# Crash Modeling Using Clustered Data from Washington State:

# Prediction of Optimal Speed Limits

By

**Jianming Ma**, Graduate Student Researcher, The University of Texas at Austin.
6.9 E. Cockrell Jr. Hall, Austin, TX 78712-1076, mjming@mail.utexas.edu

**Kara M. Kockelman**, Clare Boothe Luce Assistant Professor of Civil Engineering
The University of Texas at Austin, 6.9 E. Cockrell Jr. Hall, Austin, TX 78712-1076
kkockelm@mail.utexas.edu, Phone: 512-471-0210, FAX: 512-475-8744
(Corresponding Author)

Submitted for Presentation at the 2005 Annual Meeting of the Transportation Research Board

Word Count: 4,788 + 6 figures & tables = 6,288 word-equivalents

## ABSTRACT

This study investigates the relationship between crash frequencies, roadway design and use features by utilizing the benefits of clustered panel data. Homogeneous high-speed roadway segments across the State of Washington were grouped using TwoStep Cluster Analysis technique, resulting in grouped observations with reasonably continuous crash count values. This permitted application of both fixed- and random-effects linear regression models for the total number of crashes per million vehicle miles traveled (VMT). A crash severity model also was estimated, using an ordered logistic regression, allowing transformation of total crash counts into counts by severity. Speed limit information is found to be very valuable in predicting crash rates, and the models are seemingly able to predict "optimal" speed limits in order to minimize crash rates and crash costs. However, speed limits may have biased coefficients, most likely attributable to unobserved safety-related effects. For the "average" high-speed segment in the data set, a minimum expected crash cost is achieved at a speed limit of 70 mi/h, while the maximum crash rate is predicted to occur at a speed limit of 43.5 mi/h. While these calculations may not be realistic, the models appear to accurately predict crash rates ($R^2$ of 0.90 for total crash count) and the results provide useful information for a variety of design and use effects. For example, crashes are more frequent on shorter horizontal curves, while uphill segments with wider medians are found to experience less severe crashes.

Keywords: TwoStep cluster analysis, crash panel data, fixed-effects models, ordered logistic regression, safe speed limits.

## INTRODUCTION

Traffic crashes remain a major public health problem. In 2002 42,815 persons died on U.S. roads, and almost 3 million were injured in over 6 million police-reported motor vehicle traffic crashes (NHTSA 2003). NHTSA (2003) estimates the total cost at $230 billion, or over $800 per person annually. Through a better understanding of what impacts crash frequency and crash severity, effective life- and cost-saving measures can be pursued.

Roadway segments vary dramatically in their design and use levels, even during the course of a mile along a single routing. In modeling crash counts as a function of design details, like vertical grade and horizontal curvature, modelers often must analyze data from very short segments in order to obtain uniformity in design and use characteristics. Over the course of a year, crashes on short segments are typically few, particularly fatal crashes. Under these conditions, classical linear regression models are not workable. Discrete models of counts are used. For example, Miaou and Lum (1993) employed Poisson regression models to investigate the relationship between crash occurrence and highway geometric design features in Utah. Shankar et al. (1995) used negative binomial models to investigate the effects of roadway geometrics and environmental factors on rural freeway crash occurrence. Poch and Mannering (1996) and Milton and Mannering (1998) also used negative binomial regression models. And Shankar et al. (1997) have used zero-inflated Poisson (ZIP) models. McCarthy (1999) employed fixed-effects negative binomial models to examine fatal crash counts using 9 years of panel data for 418 cities and 57 areas in the U.S. Noland (2003) used fixed- and random-effects negative binomial models to investigate the effects of roadway improvements on traffic safety using 14 years of data for all 50 U.S. states. Kweon and Kockelman (2004) also used such models (along with zero-inflated and simpler, pooled models) to study the effects of speed limits, design, and use on crash occurrence in Washington State.

There are certain drawbacks to using count models, such as possibly inconsistent estimation and inference due to potentially inappropriate distributional assumptions, together with difficulty in quantitative interpretation of parameters (due to exponential or other transformations of the rate equation in order to ensure positive predictions). In order to account for these issues, another approach is taken here, which first classifies roadway segment observations into relatively homogeneous clusters. Others in transportation safety analysis have used cluster analysis, but for different applications. Almost three decades ago Moellering (1976) studied the patterns of traffic crashes using geographical cluster analysis. Golob and Recker (2004) utilized cluster analysis to classify traffic flow regimes for different freeway crash types. Le Blanc and Rucks (1996) clustered crashes occurring on the lower Mississippi River by crash type, traffic level, and location. Wong et al. (2004) grouped safety projects for Hong Kong roadways and examined the relationship with crash rates using linear regression models with statistically significant time trends. Wells-Parker and Cosby (1986) clustered DUI (driving while under the influence of alcohol or drugs) offenders into five subgroups based on number of traffic violations and other characteristics and examined relationships between variables like alcohol consumption and accident risk. Gregersen and Berg (1994) clustered young persons by lifestyle and examined crash risk. Ulleberg (2001) also clustered young drivers and investigated their responses to a traffic safety campaign. Finally, Sohn (1999) investigated the relationships between crash counts, roadway design, and other factors using Poisson regression models for clustered Korean

crash data. Among all these, only Wong et al. (2004) and Sohn (1999) used regression analysis based on their clustered data, resulting in applications most similar to those pursued here.

## DATA SETS AND CLUSTER ANALYSIS

The crash data sets used here were collected from Washington State through the Highway Safety Information System (HSIS). A total of 396,925 occupants were involved in 151,697 reported crashes, resulting in 2,909 fatalities from 1993 to 1996 on Washington State highways. These data contain information on occupants' demographics, roadway design features (including speed limits[1]), vehicle characteristics, environmental conditions (at the time of crash), and basic crash information (such as crash severity, time, locations and type).

Speed limits are a key variable for this work, so the data emphasize the years 1993 through 1996, which bracket the repeal of the National Maximum Speed Limit (NMSL). This work also emphasizes high-speed roadways, so roadway segments having speed limits less than 50 mi/h were removed from the data set.

Cluster analysis groups observations into relatively homogenous collections (i.e., clusters) by essentially minimizing the variance or spread across defining variables of interest within the clusters, and maximizing that between clusters. Chiu et al.'s (2001) two-step cluster analysis routine was employed here. It is effective for very large datasets with both continuous and categorical variables, as used here. Observations are pre-clustered using log-likelihood distances, creating a cluster feature "tree." The resulting subclusters are further grouped, by comparing their distances to a specified threshold. If the distance is *larger* than the threshold, the two clusters are merged. The distance between two clusters $j$ and $s$ is defined as the *de*crease in log-likelihood due to *merging* the two clusters:

$$d(j,s) = \xi_j + \xi_s - \xi_{<j,s>} \tag{1}$$

where

$$\xi_v = N_v\left(\sum_{k=1}^{K^A}\frac{1}{2}\log\left(\hat{\sigma}_k^2 + \hat{\sigma}_{vk}^2\right) + \sum_{k=1}^{K^B}\hat{E}_{vk}^2\right) \text{ and } \hat{E}_{vk} = -\sum_{l=1}^{L_k}\frac{N_{vkl}}{N_v}\log\frac{N_{vkl}}{N_v},$$

and where $K^A$ is the total number of continuous variables used, $K^B$ is the total number of categorical variables used, $L_k$ is the umber of categories for the $k$-th categorical variable, $N_j$ is the number of observations in cluster $j$, $\hat{\sigma}_k^2$ is the variance of the $k$-th continuous variable in the original data set, $\hat{\sigma}_{jk}^2$ is the variance of the $k$-th continuous variable in cluster $j$,

[1] The HSIS speed limit information is routinely provided off cycle from the other data, so correct speed limit information was obtained from Washington DOT's Bob Howden.

$N_{jkl}$ is the number of observations in cluster $j$ whose $k$-th categorical variable takes the $l$-th category, and $< j, s >$ represents the cluster formed by merging clusters $j$ and $s$.
In calculating the log-likelihood, continuous variables are assumed to be normally distributed, and categorical variables are assumed to follow multinomial distributions. Chiu et al.'s first step, pre-clustering, adopts the clustering approach used in BIRCH, as developed by Zhang et al. (1996). The typical cluster feature $CF_j$ for a cluster $C_j$ is as follows (Chiu et al., 2001):

$$CF_j = \left\{N_j, s_{Aj}, s_{Aj}^2, N_{Bj}\right\}$$

where $s_{Aj}$ is the sum of continuous variables in cluster $C_j$, $s_{Aj}^2$ is the sum of squared continuous variables in cluster $C_j$, and $N_{Bj} = \left(N_{Bj1}, N_{Bj2}, \Lambda\ , N_{BjK^B},\right)$ is an $\sum_{k=1}^{K^B}\left(L_k - 1\right)$-dimensional vector whose $k$-th sub-vector is of dimension/length $\left(L_k - 1\right)$.

When two clusters $C_j$ and $C_s$ are merged, the merged cluster feature $CF_{<j,s>}$ can be obtained using the equation below (Chiu et al., 2001):

$$CF_{<j,s>} = \left\{N_j + N_s, s_{Aj} + s_{As}, s_{Aj}^2 + s_{As}^2, N_{Bj} + N_{Bs}\right\}$$

Compared to K-means and hierarchical clustering techniques (Chiu et al., 2001), the CF structure saves a great amount of time for TwoStep cluster analysis.
The optimal number of clusters can be determined using either a Bayesian or Akaike Information Criterion (BIC and AIC). For $J$ clusters, these can be obtained as follows (Chiu et al., 2001):

$$BIC(J) = -2\sum_{j=1}^{J}\xi_j + m_J \log(N) \quad (2)$$

$$AIC(J) = -2\sum_{j=1}^{J}\xi_j + 2m_J \quad (3)$$

where

$$m_J = J\left\{2K^A + \sum_{k=1}^{K^B}\left(L_k - 1\right)\right\}.$$

Segments were clustered on the basis of their design attributes (including terrain, vertical grade, and horizontal curvature).[2] Their associated dependent variable information (i.e., crash counts and VMT) were summed to create overall crash rates for each cluster in each year of the data set.

[2] AADT per lane also is an important variable that may be of value for clustering (so that high-demand roadways are not grouped with low-demand roadways). However, it is not certain that the panel of clustered segments will remain stable in this attribute over time. Therefore, this variable was not used for clustering purposes here.

For their associated independent variables, values were averaged In order to maintain a true panel after clustering, only the 1993 data were clustered. The membership obtained from their clustering was employed to cluster the remaining observations, from 1994 through 1996. The result is a panel data set for clustered segments. Descriptive statistics for this final data set are shown in Tables 1 and 2. Before clustering, there were around 59,500 segments, the average crash count was 0.3 crashes (per year per segment), and the average segment length was just 0.09 miles. 812 clusters were created, resulting in average crash counts and lengths climbing to 21.7 crashes (per year per cluster) and 6.6 miles (per cluster). The average VMT before clustering was just 965 miles (per year per segment) before clustering, and rose to 69,793 after clustering. Clearly, the data have been made much more continuous in nature, permitting application of more standard – and easier to interpret – linear models.

## CRASH OCCURRENCE MODEL

Panel data permit identification of variations across individual roadway segments and variations over time. Accommodation of observation-specific effects also mitigates omitted-variables bias, by implicitly recognizing segment-specific attributes that may be correlated with control variables. The results of pooled data models, which ignore the panel nature of the data, may be compared to those allowing for fixed and random effects. In the fixed-effects (FE) linear model, the specification is as follows (Greene, 2002):

$$y_{it} = x'_{it}\beta + \alpha_i + \varepsilon_{it} \quad \text{for } i = 1,2,\Lambda\ ,N \text{ and } t = 1,2,\Lambda\ ,T \tag{4}$$

where $\alpha_i$ is the roadway segment's specific effect, a constant term that does not vary over time.

The segment-specific constant term in the regression model can be computed using the following formula (Greene, 2002):

$$\alpha_i = \overline{y}_i - b'\overline{x}_i \tag{5}$$

where $\overline{y}_i$ is the response variable mean over the $T$ observations for segment $i$, $\overline{x}_i$ is the means of control variables over the $T$ observations for segment $i$, and $b$ is the least squares dummy variable (LSDV) estimator.

In random-effects (RE) linear panel models, the specification is as follows (Greene, 2002):

$$y_{it} = x'_{it}\beta + \left(\alpha + u_i\right) + \varepsilon_{it} \tag{6}$$

where $u_i$ is the roadway segment's specific random element. The only difference between $u_i$ and $\varepsilon_{it}$ is that there is only one draw ($u_i$) for each segment which remains constant over time, while the conventional random term ($\varepsilon_{it}$) varies both across segments and across time periods, for each segment.

The FE linear panel models can be estimated using a least squares dummy variable (LSDV) model (Hsiao, 2003). The RE linear panel models can be estimated using a generalized least squares (GLS) approach, by assuming an appropriate distribution for the compound error term. Usually, RE estimates are more efficient than FE estimates since they are obtained by making use of both within-group and between-group variations (rather than only within-group variations) (Hsiao, 2003). However, when there is correlation between unobserved, omitted variables and included control variables, the RE estimates are biased, while the FE estimates are unbiased. (Hsiao, 2003) The question arises as to which model should be used in practice. If the FE models are used, there will be a loss of $N-1$ degrees of freedom in estimating the segments' specific effects. If the RE models are used, one must assume that segment-specific effects are uncorrelated with other, included variables. Hausman's test for such correlation can be performed by calculating the following chi-squared statistic (Greene 2002):

$$W = \chi^2[K-1] = \left[b_{FE} - \hat{\beta}_{RE}\right]' \hat{\psi}^{-1} \left[b_{FE} - \hat{\beta}_{RE}\right] \quad (7)$$

$$\text{where } \psi = Var\left[b_{FE} - \hat{\beta}_{RE}\right] = Var\left[b_{FE}\right] - Var\left[\hat{\beta}_{RE}\right] \quad (8)$$

where $b_{FE}$ is the LSDV estimator for the FE panel model, and $\hat{\beta}_{RE}$ is the GLS estimator for the RE panel model. Greene (2002) notes that Hausman's implicit assumption for calculating $\psi$ is that the covariance of the difference in these estimators is zero and provides the proof behind Equation 8.

Hsiao (2003) argues that a FE model is more appropriate when the investigator only aims to make infer results for individuals in the sample, while the RE model is preferred for inferences relating to the larger population. However, which specification gets used depends more on whether there exist correlations between omitted variables and the included control variables. Both the FE and RE model forms are estimated here, and Hausman's test is applied to evaluate the possibility of error-term correlation with control variables.

## CRASH SEVERITY MODEL

To estimate the ordered response of crash severity, an ordered logistic regression was used. Three general crash categories exist here; these are property damage only (PDO), injury, and fatal crashes, which are labeled 1 through 3. Formally, an ordered logistic regression model's specification can be expressed as follows (McCullagh, 1980, 1989):

$$\text{logit}\left(\gamma_j\right) = \log\frac{\gamma_j}{1-\gamma_j} = \theta_j - \beta' x,\ j = 1,\text{K}\ ,k-1 \quad (9)$$

where $\gamma_j = Pr\left(Y \le j \mid x\right)$ is the cumulative probability up to and including category $j$, given $x$, $k$ is the number of categories for the response variable (3 in this case), and $\theta_j$ are the odds of the outcome if the clustered segment had a 0 for every variable in vector $x$, also known as the

thresholds. This model is also known as the proportional odds model, which is based on cumulative response probabilities rather than category probabilities. McCullagh (1980) proved that these two kinds of probabilities are equivalent, but that cumulative probabilities are more likely to work well with ordered data than the models based on the category probability in practice.

## MODEL RESULTS

### Crash Occurrence

Both FE and RE linear models were estimated for total crashes per million VMT (where VMT is the multiple of segment length and Washington DOT AADT estimates for each segment). Hausman test results suggest that correlation does exist between the RE models' random error terms and included variables, so the RE estimates are expected to be biased. F tests also were conducted, to compare the FE model results with their corresponding pooled models' estimates. These all indicate the FE models to be preferred, suggesting that the segment-specific effects cannot be ignored. The final estimation results for the FE model are shown in Table 4. The R-square goodness of fit statistic suggests that virtually 90% of the variation in crash rates is explained by the model's control variables. Segment-specific constant terms can be obtained using the Equation 5, as needed.

Table 4's results suggest that roadway design and speed limits play important roles. For example, crashes are more frequent on shorter horizontal curves and on roads with more lanes, which is consistent with Noland's (2003) results. Steeper sections also incur more crashes (per mile traveled), which is consistent with Milton and Mannering's findings (1998), while sections with wider medians and more traffic per lane experience lower crash rates.

Crash rates are predicted to be highest in 1995, which may be partly due to drivers' lowest compliance rates before the November 1995 repeal of the NMSL. There is no statistically significant difference among the other three years. Freedman and Esterlitz (1990) found that drivers' compliance rates with the 55 mi/h NMSL had been falling over time. Before the repeal of 55 mi/h NMSL in November 1995, it is plausible that a relatively heterogeneous pattern of chosen speeds existed, with many drivers exceeding and others obeying the 55 mi/h speed limit. This situation may have abated after speed limits were raised, with drivers tending to choose more similar travel speeds, thereby interacting less often and having fewer collisions in 1996.

Crash severity is predicted to exhibit a concave relationship with respect to speed limits. One can therefore estimate the speed limit that is expected to maximize the number of crashes per million VMT, based on a given segment's attributes. Here, the average roadway cluster has the following characteristics: horizontal curve length of 811 feet, no vertical grade, median width of 38 feet, 9 ft total shoulder width, 3 lanes, and 4,486 AADT per lane. Given these attributes, the maximum crash rate is predicted to occur at a speed limit of 43 mi/h, on the average roadway segment, which lies below the range of speed limit data used to calibrate the model (i.e., 50 to 70 mi/h). Thus, the model implies that crash rates on the average segment are strictly falling for any

reasonable speed limit choices. This finding is consistent with Milton and Mannering's (1998) study of Washington State crashes from 1992 through 1993. However, the speed limit cannot be increased beyond the standards of roadway geometric design, and other limitations such as restricted rights of way, urban designation. In practice, it is a trade-off high-speed design and the cost of constructing such roads. According to the standardized coefficients shown in Table 4, median width is predicted to be the most important factor affecting crash frequency, with a standardized coefficient of –5.3. This means that one standard deviation increase in median width (or a sizable 108 feet) is expected to result in crash rate drop of 5.3 standard deviations (or 5.3 times 1296 crashes per million VMT). Other key design factors are the number of lanes (with a standardized coefficient of 1.03) and shoulder width (-1.08). Speed limit and speed limit squared also carry very high standardized coefficients, reinforcing the power of this variable to predict crash rates.

## Crash Severity

An ordered logistic regression model was estimated using the pooled data after filtering the zero crash observations. The estimation results are shown in Table 5. The table only includes a final model, wherein control variables not exhibiting statistical significance at the 0.05 level have been removed, via a process of step-wise deletion (Greene, 2002). The descriptions of all variables can be found in Table 3.

Variables of every type were found to be informative in the final model. Crashes occurred on longer and sharper horizontal curve are found to be more severe. This may be because road environment is more likely affected by the long curve such as presence of limited sight distance while driving speeds remain high on such long but sharp curves. Drivers are probably not expected such limitations confronted by the long and sharp curves. The crash rate is predicted to increase as increases in the number of lanes. This might be attributed to more chances available on multilane roads for drivers to overtake each other, resulting in more interactions. Downhill driving is more deadly, most surely due to higher speeds and any needed braking having to overcome the added effects of gravity. Wider shoulder is predicted to incur more severe crashes, but wider median help alleviate the severity of crashes. The safety effects of speed limits exhibit a convex relationship in terms of crash severity. Weighting each crash rate by the predicted cost of such crashes[3], where each fatal crash is valued at $951,875, each injury crash at $296,275, and each PDO crash at $1663, the relationship between total crash cost per VMT and speed limit remains convex, and the optimal speed limit is predicted to be 70 mi/h for the average roadway segment. Using this same technique, one finds that speed limits of 70 mi/h are only optimal for roadways in the data set with the following design characteristics: no horizontal curvature, uphill slope, 1000-foot median, no shoulders, 7,524 AADT per lane, rolling terrain, and no access control. The highest optimal speed limit in the data set (65.7 mi/h) was found for roadway segments that had a 1941–foot horizontal curve of 2.6º curvature, -4.1% vertical grade, 215 feet median, 20 feet of total shoulder area, 4 lanes, 11,600 AADT per lane, and controlled access in

[3] In Washington State data set, the ratio of injury crashes to the number of persons injured is 1.642; for fatal crashes, this number is 1.144. The cost of each crash can be approximately calculated by multiplying the ratios with their corresponding costs shown in Table 6.

mountainous terrain. This sort of design is unlikely to be feasible for most locations because of restricted rights of way, travel demand, and construction costs.

A more realistic scenario is probably the following: 500-foot horizontal curve with 10-degree curvature, -5% vertical grade, 10-foot median width, 20-foot total (2-side) shoulder width, 4 lanes, 10,000 AADT per lane, controlled access, and mountainous terrain, in 1996. This situation is predicted to result in 4,994, 2,886, and 139 PDO, injury, and fatal crash counts per million VMT, respectively .

## CONCLUSIONS

Traffic crashes remain a major health problem for the U.S., as well as for other countries. Roadway design and speed limit policies are important determinants of crash outcomes. The models estimated here first employ cluster analysis, to create 812 groups of what were originally 59,500 homogeneous high-speed roadway segments throughout the State of Washington. These clustered data points then provide relatively continuous crash count and, therefore, crash rate data, permitting use of linear models and straightforward estimates of speed, use and design effects. The four-year panel data sets were analyzed using three model specifications. The FE models were preferred to both RE and pooled models, based on Hausman tests (for correlation between random effects and control variables) and F tests (for model fit), respectively. Additionally, a crash severity model was estimated using an ordered logistic regression. The models reveal that speed limit information is highly valuable in predicting both crash rates and crash severity. Given roadway design and use characteristics, the models predict optimal speed limits to minimize crash rates; these conflict, however, with the direction of impacts when examining crash severity as a function of speed limits. While higher limits are estimated to reduce crash rates on this data set of high-speed roads, they also are predicted to result in more severe crashes.

As expected, many design, use, and speed limits variables are highly statistically and practically significant. For roadways with average design and use attributes, speed limits at 44 mi/h are estimated to maximize total crash rates. An ordered logistic regression model was estimated to examine the effects of speed limits as well as various geometric design features on crash severity. Speed limit information was found to be highly predictive of crash rates, and the models are seemingly able to predict "optimal" speed limits in order to minimize crash rates and crash costs. However, it seems speed limits have biased coefficients, most likely due to unobserved safety-related effects. For the "average" high-speed segment in the data set, a minimum expected crash cost[4] is achieved at a speed limit of 70 mi/h (using Blincoe's [1994] crash costs estimates for NHTSA, as shown in Table 6), while the maximum crash rate is predicted to occur at a speed limit of 43.5 mi/h. While these calculations may not be realistic, the models accurately predict crash rates ($R^2$ of 0.90 for total crash count) and the results also provide useful information for a variety of design and use effects. For example, highly practically (and statistically) significant

[4] The optimal speed limit is calculated by minimizing the total cost incurred for all kinds of crashes. The number of crashes for different severity levels is obtained by multiplying the total number of crashes predicted using the crash occurrence model with the probability predicted by the ordered logistic model.

variables are horizontal curve length, median width, total shoulder width, speed limit, total number of lanes and AADT per lane – each estimated to have total-crash-rate elasticities over100% (absolute value) when evaluated at their averages.

There is room for improvement here. Many variables of interest are not available in the HSIS data set, but may be available for future analyses, including climate (e.g., annual rain and snowfall), pavement type and condition, police presence (for enforcement of speed limits and other roadway policies), truck use (as a fraction of AADT), proximity to a hospital (for life-saving treatments), average travel speeds and speed variance among vehicles (which also are functions of design, use and speed limit variables). A longer data panel would be useful. In order to improve estimator efficiency, a single likelihood function combining both crash occurrence and crash severity models is felt to be superior.

In sum, this work appears to be the first of its kind: clustering roadway segments in order to permit linear model calibration and estimating crash-rate- and crash-cost-minimizing speed limits. Such estimates should prove useful in the design of new roadways and speed limit policies on new and existing roadways.

## ACKNOWLEDGEMENTS

The authors thank Mr. Bob Howden and Mr. Christian Cheney of the Washington State DOT for generously sharing the updated speed limit information and loop detector data, and the National Cooperative Highway Research Program for funding this research under contract number 17-23. While the National Cooperative Highway Research Program (NCHRP) sponsored this research project, the opinions expressed here do not necessarily reflect the policy of NCHRP. The results of this work are preliminary and have not yet been approved by the NCHRP 17-23 project panel for publication as an NCHRP report. The authors also are grateful to the FHWA's Yusuf Mohamedshah for provision of the crash data sets, to Young-Jun Kweon and Xiaokun Wang for offering useful discussions related to the data sets and analytical methods, and to Ms. Annette Perrone for editorial assistance.

## REFERENCES


Allison, P. D. (1999) *Logistic Regression Using the SAS System: Theory and Application*. SAS Institute, Cary, N.C.

Blincoe, L.J. (1994) The Economic Cost of Motor Vehicle Crashes, 1994. NHTSA Technical Report. National Highway Transportation Safety Administration, Washington, D.C.

Chiu, T., Fang, D., Chen, J., Wang, Y., and Jeris, C. (2001) A Robust and Scalable Clustering Algorithm for Mixed Type Attributes in Large Database Environment. *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 263-268.

Everitt, B., Landau, S. and Leese, M. (2001) *Cluster Analysis, Fourth Edition*. Arnold, a member of the Hodder Headline Group, London.

Freedman, M., Esterlitz, J. R. (1990) Effects of the 65-mph Speed Limit on Speeds in Three States. *Transportation Research Record 1281*, 52-61.

Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. (2003) *Bayesian Data Analysis, Second Edition*. Cahpman & Hall, Florida.

Golob, T.F. and Recker, W.W. (2004) A Method for Relating Type of Crash to Traffic Flow Characteristics on Urban Freeways. *Transportation Research Part A*, 38(1), 53-80.

Greene, W.H. (2002) *Econometric Analysis, Fifth Edition*. Prentice Hall, New Jersey.

Gregersen, N.P. and Berg, H.Y. (1994) Lifestyle and Accidents Among Young Drivers. *Accident Analysis and Prevention*. 26(3), 297-303.

Guo, G. (1996) Negative Multinomial Regression Models for Clustered Event Counts. *Sociological Methodology*, 26, 113-132.

Hsiao, C. (2003) *Analysis of Panel Data*. Cambridge University Press, Cambridge.

Johnson, S.W. and Walker, J. (1996) The Crash Outcome Data Evaluation System (CODES). NHTSA Technical Report, DOT HS 808 338.

Karlaftis, M.G. and Tarko, A.P. (1998) Heterogeneity Considerations in Accident Modeling. *Accident Analysis and Prevention*, 30(4), 425-433.

Kweon, Y.J. and Kockelman, K. (2004) Spatially Disaggregate Panel Models of Crash and Injury Counts: The Effect of Speed Limit and Design. *Proceedings of the 83rd TRB Annual Meeting*. Washington D.C.

Le Blanc, L.A. and Rucks, C.T. (1996) A Multiple Discriminant Analysis of Vessel Accidents. *Accident Analysis and Prevention*, 28(4), 501-510.

McCarthy, P. S. (1999) Public Policy and Highway Safety: A City-wide Perspective. *Regional Science and Urban Economics*, 29, 231-244.

McCullagh, P. (1980) Regression Models for Ordinal Data. *Journal of the Royal Statistical Society B*, 42 (2), 109-142.

McCullagh, P. (1989) *Generalized Linear Models, Second Edition*. Chapman and Hall, Cambridge, U.K.

Miaou, S.P. and Lum, H., (1993) Modeling Vehicle Accidents and Highway Geometric Design Relationships. *Accident Analysis and Prevention*, 25(6), 689-709.

Milton, J. and Mannering, F. (1998) The Relationship Among Highway Geometrics, Traffic-Related Elements and Motor-Vehicle Accident Frequencies. *Transportation*, 25(4), 395-413.

Moellering, H. (1976) The Potential Uses of A Computer Animated Film in the Analysis of Geographical Patterns of Traffic Crashes. *Accident Analysis and Prevention*, 8(4), 215-277.

NHTSA (2003) *Traffic Safety Facts 2002*. Retrieved June 16, 2004, from National Center for Statistics and Analysis (NHTSA/USDOT) Web site: http://www-nrd.nhtsa.dot.gov/departments /nrd-30/ncsa/AvailInf.html

Noland, R. B. (2003) Traffic fatalities and injuries: the effect of changes in infrastructure and other trends. *Accident Analysis and Prevention*, 35(4), 599-611.

Poch, M. and Mannering, F. (1996) Negative Binomial Analysis of Intersection Accident Frequencies. *Journal of Transportation Engineering*, 122(2), 105-113.

Shankar, V., Mannering, F., and Barfield, W. (1995) Effect of Roadway Geometrics and Environmental Factors on Rural Freeway Accident Frequencies. *Accident Analysis and Prevention*, 27(3), 371-389.

Shankar,V., Milton, J. and Mannering F. (1997) Modeling Accident Frequencies as Zero-altered Probability Processes: An Empirical Inquiry. *Accident Analysis and Prevention*, 29(6), 829-837.

Sohn, S.Y. (1999) Quality Function Deployment Applied to Local Traffic Accident Reduction. *Accident Analysis and Prevention*, 31(6), 751-761.

SPSS Inc. (2001) The SPSS TwoStep Cluster Component. Retrieved June 19, 2004, from SPSS Inc.: Leading predictive analytics through analytical applications, data mining, text mining, market research, and statistical software. Web site: http://www.spss.com/.

Ulleberg, P. (2001) Personality Subtypes of Young Drivers: Relationship to Risk-taking Preferences, Accident Involvement, and Response to a Traffic Safety Campaign. *Transportation Research Part F*, 4(4), 279-297.

Wells-Parker, E. Cosby, P.J. and Landrum, J.W. (1986) A Typology for Drinking Driving Offenders: Methods for Classification and Policy Implications. *Accident Analysis and Prevention*, 18(6), 443-453.

Wong, S.C., Leung, B.S.Y., Loo, B.P.Y., Hung, W.T., and Lo, H.K. (2004) A Qualitative Assessment Methodology for Road Safety Policy Strategies. *Accident Analysis and Prevention*, 36(2), 281-293.

Zhang, T., Ramakrishnon, R., and Livny M. (1996) BIRCH: An Efficient Data Clustering Method for Very Large Databases. *Proceedings of the ACM SIGMOD Conference on Management of Data*, 103-114.

**List of Tables**

Table 1. Descriptive statistics of dependent variables after clustering
Table 2. Descriptive statistics of independent variables after clustering
Table 3. Descriptive statistics for variables used in the Ordered Logistic Regression
Table 4. Final crash occurrence model (linear fixed effects)
Table 5. Final crash severity model (ordered logistic regression)
Table 6. The cost experienced by one person for levels of injury crashes (in 1994 dollars)

**Table 1. Descriptive statistics of dependent variables after clustering**

| Variable | Obs | Mean | Std. Dev. | Min | Max |
|---|---|---|---|---|---|
| Number of persons injured | 3248 | 16.21 | 32.50 | 0 | 398 |
| Number of persons killed | 3248 | 0.293 | 0.894 | 0 | 13 |
| Number of PDO crashes | 3248 | 11.55 | 22.88 | 0 | 270 |
| Number of injury crashes | 3248 | 9.906 | 19.75 | 0 | 240 |
| Number of fatal crashes | 3248 | 0.254 | 0.764 | 0 | 12 |
| Number of total crashes | 3248 | 21.71 | 42.66 | 0 | 512 |
| Number of persons injured per million VMT | 3248 | 444.9 | 1430 | 0 | 30915 |
| Number of persons killed per million VMT | 3248 | 26.81 | 493.6 | 0 | 14008 |
| Number of PDO crashes per million VMT | 3248 | 333.6 | 893.5 | 0 | 18300 |
| Number of injury crashes per million VMT | 3248 | 254.5 | 581.5 | 0 | 11009 |
| Number of fatal crashes per million VMT | 3248 | 14.15 | 170.2 | 0 | 4669 |
| Number of total crashes per million VMT | 3248 | 602.2 | 1296 | 0 | 21960 |
| Aggregated roadway section length (miles) | 3248 | 6.611 | 12.78 | 0.008 | 134 |
| Vehicle miles traveled (VMT) | 3248 | 69793 | 168428 | 9.829 | 2061947 |

**Table 2. Descriptive statistics of independent variables after clustering**

| Variable | Obs | Mean | Std. Dev. | Min | Max |
|---|---|---|---|---|---|
| Horizontal curve length (feet) | 3248 | 810.9 | 1310 | 0 | 12683 |
| Degree of curvature (degrees per 100 ft arc) | 3248 | 4.789 | 14.56 | 0 | 164 |
| Vertical grade (%) | 3248 | 0.071 | 3.539 | -24 | 47 |
| Median width (feet) | 3248 | 37.71 | 108.0 | 0 | 999 |
| Total shoulder width (feet) | 3248 | 9.039 | 7.556 | 0 | 50 |
| Speed limit (mi/h) | 3248 | 55.09 | 4.011 | 50 | 66 |
| Squared speed limit ($mi^2/h^2$) | 3248 | 3050 | 467.1 | 2500 | 4312 |
| Total number of lanes | 3248 | 3.075 | 1.604 | 1 | 9 |
| AADT per lane | 3248 | 4486 | 5126 | 80 | 23861 |
| Indicator for an interstate highway | 3248 | 0.229 | 0.420 | 0 | 1 |
| Indicator for an access controlled segment | 3248 | 0.446 | 0.497 | 0 | 1 |
| Indicator for level terrain | 3248 | 0.219 | 0.414 | 0 | 1 |
| Indicator for rolling terrain | 3248 | 0.621 | 0.485 | 0 | 1 |
| Indicator for year 1994 | 3248 | 0.250 | 0.433 | 0 | 1 |
| Indicator for year 1995 | 3248 | 0.250 | 0.433 | 0 | 1 |
| Indicator for year 1996 | 3248 | 0.250 | 0.433 | 0 | 1 |

**Table 3. Descriptive statistics for variables used in the Ordered Logistic Regression**

| Variable | Obs.* | Mean | Std. Dev. | Min. | Max. |
|---|---|---|---|---|---|
| Number of PDO crashes | 5873 | 15.21 | 26.70 | 0 | 270 |
| Number of injury crashes | 5873 | 13.15 | 23.10 | 0 | 240 |
| Number of fatal crashes | 5873 | .410 | .945 | 0 | 12 |
| Number of total crashes | 5873 | 28.77 | 49.89 | 1 | 512 |
| Number of PDO crashes per million VMT | 5873 | 355.6 | 892.0 | 0 | 18300 |
| Number of injury crashes per million VMT | 5873 | 280.3 | 574.3 | 0 | 11009 |
| Number of fatal crashes per million VMT | 5873 | 19.57 | 181.7 | 0 | 4669 |
| Number of total crashes per million VMT | 5873 | 655.4 | 1319 | 31 | 21960 |
| Aggregated roadway section length (miles) | 5873 | 8.620 | 15.18 | .02 | 134 |
| Vehicle miles traveled (VMT) | 5873 | 92590 | 200528 | 162 | 2061947 |
| Horizontal curve length (feet) | 5873 | 776.0 | 1250 | .00 | 12683 |
| Degree of curvature | 5873 | 2.399 | 4.886 | .00 | 52.15 |
| Vertical grade (%) | 5873 | .028 | 2.608 | -7 | 7 |
| Median width (feet) | 5873 | 35.19 | 99.38 | 0 | 999 |
| Total shoulder width (feet) | 5873 | 9.523 | 7.367 | 0 | 43 |
| Speed limit (mi/h) | 5873 | 55.36 | 3.971 | 50 | 66 |
| Squared speed limit ($mi^2/h^2$) | 5873 | 3080 | 465.0 | 2500 | 4312 |
| Indicator for access controlled segments | 5873 | .460 | .499 | 0 | 1 |
| Indicator for interstate highway | 5873 | .250 | .433 | 0 | 1 |
| Indicator for level terrain | 5873 | .210 | .410 | 0 | 1 |
| Indicator for rolling terrain | 5873 | .670 | .469 | 0 | 1 |
| Total number of lanes | 5873 | 3.190 | 1.660 | 1 | 9 |
| AADT per lane | 5873 | 4863 | 5270 | 195 | 23861 |

*Note: Initially, the number of observations for 4 years was 3248. To estimate the ordered logit model, each observation in the original data has been split into three, corresponding to three categories of severity level, resulting in 9,744 observations for model estimation. However, zero crash observation likelihoods are not a function of model parameters (Allison, 1999), so those observations have been filtered, resulting in 5,873 observations for model calibration.

**Table 4. Final crash occurrence model (linear fixed effects)**

| Variable | Coef. | Std. Err. | Stdzd. Coef. | P-Value |
|---|---|---|---|---|
| Horizontal curve length (feet) | -0.898 | 0.0803 | -0.908 | 0.0000 |
| Vertical grade (%) | 273.1 | 7.819 | 0.746 | 0.000 |
| Median width (feet) | -63.74 | 0.953 | -5.313 | 0.000 |
| Total shoulder width (feet) | 176.3 | 5.502 | 1.028 | 0.000 |
| **Speed limit** | 233.5 | 70.88 | 0.722 | 0.001 |
| **Squared speed limit** | -2.687 | 0.610 | -0.968 | 0.000 |
| Total number of lanes | 1172 | 31.42 | 1.450 | 0.000 |
| AADT per lane | -0.157 | 0.0067 | -0.620 | 0.000 |
| Indicator for year 1995 | 0.443 | 0.252 | 0.000 | 0.079 |
| Constant | -5445 | 2053 | -0.908 | 0.008 |
| Number of observations | | | | 3248 |
| Number of non zero observations | | | | 2904 |
| R-squared | | | | 0.898 |

Note: The dependent variable is #crashes per million vehicle miles traveled.

**Table 5. Final crash severity model (ordered logistic regression)**

| Variable | Coef. | Std. Err. | P-Value |
|---|---|---|---|
| $\theta_1$ | -8.204 | 0.328 | 0.000 |
| $\theta_2$ | -4.667 | 0.328 | 0.000 |
| Horizontal curve length (feet) | 3.00E-05 | 1.66E-06 | 0.000 |
| Degree of curvature | -1.90E-02 | 1.38E-04 | 0.000 |
| Vertical grade (%) | -1.58E-02 | 4.25E-04 | 0.000 |
| Median width (feet) | -1.30E-05 | 2.30E-05 | 0.000 |
| Total shoulder width (feet) | 7.20E-03 | 2.98E-04 | 0.000 |
| **Speed limit** | -0.234 | 1.14E-02 | 0.000 |
| **Squared speed limit** | 1.66E-03 | 9.80E-05 | 0.000 |
| Indicator for access controlled segments | 5.81E-02 | 4.55E-03 | 0.000 |
| Indicator for level terrain | -0.2318 | 4.77E-03 | 0.000 |
| Indicator for rolling terrain | -0.3088 | 3.88E-03 | 0.000 |
| Total number of lanes | -9.14E-02 | 1.99E-03 | 0.000 |
| AADT per lane | -4.80E-05 | 1.39E-06 | 0.000 |
| Squared AADT per lane | 3.19E-09 | 6.94E-11 | 0.000 |
| Indicator for year 1994 | -3.54E-02 | 4.04E-03 | 0.000 |
| Indicator for year 1995 | -4.03E-02 | 3.99E-03 | 0.000 |
| Indicator for year 1996 | -2.52E-02 | 3.99E-03 | 0.000 |
| LogLik value at constant | | | -1,600,403 |
| LogLik value (full model) | | | -1,579,629 |
| Number of Observations | | | 5,873 |

**Table 6. Injury Costs (in 1994 dollars)** (Blincoe, 1994)

| NHTSA | PDO | MAIS 0 | MAIS 1 | MAIS 2 | MAIS 3 | MAIS 4 | MAIS 5 | Fatal |
|---|---|---|---|---|---|---|---|---|
| | $1,663 | $1,129 | $7,243 | $34,723 | $103,985 | $230,042 | $705,754 | $831,919 |
| This Work | PDO | Injury | | | | | | Fatal |
| | $1,663 | $180,479 | | | | | | $831,919 |

Notes: The final Injury value of $180,479 is obtained by averaging MAIS 1, MAIS 2, MAIS 3, and MAIS 4 values