Predicting Local Road Crashes Using Socio-economic and Land Cover Data

Kai Wang*
Graduate Research Assistant, Department of Civil and Environmental Engineering
University of Connecticut
261 Glenbrook Road, Unit 3037, Storrs, CT 06269-3037, USA
Phone: +1 (860) 486-0586
Email: kai.wang@uconn.edu

John N. Ivan, Ph.D., P.E.
Professor, Department of Civil and Environmental Engineering
University of Connecticut
261 Glenbrook Road, Unit 3037, Storrs, CT 06269-3037, USA
Phone: +1 (860) 486-0352
Fax: +1 (860) 486-2298
Email: john.ivan@uconn.edu

Amy C. Burnicki, Ph.D.
Assistant Professor in Residence, Department of Civil and Environmental Engineering
University of Connecticut
261 Glenbrook Road, Unit 3037, Storrs, CT 06269-3037, USA
Phone: +1 (860) 486-2340
Email: amy@engr.uconn.edu

Sha A. Mamun, Ph.D.
Post-Doctoral Researcher, Department of Civil and Environmental Engineering
University of Connecticut
261 Glenbrook Road, Unit 3037, Storrs, CT 06269-3037, USA
Phone: +1 (860) 486-0586
Email: msm08014@engr.uconn.edu

*Corresponding author

Submitted for presentation and publication at the 95th Annual Meeting of the Transportation Research Board (TRB)

Committee Number: ANB25
TRB Committee on Highway Safety Performance
ABSTRACT

Estimating and applying safety performance functions (SPFs), or models for predicting expected crash counts, for roads under local jurisdiction is often challenging due to the lack of vehicle count data to be used for exposure, which is a critical variable in such functions. This paper describes estimation of SPFs for local road intersections and segments in Connecticut using socio-economic and network topological data instead of traffic counts as exposure. SPFs are developed at the traffic analysis zone (TAZ) level, where the TAZs are categorized into six homogeneous clusters based on land cover intensities and population density. SPFs were estimated for each cluster to predict the number of intersection and segment crashes occurring in each TAZ. One aggregate SPF using the entire dataset was also estimated to compare with the individual cluster SPFs. The number of intersections and the total local roadway length were also used as exposure in the intersection and segment SPFs, respectively. Total population, retail and non-retail employment and average household income are found to be significant variables. Ten percent of the observed data points were reserved for out of sample testing and in all cases, these out of sample predictions were as good as the in sample predictions.

Key Words: safety performance function, crash count, local road, cluster analysis
INTRODUCTION

A Safety Performance Function (SPF) is an equation used to predict crash counts at a location as a function of exposure and other roadway characteristics (e.g. number of lanes, lane width, shoulder width) \(^{(1)}\). One of the uses for SPFs is estimating the expected number of crashes on traffic facilities to identify road locations with higher crash potential for safety improvements, select and implement cost-effective countermeasures to reduce future crashes \(^{(2)}\). SPFs are often developed for different traffic facilities such as road segments and intersections. Local roads owned and operated by local entities including towns, counties and tribal governments play an important role in the roadway network, as approximately 60 percent of all road miles in the U.S. are maintained by these jurisdictions \(^{(3)}\). A recent Iowa study \(^{(4)}\) reported that local roads had higher crash rates compared to primary roads under State jurisdiction and the reported local road crash rate was 1.5 times higher than that of primary roads from 1974 to 2000. As a result, traffic safety on local roads is important to both traffic safety organizations and engineers. Given this situation, it is important to develop accurate tools to predict the number of crashes occurred on local roads to support identifying sites with promise for safety improvements, selecting and implementing effective countermeasures to reduce future crash volume or severity.

The Highway Safety Manual (HSM) \(^{(1)}\) provides SPFs for two lane rural highways, multilane rural highways, urban and suburban arterials, freeways and freeway ramp junctions. The SPFs in HSM were estimated using data collected from a limited number of States in the USA, including Washington, California, Minnesota, Texas, Michigan, North Carolina and Illinois. Because crash relationships in these states are not necessarily representative of those in the entire country, the HSM recommends a calibration procedure to adjust the predicted crash counts for individual jurisdiction in using the prediction from the SPF. The HSM SPFs include traffic counts for intersections or roadway segments as the most critical variables in accurately predicting the number of crashes \(^{(1, 5, 6)}\). This presents a problem for roads under local jurisdiction, where traffic counts are generally not available because it is economically impractical to implement traffic counting programs for so many facilities on which the traffic volume is typically below 400 per day \(^{(4)}\). In order to implement highway safety improvement strategies on these low volume local roads, new crash prediction approaches are desirable, in which the traffic counts are not required.

The objective of this study was to estimate SPFs for both intersections and segments on roads under local jurisdiction in the State of Connecticut using demographic data as a replacement for traffic count data. The SPFs are estimated at the level of Traffic Analysis Zone (TAZ), instead of the intersection or roadway segment level. The intersection counts \((i.e.\) the number of intersections in a TAZ) and segment mileage \((i.e.\) total local roadway length in a TAZ) are used as exposure in this study in lieu of traffic volume. Demographic records such as population, total retail and non-retail employment, household income and vehicle availability work in tandem with the exposure to predict the estimated crash counts. To account for data and crash relationship heterogeneity, the TAZs in the entire state are categorized into six clusters based on the percentage of three land cover categories – high, medium and low intensities – and the population density \((i.e.\) the number of population per km\(^{2}\))\. A different SPF was estimated for each cluster, and the similarities and differences among these functions are discussed. We also discuss how to apply the functions as a network screening tool.
LITERATURE REVIEW

SPFs have been estimated for local roads by various researchers at two levels: the facility level (e.g. roadway segment and intersection) and the zonal level (e.g. TAZ). Among facility level models, Vogt (6) provides a good review of the factors associated with crashes on local roads according to past research studies. These include channelization (right and left turn lane), number of driveways, sight distance, intersection angle, median width, surface width, shoulder width, signal characteristics, lighting, roadside condition, truck percentage in the traffic volume, posted speed, and weather. Most research on two-lane roads confirms traffic volume as the major explanatory factor for traffic crashes, which is unfortunate for the cases where the traffic volume is not available (7, 8). There is little literature on investigating alternative exposure measures in addition to or in place of traffic volume for predicting crashes. Bindra et al. (9) considered the use of geographic information system (GIS) land use inventories to supplement traffic volumes as exposure for estimating SPFs for predicting segment-intersections crashes for rural two-lane and urban two-and four-lane undivided roads. They concluded that the number of trips generated and the land use data (i.e., population, retail and non-retail employment, and driveway data) were good predictors for estimating segment-intersection crashes, that is, crashes on segments located at minor roads and driveways without traffic counts.

Zonal SPFs (ZSPFs), of which the most popular is TAZ level, make use of highly available zonal-level variables (10). Among the studies focusing on developing TAZ-level SPFs, Pulugurtha et al. (11) used socioeconomic and network variables to develop TAZ level SPFs to estimate the crash counts by severity level (injury and property damage only crashes). Ladron de Guevara et al. (12), Lovegrove and Sayed (13), Lovegrove (14) and Hadayeghi et al. (15) developed TAZ level SPFs to estimate the number of both intersection and segment crashes. Factors such as population density, the number of employees and the intersection density were considered as predictors for the number of crashes. Furthermore, Khondakar et al. (16) found that TAZ level SPFs can safely be transferred both temporally and spatially. Noland and Quddus (17) showed that TAZs with high employment density had more traffic crashes, whereas in urbanized areas with more densely populated TAZs fewer crashes were observed. Jin et al. (18) identified that besides traditional variables such as segment length, structure of roadway network should be considered in developing TAZ-level SPFs to improve prediction accuracy. Several studies developed TAZ-level SPFs using number of trips generated inside of each TAZ. Naderan and Shahi (19), Abdel-Aty et al. (20) found that number of trips generated have significant impacts on TAZ-level crashes.

Recently, an analysis tool (PLANSAFE) was developed on a National Cooperative Highway Research Program (NCHRP) project (21) to predict the expected crash counts by TAZ. The predictors include population, employment and some land use intensity variables. The purpose was to use the predicted crash counts as one of the measures of effectiveness to select the most cost-effective transportation improvement plan. Another study of TAZ level SPFs by Pirdavani et al. (10) considered establishing an association between observed crashes and a set of predictor variables in each TAZ. The study compared models using two different exposures - VHT (total daily vehicle hours traveled) and VKT (total daily vehicle kilometers traveled) along with network and socio-demographic variables. The results show that the model containing the combination of two exposures outperformed the models containing only one of the exposure
variables. Except for TAZ-level SPFs, some studies have investigated SPFs on other macroscopic levels, such as block group (22, 23) and county level (24, 25, 26).

Although these zonal level SPFs are able to estimate crash counts without traffic volume, most of them were designed to estimate the number of crashes using network and social-demographic variables etc., without accounting for the data and crash heterogeneity among different types of TAZs or zones. To address this issue, our study focuses on estimating TAZ level SPFs for local roads by different TAZ type. The TAZs were clustered into different categories using a data mining technology (i.e. K-means clustering analysis), based on their land-use intensities and population density. Socio-demographic data and roadway network data such as population, employment, income, car ownership, number of local jurisdiction road intersections and total local road length inside the TAZ are used to predict crash counts. The intention is for some of the variables to serve in lieu of actual traffic counts which are generally not available for these roads.

The remainder of the paper is organized as follows. The next section presents the methodology and the process of data collection. The third section describes the estimation of SPFs and the results. The final section discusses how to use the estimated SPFs as a network screening tool.

**METHODOLOGY AND DATA PREPARATION**

Our procedure for the estimation of TAZ level SPFs for local roads requires four types of data at the TAZ level: roadway network shape features, demographic records, geographic/land cover features and crash records. Below are a brief description of the required data and data sources.

*Roadway Network Shape Features*

The number of intersections and the total length of roadway under local jurisdiction were extracted from the 2010 Census TIGER/LINE files for Connecticut (27). The original TIGER/LINE files contained some errors, such as typos for roadway name and discrepancies in the network representation of some road links. The network links were carefully checked and the records were revised accordingly. The number of intersections and the total length of roadways under local jurisdiction were calculated for each TAZ. Details about our procedures for calculating the number of intersections and the total length of roadways are provided in the Appendix to the project final report (28).

*TAZ Level Demographic Records*

TAZ level demographic records were collected from the Census Transportation Planning Package Database (CTPP, 2010) (29). They include population, retail and non-retail employment, households, vehicles and average household income summarized by TAZ and used as the independent variables in safety performance functions. In the 2010 census, 1806 TAZs were defined for the state of Connecticut. Two of these TAZs were apparently defined to represent special generators and have no population or employment, so they were eliminated from the analysis. The remaining 1804 TAZs were used to estimate the SPFs.

*TAZ Level Geographic/Land Cover Features*

Land-cover information was collected from the 2011 National Land Cover Database (NLCD) (30). We calculated the proportion of land area in three developed land-use categories – low, medium and high intensity development. These values along with the population density were
used to categorize the TAZs into homogeneous groups using K-means clustering analysis (discussed in the next section). Originally we used only the land cover intensities, but we found that adding the population density helped to correct aberrant cluster assignments for unique development sites (e.g., airports).

**Crash Records and Integration of Crash to TAZ**

Intersection and segment crash records were collected from the Connecticut Crash Data Repository (CTCDR) (31). As more severe crashes lead to more serious consequence and generate more interest (particularly among the members of the steering committee for this project), and the underreporting issue of property damage only (PDO) crashes, only K (fatal injury), A (incapacitating injury) and B (non-incapacitating injury) intersection and segment crashes occurring on roads under local jurisdiction in Connecticut from 2010 to 2012 were considered. In total, 5403 intersection crashes and 5502 segment crashes were extracted.

Intersection and segment crashes were assigned to TAZs based on their locations. If the crash was located inside the boundary of a single TAZ, the crash was assigned to this TAZ. If the crash was located on the boundary of more than one TAZ, it was evenly assigned among the TAZs. Details about our procedures for assigning crashes are provided in the Appendix to the project final report (28).

**Clustering of TAZs**

K-means clustering analysis (32) was used to categorize the TAZs into homogeneous groups using the three land cover intensities and the population density. K-means clustering analysis categorizes data by maximizing the variation among clusters while minimizing the variation within each cluster (33, 34). Different numbers of clusters were respectively tested, and the Calinski and Harabase pseudo-F index (35) was used to select the final number of clusters. The larger the Calinski and Harabase pseudo-F index, the more accurate is the clustering analysis.

The optimum number of clusters was found to be six. Figure 1(a) shows the distributions of the three land-use intensities and the population density among the six clusters. The overall land-use intensity and the population density decrease from cluster 1 to cluster 6. The number of TAZs assigned into cluster 1 through cluster 6 is 80, 161, 270, 284, 382 and 627, respectively. Figure 1(b) shows the distribution of the six clusters across the state. Note that two TAZs with legend 0 in the western and southeastern areas were eliminated in estimating the safety performance functions, as these two TAZs have no population. Cluster 1 has the lowest number of TAZs, which is generally urbanized area, and cluster 6 is the most common cluster type and is generally rural in nature. The areas with higher land-use intensities (red and orange on the map) are mainly located in the central and southern parts of the state.
Figure 2 illustrates the distribution of KAB crashes by cluster. Comparing the two types of crashes, there are substantially more intersection crashes than segment crashes in clusters 1, 2 and 3, but fewer intersection crashes than segment crashes in clusters 5 and 6. The two types of
crashes have nearly the same distributions in cluster 4. Figure 3 and Figure 4 display the distributions of the number of intersections, local roadway mileage and demographic variables by cluster. The number of intersections increases from cluster 1 to cluster 5, and then decreases to cluster 6. The roadway mileage increases consistently from cluster 1 to cluster 6. The average household income slightly increases from cluster 1 to cluster 6. Cluster 1 has the highest average numbers for both retail and non-retail employment, and cluster 6 has the lowest numbers. One important finding is that the distribution patterns are similar among population (Figure 3(c)), households (Figure 3(d)) and vehicles (Figure 4(a)). This is caused by the high correlation among these three factors, which was also verified by a correlation test. The selection and application of these three correlated variables will be discussed under SPF development.

![FIGURE 3 Distributions of Independent Variables by Cluster](image-url)
Safety performance functions were estimated to predict the number of intersection and segment crashes in each TAZ. The number of crashes is estimated by count regression models, such as the Poisson regression model, formulated as (36):

\[
Prob[y_i|\mu_i] = \frac{\exp(-\mu_i)\mu_i^{y_i}}{y_i!}
\]  

(1)

where \(Prob[y_i|\mu_i]\) is the probability of \(y\) crashes occurring at TAZ \(i\) and \(\mu_i\) is the expected number of crashes at TAZ \(i\). Given a vector of covariates \(X_i\), which describes the demographic and roadway characteristics of a TAZ \(i\), and a vector of estimable coefficients \(\beta\), the \(\mu_i\) can be estimated by the equation:

\[
\ln(\mu_i) = \beta X_i
\]  

(2)
The limitation of the Poisson model is that the variance of the data is constrained to be equal to the mean, *i.e.*:

\[ \text{Var}(y_i) = E(y_i) = \mu_i \]  

This constraint might be questionable as the variance of crash data is usually greater than the mean, which is known as over-dispersion \((36)\). The negative binomial regression model addresses this issue, which is derived by rewriting Equation 2 such that:

\[ \mu_i = \exp(\mathbf{\beta} \mathbf{X}_i + \varepsilon_i) \]  

where \(\exp(\varepsilon_i)\) is an error term assumed to follow a gamma distribution with mean 1 and variance \(\sigma^2\). The distribution of the negative binomial model has the form \((36)\):

\[
\text{Prob}[y_i | \mu_i] = \frac{\Gamma \left[ \left( \frac{1}{\sigma} \right) + y_i \right] \left[ \frac{1}{\sigma} \right]^{\frac{1}{\sigma}} \left[ \frac{1}{\sigma} \left( \frac{1}{\sigma} + \mu_i \right) \right]^{\mu_i}}{\Gamma \left( \frac{1}{\sigma} \right) y_i!} \left[ \frac{1}{\sigma} + \mu_i \right] \]  

where \(\Gamma\) is a gamma function; the variance of the negative binomial model can be written as follows:

\[ \text{Var}(y_i) = \mu_i (1 + \sigma \mu_i) = \mu_i + \sigma \mu_i^2 \]  

We define the function for the predicted intersection crashes at TAZ \(i\) as follows:

\[ \mu_{\text{int},i} = Y I_i \exp(\mathbf{\beta} \mathbf{X}_i + \varepsilon_i) \]  

Where

- \(\mu_{\text{int},i}\) = predicted intersection crashes in TAZ \(i\)
- \(Y\) = the number of years in the time period
- \(I_i\) = the number of intersections in TAZ \(i\)
- \(P_i\) = the population of TAZ \(i\)
- \(R_i\) = the total retail employment of TAZ \(i\)
- \(N_i\) = the total non-retail employment of TAZ \(i\)
- \(V_i\) = the number of vehicles in TAZ \(i\)
- \(C_i\) = the average income in TAZ \(i\)
- \(H_i\) = the number of households in TAZ \(i\)
- \(\beta s\) = the estimated parameters

We define the function for the predicted segment crashes at TAZ \(i\) as follows:

\[ \mu_{\text{seg},i} = Y L_i \exp(\mathbf{\beta} \mathbf{X}_i + \varepsilon_i) \]  

Where

- \(\mu_{\text{seg},i}\) = predicted segment crashes in TAZ \(i\)
- \(L_i\) = the number of segments in TAZ \(i\)
\[ \mu_{seg,i} = \text{predicted segment crashes in TAZ } i \]
\[ L_i = \text{the mileage of roadways under local jurisdiction in TAZ } i \]

and the remaining variables are as defined above.

**VARIABLE SELECTION AND SPF RESULTS**

The SPFs were estimated at the TAZ level for each cluster type. One statewide SPF using the aggregate data (*i.e.*, for all TAZ’s without splitting by cluster) was also estimated for comparison purposes. When estimating each function, the crash records were randomly divided into two parts: one part including ninety percent of the observations was used to estimate the function; and the other part including ten percent of the observations was used to evaluate the function prediction performance. Three functions, each using one of the correlated independent variables at a time (population, number of households and number of vehicles), were estimated for both intersection and segment crashes. These three functions were compared according to the model goodness-of-fit (Akaike Information Criterion-AIC and Bayesian Information Criterion-BIC). The number of crashes was predicted using both estimation and prediction datasets for the entire state using the cluster-based functions and the statewide function to test the efficacy of each approach. Function performance was compared using two measures of effectiveness (MOEs), Mean Absolute Deviation (MAD) and Mean Squared Predictor Error (MSPE), proposed by Oh *et al*. (37). These criteria are calculated as:

AIC = \(2K - 2 \ln(LL)\)  

BIC = \(K \times \ln(N) - 2\ln(LL)\)  

Mean Absolute Deviation (MAD) = \(\frac{1}{N} \sum_{i=1}^{N} |\hat{y}_i - y_i|\)  

Mean Squared Predictor Error (MSPE) = \(\frac{1}{N} \sum_{i=1}^{N} (\hat{y}_i - y_i)^2\)  

Where  
\(K\) = the number of estimated parameters  
\(LL\) = the maximized value of model likelihood function  
\(N\) = the number of observations  
\(\hat{y}_i\) = the predicted number of crashes at TAZ \(i\)  
\(y_i\) = the observed number of crashes at TAZ \(i\)

The smaller the AIC, BIC, MAD or MSPE value, the better is the function performance. Table 1 shows the goodness-of-fit of the cluster based SPFs and Statewide SPFs including one of the correlated variables at a time. Due to the poorer performance of the function using the number of vehicles, only the functions including population or the number of households are presented here. For the statewide SPF, both intersection and segment SPFs have lower AIC and BIC values using population than using households. For the intersection SPF, the function for clusters 2, 3 and 4 have a lower AIC or BIC value using population as an independent variable than that using
the number of households, while the reverse is observed for clusters 1, 5 and 6. The segment SPFs for all clusters have lower AIC and BIC values using population than using households.

Table 1 Goodness-of-fit of the Cluster Based SPF

<table>
<thead>
<tr>
<th>Cluster SPF</th>
<th>Intersection SPF</th>
<th>Segment SPF</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Population SPF</td>
<td>Households SPF</td>
</tr>
<tr>
<td></td>
<td>AIC</td>
<td>BIC</td>
</tr>
<tr>
<td>1</td>
<td>432</td>
<td>448</td>
</tr>
<tr>
<td>2</td>
<td>887</td>
<td>908</td>
</tr>
<tr>
<td>3</td>
<td>1,231</td>
<td>1,256</td>
</tr>
<tr>
<td>4</td>
<td>1,110</td>
<td>1,135</td>
</tr>
<tr>
<td>5</td>
<td>1,220</td>
<td>1,247</td>
</tr>
<tr>
<td>6</td>
<td>1,247</td>
<td>1,278</td>
</tr>
<tr>
<td>Statewide SPF</td>
<td>6,935</td>
<td>6,972</td>
</tr>
</tbody>
</table>

Table 2 displays the SPF prediction performance for the statewide and cluster-based functions using both estimation data and prediction data. The cluster-based SPFs using either population or households outperform the statewide SPF in crash prediction, as they have a lower MAD or MSPE value for both estimation data and prediction data. This is to be expected, as it has the possibility of accounting for heterogeneity related to land cover intensity. Furthermore, comparing the cluster-based SPF including population with the one including the number of households, the cluster-based SPF with population slightly outperforms the one with the number of households. Additionally, it seems that the SPF performance using the prediction data are even better than those using the estimation data. This may be due to the smaller size of the prediction data set, but it also demonstrated that there is no over-fitting to the estimation data, and that the functions are transferable within Connecticut. Therefore, considering all of these MOEs (model fit and prediction), the cluster-based SPFs with population were selected.

Table 2 SPF Prediction Performance

<table>
<thead>
<tr>
<th>MOEs</th>
<th>Intersection SPF</th>
<th>Segment SPF</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Statewide SPF</td>
<td>Cluster-based SPF</td>
</tr>
<tr>
<td></td>
<td>(Population)</td>
<td>(Population)</td>
</tr>
<tr>
<td></td>
<td>(Households)</td>
<td>(Households)</td>
</tr>
<tr>
<td>MAD Estimation</td>
<td>2.65</td>
<td>1.95</td>
</tr>
<tr>
<td>MAD Prediction</td>
<td>2.65</td>
<td>1.62</td>
</tr>
<tr>
<td>MSPE Estimation</td>
<td>18.25</td>
<td>11.14</td>
</tr>
<tr>
<td>MSPE Prediction</td>
<td>13.29</td>
<td>6.41</td>
</tr>
<tr>
<td>MAD Estimation</td>
<td>2.00</td>
<td>1.77</td>
</tr>
<tr>
<td>MAD Prediction</td>
<td>1.52</td>
<td>1.30</td>
</tr>
<tr>
<td>MSPE Estimation</td>
<td>8.28</td>
<td>7.55</td>
</tr>
<tr>
<td>MSPE Prediction</td>
<td>4.00</td>
<td>3.51</td>
</tr>
</tbody>
</table>

Table 3 shows the coefficient estimates for the intersection SPFs using population as a predictor. Coefficients for all models are provided in the Appendix to the Final Report (28). The first row
in each table cell is the coefficient, the second row is the p-significance, and coefficients shown in bold are statistically significant with 95% confidence. With respect to the six cluster-based functions, the number of intersections (exposure surrogate for intersection SPFs) was not statistically significant in the cluster 2, 3 and 4 functions. The effect of total population on number of intersection crashes is shown to be positive in all functions (as expected), except for clusters 5 and 6, in which it was not statistically significant. The amount of retail employment is positively associated with the number of intersection crashes in the functions for cluster 4, 5 and 6. The amount of non-retail employment is positively associated with the number of intersection crashes for cluster 1, 2 and 6. The number of intersection crashes decreases with the increase of average household income in the first five cluster functions, but increases in the cluster 6 function.

Table 3 Coefficient Estimates for KAB Intersection Crashes

<table>
<thead>
<tr>
<th>Variables</th>
<th>Coefficient Estimates by Cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Intercept</td>
<td>-1.275 (0.001)</td>
</tr>
<tr>
<td>Log (number of intersections)</td>
<td>0.682 (0.000)</td>
</tr>
<tr>
<td>Population (*1000)</td>
<td>0.161 (0.014)</td>
</tr>
<tr>
<td>Retail employment (*1000)</td>
<td>0.196 (0.530)</td>
</tr>
<tr>
<td>Non-retail employment (*1000)</td>
<td>0.090 (0.003)</td>
</tr>
<tr>
<td>Average household income</td>
<td>-0.005 (0.067)</td>
</tr>
<tr>
<td>Overdispersion</td>
<td>0.258 (0.000)</td>
</tr>
</tbody>
</table>

Notes: first row is the coefficient, second row is the p-significance, and bold coefficients are statistically significant at 5% level of significance.

Table 4 shows the coefficient estimates for the segment SPFs. Similar to the intersection SPFs, the association between the exposure surrogate, i.e. local roadway length and the number of segment crashes, is positive in all six functions, but is only statistically significant in clusters 1, 5 and 6. The coefficient for population is positive and significant in all six cluster-based functions. The retail employment is statistically significant in clusters 3, 4 and 5, and the non-retail employment is statistically significant in clusters 1, 2 and 3. The number of segment crashes decreases with the increase of average household income in the first five cluster functions, but increases in cluster 6 function, which is consistent with the intersection SPFs.

Table 4 Coefficient Estimates for KAB Segment Crashes

<table>
<thead>
<tr>
<th>Variables</th>
<th>Coefficient Estimates by Cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Intercept</td>
<td>-3.648 (0.000)</td>
</tr>
</tbody>
</table>
Log (roadway length in miles) | (0.008) | (0.213) | (0.305) | (0.265) | (0.000) | (0.000)  
Population (*1000) | 0.403 | 0.248 | 0.160 | 0.100 | 0.539 | 0.504  
(0.020) | (0.161) | (0.297) | (0.552) | (0.000) | (0.000)  
Retail employment (*1000) | 0.166 | 0.188 | 0.239 | 0.311 | 0.165 | 0.301  
(0.030) | (0.001) | (0.000) | (0.000) | (0.005) | (0.000)  
Non-retail employment (*1000) | 0.446 | -0.442 | 0.256 | 0.587 | 0.477 | 0.376  
(0.185) | (0.268) | (0.039) | (0.003) | (0.003) | (0.090)  
Average household income (*1000) | 0.066 | 0.100 | 0.126 | 0.001 | -0.037 | 0.029  
(0.030) | (0.044) | (0.050) | (0.533) | (0.392) | (0.697)  
Overdispersion | 0.263 | 0.178 | 0.264 | 0.338 | 0.381 | 0.175  
(0.000) | (0.000) | (0.000) | (0.000) | (0.000) | (0.000)  

Notes: first row is the coefficient, second row is the p-significance, and bold coefficients are statistically significant at 5% level of significance.

APPLICATIONS FOR NETWORK SCREENING

To apply these models, we predicted the number of crashes using the cluster-based SPFs, and estimated the expected number of crashes if no countermeasure had been implemented in the future using the Empirical Bayes (EB) method as prescribed in the HSM (1) for all TAZs in the State. The EB method increases the precision of predictions for the future when only limited historical crash data are available, and it corrects for the regression-to-mean bias (38). Details about our procedures for applying the EB method and developing the network screening application tool are provided in the Appendix to the project final report (28). The resulting EB Expected Crash Counts are added to a GIS layer along with the other data for each TAZ. The resulting GIS layer can be used for reporting and manipulation within a GIS environment by road safety analysts in ConnDOT and regional or local government to identify locations that have promise for implementing road safety interventions according to HSM procedures (1).

CONCLUSIONS AND FUTURE RESEARCH

This study demonstrates an alternative in predicting the number of crashes on local roads where the traffic volumes are not available. Both the intersection SPFs and segment SPFs were estimated at the TAZ level. The TAZs were categorized into six clusters based on land cover intensities and population density, using the K-means clustering approach. Cluster-based SPFs were estimated for predicting local road intersection and segment crash counts using, respectively, the number of intersections and the total local roadway length. Demographic variables such as population, retail and non-retail employment, total households, and average household income were used as covariates to predict the crash counts.

Due to the high correlation between population and the number of households, two cluster-based SPFs including either population or the number of households were estimated for both intersection and segment crashes. Additionally, an aggregate function using the entire dataset was also developed for comparison. Based on the goodness-of-fit (AIC and BIC values) and prediction performances (MAD and MSPE values), the cluster-based SPFs outperform the
aggregate SPFs. The cluster-based SPFs with population perform better than those with the number of households for both intersection and segment crashes.

Finally, the cluster-based SPFs were applied and adjusted using the EB method to produce expected annual crash counts for all TAZs in the State. It is anticipated that the example applications can help local agencies identify areas of town with higher penitential for safety improvements, and develop cost-effective countermeasures to improve safety for local roads.

This study has demonstrated an initial exploration into developing TAZ level SPFs using demographic variables for local roads when the traffic volumes are not available, by clustering TAZs into different types to account for the data heterogeneity. These cluster based TAZ level SPFs can be used to predict the average annual intersection and segment crashes in a TAZ in the context of HSM analyses. They also might be used to help agencies evaluate alternative options for future roadway network and economic development, by identifying the effects of roadway geometric and socio-economic factors on crash counts. However, it is likely to be more difficult to transfer these models to other jurisdictions compared with facility level SPFs (e.g. roadway segment and intersection). These TAZ level SPFs are highly dependent upon not only the clustering of the TAZs, but also the definitions of the TAZs themselves, as well as the character of land development. The relationship between these factors and crash occurrence is likely to vary much more from one place to another than would the relationship between road characteristics and traffic volume. As a consequence, attempts to calibrate these models to another State are not likely to be successful. To use the cluster based TAZ level SPFs, we recommend users to collect their own data and estimate the SPFs.

One significant challenge in conducting this study was to geo-locate crashes on local roads, as the Connecticut crash data set included only route and milepost at the time of data collection. Having geocoded crash records would substantially simplify the process. Other relevant variables (e.g. trip distance and trip duration for a TAZ) that were not available when conducting this study may also affect the roadway safety, as the crash counts are expected to increase with the increase of trip distance and duration in a TAZ. It is recommended future research focus on collecting these variables in TAZ level, and then estimate the new SPFs to improve the prediction accuracy.

It is also noted that the observed, predicted and expected annual crash counts for many TAZs were quite small (less than 3). Because each TAZ contains dozens of road segments and intersections, this indicates that the annual crash counts at each individual segment or intersection would be so small as to preclude successful estimation of crash prediction models by segment or intersection. This data condition is further justification for using an area based approach for predicting crashes on local roads.

ACKNOWLEDGMENTS

This research was sponsored by the Joint Highway Research Advisory Council of the University of Connecticut and the Connecticut Department of Transportation through Project 14-1 of the Connecticut Cooperative Transportation Research Program. The contents reflect the views of the authors who are responsible for the accuracy of the information presented herein. The contents
do not necessarily reflect the official views or policies of the University of Connecticut or the Connecticut Department of Transportation. The authors would like to thank Mrs. Judy B. Raymond of the Connecticut Department of Transportation for kindly providing demographic data to support this effort.

REFERENCES

https://www.census.gov/geo/maps-data/data/tiger-line.html


