A CLUSTERWISE REGRESSION APPROACH FOR THE ESTIMATION OF CRASH FREQUENCIES

Naveen Veeramisti, Ph.D., E.I.
CA Group, Inc.
Transportation Engineer
Cell: (702) 324-4790
Email: naveen.veeramisti@c-agroup.com

Alexander Paz, Ph.D., P.E., Corresponding Author
Professor and Transport Main Roads Chair
Civil Engineering and Built Environment
Science and Engineering
Queensland University of Technology
2 George St
Brisbane, QLD 4000
Phone: +61-731-381-157
Email: alexander.paz@qut.edu.au

Mukesh Khadka, Ph.D.
Louis Berger, Inc.
Cell: (702) 683-2722
Email: imukeshkhadka@gmail.com

Cristian Arteaga, M.S.
Graduate Student
University of Nevada, Las Vegas
Howard R. Hughes College of Engineering
Civil and Environmental Engineering
4505 Maryland Parkway, PO Box 454007
Las Vegas, NV 89154-4007
Cell: (702) 427-4847
Email: arteagas@unlv.nevada.edu
A CLUSTERWISE REGRESSION APPROACH FOR THE ESTIMATION OF CRASH FREQUENCIES

ABSTRACT

In the current literature, data is aggregated for the estimation of functions to explain or predict crash patterns using either clustering analysis, regression analysis, or stage-wise models. Typically, analysis sites are grouped into site subtypes based on pre-defined characteristics. The assumption is that sites within each subtype experience similar crash patterns as a function of pre-specified explanatory characteristics. To develop functions to estimate crashes, all data points are clustered only as a function of associated site characteristics. As a consequence, estimated parameters may be based on different crash patterns that represents various trends which could be better captured by using multiple functions. To address this limitation, this study proposes a mathematical program utilizing clusterwise regression to assign sites to clusters, and simultaneously seek sets of parameter values for the corresponding estimation functions, so as to maximize the probability of observing the available data. A simulated annealing, coupled with maximum likelihood estimation, was used to solve the mathematical program. Results were analyzed for two site subtypes with fatal and all injury crashes: 1) roadway segments for urban multi-lane divided segments, and 2) urban 4-leg signalized intersections. Clusterwise regression improved the predicted number of crashes with multiple estimation functions within the same site subtype.

Keywords: Clusterwise Regression; Negative Multinomial; Log-likelihood; Traffic Safety; Network Screening; Crash Frequency; Accident Prediction Model
1. INTRODUCTION

Network screening for sites with the potential for safety improvements is a critical first step in a roadway safety management (RSM) process (Hauer, Harwood, Council, & Griffith., 2002; Montella, 2010; AASHTO, 2010). Network screening can be performed using either: 1) traditional methods, such as crash frequencies, rates, and proportions; 2) the empirical Bayes (EB) method; 3) the continuous risk profiling (CRP) method; or 4) crash estimation methods capable of handling unobserved heterogeneity, temporal and spatial instability, and self-selectivity issues, (Mannering, Shankar, & Bhat, 2016; Mannering, 2018) coupled with screening capabilities. Traditional methods have limitations, including bias associated with traffic volume, segment length, and regression-to-the-mean (Montella, 2010; AASHTO, 2010).

The EB method seeks to address some limitations of traditional methods by combining the results obtained, using an estimation function with observed crashes, in order to obtain a better estimate of the expected number of crashes (AASHTO, 2010). Estimation functions are crash prediction models that provide estimates of the number of crashes and the associated severity as a function of site characteristics. In some literature, these functions are called Safety Performance Functions (SPFs) (AASHTO, 2010). However, the statistical and econometrics community do not use this terminology because the name SPFs does not provide clear information about what exactly is estimated, nor the frequency or severity. In addition, it is important to state that traditional SPFs are not able to handle unobserved heterogeneity, temporal and spatial instability, or self-selectivity issues present in commonly available crash data (Mannering et al., 2016; Mannering, 2018). A comparative study by Montella (2010) recommends the EB method as the best for network screening, among other existing alternatives.
The CRP method uses a weighted moving average technique (Chung, Ragland, Madanat, & Oh., 2009) to continuously plot a collision risk profile. However, this method has several disadvantages relative to the EB method. CRP uses only observed crashes; thus, it does not seek to account for regression-to-the-mean or exposure, such as traffic volumes or other variables affecting expected crash frequencies. In a recent study by Grembek, Kim and Kwon (2012), crash estimation functions were used with CRP to estimate excess crash frequency. The predicted crash frequency, based on the annual average daily traffic (AADT) for the segment, was obtained from the corresponding estimation function. The estimation value, output from a SPF, is then transferred to the CRP profile on the same plot. The authors of the study mentioned, “the unit of the value from SPF is converted to the unit comparable to CRP to be plotted together” for estimating excess crash frequency. More details about how these estimation functions are combined with the collision risk profile could provide better insight for estimation of excess crash frequency.

In all of the mentioned methods, appropriate functions to estimate crashes are essential to determine reliable estimates and avoid bias. In addition, these functions play a key role in economic analysis, as well as priority-ranking, as part of the RSM process (AASHTO, 2010). A single estimation function cannot be used for an entire region or jurisdiction, or for all crash types and severities. Similarly, developing estimation functions for each possible combination of crash severity, crash type, facility type, and range of explanatory characteristics would require large amounts of data to obtain statistical significance/reliability. It is impractical to take into consideration all potential explanatory characteristics because some of them could be expensive to collect, unobservable, or difficult to quantify. The existing literature classifies sites into several predefined groups, as well as site subtypes, with measurable and available homogeneous
characteristics, such as area type, number of lanes, access control, and median type (e.g., rural two-lane, or urban 4-lanes divided.) (AASHTO, 2010). Hence, estimation functions for fatal and all injury crash severity, and site subtypes are developed. The assumption is that sites within each subtype experience similar crash patterns as a function of pre-specified explanatory characteristics. That is, the observed pattern of the dependent variable, observed crashes, is not considered explicitly to create site subtypes. For the development of the estimation functions, all data points are clustered based only on the associated site characteristics.

The consequence could be estimation functions with parameters estimated using very different crash patterns. That is, a single function is estimated to represent various distinct trends that are captured more accurately by using multiple functions. There are a number of scenarios; for example, crash trends for low-volume roads may be different than those for high-volume roads; similarly, crash trends for low-volume, high-speed-limit roads are different than those for high-volume, low-speed-limit roads. Hence, we propose to create clusters of sites within each predefined site subtype, based on the observed crash trends, so as to develop superior estimation functions, when compared to those developed using all data together as a single site subtype. This may lead to an optimal number of estimation functions, further classifying site subtypes into various sub-groups (clusters) to provide better crash estimates that minimize the overall estimation error. To some extent, estimation functions specified using approaches such as random parameters, seek to capture some of the above described potential heterogeneity in the data. A limitation of these approaches, relative to the proposed approach, is that although the parameters are random, a single set of explanatory variables is used for the entire model. Our proposed approach provides an opportunity to have multiple models with different sets of explanatory variables. Future research could involve using the proposed approach to estimate
superior functions that explicitly seek to handle unobserved heterogeneity, temporal and spatial
instability, and self-selectivity issues.

2. Literature Review

2.1. Background on Clusterwise Regression

With advances in data warehousing technology, multiple years of crash data along with
associated traffic and roadway characteristics are readily available. The yearly crash and AADT
data, along with other roadway characteristics, constitute panel count data. Karlaftis and Tarko
(1998) used a clustering technique on panel crash datasets to account for heterogeneity. Clusters
were developed to identify the homogeneous data, and then separate Negative Binomial models
were applied to each cluster. Separate models of each cluster provided better results than a joint
model. Data was segmented for each of the cluster-based analyses conducted by the authors.
However, this technique may not guarantee that each cluster consists of homogeneous sites in
terms of crash trends. Many previous studies used clustering analysis to segment crash data into
various clusters so that it could reveal hidden variables that influence crash severity (Depaire,
Wets, & Vanhoof, 2008; Mohamed, Saunier, Miranda-Moreno, & Ukkusuri, 2013; Sasidharan,
Wu, & Menendez, 2015). These studies used latent-class cluster analysis to identify clusters, and
then used various types of logit models for modeling crash severity outcomes.

A few studies have estimated the frequency of total crashes using count regression
models. The proportion of observed crashes was used to estimate the severity of crashes, crash
types, light conditions, or vehicles involved in crashes (Geedipally & Lord, 2010; Milton,
Shankar, & Mannering, 2008). Wang (2011) used two-stage regression models, first to estimate
crash frequency and then to estimate crash severity. The advantage of this approach is that traffic
and road characteristic data are first used to identify crash frequency based on a full Bayesian
approach. Then, more detailed data for individual crashes are used to find the proportion of crashes at various severity levels based on a mixed-logit model. It is clear that clustering was performed to develop estimation functions using either: 1) clustering analysis, 2) regression analysis, or 3) stage-wise models; then, regression models of crash frequencies were estimated for each cluster.

Most often, segmentation of crash data and classification of estimation functions are based on expert knowledge, modeling needs, or the desire to study a specific problem (AASHTO, 2010; Depaire, 2008; Srinivasan, 2013). The selection of a group of sites affects the estimation and reliability of an estimation function (Hauer, 2015). The use of clusterwise regression analysis to simultaneously perform clustering and the generation of the corresponding estimation functions, so as to minimize the estimation error, is lacking in existing traffic-safety literature. Clusterwise regression, introduced by Spath (1979) and extended by Lau, Leung, and Tse (1999), has been used in modeling pavement and environment performance, as well as in business applications (Luo & Chou, 2006; Lu, Huang, Li, & Yang, 2014; Poggi & Portier, 2011).

Spath (1979) introduced the mathematics behind clusterwise regression problems and proposed a solution approach using an Expectation-Maximization algorithm. Since then, clusterwise regression has been used in multiple applications. For example, Luo and Chou (2006) used clusterwise regression to model the deterioration of pavement conditions. A hypothetical example was used to explain clusterwise regression, where pavement condition rating data was plotted as a function of pavement age. Linear regression was used to fit one equation (single cluster), while clusterwise regression was used to generate two equations (two clusters). Two equations provided better fit in terms of $R^2$. Luo and Chou (2006) explained the fundamentals of clusterwise regression by minimizing the total sum of squared errors to fit data
using a heuristic algorithm. Poggi and Portier (2011) used clusterwise linear regression models to predict the daily mean PM10 concentration as a function of meteorological variables and the average measurement from the previous day. The Expectation-Maximization algorithm was used to minimize the total Bayesian Information Criteria. DeSarbo, Oliver, and Rangaswamy (1989) used clusterwise linear regression to develop performance functions of trade shows. A Maximum Likelihood methodology estimation approach was used. A solution to the clusterwise regression problem was obtained using the E-M algorithm. In contrast to our specific problem of estimating nonlinear crash estimation functions, all of these previous studies estimated linear regression models.

This study proposes clusterwise regression to assign sites to clusters, and simultaneously seek sets of parameter values for corresponding crash frequency estimation functions, so as to maximize the probability of observing the available data. Site membership to clusters and regression parameters are estimated simultaneously in order to improve the predictive capability of the estimation functions. A mathematical programming model is provided in Section 3 to describe the proposed approach in detail.

### 2.2. Background on Count Regression Models

Various count data models are available for development and estimation of crash frequency functions. The most common models are either Poisson or Negative Binomial (NB). If the appropriate data count model is not considered, the statistical validity of the analysis is compromised (Lord & Mannering, 2011; Mannering & Bhat, 2014). Recent literature has described other models, including Poisson-lognormal, Zero-inflated, Conway-Maxwell-Poisson, Negative Multinomial, and Random Parameters (Lord & Mannering, 2011; Mannering & Bhat, 2014).
Poisson regression models are estimated by specifying the expected number of crashes per time period as a function of explanatory variables. The Poisson model assumes an equal mean and the variance of the crash counts, which often is not correct. In such cases where variance exceeds the mean, referred to as overdispersion, NB models are expected to provide better parameter estimates. The NB model assumes that the expected number of crashes per period follows a gamma probability distribution, and an associated error term is gamma-distributed. In Poisson-lognormal models, the error term is assumed as lognormal. Poisson-lognormal models often are influenced by small sample sizes, and do not have a closed form solution (Miaou, Song, & Mallick, 2003). Zero-inflated Poisson or NB models are used when the data is characterized by a substantial number of zeros. Segments having zero numbers of observed crashes are assumed to have a long-term mean equal to zero, which is not correct; that is, there may be a crash, in the process generating crash data over the long term (Lord & Mannering, 2011).

The Conway-Maxwell-Poisson model was proposed to handle over- and under-dispersion (mean exceeds variance) characteristics of the crash data. A comparative study by Lord, Guikema, and Gredipally (2008) found that the results obtained by using NB and Conway-Maxwell-Poisson models were similar. Although Conway-Maxwell-Poisson models can handle under-dispersion of data, the results are influenced by low sample means and small sample bias (Lord & Mannering, 2011). In fixed-parameter models, parameter estimates are fixed across observations. If some parameters vary across observations, then a random parameter model can be used to estimate parameters that vary across observations, based on a pre-specified distribution. These models provide a better statistical fit than previous models; however, a more complex model estimation is required. A few studies have illustrated that this type of model may
not improve predictive capability (Shugan, 2006; Washington, Karlaftis, & Mannering, 2010; Lord & Mannering, 2011; Mannering & Bhat, 2014).

When data is associated with multiple time periods, the regression models mentioned above cannot be used. For example, crash and AADT data available for five years are aggregated over time. The estimation functions developed using aggregated data could underestimate the over-dispersion parameter (Kweon & Lim, 2012). With the use of a Negative Multinomial (NM) model, yearly crash data and AADT data can be used as multiple time period data (panel data). This prevents the loss of information by not smoothing the AADT over a certain time period (Hauer, 2015). Ulfarsson and Shankar (2003) compared NM models with NB models for predictive modeling and reported that NM models converge with significant higher log-likelihood, providing better fit in terms of log-likelihood ratio and outperforming NB models. In terms of over-dispersion, NB models provide better behavior than NM models. The reason could be that some of the over-dispersion may be captured by the NM model’s temporal serial correlation.

For further details and applications related to these models, for the development of crash frequency models, readers are referred to the studies by Lord and Mannering (2010), and by Mannering and Bhat (2014). Considering the multi-year data being used in this research, the Negative Multinomial model specification is proposed for the development of the estimation functions of crash frequency.

Ulfarsson and Shankar (2003), and Hauer (2004, 2015) explained in detail the theory of NM models, and suggested a log-likelihood distribution function, as shown in Equation 1 (Hauer, 2015). Consider a panel data with $i$ sites and $m$ time periods, with Poisson distributed observed crash counts $N$. Assuming the distribution as Gamma, an extension of the NB can be
applied to panel data, which is the NM distribution. The contribution of site $i$ to the log-likelihood is:

$$
\ln[L_i^*(\alpha, \beta_1, ..., b)] = bl_i \ln(b_l_i) + \sum_{t=1}^{m_i} Y_{i,t} \ln(\hat{E}\{\mu_{i,t}\}) + \ln\Gamma(\sum_{t=1}^{m_i} Y_{i,t} + bl_i) -
$$

$$
\ln\Gamma(b_l_i) - (\sum_{t=1}^{m_i} Y_{i,t} + b_l_i)\ln(\hat{E}\{\mu_{i,t}\}) + bl_i]
$$

(1)

where,

$i =$ number of sites

$m =$ number of time periods starting $t =$1

$Y =$ observed crash counts for site $i$ with $m$ time periods

$\mu =$ mean of Poisson distributed crash counts

$\hat{E}\{\mu_{i,t}\} =$ predicted number of crashes at site $i$ for time period $t$

$l =$ site length

$b =$ shape parameter

3. METHODOLOGY

3.1. Mathematical Program - Problem Formulation

Notation and Definitions:

The following notations and definitions are used in describing the proposed problem formulation:

$i =$ Subscript for a site

$I =$ Set of roadway sites (segments, intersections or ramps) to be clustered for a crash estimation function, indexed $1 \leq i \leq I$

$j =$ Subscript for an explanatory variable

$J =$ Set of explanatory variables of a crash estimation function, indexed $1 \leq j \leq J$

$X =$ An $I \times J$ matrix with elements $x_{ij}$, which are the measurements of explanatory variables for site $i \forall i \in I, j \in J$
Subscript for a cluster of sites

SPF<sub>k</sub> Crash frequency estimation function for cluster <i>k</i>

<i>K</i> Number of crash frequency estimation functions of sites, indexed 1 ≤ <i>k</i> ≤ <i>K</i>

<i>p</i><sub><i>lk</i></sub> Membership assignment of a site <i>i</i> to a cluster <i>k</i>

<i>P</i> An I x K binary matrix with elements <i>p</i><sub><i>lk</i></sub> ∀ <i>i</i> ∈ I, <i>k</i> ∈ <i>K</i>

<i>α</i><sub><i>k</i></sub> Intercept for SPF<sub>k</sub>, ∀ <i>k</i> ∈ <i>K</i>

<i>β</i><sub><i>jk</i></sub> Coefficient for explanatory variable <i>j</i> in SPF<sub>k</sub>, ∀ <i>j</i> ∈ J, <i>k</i> ∈ <i>K</i>

<i>Y</i><sub><i>ik</i></sub> Observed number of crashes for site <i>i</i> assigned to cluster <i>k</i> ∀ <i>i</i> ∈ I, <i>k</i> ∈ <i>K</i>

<i>𝔼</i><sub>{</sub><i>μ</i><sub><i>lk</i></sub><i>}</sub> Predicted number of crashes of a site <i>i</i> using a SPF in a cluster <i>k</i> ∀ <i>i</i> ∈ I, <i>k</i> ∈ <i>K</i>

<i>b</i><sub><i>k</i></sub> Shape parameter or inverse over-dispersion parameter of the underlying gamma pdf in a SPF<sub><i>k</i></sub>, ∀ <i>k</i> ∈ <i>K</i>

**Objective function**

Max. \( \ln (L^*) = \sum_k \sum_{i|p_{ik}=1} \ln [L^*_{ik} (\alpha_k, \beta_{jk}, \ldots, b_k)] * p_{ik} \)

\( (2) \)

Subject to

**Constraints**

**Log-Likelihood Function**

\( \sum_k \sum_{i|p_{ik}=1} f(Y_{ik}, \hat{E}\{\mu_{lk}\}, b_k) = \sum_k \sum_{i|p_{ik}=1} \ln [L^*_{ik} (\alpha_k, \beta_{jk}, \ldots, b_k)] \)

\( (3) \)

\( \ln (\hat{E}\{\mu_{lk}\}) = \alpha_k + \sum_{j=1}^{J} \beta_{jk} \ln (x_{ij}) \)

\( (4) \)

**Membership constraints**

\( p_{ik} = \begin{cases} 1, & \text{if and only if site } i \text{ is assigned to safety performance function } k; \\ 0, & \text{Otherwise} \end{cases} \)

\( (5) \)

\( \sum_k p_{ik} = 1 \forall \ i \in I, k \in K \)

\( (6) \)
For this mathematical program, the sites in the entire data are associated with $K$ crash frequency estimation functions by maximizing the log-likelihood of a NM distribution function. The decision variables are the number of SPFs, $K$; the parameters of NM models, $\alpha_k, \beta_{jk}, \ldots, b_k$; and the cluster membership, $p_{ik}$.

Maximization of log-likelihood is used as the objective function, as shown in Equation 2. The objective function finds the set of parameter values that maximize the probability to observe the available data. The constraint, Equation 3, provides the log-likelihood distribution function of the count regression model. The constraint, Equation 4, provides the count regression model. In this research, the NM log-likelihood distribution function, as shown in Equation 1, is used for the reasons explained in Section 2. In order to find the parameter values that maximize the log-likelihood function, the predicted number of crashes are estimated by fitting the distribution function of the model to the data, using Equation 4. This is performed for the identified number of crash frequency estimation functions, $K$.

The constraints, Equations 5 and 6 ensure that each site is assigned to exactly one cluster (or crash estimation function). Membership $p_{ik}$ takes value of ‘1’ if and only if a site $i$ belongs to crash estimation function $k$. Otherwise, it takes value ‘0’.

3.2. Solution Algorithm to the Mathematical Program

In order to solve the above mathematical program, a simulated annealing (SA), combined with the maximum likelihood estimation (MLE) algorithm, was implemented using the statistical and mathematical programming language R. SA was used for clustering the data to estimate membership of clusters, $p_{ik}$. For each accepted neighborhood clusters, an MLE was employed to estimate the parameters of the crash estimation functions, $\alpha_k, \beta_{jk}$, and $b_k$. The ‘mle2’ function available in R was used to estimate these parameters (Bolker, 2016). Román-Román, Romero,
Rubio, and Torres-Ruiz (2012) successfully implemented a SA algorithm for the MLE of the parameters of a Gompertz-type process, which assessed the behavior patterns in several fields of application. DeSarbo et al. (1989) applied such an algorithm to solve a clusterwise linear regression problem.

The algorithm developed to solve the clusterwise regression for this study is illustrated in Figure 1, and is described as follows:

Step 1. Initialization

Step 1.1 For a given number of clusters (K), randomly assign cluster memberships to sites.

Step 1.2 Set values of initial temperature ($T_0$), final minimum temperature ($T_{min}$), cooling rate ($\phi$), and the maximum number of neighbors to be generated ($N_{max}$) at each temperature level. Set iterator to $N = 0$.

Step 1.3 Count the number of observations of all sites assigned to each cluster. If all the clusters have at least the minimum number of sites, then set $\alpha_k$, $\beta_{jk}$, and $b_k = 1$, and go to step 2; otherwise, reassign the cluster memberships until all clusters have at least the minimum number of sites. Let $C_N$ be the valid initial clusters. Set $\alpha_k$, $\beta_{jk}$ and $b_k = 1$.

Step 2. Objective function evaluation and initial parameters estimation

Step 2.1 Estimate the predicted number of crashes with default $\alpha_k$ and $\beta_{jk}$.

Step 2.2 Estimate the log-likelihood function using the observed and predicted number of crashes, and parameter $b_k$. Predicted number of crashes for site ‘$i$’ in cluster ‘$k$’ is estimated using Equations 8 and 9 for roadway segments and intersections, respectively.

$$\hat{E}\{\mu_{ik}\} = \alpha_{Scale} * \text{SegmentLength}_{i}^{(\beta_{1k})} * \text{AADT}_{i}^{(\beta_{2k})} * f_{ik}(\text{RoadwayType}) * f_{ik}(\text{FunctionalClass}) * f_{ik}(\text{NumberOfLanes}) * f_{ik}(\text{MedianType}) *$$
\[ f_{ik}(\text{MedianWidth}) \times f_{ik}(\text{PostedSpeed}) \]  \hspace{1cm} (8)

where

- \( f_{ik}(\text{RoadwayType}) = \beta_{(\text{non-StateRoad})} \)
- \( f_{ik}(\text{FunctionalClass}) = \beta_{(\text{FunctionalClass 4})} \) or \( \beta_{(\text{FunctionalClass 6})} \) or \( \beta_{(\text{FunctionalClass 7})} \)
- \( f_{ik}(\text{NumberOfLanes}) = \beta_{(\text{NumberOfLanes 5})} \) or \( \beta_{(\text{NumberOfLanes 6})} \) or \( \beta_{(\text{NumberOfLanes 7})} \) or \( \beta_{(\text{NumberOfLanes 8})} \)
- \( f_{ik}(\text{MedianType}) = \beta_{(\text{Flushpaved})} \) or \( \beta_{(\text{OtherDivided})} \)
- \( f_{ik}(\text{PostedSpeed}) = \beta_{(\text{PostedSpeed 30 & 35})} \) or \( \beta_{(\text{PostedSpeed 40 & 45})} \) or \( \beta_{(\text{PostedSpeed 50 & 55})} \) or \( \beta_{(\text{PostedSpeed > 55})} \)

\[ \hat{E}\{\mu_{ik}\} = \alpha_{\text{Scalek}} \times \text{MajorRoadAADT}^{\beta_{1}}_{ik} \times \text{MinorRoadAADT}^{\beta_{2}}_{ik} \times f_{ik}(\text{MajorRoadLanes}) \times f_{ik}(\text{MinorRoadLanes}) \times f_{ik}(\text{MajorRoadMedianType}) \times f_{ik}(\text{MinorRoadMedianType}) \times f_{ik}(\text{MajorRoadPostedSpeed}) \times f_{ik}(\text{MinorRoadPostedSpeed}) \]  \hspace{1cm} (9)

where

- \( f_{ik}(\text{MajorRoadLanes}) = \beta_{(\text{MajorRoadLanes=3,4})} \) or \( \beta_{(\text{MajorRoadLanes>4})} \)
- \( f_{ik}(\text{MinorRoadLanes}) = \beta_{(\text{MinorRoadLanes=3,4})} \) or \( \beta_{(\text{MinorRoadLanes>4})} \)
- \( f_{ik}(\text{MajorRoadMedian}) = \beta_{(\text{MajorRoadMedian-Divided})} \)
- \( f_{ik}(\text{MinorRoadMedian}) = \beta_{(\text{MinorRoadMedian-Divided})} \)
- \( f_{ik}(\text{MajorRdPostedSpeed}) = \beta_{(\text{MajorRd PostedSpeed 30 & 35})} \) or \( \beta_{(\text{MajorRd PostedSpeed 40 & 45})} \) or \( \beta_{(\text{MajorRd PostedSpeed > 55})} \)
- \( f_{ik}(\text{MinorRdPostedSpeed}) = \beta_{(\text{MinorRd PostedSpeed 30 & 35})} \) or \( \beta_{(\text{MinorRd PostedSpeed 40 & 45})} \) or \( \beta_{(\text{MinorRd PostedSpeed > 55})} \)

Step 2.3 Evaluate the objective function; maximize \( \ln[L^*_i(\alpha_k, \beta_{jk}, \ldots, b_k)] \) using MLEm and set this value as MLE_N.

Step 2.4 For \( C_N \), obtain \( \alpha_k, \beta_{jk}, \) and \( b_k \) for all \( K \) clusters from MLE.

Step 3. Set of neighborhood clusters generation

Create a set of neighborhood clusters randomly near to the previous cluster, using the following steps:
Step 3.1 Randomly select a pre-specified number of sites to change their memberships.

Step 3.2 For each of the sites selected, assign a new membership by generating a random number $r \sim R(1, K)$. If the new membership is the same as the previous one, regenerate a random number $r' \sim R(1, K)$ until it is different. Repeat this process until the memberships of all selected sites are different than previously assigned.

Step 3.3 Count the total number of sites assigned to each cluster and calculate the minimum sample size using Equation 10 (Berenson, 2005).

$$n = \frac{Z^2W}{(Z^2P(1-P)/\varepsilon^2) + (W-1)}$$

(10)

where,

$n$ = the number of samples,

$Z$ = the Z value for confidence interval,

$P$ = the true parameter, the maximum variance of distribution, 0.5

$\varepsilon$ = the significance level,

$W$ = the population size

Step 3.4 If all clusters have at least the minimum number of sites, go to Step 5; otherwise, repeat steps 3.1., 3.2., and 3.3., until all clusters have at least the minimum number of sites. Let $C_{N+1}$ be the new set of valid neighborhood clusters.

Step 4. Solution search

Step 4.1 Estimate the predicted number of crashes with default $\alpha_k$ and $\beta_{jk}$.

Step 4.2 Estimate the log-likelihood function using the observed number of crashes, predicted number of crashes, and parameter $b_k$.

Step 4.3 Evaluate the objective function; maximize $\ln[L_i^*(\alpha_k, \beta_{jk}, ..., b_k)]$ using MLE, and set this value as $MLE_N$. 
Step 4.4 For $C_{N+1}$, obtain new $\alpha_k$ and $\beta_{jk}$ for all $K$ clusters from MLE.

Step 4.5 Calculate $\Delta MLE = MLE_{N+1} - MLE_N$.

Step 4.6 Check the following two conditions:

a. If $\Delta MLE > 0$, accept the current solution, $C_{N+1}$ and the corresponding $\alpha_k$, $\beta_{jk}$, and $b_k$; go to Step 5; otherwise, go to Step 4.6b.

b. Generate a random number $r'' \sim R(0,1)$. Calculate the acceptance probability, $p_{accept} = exp\left(\frac{-\Delta MLE}{B^*T}\right)$, where $B$ is a Boltzmann’s constant. If $r'' > p_{accept}$, accept the current solution, $C_{N+1}$, and corresponding $\alpha_k$, $\beta_{jk}$, and $b_k$; go to Step 6; otherwise, return to Step 3 to generate a set of new neighborhood clusters.

Step 5. Stopping Criteria

Step 5.1 Repeat Steps 3 and 4 for $N_{max}$ times.

Step 5.2 If $T < T_{min}$, stop the algorithm. Otherwise, multiply the current temperature by the prespecified cooling rate, $\phi$, set $N=1$, and go to Step 2.

The SA algorithm seeks an optimum solution using a probabilistic approach for a given function. The logic behind this algorithm is as follows. Annealing corresponds to progressing a material to its equilibrium state, a process that causes the diffusion of atoms by heating followed by cooling; SA works using a similar technique. Initially, the probability of accepting a bad solution is high. In annealing, this is equivalent to having a high temperature, $T$. This enables the algorithm to escape from local maxima, moving downhill as it explores the solution search space vertically, as well as horizontally, with big step lengths. As temperature cools down, the probability of accepting bad solutions drops; at this stage, the algorithm uses small step lengths to search heuristically for an optimum solution on the most promising search space. Román-Román et al.
(2012) illustrated that the algorithm converges to a global minimum with a substantially slow cooling rate.

4. EXPERIMENT AND RESULTS

4.1. Data Resources and Preparation

Development of crash estimation functions requires key data, including information about crashes, traffic flow, traffic control, and roadway characteristics. The data used in this study were extracted from the Nevada Citation and Accident Tracking System (NCATS) database, the Highway Performance Monitoring System (HPMS), and Traffic Records and Information Access (TRINA) of the Nevada Department of Transportation (NDOT) (NDOT, 2016). In addition, the Travel Demand Model (TDM) and Intersections database from the Regional Transportation Commission of Southern Nevada (RTC-SN) were accessed. The data consisted of roadway, traffic, intersection, and crash characteristics collected in Clark County, the largest region of the State of Nevada.

A comprehensive database was developed by integrating all the data sources listed above. Various issues were encountered during the development of the database, including the availability of data, requirement of data from multiple agencies, and consistency of the collected data. It was a substantial task to identify and integrate data, and develop the database.

Various tools were used to integrate the data. Some data were integrated using a location reference system, such as county, route, and milepost. For example, a site was represented as a roadway segment in Clark County on Route 123 from Milepost 1.300 to 2.500. Crashes were mapped onto this site with the same county and route information for mileposts between 1.300 and 2.500. Other sources, where such information was not available, were integrated using spatial integration with the help of ArcGIS (ESRI®). For example, crashes mapped onto an...
intersection specifically required such GIS operations as buffering the signal-control intersection with a 200-ft radius, and mapping crashes within that buffer as intersection-related crashes. The details related to the development of this comprehensive database are provided by Paz, Veeramisti, Khanal, Baker, and de la Fuente-Mella (2015c).

The database that was developed contains five years of data from 2007 to 2011; it contains 10,287 roadway segments, 963 signal and stop-control intersections, and 763 ramp segments. Few of the consecutive roadway segments indicated homogeneous characteristics. As a result, the data was post-processed to combine the roadway segments containing homogeneous characteristics. The characteristics considered were functional class, number of lanes, median type, median width (with 5-ft thresholds), speed limits (with 5-mph thresholds), and AADT (with 20% thresholds). The data were classified based on site subtypes that were regularly used in the literature (Hauer 2002; AASHTO, 2010; Hauer, 2015; Paz et al., 2015c).

In this study, fatal and all injury crashes were used to develop SPFs. Site subtypes, with a minimum of 70 sites, were chosen. Explanatory variables considered in the analysis were segment length, AADT, number of lanes, access control, functional class, median width, median type, posted speed, and terrain type. Site subtypes identified in the data, based on the literature, are shown in Table 1. The post-processing of data and the identification of site subtypes were generated systematically by R code. In this paper, the results of two site subtypes were analyzed: urban multilane divided segments (SS1), and urban 4-leg signalized intersections (SS2). SPFs of fatal and all injury crashes within these site subtypes were estimated. The descriptive statistics of crash data and other explanatory variables are provided for site subtype SS1 in Table 2, and SS2 in Table 3.
4.2. Parameters of the search algorithm

Performance of a SA algorithm is affected by the optimization parameters used to solve a specific problem. Selecting an appropriate strategy to generate a neighborhood solution and appropriate annealing parameters to attain an optimal solution is critical (AASHTO, 2011; Roshan et al., 2013). The literature provides various methods to determine annealing parameters, including sensitivity analysis (Park, & Kim, 1998; Kirkpatrick, Gelatt, & Vecchi, 1983; Collins, Eglese, & Golden, 1988; Rose, Klebsch, & Wolf, 1990; Selim & Alsultan 1991; Guo & Zheng, 2005). Previous experience by this research team solving relevant problems was used to choose the initial set of optimization parameters (Paz, Molano, Martinez, Gaviria, & Arteaga, 2015a; Paz, Molano, & Sanchez, 2015b). Sensitivity analysis was performed to adjust the appropriate annealing parameters, taking into consideration a reasonable amount of computation time to reach the optimum solution (Park & Kim, 1998). SA parameter values used in this study are shown in Table 4.

4.3. Results and Discussion

In this paper, results for two site subtypes were analyzed: urban multilane divided segments (SS1), and urban 4-leg signalized intersections (SS2). Crash frequency functions for fatal and all injury crashes within these site subtypes were estimated. These site subtypes were classified into various subgroups (clusters), using clusterwise regression to determine the parameters of the crash frequency functions (fatal and all injury) that could provide better estimates. Preliminary investigation of crash data – along with the explanatory variables of AADT, speed limit, and median type – determined the two clusters that could provide the optimal number of clusters hypothesized for an urban multilane divided segments, SS1. For urban 4-leg signalized intersections, SS2, based on AADT, median type, and speed limit, four clusters could be the
hypothesized as the optimal number of clusters. The number of clusters hypothesized were confirmed based on the sensitivity analysis.

The algorithm partitioned the data and provided memberships for sites in these clusters. Figures 2a and 2b indicate the trajectory of the objective function (MLE) when the clusterwise regression models were used for SS1 and SS2, respectively. Figures 2c and 2d show results from sensitivity analyses for determining the optimum number of clusters for SS1 and SS2, respectively. For SS1, the initial value of maximum likelihood was 12585. After 1,146 iterations, the final value increased to 12736. In the case of SS2, the initial value of maximum likelihood was 12907; the final value increased to 13004.

In addition, to investigate the performance of the proposed method, the accuracy of the clusterwise models was compared to models developed using the traditional cluster method (hereafter called “single cluster”). In the single cluster method, all the data were used to estimate a single crash frequency function for SS1, and single crash frequency function for SS2.

In both SS1 and SS2, as shown in Table 5 and Table 6, there are substantial differences in the coefficients of the parameters of sites assigned to clusters. In the case of SS1, the coefficients of AADT across clusters are significantly different in magnitude. In the case of SS2, Major Rd AADT and Minor Rd AADT show differences in magnitude across clusters. In addition, other parameters contribute in clustering the sites with minor differences in magnitude across clusters. The significance level for the parameters was set as 5%. In Table 5 and Table 6, it can be observed that not all of the parameters were significant, i.e., with p-values less than 0.05. However, for SS1, segment length, AADT, speed limit, and functional class (other than functional class 7) were significant variables in all clusters. A divided median type was not significant in any clusters. This could be due to the collinearity of median width with median
type. Functional Class 7 (local road) is not significant, which could be due to the very few local road segments in the data. All the parameters in a single cluster model are significant. In the case of SS2, the parameter associated with major road speed in cluster 4 had p-values larger than 0.05.

The accuracy of the clusterwise models were compared to models developed using the single cluster method. All the sites (as single cluster) in the SS1 were used in the regression analysis to estimate the parameters of the crash estimation function. Results obtained from the single cluster are presented in Table 5 and Table 6 for SS1 and SS2, respectively.

The optimal number of clusters were verified using the Bayesian Information Criteria (BIC), which penalizes the inclusion of additional parameters. Results with BIC values close to negative infinity are categorized as optimal in terms of the number of clusters (Schwarz, 1978; R-Language, 2011). Table 7 shows the number of clusters and the corresponding BIC values. The lowest BIC values are associated with the same optimum number of clusters as identified using sensitivity analysis in Figure 2(c) and 2(d) for SS1 and SS2, respectively.

The number of sites (samples) in each cluster should be sufficient to obtain a statistically reliable estimate of parameters. A larger number of samples in a cluster leads to increased precision during the estimation of parameters using regression. Sample size is estimated using Equation 10 (Berenson, 2005). The equation is most suitable for data assumed to be normally distributed. There is no ideal equation for non-normally distributed data and the results using Equation 10 tend to be conservative. For SS1 and SS2, the minimum number of samples required with a 95% confidence interval and 0.05 significance level, are 91 and 75, respectively. The number of sites assigned to each cluster for both SS1 and SS2 are higher than the minimum number of samples required.
Model fit for the crash frequency functions was assessed with a Goodness-of-Fit (GOF) statistic, Freeman-Tukey $R^2$ coefficient of determination ($R^2_{FT}$) using Equation 11. Freeman and Tukey (1950) developed the variance stabilizing transformation for the Poisson distribution as shown in Equations 12, 13, 14, and 15. Fridstrøm, Ifver, Ingebritsen, Kulmala, and Thomsen (1995) applied this transformation when developing generalized regression models for crash data by using Equation 11. For the assessment of GOF for models associated with crash data, this statistic has been used by previous studies.

$$R^2_{FT} = 1 - \frac{\sum_i e_i^2}{\sum_i (Y_i - \hat{Y}_i)^2}$$  \hspace{1cm} (11)$$

$$f_i = \sqrt{N_i} + \sqrt{Y_i + 1}$$  \hspace{1cm} (12)$$

$$\bar{f} = \frac{\sqrt{Y_i+1} + \sqrt{Y_i+1}}{\eta}$$  \hspace{1cm} (13)$$

$$\hat{f}_i = \sqrt{4N_i + 1}$$  \hspace{1cm} (14)$$

$$\hat{e}_i = f_i - \hat{f} = \sqrt{Y_i} + \sqrt{N_i + 1} - \sqrt{4N_i + 1}$$  \hspace{1cm} (15)$$

where,

$N_i$ = the observed crashes of site $i$ in a cluster,

$\hat{N}_i$ = the predicted crashes of site $i$ in a cluster, and

$\eta$ = the number of sites in a cluster.

Membership of the sites, determined by the algorithm for clusterwise models as well as the associated parameters of the crash frequency function, were used to estimate the predicted number of crashes for a clusterwise model. Parameters of the crash frequency function from a single cluster model were used to estimate the predicted number of crashes for a single cluster model. Freeman-Tukey $R^2$ ($R^2_{FT}$) results are provided in Table 5 and Table 6. For both SS1 and SS2, $R^2_{FT}$ of clusterwise regression functions are higher than that of a single cluster function.
4.3.1. Discussion on Model Overfitting

In addition to the GOF statistic, potential overfitting needs to be investigated. In clusterwise regression modeling, overfitting is a potential issue, which was first highlighted and analyzed by Brusco, Cradit, Steinley, and Fox (2008). The study analyzed the effect of the use of explanatory variables to explain variation in the response variable.

This research adopted the approach developed by Brusco et al. (2008) to investigate the presence of overfitting. The total sum of squares (TSS), which is the variation of the response variable about its mean, should be equal to a between-cluster sum of squares (BCSS) and a within-cluster sum of squares (WCSS). The WCSS is equal to the sum of the regression sum of squares (SSR), and the sum of squared error of prediction (SSE). The SSR is within-cluster variation explained by regression models, and the SSE is the residual error in the clusters. As this study used count data, the TSS, BCSS, SSR and SSE were calculated using Equations 12, 13 and 15. Based on these transformations, the TSS, BCSS, WCSS, SSR and SSE are given by Equations 17, 18, 19, 20 and 21 respectively:

\[
TSS = \sum_{i=1}^{n}(f_i - \bar{f})^2
\]  
(17)

\[
BCSS = \sum_{k=1}^{K} \eta_k (\bar{f}_k - \bar{f})^2
\]  
(18)

\[
WCSS = \sum_{k=1}^{K} \sum_{i \in k} (f_i - \bar{f}_k)^2
\]  
(19)

\[
SSR = \sum_{k=1}^{K} \sum_{i \in k} (\bar{f}_k - \hat{f}_k)^2
\]  
(20)

\[
SSE = \sum_{k=1}^{K} \sum_{i \in k} (f_i - \hat{f}_k)^2
\]  
(21)

The TSS, BCSS, WCSS, SSR, and SSE components were calculated for the optimum number of clusters for SS1 and SS2, as shown in Table 8. The results illustrated that, for SS1, BCSS is equal to 0.15% of TSS, and SSR is equal to 75.85% of WCSS. For SS2, BCSS is equal to 1.00% of TSS, and SSR is equal to 78.05% of WCSS. The BCSS accounts only for variation
in the response variable by clustering, and the SSR accounts for variation in the response variable due to explanatory variables. Obtaining a lower percentage for the BCSS and a higher percentage for SSR in WCSS indicates that there is no overfitting, as most of the variation in the response variable is explained by clustering of the response variable with the use of explanatory variables.

4.3.2. Discussion on Prediction Accuracy

Predicted crashes then were compared with observed crashes, and the RMSE for predictions were calculated for both models (Figures 3 and 4). The proposed clusterwise regression method was found to perform better than the single cluster method. Crash frequency functions estimated using the clusterwise method had a lower value of RMSE in prediction, compared to that of the single cluster method for both SS1 and SS2. In addition, the results showed that when using the proposed clusterwise method, the predicted crashes were closer to the 45-degree line, compared with the corresponding prediction using the single cluster method. This indicates that predicted crashes were closer to observed crashes using the proposed method. This trend is clearly seen for higher numbers of observed crashes.

4.4. Network Screening

For the estimated crash frequency function, a single cluster model and clusterwise models were used to estimate predicted crash frequency. The predicted crash frequency of sites was combined with the observed crash frequency in order to obtain a better estimate of the expected and excess crash frequency, using the empirical Bayes method. To identify the sites with potential for safety improvements, network screening analyses for arterial roadway segments and intersections were performed, using excess crash frequency. This provides a measure of crash frequency at sites where crashes were reduced if a safety improvement was implemented. Excess crash frequency
was estimated for total crashes, with peak searching on roadway segments having coefficient-of-
variation limits for the entire network. With peak searching screening type, a minimum window
length of a 0.1 mi segment of the site that had the potential for safety improvement could be
determined in order to deploy a countermeasure. Table 9 shows the results of the top 15 sites (the
first 15 ranks) having the potential for safety improvements for the estimated crash frequency
function, using single cluster and clusterwise regression. It is clear that sites were ranked in
different orders when using a crash frequency function derived from clusterwise regression
versus single cluster. In addition, one roadway segment site (S5658), in Table 9, is not ranked by
the single cluster method within the top 15 sites, as it was with clusterwise regression ranking.
Similarly, four intersection sites (2195, 3440, 3349 and 3301), in Table 9, are not ranked within
the top 15 sites by the single cluster method, as they were with clusterwise regression method.
Considering budget constraints, these results could make a significant difference in choosing
sites with potential for safety improvements.

5. CONCLUSIONS

This study proposed and implemented a clusterwise regression model to develop functions in
order to estimate crash frequencies. The objective was to minimize estimation error by
considering multiple functions, rather than a single function, for a sample of similar sites. A
combinatorial nonlinear mathematical program was formulated. The clusterwise method
simultaneously segmented the roadway sites into a number of clusters, and estimated the
parameters of the crash frequency functions for each cluster. A simulated annealing, coupled
with maximum likelihood, was used to solve the mathematical program. Considering the data
characteristics, a Negative Multinomial count model was used for regression, which took into
account temporal factors by not aggregating traffic and crashes over the period. The algorithm
was tested for number of clusters, and a sensitivity analysis was performed to validate that it was an optimum solution for the provided dataset.

The results obtained from the proposed clusterwise models were compared with the results obtained using a traditional single cluster method. The comparison showed that the proposed clusterwise regression method performed slightly better than a single cluster method in predicting crash estimates. Overfitting in clusterwise models was investigated and verified that the clusterwise models do not have an issue. Benefits of the proposed approach were illustrated with network screening using clusterwise regression as well as a single cluster method. The gain in predicting crashes could translate into significant savings in terms of lives and societal costs.

The clusterwise model could be improved by determining the optimum number of clusters as one of the decision variables. The significance of explanatory variables needs to be evaluated before assigning cluster memberships to roadway sites. This would exclude variables without explanatory power and reassign cluster memberships, which may result in improved estimates. In addition, correlation of explanatory variables should be investigated during the optimization process. Various alternate objective functions such as Bayesian information criteria, total absolute bias, and sum of absolute residuals could be tested. This could provide better goodness of fit and parameter estimates. In addition, future research could involve using the proposed clusterwise approach to estimate superior crash frequency estimation functions that explicitly seek to handle unobserved heterogeneity, temporal and spatial instability, and self-selectivity issues.

ACKNOWLEDGEMENTS
Many thanks to the Nevada Department of Transportation for providing the data that was used for analysis and to our Technical Writer at the Howard R. Hughes College of Engineering, Mrs. Meagan Madariaga-Hopkins, for her help in improving this manuscript.

REFERENCES

American Association of State Highway and Transportation Officials (2010). *Highway Safety Manual*.

American Association of State Highway and Transportation Officials (2011). *Safety Analyst*. Retrieved from http://www.safetyanalyst.org.

Berenson, M. L., Levine, D.M., & Krehbiel, T.C. (2005). *Basic Business Statistics: Concepts and Applications and CD package, 10th Edition, Chapter 8, CD8-1*. Prentice Hall.

Bolker, B. (2016). Tools for general maximum likelihood function – Package ‘bbmle’. *CRAN*. Retrieved from https://cran.r-project.org/web/packages/bbmle/bbmle.pdf

Brusco, M. J., Cradit, J.D., Steinley, D., & Fox, G.L. (2008). Cautionary remarks on the use of clusterwise regression. *Multivariate Behavioral Research, 43* (1), 29-49.

Chung, K., Ragland, D.R., Madanat, S., & Oh, S.M. (2009). The continuous risk profile approach for the identification of high collision concentration locations on congested highways. *Transportation and Traffic Theory, Springer Science, Chapter 23, 463-480*.

Collins, N. E., Eglese, R.W., & Golden, B.L. (1988). Simulated annealing – An annotated bibliography. *American Journal of Mathematical and Management Sciences, 8* (3–4), 209-307.

Depaire, B., Wets, G., & Vanhoof, K. (2008). Traffic accident segmentation by means of latent class clustering. *Accident Analysis and Prevention, 40*, 1257-1266.
DeSarbo, W. S., Oliver, R.L., & Rangaswamy, A. (1989). A simulated annealing methodology for clusterwise linear regression. *Psychometrika, 54* (4), 707–736.

Freeman, T. & Tukey, J. (1950). Transformations related to the angular and the square root. *The Annals of Mathematical Statistics, 21* (4), 607-611.

Fridstrøm, L., Ifver, J., Ingebrigtsen, S., Kulmala, R., & Thomsen, L. (1995). Measuring the contribution of randomness, exposure, weather, and daylight to the variation in road accident counts. *Accident Analysis and Prevention, 27* (1), 1-20.

Geedipally, S. & Lord, D. (2010). Investigating the effect of modelling single-vehicle and multi-vehicle crashes separately on confidence intervals of Poisson-gamma models. *Accident Analysis and Prevention, 42* (4), 1273-1282.

Grembek, O., Kim, K., & Kwon, O.H. (2012). *Experimental evaluation of the continuous risk profile (CRP) approach to the current Caltrans methodology for high collision concentration location identification.* Research Reports, University of California, Berkeley. Retrieved from https://escholarship.org/uc/item/6sg5c0ng

Guo, J. Q. & Zheng, L. (2005). A modified simulated annealing algorithm for estimating solute transport parameters in streams from tracer experiment data. *Environmental Modelling & Software, 20*, 811–815.

Hamidi, A., Fontaine, M.D., & Demetsky, M.J. (2010). *A planning-level methodology for identifying high-crash sections of virginia’s primary system.* Virginia Transportation Research Council, Final Report VTRC 11-R4, VA. Retrieved from http://www.virginiadot.org/vtrc/main/online_reports/pdf/11-r4.pdf.

Hauer, E. (2004). Statistical road safety modeling. *Transportation Research Record: Journal of the Transportation Research Board, 1974*, 799-808.
Hauer, E. (2015). *The Art of Regression Modeling in Road Safety*. New York: Springer.

Hauer, E., Harwood, D., Council, F., & Griffith, M. (2002). Estimating safety by the Empirical Bayes Method: A tutorial. *Transportation Research Record 1784*, 126-131.

Karlaftis, M.G. & Tarko, A.P. (1998). Heterogeneity considerations in accident modelling. *Accident Analysis and Prevention, 30* (4), 425-433.

Kirkpatrick, S., Gelatt, C., & Vecchi M. (1983). Optimization by simulated annealing. *Science, 220* (4598), 671–680.

Kweon, Y. & Lim, I. (2012). Appropriate regression model types for intersections in safety analyst. *Journal of Transportation Engineering, American Society of Civil Engineers, 138* (10), 1250-1258.

Ling-r-Lang-L. (2011). *A Mailing list for language researchers who use R statistical programming language for modeling and data analysis*. Retrieved from https://mailman.ucsd.edu/pipermail/ling-r-lang-l/2011-August/000282.html.

Lau, K., Leung, P., & Tse, K. (1999). A mathematical programming approach to clusterwise regression model and its extensions. *European Journal of Operational Research, 116*, 640-652.

Lord, D., Guikema, S., & Geedipally, S.R. (2008). Application of the Conway–Maxwell–Poisson generalized linear model for analyzing motor vehicle crashes. *Accident Analysis and Prevention, 40* (3), 1123–1134.

Lord, D. & Mannering, F. (2010). The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives. *Transportation Research Part A: Policy and Practice, 44* (5), 291-305.

Lu, H., Huang, S., Li, Y., & Yang, Y. (2014). Panel Data
Analysis via variable selection and subject clustering. In *Data Mining for services, studies in Big Data* 3. Springer-Verlag.

Luo, Z. & Chou, E.Y.J. (2006). Pavement condition prediction using Clusterwise Regression. *Transportation Research Record: Journal of the Transportation Research Board, 1974*, 70–77.

Mannering, F. (2018). Temporal instability and the analysis of highway accident data. *Analytic Methods in Accident Research, 17*, 1-13.

Mannering, F.L., & Bhat, C.R. (2014). Analytic methods in accident research: Methodological frontier and future directions. *Analytic Methods in Accident Research, 1*, 1-22.

Mannering, F., Shankar, V., & Bhat, C.R. (2016). Unobserved heterogeneity and the statistical analysis of highway accident data. *Analytic Methods in Accident Research, 11*, 1-16.

Miaou, S.P., Song, J.J., & Mallick, B.K. (2003). Roadway traffic crash mapping: a space-time modeling approach. *Journal of Transportation and Statistics, 6* (1), 33–57.

Milton, J., Shankar, V., & Mannering, F. (2008). Highway accident severities and the mixed logit model: an exploratory empirical analysis. *Accident Analysis and Prevention, 40* (1), 260-266.

Mohamed, M.G., Saunier, N., Miranda-Moreno, L.F., & Ukkusuri, S.V. (2013). A clustering regression approach: A comprehensive injury severity analysis of pedestrian-vehicle crashes in New York, US and Montreal, Canada. *Safety Science, 54*, 27-37.

Montella, A. (2010). A comparative analysis of hotspot identification methods. *Accident Analysis and Prevention, 42*, 571–581.
Nevada Department of Transportation (2016). *Traffic Safety Engineering*. Retrieved from http://www.nevadadot.com/About_NDOT/NDOT_Divisions/Planning/Traffic_Safety_Engineering/Traffic_Safety_Engineering.aspx

Park, M. W. & Kim, Y.D. (1998). A systematic procedure for setting parameters in simulated annealing algorithms. *Computers & Operations Research, 25* (3), 207–217.

Paz, A., Molano, V., Martinez, E., Gaviria, C., & Arteaga, C. (2015a). Calibration of traffic flow models using a memetic algorithm. *Transportation Research Part-C: Emerging Technologies, 55*, 432-443. Paz, A., Molano, V., & Sanchez, M. (2015b). Holistic calibration of microscopic traffic flow models: Methodology and real world application studies. In N.D. Largos & M. Papadrakakis (Eds.), *Engineering and Applied Sciences Optimization: Dedicated to the memory of Professor M.G. Karlaftis. 38, Ed. 1*. Springer International Publishing.

Paz, A., Veeramisti, N., Khanal, I., Baker, J., & de la Fuente-Mella, H. (2015c). Development of a comprehensive database system for safety analyst. *The Scientific World Journal, 2015, Article ID 636841*, 14 pages.

Poggi J. & Portier, B. (2011). PM10 forecasting using clusterwise regression. *Atmospheric Environment, 45*, 7005-7014.

Román-Román, P., Romero, D., Rubio, M.A., & Torres-Ruiz, F. (2012). Estimating the parameters of a Gompertz-type diffusion process by means of simulated annealing. *Applied Mathematics and Computation, 218* (9), 5121-5131.

Rose, J., Klebsch, W., & Wolf, J. (1990). Temperature measurement and equilibrium dynamics of simulated annealing placement. *IEEE Transactions on Computer Aided Design, 9* (3), 253–259.
Roshan S. B., Jooibari, M.B., Teimouri, R., Asgharzade-Ahmadi, G., Falahati-Naghibi, M., & Sohrabpoor, H., 2013. Optimization of friction stir welding process of AA7075 aluminum alloy to achieve desirable mechanical properties using ANFIS models and simulated annealing algorithm. *International Journal of Advanced Manufacturing Technology, 69* (5-8), 1803–1813.

R-Language. A Mailing list for language researchers who use R statistical programming language for modeling and data analysis. Accessed Online https://mailman.ucsd.edu/pipermail/ling-r-lang-l/2011-August/000282.html

Sasidharan, L., Wu, K., & Menendez, M. (2015). Exploring the application of latent class cluster analysis for investigating pedestrian crash injury severities in Switzerland. *Accident Analysis and Prevention, 85*, 219-228.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6* (2), 461-464.

Selim, S. Z. & Alsultan, K. (1991). A simulated annealing algorithm for the clustering problem. *Pattern Recognition, 24* (10), 1003–1008. Shugan, S. (2006). Editorial: errors in the variables, unobserved heterogeneity, and other ways of hiding statistical error. *Marketing Science, 25* (3), 203–216.

Spath, H. (1979). Clusterwise linear regression. *Computing, 22* (4), 367–373.

Srinivasan, R., Carter, D., & Bauer, K. (2013). *Safety Performance Function decision guide: SPF calibration vs SPF development*. Federal Highway Administration – Office of Safety Report, Washington, DC.

Ulfarsson, G.F. & Shankar, V.N. (2003). An accident count model based on multi-year cross-sectional roadway data with serial correlation. *Transportation Research Record: Journal of the Transportation Research Board, 1840*, 193-197.
Wang, C., Quddus, M.A., & Ison, S.G. (2011). Predicting accident frequency at their severity levels and its application in site ranking using a two-stage mixed multivariate model. *Accident Analysis and Prevention, 43*, 1979-1990.

Washington, S.P., Karlaftis, M.G., & Mannering, F.L. (2010). *Statistical and Econometric Methods for Transportation Data Analysis, 2nd ed.* Boca Raton, FL: Chapman Hall/CRC.
| No. | Site Subtype                  | Description                                                                 | Number of sites |
|-----|------------------------------|-----------------------------------------------------------------------------|-----------------|
| 1   | Rural 2-lane                 | Segments on rural 2 lane road                                               | 122             |
| 2   | Rural multilane undivided    | Segments on rural 4+ lanes with no median                                    | 45              |
| 3   | Rural freeway 4 lane         | Segments on rural 4+ lane with no access control                            | 55              |
| 4   | Rural freeway in interchange area | Segments on rural 4+ lane with no access control and in interchange influence area | 98              |
| 5   | Urban 2-lane arterial        | Segments on urban 2 lane arterial                                           | 1788            |
| 6   | Urban multilane undivided    | Segments on urban 4+ lanes with no median                                    | 1227            |
| 7   | Urban multilane divided      | Segments on urban 4+ lanes with median                                       | 1891            |
| 8   | Urban One-way arterial       | Segments on urban one-way arterial                                          | 70              |
| 9   | Urban freeway 4 lane         | Segments on urban 4+ lane with no access control                            | 47              |
| 10  | Urban freeway in interchange area | Segments on urban 4+ lane with no access control and in interchange influence area | 289             |
| 11  | Urban 3-leg signalized       | Urban 3-leg Intersections with signal control                               | 125             |
| 12  | Urban 4-leg signalized       | Urban 4-leg Intersections with signal control                               | 340             |
| 13  | Urban 3-leg minor road stop  | Urban 3-leg Intersec with stop control on minor roads                       | 155             |
| 14  | Urban 4-leg minor road stop  | Urban 4-leg Intersec with stop control on minor roads                       | 74              |
| 15  | Urban 4-leg all-way stop     | Urban 4-leg Intersections with stop control on all roads                    | 78              |
| 16  | Urban Diamond off-ramp       | Diamond off-ramps in urban area                                             | 233             |
| 17  | Urban Diamond on-ramp        | Diamond on-ramps in urban area                                              | 229             |
### TABLE 2 Descriptive Statistics of Urban Multilane Divided Segments (SS1) with Fatal and all Injury Crashes

| Variables                  | Count | Min | Max  | Mean   | Median | Mode | Std Dev   | Range  | Sum   |
|----------------------------|-------|-----|------|--------|--------|------|-----------|--------|-------|
| **Continuous Variables**   |       |     |      |        |        |      |           |        |       |
| Observed Crash 2007        | 1891  | 0   | 77   | 2.03   | 0      | 0    | 5.37      | 77     | 3835  |
| Observed Crash 2008        | 1891  | 0   | 91   | 1.88   | 0      | 0    | 5.07      | 91     | 3561  |
| Observed Crash 2009        | 1891  | 0   | 58   | 1.84   | 0      | 0    | 4.64      | 58     | 3483  |
| Observed Crash 2010        | 1891  | 0   | 58   | 1.71   | 0      | 0    | 4.51      | 58     | 3235  |
| Observed Crash 2011        | 1891  | 0   | 57   | 1.75   | 0      | 0    | 4.56      | 57     | 3309  |
| AADT 2007                  | 1891  | 1148| 12600| 22752.77| 19000  | 17000| 15917.57  | 124852 | 16852 |
| AADT 2008                  | 1891  | 702 | 10800| 21672.84| 18217  | 18000| 15082.60  | 107298 | 12852 |
| AADT 2009                  | 1891  | 1148| 10800| 21272.14| 18000  | 30000| 14599.30  | 106852 | 11005 |
| AADT 2010                  | 1891  | 1000| 10800| 20282.38| 17000  | 14000| 14225.10  | 107000 | 15825 |
| AADT 2011                  | 1891  | 880 | 11000| 20111.56| 16902  | 16000| 14200.58  | 109120 | 14000 |
| Segment Length             | 1891  | 0.005 | 3.559 | 0.246   | 0.166  | 0.125| 0.250     | 3.553  | 6000  |
| **Categorical Variables**  |       |     |      |        |        |      |           |        |       |
| Road Type - State Road      | 285   |     |      |        |        |      |           |        |       |
| Road Type - Non State Road  | 1606  |     |      |        |        |      |           |        |       |
| Functional Class 3         | 232   |     |      |        |        |      |           |        |       |
| Functional Class 4         | 1021  |     |      |        |        |      |           |        |       |
| Functional Class 6         | 636   |     |      |        |        |      |           |        |       |
| Functional Class 7         | 2     |     |      |        |        |      |           |        |       |
| Number of Lanes 4          | 1090  |     |      |        |        |      |           |        |       |
| Number of Lanes 5          | 40    |     |      |        |        |      |           |        |       |
| Number of Lanes 6          | 718   |     |      |        |        |      |           |        |       |
| Number of Lanes 7          | 21    |     |      |        |        |      |           |        |       |
| Number of Lanes 8          | 208   |     |      |        |        |      |           |        |       |
| Median Divided             | 1877  |     |      |        |        |      |           |        |       |
| Flush Paved                | 12    |     |      |        |        |      |           |        |       |
| other Divided              | 2     |     |      |        |        |      |           |        |       |
| Median Width < 4           | 672   |     |      |        |        |      |           |        |       |
| Median Width 4 – 14 ft     | 818   |     |      |        |        |      |           |        |       |
| Median Width > 14 ft       | 401   |     |      |        |        |      |           |        |       |
| Posted Speed < 25          | 105   |     |      |        |        |      |           |        |       |
| Posted Speed 30 & 35 mph   | 954   |     |      |        |        |      |           |        |       |
| Posted Speed 40 & 45 mph   | 793   |     |      |        |        |      |           |        |       |
| Posted Speed 50 & 55 mph   | 37    |     |      |        |        |      |           |        |       |
| Posted Speed > 55 mph      | 7     |     |      |        |        |      |           |        |       |
| Variables | Count | Min | Max | Mean | Median | Mode | Std Dev | Range | Sum |
|-----------|-------|-----|-----|------|--------|------|---------|-------|-----|
| **Continuous Variables** |       |     |     |      |        |      |         |       |     |
| Observed Crash 2007 | 340 | 0  | 30  | 7.08 | 5.5    | 0    | 6.12    | 30    | 2407|
| Observed Crash 2008 | 340 | 0  | 36  | 6.13 | 5      | 1    | 5.69    | 36    | 2084|
| Observed Crash 2009 | 340 | 0  | 46  | 6.63 | 5      | 1    | 5.93    | 46    | 2255|
| Observed Crash 2010 | 340 | 0  | 31  | 6.62 | 5      | 4    | 5.83    | 31    | 2250|
| Observed Crash 2011 | 340 | 0  | 41  | 6.81 | 5      | 7    | 6.35    | 41    | 2316|
| Major AADT 2007 | 340 | 2942 | 212000 | 60058.24 | 56038.0 | 50000 | 31074.07 | 209058 |
| Major AADT 2008 | 340 | 2578 | 184000 | 57347.09 | 54803.5 | 50000 | 29391.63 | 178276 |
| Major AADT 2009 | 340 | 2360 | 182000 | 53855.07 | 50415.5 | 42000 | 27825.17 | 179640 |
| Major AADT 2010 | 340 | 2252 | 188000 | 53611.38 | 51685.5 | 42000 | 27888.77 | 185748 |
| Minor AADT 2007 | 340 | 1970 | 133486 | 31645.81 | 25969.0 | 28000 | 23256.78 | 131516 |
| Minor AADT 2008 | 340 | 2800 | 123000 | 29877.29 | 24807.5 | 26000 | 21974.17 | 120200 |
| Minor AADT 2009 | 340 | 2280 | 125000 | 29395.20 | 24430.0 | 33000 | 21531.41 | 122720 |
| Minor AADT 2010 | 340 | 2400 | 123000 | 27656.12 | 21880.5 | 39000 | 20375.66 | 120600 |
| Minor AADT 2011 | 340 | 2400 | 124000 | 27411.66 | 21360.0 | 38000 | 20526.22 | 121600 |
| **Categorical Variables** |       |     |     |      |        |      |         |       |     |
| Major Road Lanes = 2 | 25 |     |     |      |        |      |         |       |     |
| Major Road Lanes = 3-4 | 143 |     |     |      |        |      |         |       |     |
| Major Road Lanes > 4 | 172 |     |     |      |        |      |         |       |     |
| Minor Road Lanes = 2 | 68 |     |     |      |        |      |         |       |     |
| Minor Road Lanes = 3-4 | 173 |     |     |      |        |      |         |       |     |
| Minor Road Lanes > 4 | 99 |     |     |      |        |      |         |       |     |
| Major Road Median – UnDivided | 57 |     |     |      |        |      |         |       |     |
| Major Road Median - Divided | 283 |     |     |      |        |      |         |       |     |
| Minor Road Median – UnDivided | 108 |     |     |      |        |      |         |       |     |
| Minor Road Median - Divided | 233 |     |     |      |        |      |         |       |     |
| Major Rd Posted Speed 15 & 25 mph | 17 |     |     |      |        |      |         |       |     |
| Major Rd Posted Speed 30 & 35 mph | 145 |     |     |      |        |      |         |       |     |
| Major Rd Posted Speed 40 & 45 mph | 165 |     |     |      |        |      |         |       |     |
| Major Rd Posted Speed 50 & 55 mph | 13 |     |     |      |        |      |         |       |     |
| Minor Rd Posted Speed 15 & 25 mph | 47 |     |     |      |        |      |         |       |     |
| Minor Rd Posted Speed 30 & 35 mph | 160 |     |     |      |        |      |         |       |     |
| Minor Rd Posted Speed 40 & 45 mph | 122 |     |     |      |        |      |         |       |     |
| Minor Rd Posted Speed 50 & 55 mph | 12 |     |     |      |        |      |         |       |     |
| Parameter | Value     | Description                                                                                                                                 |
|-----------|-----------|------------------------------------------------------------------------------------------------------------------------------------------------|
| $T_0$     | 1         | Initial temperature                                                                                                                        |
| $T_{\text{min}}$ | 0.000001 | Minimum temperature                                                                                                                        |
| $k$       | 4         | Boltzmann constant                                                                                                                         |
| $\Phi$    | 0.985     | Cooling rate                                                                                                                               |
| $N_{\text{max}}$ | 10       | Maximum number of neighborhood solutions to be generated at each temperature level                                                        |
TABLE 5 Estimated Parameters Using the Proposed Clusterwise Regression and the Single-Cluster Method – Urban Multilane Divided Segments (SS1)

| Parameters | Clusters from Clusterwise Regression Method | Single Cluster (Traditional Approach) |
|------------|---------------------------------------------|---------------------------------------|
|            | Cluster 1 (η = 941)                         | Cluster 2 (η = 950)                   |
|            | Estimate | Std Error | Estimate | Std Error | η = 1891 Estimate | Std Error |
| Intercept (Scale) | 37.754 | 1.09E+01 | 675.250 | 3.30E-04 | 272.283 | 5.19E+01 |
| length | 0.958 | 4.72E-02 | 0.984 | 3.77E-02 | 0.977 | 3.34E-02 |
| AADT | 0.677 | 5.46E-02 | 1.530 | 5.11E-02 | 1.079 | 4.01E-02 |
| Road Type – State Rd | 0.526 | 4.79E-02 | 1.111 | 9.39E-02 | 0.682 | 4.63E-02 |
| [2] Functional Class 4 | 0.952 | 9.34E-02 | 0.740 | 6.35E-02 | 0.891 | 6.33E-02 |
| [3] Functional Class 6 | 0.397 | 5.32E-02 | 1.133 | 1.28E-01 | 0.739 | 6.87E-02 |
| [4] Functional Class 7 | 0.794 | 1.43E+00 | 0.093 | 1.17E-01 | 0.217 | 2.49E-01 |
| [2] Number of Lanes 5 | 1.353 | 3.52E-01 | 1.030 | 2.00E-01 | 1.283 | 2.23E-01 |
| [3] Number of Lanes 6 | 1.184 | 9.99E-02 | 1.317 | 9.32E-02 | 1.233 | 7.43E-02 |
| [4] Number of Lanes 7 | 1.394 | 3.24E-01 | 0.822 | 2.10E-01 | 1.228 | 2.29E-01 |
| [5] Number of Lanes 8 | 0.063 | 2.02E-02 | 0.578 | 9.06E-02 | 0.541 | 8.20E-02 |
| [2] Median–Flush Paved | 1.411* | 8.33E-01 | 0.787 | 3.47E-01 | 0.838 | 3.23E-01 |
| [3] Median–Other Divided | 1.000 | 1.62E-12 | 0.130* | 1.03E-01 | 0.188* | 1.64E-01 |
| [2] Median Width 4 – 14 ft | 0.782 | 5.90E-02 | 0.680 | 4.52E-02 | 0.737 | 4.06E-02 |
| [3] Median Width > 14 ft | 0.450 | 4.58E-02 | 1.139 | 9.31E-02 | 0.834 | 5.86E-02 |
| [2] Speed Limit <= 25 | 2.649 | 6.59E-01 | 0.521 | 6.08E-02 | 0.957 | 1.41E-01 |
| [3] Speed Limit <= 45 | 2.185 | 5.57E-01 | 0.436 | 5.05E-02 | 0.792 | 1.20E-01 |
| [4] Speed Limit <= 55 | 1.015 | 3.26E-01 | 0.436 | 9.07E-02 | 0.535 | 1.11E-01 |
| [5] Speed Limit > 55 | 3.091 | 1.14E+00 | 0.655 | 2.67E-01 | 0.903 | 2.56E-01 |
| b-shape parameter (scaled) | 11.236 | 7.46E-01 | 11.236 | 7.46E-01 | 25.323 | 1.69E+00 |
| Log Likelihood | 5969 | 6767 | 12549 | 6767 |
| Freeman-Tukey R² (R²FT) | 0.87 | 0.88 | 0.84 | 0.88 |

* coefficient with p-value > 0.05; η represents the total number of sites in a cluster; [] base levels for all categorical variables; speed limit unit is mph
### TABLE 6 Estimated Parameters Using the Proposed Clusterwise Regression and the Single-Cluster Method – Urban 4-leg Signalized Intersections (SS2)

**Urban 4-leg Signalized Intersections (SS2)**

| Parameters                                      | Clusters from Clusterwise Regression Method | Single Cluster (Traditional Approach) |
|-------------------------------------------------|--------------------------------------------|---------------------------------------|
|                                                 | Cluster 1 ($\eta = 89$)                  | Cluster 2 ($\eta = 88$)               | Cluster 3 ($\eta = 94$) | Cluster 4 ($\eta = 69$) | $\eta = 340$ |
|                                                 | Estimate        | Std Error | Estimate        | Std Error | Estimate        | Std Error | Estimate        | Std Error | Estimate        | Std Error |
| Intercept                                       | 13.195          | 4.015     | 20.687          | 5.448     | 8.028           | 2.898     | 8.171           | 3.127     | 17.780          | 3.492     |
| Major Road AADT                                 | 0.361           | 0.084     | 1.115           | 0.094     | 0.418           | 0.109     | 0.350           | 0.103     | 0.547           | 0.057     |
| Minor Road AADT                                 | 0.254           | 0.067     | 0.202           | 0.066     | 0.441           | 0.093     | 0.104           | 0.085     | 0.347           | 0.047     |
| [2] Major Road Lanes = 3-4                      | 1.478           | 0.373     | 1.562           | 0.400     | 1.611           | 0.370     | 0.876           | 0.177     | 1.348           | 0.180     |
| [3] Minor Road Lanes > 4                        | 2.739           | 0.721     | 1.371           | 0.383     | 1.543           | 0.398     | 0.966           | 0.191     | 1.541           | 0.219     |
| [2] Minor Road Lanes = 3-4                      | 1.309           | 0.152     | 1.625           | 0.229     | 0.599           | 0.104     | 1.645           | 0.270     | 0.938           | 0.083     |
| [3] Minor Road Lanes > 4                        | 2.637           | 0.408     | 1.610           | 0.245     | 0.955           | 0.197     | 2.715           | 0.540     | 1.147           | 0.124     |
| [2] Major Road Median - Divided                 | 1.229           | 0.188     | 1.174           | 0.209     | 0.553           | 0.111     | 0.353           | 0.082     | 0.859           | 0.093     |
| [2] Minor Road Median - Divided                 | 0.855           | 0.100     | 0.815           | 0.091     | 1.172           | 0.173     | 1.224           | 0.216     | 0.894           | 0.071     |
| [2] Major Rd Posted Speed 26 to 35              | 0.330           | 0.097     | 0.767           | 0.188     | 1.326           | 0.470     | 0.741           | 0.201     | 0.895           | 0.147     |
| [3] Major Rd Posted Speed 36 to 45              | 0.584           | 0.175     | 0.650           | 0.169     | 1.087           | 0.422     | 0.448           | 0.129     | 0.902           | 0.159     |
| [4] Major Rd Posted Speed 46 to 65              | 0.517           | 0.255     | 0.282           | 0.098     | 1.082           | 0.504     | 0.059           | 0.037     | 0.747           | 0.186     |
| [2] Minor Rd Posted Speed 26 to 35              | 0.799           | 0.105     | 1.181           | 0.180     | 1.343           | 0.379     | 1.895           | 0.292     | 1.102           | 0.117     |
| [3] Minor Rd Posted Speed 36 to 45              | 0.437           | 0.070     | 1.245           | 0.217     | 1.793           | 0.591     | 1.940           | 0.397     | 1.026           | 0.128     |
| [4] Minor Rd Posted Speed 46 to 65              | 0.756           | 0.259     | 2.538           | 0.701     | 2.429           | 1.356     | 3.340           | 1.535     | 1.323           | 0.296     |
| b-shape parameter                               | 14.088          | 3.470     | 13.496          | 3.040     | 5.181           | 1.029     | 14.693          | 3.796     | 4.397           | 0.409     |
| Log Likelihood                                  | 3350            | 3371      | 3335            | 2948      | 1.025           | 0.509     | 2948            | 1.2882    | 0.98            | 0.94      |

\* coefficient with p-value > 0.05; \( \eta \) represents the total number of sites in a cluster; \([\] base levels for all categorical variables; speed limit unit is mph
| Clusters | BIC    | Clusters | BIC    |
|----------|--------|----------|--------|
| 1        | -25032 | 1        | -52339 |
| 2        | -25338 | 2        | -52644 |
| 3        | -25261 | 3        | -52786 |
| 4        | -25174 | 4        | -52835 |
| 5        | -25113 | 5        | -52798 |
| 6        | -24982 |          |        |
### Table 8 Measures of Overfitting Components Associated with Clusterwise Regression

| Measure                                      | Clusterwise Regression - SS1 | Clusterwise Regression - SS2 |
|----------------------------------------------|------------------------------|------------------------------|
| Total sum of squares (TSS)                   | 33,915 (100%)                | 6,520 (100%)                 |
| Between-clusters sum of squares (BCSS)       | 48 (0.15% of TSS)            | 65 (1.00% of TSS)            |
| Within-clusters sum of squares (WCSS)        | 33,867 (99.85% of TSS)       | 6,455 (99.00% of TSS)        |
| Sum of squares due to regression (SSR)       | 25,691 (75.85% of WCSS)      | 5,038 (78.05% of WCSS)       |
| Sum of squared error of prediction (SSE)     | 8,176 (24.15% of WCSS)       | 1,417 (21.95% of WCSS)       |
### TABLE 9 Network Screening Results using the Proposed Clusterwise Regression and the Single-Cluster Method for SS1 and SS2

| Segment/Intersect ID | Observed Crash Frequency * | Predicted Crash Frequency * | Excess Crash Frequency * | Segment/Intersect ID | Observed Crash Frequency ** | Predicted Crash Frequency ** | Excess Crash Frequency ** |
|----------------------|-----------------------------|----------------------------|--------------------------|----------------------|----------------------------|----------------------------|--------------------------|
| **Urban Multilane Divided Segments (SS1)** |                             |                            |                          | **Urban 4-leg Signalized Intersections (SS2)** |                             |                            |                          |
| S5700a               | 164.65                      | 35.05                      | 129.31                   | S5700a               | 164.65                      | 25.90                      | 138.59                   |
| S4438                | 116.26                      | 20.68                      | 95.22                    | S4438                | 116.26                      | 22.86                      | 93.11                    |
| S4459                | 100.94                      | 17.58                      | 82.52                    | S7758                | 105.18                      | 16.76                      | 88.09                    |
| S7758                | 105.18                      | 26.59                      | 78.36                    | S4459                | 100.94                      | 15.42                      | 85.06                    |
| S9695                | 104.01                      | 26.08                      | 77.40                    | S7736a               | 89.88                       | 11.01                      | 78.78                    |
| S7736a               | 89.88                       | 13.55                      | 75.12                    | S7693a               | 86.09                       | 10.23                      | 76.92                    |
| S7693a               | 86.09                       | 13.40                      | 72.22                    | S9695                | 104.01                      | 27.93                      | 75.95                    |
| S3129a               | 83.14                       | 13.40                      | 68.83                    | S5798                | 88.84                       | 14.78                      | 73.42                    |
| S5798                | 88.84                       | 25.93                      | 62.72                    | S8070                | 103.48                      | 31.58                      | 71.58                    |
| S8070                | 103.48                      | 45.53                      | 57.85                    | S5619a               | 83.47                       | 12.92                      | 70.92                    |
| S918a                | 81.77                       | 24.23                      | 57.53                    | S3346                | 83.47                       | 12.92                      | 69.92                    |
| S5619a               | 84.06                       | 27.13                      | 56.76                    | S5619a               | 83.14                       | 15.00                      | 67.81                    |
| S3346                | 83.47                       | 30.04                      | 53.29                    | S3346                | 83.47                       | 15.00                      | 67.14                    |
| S6570                | 87.29                       | 36.13                      | 51.05                    | S918a                | 81.77                       | 17.35                      | 64.07                    |
| S9678a               | 81.89                       | 30.82                      | 50.78                    | S9678a               | 81.89                       | 21.43                      | 63.28                    |

*crashes/mi/year; **crashes/year
FIGURE 1 Algorithm for clusterwise regression to estimate parameters of the crash frequency estimation functions.
(a) MLE during optimization for SS1  
(b) MLE during optimization for SS2  
(c) Sensitivity Analysis for SS1  
(d) Sensitivity Analysis for SS2  

FIGURE 2 (a and b) Evolution of MLE during optimization for a clusterwise regression model and (c and d) sensitivity analyses for the number of clusters.
FIGURE 3 Comparison of the predicted and observed number of crashes, using the proposed methods for SS1.
FIGURE 4 Comparison of the predicted and observed number of crashes, using the proposed methods for SS2