	67.28	5.82	0.25	58	0.23	58
4							--- N/A ----						
5							--- N/A ----						
6	-343.20	54.88	0.10	12.98	77.10	0.10	0.002	74.09	6.96	0.08	75	0.08	70
7	-308.09	4.08	0.03	12.05	60.02	0.03	0.007	57.53	5.80	0.03	58	0.02	55
8	-298.09	0.85	0.27	23.10	52.64	0.29	0.349	51.42	2.77	0.27	53	0.15	50

Chapter 4. Concluding remarks

This study analyses the probability of breakdown using bivariate Weibull distribution. Contrasting with previous studies in the literature, the probability of breakdown is considered here to be a function of the combination of mean speed and occupancy of the approaching traffic, which is an extension. Such extension could benefit the development of on-line calibrated probability distribution for traffic management and control.

The bivariate Weibull distribution is derived from its univariate components by using a Copula function approach. Copula functions (Sklar, 1973) are regarded as a general way to construct multivariate distributions. The structure of a Copula function consists of two parts: one representing the marginal distributions; and one representing the correlation of the two marginal variables. The bivariate Weibull distribution is calibrated using a two-stage maximum likelihood estimation. In the first stage, the parameters of the marginal distributions are determined from the closed form expressions that we have derived. In the second stage, the correlation, which is known as Copula parameter, is determined by a line search such that a predefined log-likelihood function is maximized.

The proposed methodology is applied to three case studies with three datasets: PORTAL, PeMS, and BHL. PORTAL data is collected from a bottleneck near Terwilliger Avenue on I-5N in Portland, Oregon; PeMS and BHL data are collected from a 4-mile section of Freeway I-80W in West Berkeley, California. Contour plots of breakdown probability are derived at the locations where the data are measured. The contour plots are useful information for developing various control strategies including variable speed control and ramp metering. An optimal traffic control problem typically can be formulated as an optimization problem with an objective to maximize total benefit of the system (e.g. minimizing the total system delay) subject to a set of constraints including traffic dynamics and constraints on control variables. This study can be used to construct the second kind of constraint as bounds on occupancies and speeds such that the traffic breakdown probability is less than a specified threshold, say 10%. Persistence of the distribution function for the same location over time has been investigated as well empirically, which is a necessary condition for correctness of the approach. Empirical analysis shows that the bivariate approach and the method for calibration of the distribution function traffic breakdown proposed in Chapter 2 is stable and persistent over time for the selected location. Effect of data resolution on the breakdown probability is also investigated. It is found that lower breakdown probability is obtained with finer data. On the requirement of data resolution, Brilon et al. (2005) suggested that a 5-min data resolution should be enough for operational purpose. Moreover, consider that BHL data are expensive and not readily available, we suggest the use of PeMS data for practical purpose.

For real time deployment, the parameters of the Weibull distribution can be updated with real time traffic data by using the rolling horizon concept. Progress and results will be reported in the future.

Chapter 5. Further studies

Further study will investigate other factors influencing the breakdown in addition to the characteristics of traffic covered in this study. The additional information to consider includes:

- Geometric variation effect
  - including lane configurations, curvature, and grades. These data are obtainable from the PeMS.

- Truck volume effect
  - PeMS is able to estimate truck volumes at each detector based on the measured 5-minute, lane-by-lane values of flow and occupancy (see Kwon et al., 2003). If there are sensors that do report truck volumes then PeMS will use those directly.

    The algorithm attempts to break down the total flow into passenger cars and large trucks. The algorithm makes the following assumptions:

    - There are no heavy trucks on the inner lanes.
    - For multi-lane freeways, the vehicle speeds over different lanes are synchronized.
    - The traffic volume consists mostly of short passenger cars and long trucks.
    - The average length of passenger cars is 16 feet and the average length of trucks is 60 feet.

    With these assumptions, the algorithm estimates the proportion of trucks in each lane, starting with the first lane (where by assumption that the proportion of trucks is zero), and working to the outer lanes.

- Weather effect such as rainfall
  - Rainfall data consisting of hourly precipitation records can be obtained from the California Data Exchange Center (California Department of Water Resources, 2008).

- Incident effect
  - Incident data will be used to identify periods during which traffic may have been affected by incidents. The log records of incidents are generated by the California Highway Patrol (CHP) Computer-Aided Dispatch System and can be downloaded through PeMS. The log records incident types, post mile location, direction of traffic, start times and durations of the incidents. Types of incidents recorded include collisions, debris, and breakdowns of vehicles. However, the log does not include incidents related to work zones and adverse weather conditions.

References

Balakrishnan, N and Kateri, M (2008) On the maximum likelihood estimation of parameters of Weibull distribution based on complete and censored data. *Statistics and Probability Letters.* **78**, 2971 – 2975.

Banks, J (2006) New approach to bottleneck capacity analysis: Final Report. UCB-ITS-PRR-2006-13. California PATH, Institute of Transportation Studies, University of California, Berkeley.
(Accessible on http://database.path.berkeley.edu/reports/index.cgi).

BHL (2008) *The Berkeley Highway Laboratory*. University of California at Berkeley. Accessible on: http://bhl.calccit.org:9006/bhl/.

Brilon,W, Geistefeldt, J and Regler, M (2005) Reliability of freeway traffic flow: a stochastic concept of capacity. *Proceedings of 16th International Symposium of Transportation and Traffic Theory*, 125 -144.

Brilon, W and Zurlinden, H (2003) Ueberlastungswahscheinlichkeiten und Verkehrsleistung als Bemessungskriterium fuer Strassenverkehrsanlagen (Breakdown probability and traffic efficiency as design criteria for freeways). *Forschung Strassenbau and Strassenverkehrstechnik*, **870**. Bonn.

California Department of Water Resources (2008) *California Data Exchange Centre*. Accessible on: http://cdec.water.ca.gov/snow_rain.html.

Choe, T., Skabardonis, A., Varaiya, P. (2002) Freeway Performance Measurement System (PeMS): an operational analysis tool, *Proceedings of 81st TRB Annual Meeting*, Jan. 13-17, Washington, D.C.

Dowling, R, Skabardonis, A, Reinke, D (2008) Predicting impacts of intelligent transportation systems on freeway queue discharge flow variability. *Transportation Research Record*, **2047**, 49-56.

Elefteriadou, L, Roess, RP and McShane, WR (1995) The probabilistic nature of breakdown at freeway – Merge junctions. *Transportation Research Record*, **1484**, 80-89.

Evans, J, Elefteriadou, L and Natarajan, G (2001) Determination of the probability of breakdown on a freeway based on zonal merging probabilities. *Transportation Research Part B*, **35**, 237-254.

Frank, MJ (1979) On the simultaneous associativity of $F(x,y)$ and $x+y-F(x,y)$. *Aequationes Mathematics*, **19**, 194-226.

Geistefeldt, J (2008) Empirical relation between stochastic capacities and capacities obtained from the speed-flow diagram. *Proceedings of the Symposium on the Fundamental Diagram: 75 years.* Woods Hole, MA, July 8 -10.

Genest, C (1987) Frank's family of bivariate distributions. *Biometrika*, **74**(3), 549 – 555.

Genest, C and Rivest, LP (1993) Statistical inference procedures for bivariate Archimedean copulas. *Journal of the American Statistical Association*, **88**(423), 1034 – 1043.

Jia, Z, Varaiya, P, Chen, C, Petty, K and Skabardonis, A (2000) Congestion, excess demand, and effective capacity in California freeways. Online at pems.eecs.berkeley.edu.

Jia, Z, Chen, C, Coifman, B, and Varaiya, P (2001) The PeMS algorithms for accurate, real-time estimates of g-factors and speeds from single-loop detectors. *Proceedings of the 4th IEEE Intelligent Transportation System Conference*. Oakland, CA, August 25-29.

Kaplan, EL and Meier P (1958) Non-parametric estimation from incomplete observations. *Journal of the American Statistical Association*, **53**, 457 – 481.

Kerner, BS (2004) The Physics of traffic: Empirical freeway pattern features, engineering applications, and theory. Berlin, Germany: Springer-Verlag.

Kwon, J, Varaiya, P, Skabardonis A (2003) Estimation of truck traffic volume from single loop detectors using lane-to-lane speed correlation. *Proceedings of 82nd Transportation Research Board*, Washington DC.

Lorenz, M and Elefteriadou, L (2001) Defining freeway capacity as a function of breakdown probability. *Transportation Research Record*, **1776**, 43-51.

PeMS (2008) *Freeway Performance Measurement System 9.4*. University of California at Berkeley. Accessible on: http://pems.eecs.berkeley.edu.

PORTAL (2008) *Portland Oregon Regional Transportation Archive Listing*. Portland State University. Accessible on: http://portal.its.pdx.edu/.

Sklar (1973) Random variables, joint distribution functions and copulas. *Kybernetika*, **9**, 449 – 460.

Transportation Research Board (2000) *Highway Capacity Manual 2000.* National Research Council, Washington, DC.

Wikipedia (2008) *Weibull distribution*. Accessed on 22 November from: http://en.wikipedia.org/wiki/Weibull_distribution.

Zhang, L and Levinson, D (2004) Some properties of flows at freeway bottlenecks. *Transportation Research Record*, **1883**, 122-131.

Appendix A - Derivation of Closed Form Solutions for Parameters of Weibull Distribution

This appendix derives the closed form solutions (2-16) and (2-17) for the parameters of univariate Weibull distribution based on maximum likelihood estimation. Nevertheless, it is noted that the closed form solutions derived are applicable for marginal distributions in other variables, such as occupancy and speed.

The likelihood function is defined as

$$L_q = \prod_{i=1}^{n} \left\{ \left[f_c(q_i) \right]^{\theta_i} [1 - F_c(q_i)]^{(1-\theta_i)} \right\}, \tag{A-1}$$

The corresponding log-likelihood function is

$$\ln L_q = \sum_{i=1}^{n} \left\{ \theta_i \ln f_c(q_i) + (1-\theta_i) \ln[1 - F_c(q_i)] \right\}, \tag{A-2}$$

The parameters α_q and β_q in the distribution function are determined such that the above log-likelihood functions are maximized. It is achieved by setting the partial derivatives equal to zero:

$$\frac{\partial \ln L_q}{\partial \beta_q} = -\frac{\alpha_q}{\beta_q} \sum_{i=1}^{n} \theta_i + \frac{\alpha_q}{\beta_q^{\alpha_q+1}} \sum_{i=1}^{n} q_i^{\alpha_q} \theta_i + \sum_{i=1}^{n} \frac{\alpha_q}{\beta_q} \left(\frac{q_i}{\beta_q} \right)^{\alpha_q} (1-\theta_i) = 0 \tag{A-3}$$

$$\frac{\partial \ln L_q}{\partial \alpha_q} = \frac{1}{\alpha_q} \sum_{i=1}^{n} \theta_i - \ln \beta_q \sum_{i=1}^{n} \theta_i + \sum_{i=1}^{n} \ln q_i \theta_i + \frac{\ln \beta_q}{\beta_q^{\alpha_q}} \sum_{i=1}^{n} q_i^{\alpha_q} \theta_i - \frac{1}{\beta_q^{\alpha_q}} \left[\sum_{i=1}^{n} q_i^{\alpha_q} \ln q_i \right] \theta_i - \sum_{i=1}^{n} \left[\ln\left(\frac{q_i}{\beta_q} \right) \right] \left(\frac{q_i}{\beta_q} \right)^{\alpha_q} (1-\theta_i) = 0 \tag{A-4}$$

Following Balakrishnan and Kateri (2008), we let r, where $1 \le r \le n$, be the number of time intervals with $\theta_i = 1$, $(q_{1:n}, q_{2:n}, ..., q_{r:n})$ be the associated flow rates in those time intervals. The above expressions can then be rewritten as:

$$\frac{\partial \ln L_q}{\partial \beta_q} = -\frac{\alpha_q}{\beta_q} \sum_{i=1}^{n} \theta_i + \frac{\alpha_q}{\beta_q^{\alpha_q+1}} \sum_{i=1}^{n} q_i^{\alpha_q} \theta_i + \sum_{i=1}^{n} \frac{\alpha_q}{\beta_q} \left(\frac{q_i}{\beta_q} \right)^{\alpha_q} (1-\theta_i) = 0 \tag{A-5}$$

$$\Rightarrow \frac{\partial \ln L_q}{\partial \beta_q} = -\frac{r\alpha_q}{\beta_q} + \frac{\alpha_q}{\beta_q^{\alpha_q+1}} \sum_{i=1}^{n} q_{i:n}^{\alpha_q} + \frac{\alpha_q}{\beta_q^{\alpha_q+1}} \sum_{i=1}^{n} q_i^{\alpha_q} - \frac{\alpha_q}{\beta_q^{\alpha_q+1}} \sum_{i=1}^{n} q_{i:n}^{\alpha_q} = 0 \tag{A-6}$$

$$\Rightarrow \frac{\partial \ln L_q}{\partial \beta_q} = -\frac{r\alpha_q}{\beta_q} + \frac{\alpha_q}{\beta_q^{\alpha_q+1}} \sum_{i=1}^{n} (q_i)^{\alpha_q} = 0 \tag{A-7}$$

$$\Rightarrow \beta_q = \left\{ \frac{1}{r} \sum_{i=1}^{n} (q_i)^{\alpha_q} \right\}^{\frac{1}{\alpha_q}} \tag{A-8}$$

Furthermore,

$$\frac{\partial \ln L_q}{\partial \alpha_q} = \frac{1}{\alpha_q} \sum_{i=1}^{n} \theta_i - \ln \beta_q \sum_{i=1}^{n} \theta_i + \sum_{i=1}^{n} \ln q_i \theta_i + \frac{\ln \beta_q}{\beta_q^{\alpha_q}} \sum_{i=1}^{n} q_i^{\alpha_q} \theta_i - \frac{1}{\beta_q^{\alpha_q}} \left[\sum_{i=1}^{n} q_i^{\alpha_q} (\ln q_i) \theta_i \right] - \sum_{i=1}^{n} \left[\ln\left(\frac{q_i}{\beta_q} \right) \right] \left(\frac{q_i}{\beta_q} \right)^{\alpha_q} (1 - \theta_i) = 0$$

$$\Rightarrow \frac{\partial \ln L_q}{\partial \alpha_q} = \frac{r}{\alpha_q} - r \ln \beta_q + \sum_{i=1}^{r} \ln q_{i:n} + \frac{\ln \beta_q}{\beta_q^{\alpha_q}} \sum_{i=1}^{r} q_{i:n}^{\alpha_q} - \frac{1}{\beta_q^{\alpha_q}} \left[\sum_{i=1}^{r} q_{i:n}^{\alpha_q} \ln q_{i:n} \right] - \sum_{i=1}^{n} \left[\ln\left(\frac{q_i}{\beta_q} \right) \right] \left(\frac{q_i}{\beta_q} \right)^{\alpha_q} + \sum_{i=1}^{r} \left[\ln\left(\frac{q_{i:n}}{\beta_q} \right) \right] \left(\frac{q_{i:n}}{\beta_q} \right)^{\alpha_q} = 0 \tag{A-9}$$

$$\Rightarrow \frac{\partial \ln L_q}{\partial \alpha_q} = \frac{r}{\alpha_q} - r \ln \beta_q + \sum_{i=1}^{r} \ln q_{i:n} - \left(\frac{1}{\beta_q} \right)^{\alpha_q} \sum_{i=1}^{n} \left[\ln\left(\frac{q_i}{\beta_q} \right) \right] q_i^{\alpha_q} = 0 \tag{A-10}$$

$$\Rightarrow \frac{\partial \ln L_q}{\partial \alpha_q} = \frac{r}{\alpha_q} - r \ln \beta_q + \sum_{i=1}^{r} \ln q_{i:n} - \left(\frac{1}{\beta_q} \right)^{\alpha_q} \sum_{i=1}^{n} (\ln q_i) q_i^{\alpha_q} + \left(\frac{1}{\beta_q} \right)^{\alpha_q} (\ln \beta_q) \sum_{i=1}^{n} q_i^{\alpha_q} = 0 \tag{A-11}$$

Substituting $\beta_q = \left\{ \frac{1}{r} \sum_{i=1}^{n} (q_i)^{\alpha_q} \right\}^{\frac{1}{\alpha_q}}$ gives

$$\frac{\partial \ln L_q}{\partial \alpha_q} = \frac{r}{\alpha_q} - r \ln \beta_q + \sum_{i=1}^{r} \ln q_{i:n} - r \frac{\sum_{i=1}^{n} (\ln q_i) q_i^{\alpha_q}}{\sum_{i=1}^{n} (q_i)^{\alpha_q}} + r \frac{(\ln \beta_q) \sum_{i=1}^{n} q_i^{\alpha_q}}{\sum_{i=1}^{n} (q_i)^{\alpha_q}} = 0 \tag{A-12}$$

$$\Rightarrow \frac{\partial \ln L_q}{\partial \alpha_q} = \frac{r}{\alpha_q} + \sum_{i=1}^{r} \ln q_{i:n} - r \frac{\sum_{i=1}^{n} (\ln q_i) q_i^{\alpha_q}}{\sum_{i=1}^{n} (q_i)^{\alpha_q}} = 0 \tag{A-13}$$

$$\Rightarrow \frac{1}{\alpha_q} = \frac{\sum_{i=1}^{n} (\ln q_i) q_i^{\alpha_q}}{\sum_{i=1}^{n} (q_i)^{\alpha_q}} - \frac{1}{r} \sum_{i=1}^{r} \ln q_{i:n} \,. \tag{A-14}$$

Appendix B – MATLAB Codes

This appendix includes the MATLAB codes developed for this study. The set of codes depicted herein is for PeMS dataset (Section 3.2). Nevertheless, the only difference between codes for different datasets is on reading the input files whose formats are different for different sources.

There are in total six codes:

1. 'capacity_pdf.m'

   – it is the main program. It first reads the input file, which is downloaded from PeMS website in spreadsheet (*.xls) format. The program then categorizes the traffic data according to the classification in Section 2.1. The speed threshold, v^*, is predefined at 45 mph as discussed in Section 3.2.

   After doing a series of numerical experiments, it is found that we need to specify a 'maximum censored value' for both speed and occupancy in the set of 'non-breakdown flow', otherwise the breakdown probability will be significantly underestimated. These 'maximum censored values' are denoted by 'max_occ' and 'max_spd' in the code for occupancy and speed respectively. The associated values of them are set after investigating the empirical data. The necessity of setting those values is subject to further investigation. It may be due to the existence of some 'outliers' in the dataset that may affect the calibration result.

   After classifying the data, calibration will be carried out by calling the following subroutines.

2. 'weibull_cal.m'

   – it is a subroutine for calibrating the marginal distributions parameters (i.e. Step 1 in Section 2.3.1)

3. 'bisection_method.m'

   – it is a subroutine to calculate the value of L.H.S. of Equation (2-19) (Step 1)

4. 'H.m'

   – it is a subroutine of 'bisection method' for solving Equation (2-19) (Step 1)

5. 'golden_section.m'

   – it is a subroutine of 'golden section' for determining the Copula parameter (Step 2 in Section 2.3.2)

6. 'log_likelihood.m'

   – it is a subroutine to calculate the value of the log-likelihood function (2-24) (Step 2 in Section 2.3.2)

function capacity_pdf()

```
% This is the main code

% Read input file -  the input file is downloaded from the PeMS dataset in EXCEL:
% First column: time index (not used)
% Second column: flow (not used)
% Third column: occupancy
% Fourth column: speed
% We looked at data collected from the weekdays in Sept 2008, in which there are 21 days in total
% The data are aggregated in 5-min intervals, which gives 21x(24x60)/5 = 6048 rows for each column

datadir = pwd;
data = xlsread('vds401698_sept08');

% all weekdays in Sept 08
occ_raw = data(1:6048, 3);
speed_raw = data(1:6048, 4);

% =======================
% Classification of traffic data
% =======================
spd_threshold = 45;   % Speed threshold classifying breakdown flow and non-breakdown flow
max_occ = 0.15;                       % censored value, from observation  (subject to review)
max_spd = spd_threshold + 5;          % censored value, from observation  (subject to review)

%  uncensored data (‘breakdown’ flow; Case 1 in section 2.1 in the report)
z=1;
for c = 1:size(speed_raw,1)-1
   if  (  (  (speed_raw(c+1)  <  spd_threshold)  &&  (speed_raw(c)  >  spd_threshold)  )  &&  (speed_raw(c+1)  <
          speed_raw(c)-5 )  && (occ_raw(c) > 0) )
                speed_jam(z) = speed_raw(c);
                occ_jam(z) = occ_raw(c);
                z = z+1;
end
end

speed_jam = speed_jam'; occ_jam = occ_jam';

% censored data ('non-breakdown' flow; Case 2 in Section 2.1 in the report)
z=1;
for c = 1:size(speed_raw,1)-1
   if ( ( (speed_raw(c+1) > spd_threshold) && (speed_raw(c) > spd_threshold) ) && (occ_raw(c) > 0) )
       speed(z) = min(speed_raw(c), max_spd);
       occ(z) = min(occ_raw(c), max_occ);
       z = z+1;
   end
end

speed = speed'; occ = occ';

% data belong to Case 1 or Case 2 (i.e. Case 3 is excluded here)
speed_total = [speed; speed_jam];  occ_total = [occ; occ_jam];
```

```
% ==========================================
% 'weibull calibration' - Balakrishnan and Kateri (2008)
% ==========================================

% Step 1: Call function ‘weibull_cal’ to determine the marginal parameters
%--------------------------------------------------------------------------------------------

[alpha_speed, beta_speed] = weibull_cal(speed_jam, speed_total)
[alpha_occ, beta_occ] = weibull_cal(occ_jam, occ_total)

% Parameters of the marginal distributions

t1 = beta_speed;
b1 = alpha_speed;
t2 = beta_occ;
b2 = alpha_occ;

% Statistics of the marginal distributions
wei_mean_spd = beta_speed *  gamma(1 + (1/alpha_speed))
wei_sd_spd = beta_speed * ( gamma(1 + (2/alpha_speed)) - (gamma(1 + (1/alpha_speed)) )^2  )^.5
wei_mean_occ = beta_occ *  gamma(1 + (1/alpha_occ))
wei_sd_occ = beta_occ * ( gamma(1 + (2/alpha_occ)) - (gamma(1 + (1/alpha_occ)) )^2  )^.5


% Step 2: calculating ‘delta’ (the Copula parameter) by a golden section line search
%-----------------------------------------------------------------------------------------------------

delta = golden_section('log_likelihood',-10000,100,speed,occ,t1,b1,t2,b2,0.001,40)


% ====================
% Generating contour plots
% ====================

for e = 1:25          % index for speed
   for f = 1:55       % index for occupancy
      capacity_cdf_Weibull(e, f) = -1/delta*log(1+ (exp(-delta*(1- exp(-(e*5/t1)^(b1))))-1)*...
      (exp(-delta*(1- exp(-(f*0.01/t2)^(b2))))-1)/(exp(-delta)-1) );
   end
end

figure(1);
[C,h] = contour(0:.01:.54,0:5:120,capacity_cdf_Weibull,5);
clabel(C,h,'FontSize',25, 'LabelSpacing',1000, 'Rotation',0)
set(h,'ShowText','on','TextStep',get(h,'LevelStep')*20)
colormap cool


end
```

function [alpha, beta] = weibull_cal(u_data, c_data)

```
%  Function for calibrating  marginal Weibull distributions. (Step 1)
%   It returns the maximum likelihood estimates of the
%   parameters of the Weibull distribution given the provided data.
%

%   Reference:
%      [1]  Balakrishnan, N and Kateri, M
%      "On the MLE of parameters of WEibull dist. based on complete and
%      censored data", Statistics and Probability Letters
%      2008 p. 2971 - 2975.


num_complete = size(c_data,1);          % number of censored ('non-breakdown' flow) data
num_uncensored = size(u_data,1);        % number of uncensored ('breakdown' flow) data

alpha = bisection_method('H',u_data,c_data,0.00001,100.0,0.00001);

 t = c_data.^alpha;
 t = sum(t);

beta = ( t/num_uncensored )^(1/alpha);
```

```
function c = bisection_method(f,u_data,c_data,a,b,delta)

%     Subroutine of ‘bisection method’ used to solving ‘H = 0’
%     This subroutine is used in Step 1 – determining the marginal parameters
%     Input - f is the function input as a string 'f' (H-function in Balakrishnan and Kateri, 2008, see ‘H-function’).
%        - a and b are the left and right endpoints
%        - delta is the tolerance
%          - Output - c is the solution
%         - yc= f(c)
%          - err is the error estimate for c

ya=feval(f,a,u_data,c_data);
yb=feval(f,b,u_data,c_data);

max1 = 100;

for k=1:max1
    c=(a+b)/2;
    yc=feval(f,c,u_data,c_data);
    if yc==0
        a=c;
        b=c;
    elseif yb*yc>0
        b=c;
        yb=yc;
    else
        a=c;
        ya=yc;
    end
    if b-a < delta, break,end
end


end
```

function H_alpha = H(alpha, uncensored_data, complete_data)

```
%
% H-function in Balakrishnman and Kateri (2008) (i.e. Equation (2-19) in the report: Omega – 1/alpha)
%   Reference:
%      [1]  Balakrishnan, N and Kateri, M
%      "On the MLE of parameters of WEibull dist. based on complete and
%      censored data", Statistics and Probability Letters
%      2008 p. 2971 - 2975.
%
%   This subroutine is used in Step 1 – determining the marginal parameters

uncensored_data_log = log(uncensored_data);
complete_data_log = log(complete_data);

complete_data_power = complete_data.^alpha;
product = complete_data_log.*complete_data_power;

r =  size(uncensored_data,1);

H_alpha = sum(product) / sum(complete_data_power) - sum(uncensored_data_log)/r - 1/alpha;
```

```
function x = golden_section(f,a,b,v,pho,t1,b1,t2,b2,eps,N)
%
% Golden section search on the function f (log-likehood function)  to determine Copula parameter: delta (Step 2)
% t1,b1,t2,b2 are the marginal parameters
% Assumptions: f is continuous on [a,b]; and
% f has only one minimum (maximum for -f) in [a,b].
% N - maximum number of iterations.
% When b-a < eps, the iteration stops.
%

c = (-1+sqrt(5))/2;

x1 = c*a + (1-c)*b;
fx1 = -feval(f,x1,v,pho,t1,b1,t2,b2);

x2 = (1-c)*a + c*b;
fx2 = -feval(f,x2,v,pho,t1,b1,t2,b2);

for i = 1:N-2
   if fx1 < fx2
      b = x2;
      x2 = x1;
      fx2 = fx1;
      x1 = c*a + (1-c)*b;
      fx1 = -feval(f,x1,v,pho,t1,b1,t2,b2);
   else
      a = x1;
      x1 = x2;
      fx1 = fx2;
      x2 = (1-c)*a + c*b;
      fx2 = -feval(f,x2,v,pho,t1,b1,t2,b2);
   end;

   if (abs(b-a) < eps)
      fprintf('succeeded after %d steps\n', i);
      x = 1/2*(a+b);
      return;
   end;

end;
```

function L = log_likelihood(delta, v, pho, t1, b1, t2, b2)

```
%
% return the value of L given the delta, and the marginal parameters
% This subroutine is used in Step 2 – determining Copula function: delta
%
% Parameters of the marginal distributions:
% t1 = 'beta' of the first marginal distribution
% b1 = 'alpha' of the first marginal distribution
% t2 = 'beta' of the second marginal distribution
% b2 = 'alpha' of the second marginal distribution


for c = 1:size(v,1)-1

        L(c) = -delta*(exp(-delta)-1)*exp(-delta*( 1+(1- exp(-(v(c)/t1)^(b1)))+...
        (1- exp(-(pho(c)/t2)^(b2))) )) / ( exp(-delta)-exp(-delta*(1+1- exp(-(v(c)/t1)^(b1))))-exp(-delta*(1+1- ...
        exp(-(pho(c)/t2)^(b2))))+exp(-delta*(1- exp(-(v(c)/t1)^(b1))+1- exp(-(pho(c)/t2)^(b2))))  )^2;

end

L = log(L);
L = sum(L);
```