Development of Zonal Safety Performance Functions for Local Road Intersections and Segments in Connecticut

Ishraq Rayeed Ahmed
ishraq.ahmed@uconn.edu

Recommended Citation
https://opencommons.uconn.edu/gs_theses/1172
Development of Zonal Safety Performance Functions for Local Road Intersections and Segments in Connecticut

Ishrag Rayeed Ahmed

B.Sc., Bangladesh University of Engineering of Technology, 2016

A Thesis
Submitted in partial Fulfillment of the
Requirements of the Degree of
Master of Science
University of Connecticut
2018
Masters of Science Thesis

Development of Zonal Safety Performance Functions for Local Road Intersections and Segments in Connecticut

Presented by

Ishraq Rayeed Ahmed, B.Sc.

Major Advisor

John N. Ivan

Associate Advisor

Eric Jackson

Associate Advisor

Karthik C. Konduri

University of Connecticut

2018
# Table of Contents

ABSTRACT ........................................................................................................................................ vi

1 INTRODUCTION ................................................................................................................................. 1

2 LITERATURE REVIEW .......................................................................................................................... 4

3 METHODOLOGY ...................................................................................................................................... 7

3.1 MODEL FORM / STUDY DESIGN ........................................................................................................ 7

3.2 DATA COLLECTION ............................................................................................................................. 8

3.2.1 ROADWAY NETWORK SHAPE FEATURES .............................................................................. 9

3.2.2 TAZ LEVEL DEMOGRAPHIC RECORDS .................................................................................... 9

3.2.3 TAZ LEVEL LAND COVER FEATURES .................................................................................... 10

3.2.4 CRASH RECORDS AND INTEGRATION OF CRASH TO TAZ .............................................. 11

3.2.5 CLUSTERING OF TAZS ............................................................................................................. 12

3.2.6 K-MEANS APPROACH ............................................................................................................. 12

3.2.7 LATENT CLASS APPROACH .................................................................................................... 13

3.3 CRASH PREDICTION MODEL VARIABLE SELECTION ......................................................... 13

4 RESULTS AND DISCUSSION ............................................................................................................... 14

4.1 CLUSTERING RESULTS ................................................................................................................... 14

4.2 MODEL ESTIMATION RESULTS ................................................................................................. 25

4.3 OUT OF SAMPLE PREDICTION RESULTS .................................................................................... 33
5 CONCLUSIONS AND FUTURE RESEARCH .................................................. 35
6 REFERENCES .......................................................................................... 37
7 APPENDIX ............................................................................................. 41
Development of Zonal Safety Performance Functions for Local Road Intersections and Segments in Connecticut

Ishraq Rayeed Ahmed
ishraq.ahmed@uconn.edu

ABSTRACT
Crash counts at a location are predicted using a Safety Performance Function (SPF) equation. The expected number of crashes on a traffic facility can be estimated using SPFs and the required countermeasures can be taken to reduce crashes in future. Due to the absence of sufficient traffic count data for local roads, new crash prediction approaches are essential to implementing highway safety improvement strategies. The study focuses on developing SPFs with coefficients varying by geographic covariate for segment and intersection crashes on local roads in Connecticut. Demographic and network topology data has been used as a surrogate for traffic count data which are not available for these roads. Two clustering methods – K-Means and Latent Class Clustering (LCC) has been explored for classifying cases for varying coefficients. The variables that were used to classify into clusters were land cover, population density and employment density. The models clustered using LCC with total population, retail and non-retail employment and average household income as independent variables were found to be the best based on model fit and out of sample prediction.

KEYWORDS: SPF; Crash Prediction; Local Roads; K-Means Clustering; Latent Class Clustering;
1 INTRODUCTION

In the year 2016, there were 34,439 fatal motor vehicle crashes in United States, which caused the death of 37,461 people. This counts for 11.6 deaths per 100,000 people and 1.16 deaths per 100 million miles traveled [1]. The economic toll that highway crashes create on American lives is enormous. The annual price tag for these crashes was calculated to be $871 billion in economic loss and societal harm in 2010 [2]. This cost is increasing further every year. It is important to identify and predict these crashes accurately. Once predicted, other factors like roadway factors, weather conditions, engineering factors, and driver behavior can be studied at these locations for current conditions and necessary measures can be implemented for improving highway safety.

In order to predict the number of crashes at any specific location, an equation known as the Safety Performance Function is used. This equation is a function of roadway exposure and roadway geometric features, such as number of lanes, lane width, and shoulder width. SPFs are used for estimating crash counts on roadway segments or intersections for identifying hotspots with high crash rates and thereby selecting and implementing feasible countermeasures to mitigate the crash rates. [4]. SPFs have been developed for various traffic facilities and roadway classifications. According to the FHWA, approximately 60 percent of all road miles in the US are maintained by local jurisdictions. This includes the local roads that are owned and operated by towns, counties and tribal governments [5]. From 1974 to 2000, the local road crash rate was 1.5 times higher than that of primary roads [6]. Thus it’s extremely important for transportation engineers and traffic safety organizations to ensure traffic safety on local roads. By developing accurate crash prediction tools for local roads, it will be possible to perform safety improvement
actions at sites with high crash potential and thereby reduce volume and severity of traffic crashes in the future.

The SPFs for two-lane rural highways, multilane rural highways, urban and suburban arterials, freeways and freeway ramp junctions are available in the Highway Safety Manual (HSM) [3]. However, these SPFs were developed using data from the states of Washington, California, Minnesota, Texas, Michigan, North Carolina and Illinois. Due to the variation in geography and demographic features, the crash relationships in these states cannot be considered to be accurate representatives for all the other states. In order to perform better crash predictions, the HSM recommends calibration procedures. Traffic counts is the most significant variable while accurately predicting crash counts [3], [7], [8]. This is where the problem arises in case of local roads. Traffic counts are not available for most roads under local jurisdiction as it is not economically feasible to carry out traffic counting operations on these roads with low traffic volumes [6]. So to implement highway safety improvement strategies on local roads with low traffic volumes, alternate crash prediction models need to be estimated where traffic count data will not be required.

In this study, the main objective was to estimate crash prediction models or SPFs for local road intersections and segments in Connecticut using demographic data instead of traffic count data. The SPFs have been estimated at Traffic Analysis Zone (TAZ) levels. To partially account for the roadway exposure, the total number of intersections and the total roadway mileage in each TAZs have been used. The demographic data includes population, total retail and non-retail employment, household income and vehicle availability. To incorporate the heterogeneity in crash relationship across the TAZs, they have been categorized into clusters based on land cover, population density (number of people per square mile) and employment density (number of
employed people per square mile). Statewide SPF and cluster-based SPFs were estimated and the best variables for assigning clusters and predicting crash counts were determined in this study.
2 LITERATURE REVIEW

Various researchers have estimated SPFs at two basic levels. One is the facility level, such as intersections and roadway segments. The other is the zonal level, such as block group level or TAZ level. Vogt developed accident models for two-lane rural segments and intersections [8]. The variables he used included traffic, horizontal and vertical alignments, lane and shoulder widths, roadside hazard rating, channelization, and number of driveways from the states of Minnesota and Washington. The study recommended development of adjustment factors for different regions and times and further development of extended negative binomial models. Majority of the researches on two-lane roads have traffic volume as the prime variable for traffic crashes [8], [9]. Investigative studies on alternative exposure variable for predicting crashes in lieu of traffic volume has not been widely explored. Bindra et al. explored the use of Geographic Information System (GIS) land use data as an exposure measure for estimating crash prediction models for segments and intersections at rural two-lane and urban two-and four-lane undivided roads [10]. The study concluded that segment-intersection crashes, minor road and driveway crashes can be predicted using trip generation data and land use data.

Proactive assessment of the safety impacts of Travel Demand Management (TDM) policies is extremely important. Since TDM policies are always implemented at an aggregate level, SPFs should also be developed at an equivalent scale. As such, the model resolution should match that of the TDM. Different studies have incorporated high use of available zonal-level variables based on the applicable scenario [11]. When it comes to TAZ-level SPFs, socioeconomic and network variables were used to develop TAZ level SPFs by severity level (injury and PDO crashes) [12]. However, the most common zonal SPFs have been developed at TAZ levels [13]–[16]. For the number of crashes, different factors like population density, the number of employees and the intersection density were considered as predictors. Lovegrove et al. [17] showed that temporal and
spatial transfer of TAZ level SPFs can be very effective. Noland [18] showed that with the increase in employment density in a TAZ, the traffic crashes increased. On the other hand, fewer crashes were observed in urbanized TAZs with dense population. It was also find that common variables like segment length and geometric features of roadway networks can be important for higher prediction accuracy of TAZ-level SPFs [19]. In the study by Shahi et al. [20] and Aty et al. [21], trip generation rate was found to be a significant factor for predicting TAZ level crashes.

In the study by Pirdavani et al., an association was established between the observed crashes and predictor variables in each zone. Two different exposures such as VHT and VKT were compared in this study along with demographic parameters. Russo et al. [22] developed four sets of SPFs for undivided rural highways by SPF calibration to predict annual injury and fatality frequency. Lee et al. [23] developed a multivariate Poisson lognormal crash model that can estimate crashes on all travel modes simultaneously using TAZ based demographic data. It showed that number of households, employment and hotels had positive association with motor vehicle crashes, bicycle crashes and pedestrian crashes. Some studies [24]-[26] have also investigated crash prediction models on other microscopic levels including block group, state level, grid structure level and county level.

Although these zonal level SPFs are all able to predict expected crash frequencies without traffic volume, most of them 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, Wang et al. [27] focused on developing cluster based SPFs for local roads using K-Means Clustering, based on their land-use intensities and population density. K-means clustering analysis is a traditional distance-based technique...
which has a limitation that a distance measured objective function is required to be pre-determined. It also requires large memory demands especially for a large dataset [28].

To account for these issues, the Latent Class Clustering (LCC) analysis was also applied in this study, as it doesn’t require selecting a distance measure. Socio demographic data and roadway network data such as population, employment, income, car ownership, number of local road intersections and total local road length inside the TAZs are used in this study to predict crash counts. Some of these variables are intended to function as alternatives for actual traffic counts which are not available for the local roads under consideration.

The next sections of the thesis is organized as follows: The next section presents the process of data collection, methodology and model development and evaluation. The fourth section describes the estimation of SPFs and the results. The final section discusses the findings of the research and the potential applications of the estimated SPFs for crash reduction and policy making.
3 METHODOLOGY

3.1 Model Form / Study Design

The procedure for the estimation of TAZ level SPFs for local roads in this study uses negative binomial modeling. In addition to estimating a single model for the entire state, separate models for the intersection and segment crashes were estimated for TAZs grouped by geographic characteristics. Since the extent of travel activity at a particular location, i.e., exposure, is an indicator of a higher number of crashes, the variables used for clustering were selected to account for this variation. Two methods were explored for clustering, each considering the following variables - land cover intensity, population density and total employment density. For the model development, crash data and demographic variables were deemed essential.

Safety performance functions were estimated to predict the number of local road intersection and segment crashes in each TAZ following the procedures by Wang et al. [27]. The number of crashes is estimated by Poisson regression model which is the state of the art model for developing Safety Performance Functions. It is formulated as [35]:

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

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: \(ln(\mu_i) = \beta X_i\).

One of the key limitations of the Poisson Model is that the variance in the data is equal to the mean. Usually the variance of crash data is greater than the mean, questioning this very constraint of the model. The negative binomial regression model addresses this issue of over-dispersion. This can be derived by rewriting the first equation as follows: \(\mu_i = \exp(\beta X_i + \epsilon_i)\),

\[
\ln(\mu_i) = \beta X_i + \epsilon_i,
\]
where \( \exp(\varepsilon_i) \) is the error term assumed to follow a gamma distribution with mean 1 and variance \( \sigma^2 \). The distribution of the negative binomial will have the following form:

\[
\text{Prob} \left[ y_i | \mu_i \right] = \frac{\Gamma \left( \frac{1}{\sigma} + y_i \right)}{\Gamma \left( \frac{1}{\sigma} \right) \Gamma \left( \frac{1}{\sigma} + \mu_i \right)} \left( \frac{1}{\sigma} + \mu_i \right)^{\mu_i} \left( \frac{1}{\sigma} \right)^{y_i}, \text{where } \Gamma \text{ is a gamma function and the variance of the negative binomial model is as follows: } \text{Var}(y_i) = \mu_i(1 + \sigma \mu_i) = \mu_i + \sigma \mu_i^2
\]

The function for the predicted intersection crashes at TAZ \( i \) is defined as follows:

\[
\mu_{\text{int},i} = YI_i \beta_i \exp(\beta_0 + \beta_P P_i + \beta_R R_i + \beta_N N_i + \beta_V V_i + \beta_C C_i + \beta_H H_i), \text{ where, } \mu_{\text{int},i} \text{ is the predicted intersection crashes in TAZ } i, Y \text{ is the number of years in the time period, } I_i \text{ the number of intersections in TAZ } i, P_i \text{ is the population of TAZ } i, R_i \text{ is the total retail employment of TAZ } i, N_i \text{ is the total non-retail employment of TAZ } i, V_i \text{ is the number of vehicles in TAZ } I, C_i \text{ is the average income in TAZ } i, H_i \text{ is the number of households in TAZ } I \text{ and } \beta s \text{ are the estimated parameters.}
\]

The function for the predicted segment crashes at TAZ \( i \) is defined as follows:

\[
\mu_{\text{seg},i} = YL_i \beta_i \exp(\beta_0 + \beta_P P_i + \beta_R R_i + \beta_N N_i + \beta_V V_i + \beta_C C_i + \beta_H H_i), \text{ where, } \mu_{\text{seg},i} \text{ are the predicted segment crashes in TAZ } I \text{ and } L_i \text{ is the mileage of local roadways in TAZ } i.
\]

3.2 Data Collection

The final data preparation came down to four types of TAZ level data - roadway network shape features, demographic records, land cover features and crash records. In order to take advantage of the demographic data available by TAZ, the TAZ structure defined by CT DOT for statewide planning purposes was used for the study. The following section provides a brief description of the data and data sources used for this study.
3.2.1 Roadway Network Shape Features

The number of intersections and the total length of local roads were extracted from the CTDOT 2016 Roadway Inventory System (RIS) files for Connecticut. The number of intersections and the total length of local roadways were calculated for each TAZ. There were a number of intersections which were at the border line of two adjacent TAZs. Such intersections were equally distributed between the adjacent TAZs. Figure 1 shows the RIS 2016 Local Roads on a TAZ map of Connecticut.

![Local Roads in Connecticut (RIS Data 2016)](image)

Figure 1 Local Roads in Connecticut (RIS Data 2016)

3.2.2 TAZ Level Demographic Records

The TAZ level demographic records were compiled from the Census Transportation Planning Package Database [29]. These records included TAZ level population, retail and non-retail employment, households, vehicles and average household income. In the SPF S, these were
used as independent variables. The data comes from the decennial census, the last one being from 2010. Figure 2 shows a distribution of population in Connecticut at TAZ Level.

![Connecticut Population Map](image)

**Figure 2 Connecticut Population Map**

### 3.2.3 TAZ Level Land Cover Features

For the land cover information, the 2011 National Land Cover Database (NLCD) was used [30]. USGS defined the proportion of land area in three developed land-use categories – low, medium and high intensity development. The developed areas usually comprise of a combination of vegetation and impervious surfaces. The intensity of development represents the differences in the ratios of these covers. The 2011 NLCD has defined the low intensity areas as having 20-49% impervious cover, medium intensity areas as having 50%-79% impervious cover, and high
intensity areas as having greater than 80% impervious cover. While categorizing the TAZs into clusters, these values were used along with population density and employment density. The details of the clustering methods have been discussed in a later section.

3.2.4 Crash Records and Integration of Crash to TAZ

The crash records in this study were collected from the Connecticut Crash Data Repository [31]. This crash data included both segment and intersection crashes at local roads in Connecticut in 2015 and 2016. All crash severities were included, which are - K (fatal injury), A (incapacitating injury), B (non-incapacitating injury), C (possible injury) and O (not injured). A total of 45,305 crashes were extracted including 25,312 roadway segment crashes and 19,993 intersection crashes.

Figure 3 Heat Map of Local Road Crashes in Connecticut (2015-2016)

The extracted crashes were then assigned to the TAZs on the basis of their spatial locations. There were a number of crashes that were located at the boundary of multiple TAZs. In those cases, the total number of crashes on the boundary was equally distributed between the TAZs. In case of intersection crashes, the same procedure was followed.
3.2.5 Clustering of TAZs

The main objective of clustering analysis is to maximize the homogeneity within the same cluster and the heterogeneity between the clusters [28]. One of the most popular distance-based clustering techniques is the K-means clustering analysis ([28], [32], [33]). In this method, a distance measured objective function is determined at first. To solve this issue, many studies have used the latent class clustering (LCC) analysis or finite mixture model (FMM). The positive side of this other method of clustering is that it doesn’t require selecting a distance measure. In this study, both the K-Means Clustering and the Latent Class Clustering have been considered.

3.2.6 K-Means Approach

For the K-Means Clustering, different numbers of clusters were tested, respectively, and the Calinski and Harabase pseudo-F index [34] was used to select the final number of clusters. The pseudo F statistic describes the ratio of between-cluster variance to within cluster variance.

\[
Pseudo F = \frac{GSS}{WSS} \frac{K - 1}{N - K}
\]

where N is the number of observations, K is the number of clusters at any step in the hierarchical clustering, GSS is the between-group sum of squares, and WSS is the within group sum of squares. Large values of pseudo F indicate close-knit and separated clusters. In particular, peaks in the pseudo F statistic are indicators of greater cluster separation.
3.2.7 Latent Class Approach

The Latent Class Clustering (LCC) is a probabilistic clustering approach. In this probabilistic point of view, every cluster has a different underlying probability distribution from which its data elements are generated. When the distribution functions are known, the problem of assigning cases to the clusters reduces to a parameter estimation problem [28]. Following the maximum likelihood approach, the unknown parameter vector is often estimated by means of the expectation–maximization algorithm. Given the data elements \( Y_1, \ldots, Y_n \), each described by a set of features \( (y_1, \ldots, y_m) \), the prior probability (before estimation), \( P(z) \) for cluster \( C_z \) with \( z = 1, \ldots, K \) and the conditional multi-variate probability density \( p(Y|C_z, \theta_z) \), where \( \theta_z \) is the unknown parameter vector, the mixture probability density for the whole data set can be expressed as:

\[
p(Y|\theta) = \sum_{z=1}^{K} P(C_z)p(Y|C_z, \theta_z)
\]

3.3 Crash Prediction Model Variable Selection

The SPF's were estimated for the entire state as well as based on clusters. During model estimation, the TAZs were divided into two portions. The first part included ninety percent of the data set. This part was used to estimate the model. The remaining ten percent was used to evaluate the performance of the model prediction. The correlation was checked among the independent variables and the AIC and BIC values were compared to finally select the variables that performed the best.
4 RESULTS AND DISCUSSION

4.1 Clustering Results

In case of the K-Means Clustering, the larger the Calinski and Harabase pseudo-F index, the more accurate is the clustering analysis. For the K-Means, two types of clustering were performed. The optimum number of clusters was found to be 6 for the clustering based on land cover and population density. The optimum number of clusters was also found to be 5 for the clustering based on land cover and employment density, but there were shifts in the assignment of the TAZs to the clusters. In case of Latent Class Clustering, the optimum number of clusters was 4 for both land cover and population density, and land cover and employment density.

Under LCC, the optimum number of clusters were selected based on the AIC and BIC values. Since for both the clusters under LCC, the AIC and BIC value curves became somewhat flat after the 4th cluster, the number of optimum clusters were selected to be 4.

Figure 4 shows the distribution of KABCO crashes using the four types of clustering explored in this study. The distribution of the number of intersections, roadway mileage, population, households, number of vehicles, household income, total retail and non-retail employment in each of the clustering types are also shown hereby.
Figure 8 Distribution of Number of Intersections (K-Means with Land Cover & Population Density)

Figure 9 Distribution of Roadway Mileage (K-Means with Land Cover & Population Density)

Figure 10 Distribution of Population (K-Means with Land Cover & Population Density)

Figure 11 Distribution of Number of Households (K-Means with Land Cover & Population Density)
Figure 12 Distribution of Number of Vehicles (K-Means with Land Cover & Population Density)

Figure 13 Distribution of Household Income (K-Means with Land Cover & Population Density)

Figure 14 Distribution of Retail Workers (K-Means with Land Cover & Population Density)

Figure 15 Distribution of Non-Retail Workers (K-Means with Land Cover & Population Density)
Figure 16 Distribution of Number of Intersections (K-Means with Land Cover & Employment Density)

Figure 17 Distribution of Roadway Mileage (K-Means with Land Cover & Employment Density)

Figure 18 Distribution of Population (K-Means with Land Cover & Employment Density)

Figure 19 Distribution of Number of Households (K-Means with Land Cover & Employment Density)
Figure 20 Distribution of Number of Vehicles (K-Means with Land Cover & Employment Density)

Figure 21 Distribution of Household Income (K-Means with Land Cover & Employment Density)

Figure 22 Distribution of Retail Workers (K-Means with Land Cover & Employment Density)

Figure 23 Distribution of Non-Retail Workers (K-Means with Land Cover & Employment Density)
Figure 24 Distribution of Number of Intersections (LCC with Land Cover & Population Density)

Figure 25 Distribution of Roadway Mileage (LCC with Land Cover & Population Density)

Figure 26 Distribution of Population (LCC with Land Cover & Population Density)

Figure 27 Distribution of Number of Households (LCC with Land Cover & Population Density)
Figure 28 Distribution of Number of Vehicles (LCC with Land Cover & Population Density)

Figure 29 Distribution of Household Income (LCC with Land Cover & Population Density)

Figure 30 Distribution of Retail Workers (LCC with Land Cover & Population Density)

Figure 31 Distribution of Non-Retail Workers (LCC with Land Cover & Population Density)
Figure 32 Distribution of Number of Intersections (LCC with Land Cover & Employment Density)

Figure 33 Distribution of Roadway Mileage (LCC with Land Cover & Employment Density)

Figure 34 Distribution of Population (LCC with Land Cover & Employment Density)

Figure 35 Distribution of Number of Households (LCC with Land Cover & Employment Density)
Figure 36 Distribution of Number of Vehicles (LCC with Land Cover & Employment Density)

Figure 37 Distribution of Household Income (LCC with Land Cover & Employment Density)

Figure 38 Distribution of Retail Workers (LCC with Land Cover & Employment Density)

Figure 39 Distribution of Non-Retail Workers (LCC with Land Cover & Employment Density)
Focusing on the last cluster, LCC with land cover and employment density, we can see that the distribution of the cluster variables, the overall land-use intensity, the population density and the employment density gradually decreased from one cluster to the next. The first cluster has the lowest number of TAZs, and is the most urbanized in nature, and last cluster is the most common cluster type and is the most rural in nature. The average household income increases from the first to the last cluster.

The first cluster has the highest average numbers for both retail and non-retail employment, and last cluster has the lowest numbers. The distribution patterns showed similar patterns among population, households and vehicles. In order to avoid multi-collinearity issues in modelling, a correlation test was carried out. It was found that these three variables are highly correlated. Due to the poorer performance of the function using the number of vehicles and number of households, only the function including population is presented in this study.

Figure 40 Correlation Test of Independent Variables
4.2 Model Estimation Results

Tables 1 through 8 show the coefficient estimates for the segment and intersection SPFs using both the clusters by exposure covariates and clustering approach. 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 90% confidence.

Table 1 Coefficient Estimates for KABCO Intersection Crashes using K-Means (Land Cover & Population Density)

<table>
<thead>
<tr>
<th>Variables</th>
<th>Coefficient Estimates by Cluster &amp; State</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Intercept</td>
<td>2.623</td>
</tr>
<tr>
<td></td>
<td>(0.108)</td>
</tr>
<tr>
<td>Population</td>
<td>-0.200</td>
</tr>
<tr>
<td></td>
<td>(0.376)</td>
</tr>
<tr>
<td>Retail employment</td>
<td>-0.004</td>
</tr>
<tr>
<td></td>
<td>(0.988)</td>
</tr>
<tr>
<td>Non-retail employment</td>
<td>-0.354</td>
</tr>
<tr>
<td></td>
<td>(0.233)</td>
</tr>
<tr>
<td>Average household income</td>
<td>0.089</td>
</tr>
<tr>
<td></td>
<td>(0.633)</td>
</tr>
<tr>
<td>Log (number of intersections)</td>
<td>0.078</td>
</tr>
<tr>
<td></td>
<td>(0.865)</td>
</tr>
<tr>
<td>Overdispersion</td>
<td>1.581</td>
</tr>
<tr>
<td></td>
<td>(0.000)</td>
</tr>
</tbody>
</table>
Table 2 Coefficient Estimates for KABCO Segment Crashes using K-Means (Land Cover & Population Density)

<table>
<thead>
<tr>
<th>Variables</th>
<th>Coefficient Estimates by Cluster &amp; State</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Intercept</td>
<td>3.493</td>
</tr>
<tr>
<td></td>
<td>(0.133)</td>
</tr>
<tr>
<td>Population (*1000)</td>
<td>0.284</td>
</tr>
<tr>
<td></td>
<td>(0.000)</td>
</tr>
<tr>
<td>Retail employment (*1000)</td>
<td>-0.017</td>
</tr>
<tr>
<td></td>
<td>(0.835)</td>
</tr>
<tr>
<td>Non-retail employment (*1000)</td>
<td>0.289</td>
</tr>
<tr>
<td></td>
<td>(0.001)</td>
</tr>
<tr>
<td>Average household income (*1000)</td>
<td>-0.142</td>
</tr>
<tr>
<td></td>
<td>(0.018)</td>
</tr>
<tr>
<td>Log (roadway segment length)</td>
<td>0.028</td>
</tr>
<tr>
<td></td>
<td>(0.898)</td>
</tr>
<tr>
<td>Overdispersion</td>
<td>6.17</td>
</tr>
<tr>
<td></td>
<td>(0.029)</td>
</tr>
</tbody>
</table>

Notes: first row is the coefficient, second row is the p-significance, and bold coefficients are statistically significant at 10% level of significance.
Table 3 Coefficient Estimates for KABCO Intersection Crashes using K-Means (Land Cover & Employment Density)

<table>
<thead>
<tr>
<th>Variables</th>
<th>Coefficient Estimates by Cluster &amp; State</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Intercept</td>
<td></td>
</tr>
<tr>
<td>2.097</td>
<td>10.014</td>
</tr>
<tr>
<td>(0.00)</td>
<td>(0.007)</td>
</tr>
<tr>
<td>Population (*1000)</td>
<td></td>
</tr>
<tr>
<td>0.121</td>
<td>0.124</td>
</tr>
<tr>
<td>(0.066)</td>
<td>(0.764)</td>
</tr>
<tr>
<td>Retail employment (*1000)</td>
<td></td>
</tr>
<tr>
<td>-0.04</td>
<td>1.453</td>
</tr>
<tr>
<td>(0.414)</td>
<td>(0.006)</td>
</tr>
<tr>
<td>Non-retail employment (*1000)</td>
<td>0.112</td>
</tr>
<tr>
<td>(0.028)</td>
<td>(0.064)</td>
</tr>
<tr>
<td>Average household income (*1000)</td>
<td>0.106</td>
</tr>
<tr>
<td>(0.02)</td>
<td>(0.665)</td>
</tr>
<tr>
<td>Log (number of intersections)</td>
<td>0.086</td>
</tr>
<tr>
<td>(0.023)</td>
<td>(0.029)</td>
</tr>
<tr>
<td>Overdispersion P-Value</td>
<td>1.581</td>
</tr>
<tr>
<td>(0.000)</td>
<td>(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 10% level of significance.
Table 4 Coefficient Estimates for KABCO Segment Crashes using K-Means (Land Cover & Employment Density)

<table>
<thead>
<tr>
<th>Variables</th>
<th>Coefficient Estimates by Cluster &amp; State</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Intercept</td>
<td>0.158</td>
</tr>
<tr>
<td></td>
<td>(0.031)</td>
</tr>
<tr>
<td>Population (*1000)</td>
<td>0.253</td>
</tr>
<tr>
<td></td>
<td>(0.000)</td>
</tr>
<tr>
<td>Retail employment (*1000)</td>
<td>0.19</td>
</tr>
<tr>
<td></td>
<td>(0.000)</td>
</tr>
<tr>
<td>Non-retail employment (*1000)</td>
<td>0.214</td>
</tr>
<tr>
<td></td>
<td>(0.000)</td>
</tr>
<tr>
<td>Average household income (*1000)</td>
<td>0.008</td>
</tr>
<tr>
<td></td>
<td>(0.067)</td>
</tr>
<tr>
<td>Log (roadway segment length)</td>
<td>0.160</td>
</tr>
<tr>
<td></td>
<td>(0.005)</td>
</tr>
<tr>
<td>Overdispersion</td>
<td>4.269</td>
</tr>
<tr>
<td></td>
<td>(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 10% level of significance.
Table 5 Coefficient Estimates for KABCO Intersection Crashes
using LCC (Land Cover & Population Density)

<table>
<thead>
<tr>
<th>Variables</th>
<th>Coefficient Estimates by Cluster &amp; State</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Intercept</td>
<td>5.372</td>
</tr>
<tr>
<td></td>
<td>(0.000)</td>
</tr>
<tr>
<td>Population (*1000)</td>
<td>-0.061</td>
</tr>
<tr>
<td></td>
<td>(0.755)</td>
</tr>
<tr>
<td>Retail employment (*1000)</td>
<td>0.187</td>
</tr>
<tr>
<td></td>
<td>(0.343)</td>
</tr>
<tr>
<td>Non-retail employment (*1000)</td>
<td><strong>-0.423</strong></td>
</tr>
<tr>
<td></td>
<td>(0.055)</td>
</tr>
<tr>
<td>Average household income (*1000)</td>
<td>-0.083</td>
</tr>
<tr>
<td></td>
<td>(0.612)</td>
</tr>
<tr>
<td>Log (number of intersections)</td>
<td><strong>-0.721</strong></td>
</tr>
<tr>
<td></td>
<td>(0.049)</td>
</tr>
<tr>
<td>Overdispersion P-Value</td>
<td>1.343</td>
</tr>
<tr>
<td></td>
<td>(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 10% level of significance.
Table 6 Coefficient Estimates for KABCO Segment Crashes using LCC (Land Cover & Population Density)

<table>
<thead>
<tr>
<th>Variables</th>
<th>Coefficient Estimates by Cluster &amp; State</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Intercept</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.084</td>
</tr>
<tr>
<td></td>
<td>(0.677)</td>
</tr>
<tr>
<td>Population (*1000)</td>
<td>0.145</td>
</tr>
<tr>
<td></td>
<td>(0.091)</td>
</tr>
<tr>
<td>Retail employment (*1000)</td>
<td>0.124</td>
</tr>
<tr>
<td></td>
<td>(0.112)</td>
</tr>
<tr>
<td>Non-retail employment (*1000)</td>
<td>0.212</td>
</tr>
<tr>
<td></td>
<td>(0.007)</td>
</tr>
<tr>
<td>Average household income</td>
<td></td>
</tr>
<tr>
<td>(*1000)</td>
<td>-0.128</td>
</tr>
<tr>
<td></td>
<td>(0.066)</td>
</tr>
<tr>
<td>Log (roadway segment length)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.249</td>
</tr>
<tr>
<td></td>
<td>(0.296)</td>
</tr>
<tr>
<td>Overdispersion</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2.441</td>
</tr>
<tr>
<td></td>
<td>(0.002)</td>
</tr>
</tbody>
</table>

Notes: first row is the coefficient, second row is the p-significance, and bold coefficients are statistically significant at 10% level of significance.
Table 7 Coefficient Estimates for KABCO Intersection Crashes using LCC (Land Cover & Employment Density)

<table>
<thead>
<tr>
<th>Variables</th>
<th>Coefficient Estimates by Cluster &amp; State</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Intercept</td>
<td>2.642</td>
</tr>
<tr>
<td></td>
<td>(0.007)</td>
</tr>
<tr>
<td>Population (*1000)</td>
<td>-0.139</td>
</tr>
<tr>
<td></td>
<td>(0.449)</td>
</tr>
<tr>
<td>Retail employment (*1000)</td>
<td>0.175</td>
</tr>
<tr>
<td></td>
<td>(0.370)</td>
</tr>
<tr>
<td>Non-retail employment (*1000)</td>
<td>0.025</td>
</tr>
<tr>
<td></td>
<td>(0.895)</td>
</tr>
<tr>
<td>Average household income (*1000)</td>
<td><strong>-0.423</strong></td>
</tr>
<tr>
<td></td>
<td>(0.042)</td>
</tr>
<tr>
<td>Log (number of intersections)</td>
<td>-0.005</td>
</tr>
<tr>
<td></td>
<td>(0.985)</td>
</tr>
<tr>
<td>Overdispersion P-Value</td>
<td><strong>1.194</strong></td>
</tr>
<tr>
<td></td>
<td>(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 10% level of significance.
Table 8 Coefficient Estimates for KABCO Segment Crashes using LCC (Land Cover & Employment Density)

<table>
<thead>
<tr>
<th>Variables</th>
<th>Coefficient Estimates by Cluster &amp; State</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Intercept</td>
<td>-6.639</td>
</tr>
<tr>
<td></td>
<td>(0.000)</td>
</tr>
<tr>
<td>Population (*1000)</td>
<td>0.216</td>
</tr>
<tr>
<td></td>
<td>(0.000)</td>
</tr>
<tr>
<td>Retail employment (*1000)</td>
<td>0.028</td>
</tr>
<tr>
<td></td>
<td>(0.674)</td>
</tr>
<tr>
<td>Non-retail employment (*1000)</td>
<td>0.246</td>
</tr>
<tr>
<td></td>
<td>(0.000)</td>
</tr>
<tr>
<td>Average household income (*1000)</td>
<td>-0.042</td>
</tr>
<tr>
<td></td>
<td>(0.529)</td>
</tr>
<tr>
<td>Log (roadway segment length)</td>
<td>0.975</td>
</tr>
<tr>
<td></td>
<td>(0.000)</td>
</tr>
<tr>
<td>Overdispersion P-Value</td>
<td>5.44</td>
</tr>
<tr>
<td></td>
<td>(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 10% level of significance.
From the coefficient estimates of the eight models illustrated above, the last four models, using LCC with land cover and population density and LCC with land cover and employment density data were the new additions in this study and need to be evaluate.

For LCC with land cover and population density data, we see that the variables are not very significant even at 10% level of significance for the intersection model. However, the roadway segment model performed quite well, with highly significant values for population, retail and non-retail employment, household income and roadway segment length. For LCC with land cover and employment density data, we see a similar trend. The intersection model does not have many significant variables, but the segment model shows that the population, the roadway segment length and the non-retail employment variables are very significant in the cluster-based models.

4.3 Out of Sample Prediction Results

In order to evaluate the performance of the function, the criteria for evaluation are AIC, BIC, Mean Absolute Deviation (MAD) or Mean Square Prediction Error (MSPE). MAD provides a measure of the average misprediction of the model. A value close to 0 suggests that, on average, the model predicts the observed data well. According to Oh et al. [36], MAD is given by:

$$MAD = \frac{\sum_{i=1}^{n} |\hat{Y}_i - Y_i|}{n},$$

where $n$ is the validation data sample size.

MSPE is the sum of the squared differences between observed and predicted crash frequencies divided by sample size. MSPE is typically used to assess the error associated with a validation or external data set and is given by:

$$MSPE = \frac{\sum_{i=1}^{n_2} (\hat{Y}_i - Y_i)^2}{n_2^2},$$

where $n_2$ is the validation data sample size.
Lower values of MAD and MSPE are preferred. Here in Table 9, the MAD and MSPE values are calculated for the estimated and predicted data for the statewide model and the cluster-based models. The cluster based model K-Means with land cover and population density turned out to have lower MAD and MSPE values among all the models. The large difference between estimation and prediction values were likely caused by the wide range of values in the absolute difference of the estimation and prediction and existing outliers. Considering all model fit and predictions, the K-Means cluster based SPFs with land cover and population density as cluster variables can be selected. However, the Latent Class Cluster models performed better for roadway segment models.

So depending on the type of facility (intersection or roadway segment), the use of different cluster-based models can be explored.

Table 9 MOE Comparison for State-wide and Cluster-Based SPF for Intersection and Segment Crashes

<table>
<thead>
<tr>
<th></th>
<th>Measure of Effectiveness</th>
<th>Statewide SPF</th>
<th>KMeans (Land Cover, Population Density)</th>
<th>KMeans (Land Cover, Employment Density)</th>
<th>LCCPD (Land Cover, Population Density)</th>
<th>LCCED (Land Cover, Employment Density)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intersection</td>
<td>MAD Estimation</td>
<td>0.908</td>
<td>0.905</td>
<td>0.905</td>
<td>0.888</td>
<td>0.897</td>
</tr>
<tr>
<td></td>
<td>MAD Prediction</td>
<td>11.940</td>
<td><strong>10.046</strong></td>
<td>13.318</td>
<td>10.709</td>
<td>10.520</td>
</tr>
<tr>
<td></td>
<td>MSPE Estimation</td>
<td>1.998</td>
<td>2.25</td>
<td>2.059</td>
<td>2.113</td>
<td>2.286</td>
</tr>
<tr>
<td></td>
<td>MSPE Prediction</td>
<td>356.58</td>
<td><strong>310.219</strong></td>
<td>537.618</td>
<td>384.817</td>
<td>363.342</td>
</tr>
<tr>
<td>Segment</td>
<td>MAD Estimation</td>
<td>0.599</td>
<td>0.542</td>
<td>0.558</td>
<td>0.545</td>
<td>0.493</td>
</tr>
<tr>
<td></td>
<td>MAD Prediction</td>
<td>10.628</td>
<td><strong>3.816</strong></td>
<td>4.990</td>
<td>4.343</td>
<td>4.170</td>
</tr>
<tr>
<td></td>
<td>MSPE Estimation</td>
<td>0.733</td>
<td>0.728</td>
<td>0.642</td>
<td>0.817</td>
<td>0.474</td>
</tr>
<tr>
<td></td>
<td>MSPE Prediction</td>
<td>432.565</td>
<td><strong>30.451</strong></td>
<td>56.212</td>
<td>49.05</td>
<td>65.594</td>
</tr>
</tbody>
</table>
5 CONCLUSIONS AND FUTURE RESEARCH

This study was valuable in demonstrating ways to develop crash prediction models using demographic variables instead of traffic count data. SPFs were developed for both roadway segments and intersections at TAZ level. The TAZs were clustered into 6 and 5 clusters following the K-Means clustering method and with two sets of variables. The TAZs were clustered into 4 and 4 clusters using Latent Class Clustering method using two sets of variables. The local road intersection and segment crashes were predicted by including intersection crashes and roadway segment crashes respectively into the models. The demographic variables included population, retail and non-retail employment, total number of households and average household incomes.

There was high correlation among population, number of households and the total number of vehicles. The models were estimated using only the population variable out of the three, based on goodness of fit. It was also observed that the cluster based SPFs performed better than the statewide SPFs. It is expected that the study will contribute towards identifying crash hotspots and locations with high potential for safety improvements and feasible and economic countermeasures can be taken to make these city and town roads safer.

Although the cluster based TAZ level SPFs are able to predict crashes for intersections and segments in the local roads in Connecticut, it might be difficult to transfer these models to other jurisdictions. Apart from the unique clustering, these models are very much dependent on how the TAZs have been defined along with the land cover. So the relationship between these factors and the possibility of a crash to occur is likely to vary from place to place. As such, calibrating these models for a different location might not be effective.

The most difficult part of conducting this study was to identify significant variables for the models. There are lots of different variables such as trip distance or trip duration, which could have explained the high magnitudes of the intercepts in the models. Future research can focus on testing
other variables that might turn out to be significant for crash prediction models. Besides, probability based clustering models can also improve prediction performance of the models. This study dealt with crashes of all severities (KABCO). It would be a good idea to develop separate models for PDO, injury and fatal crashes. Instead of using TAZ level SPF, a more compact and concentrated area based approach could also be investigated for developing better crash prediction models.
6 REFERENCES


[19] S. Washington, National Cooperative Highway Research Program, National Research Council (U.S.), and American Association of State Highway and Transportation Officials,


7 APPENDIX

Clustering Codes

K6 = K-Means Clustering with Land Cover and Population Density
K5 = K-Means Clustering with Land Cover and Employment Density
LCCPD = Latent Class Clustering with Land Cover and Population Density
LCCED = Latent Class Clustering with Land Cover and Employment Density

Clustering Models

R Programming Codes for K-Means Clustering with Land Cover and Population Density:

dat<-read.csv(file="c:/Data Excel File/FinalData.csv", header=TRUE, sep="",)

dat_kmeans<-dat[,c('FID','NLCD','POPDEN')]

n<-15
crit<-c(0) # within / between
varr<-c(0) # variance explained
for(i in 2:n){
    km<-kmeans(dat_kmeans[,c(1)],i)
    crit[i]<-km$tot.withinss/km$betweenss
    varr[i]<-km$betweenss/km$totss
}

x<-c(1:n)
par(mfrow=c(1,2))
plot(x,crit,main = 'Within/Between Criterion')
plot(x,varr,main = 'Variance Explained Criterion')
# for k means cluster 6 was selected for K
clus<-kmeans(dat_kmeans[,c(1)],6)
clus$cluster
dat_kmeans$K6<-clus$cluster
plot(dat_kmeans$NLCD,dat_kmeans$POPDEN,col=clus$cluster)
table(clus$cluster)

write.csv(dat_kmeans,'c:/Data Excel File/Output/K-Means/NLCD_POPDEN.csv',row.names = F,col.names = T,quote = F)

R Programming Codes for K-Means Clustering with Land Cover and Employment Density:
dat<-read.csv(file="c:/Data Excel File/FinalData.csv", header=TRUE, sep="",)
dat_kmeans<-dat[,c('FID','NLCD','EMPDEN')]

n<-15
crit<-c(0) # within / between
varr<-c(0) # variance explained
for(i in 2:n){
  km<-kmeans(dat_kmeans[,c(1)],i)
  crit[i]<-km$tot.withinss/km$betweenss
  varr[i]<-km$betweenss/km$totss
}

x<-c(1:n)
par(mfrow=c(1,2))
plot(x,crit,main = 'Within/Between Criterion')
plot(x,varr,main = 'Variance Explained Criterion')

# for k means cluster 5 was selected for K
clus<-kmeans(dat_kmeans[,c(1)],8)
clus$cluster
dat_kmeans$K8<-clus$cluster
plot(dat_kmeans$NLCD,dat_kmeans$EMPDEN,col=clus$cluster)
table(clus$cluster)
write.csv(dat_kmeans,'c:/Data Excel File/Output/K-Means/NLCD_EMPDEN.csv',row.names = F,col.names = T,quote = F)

**SPF Negative Binomial Model**

**R Programming Codes for Intersection Models:**

library(MASS) # Need this library to run the model
dat<-read.csv(file="c:/Data Excel File/Input4NB.csv", header=TRUE, sep=",")
dat$LOG_SUM_INT<-log(dat$Sum_Intersections)
dat<-dat[dat$Sum_IC>0,]

#install.packages('corrplot')
#install.packages('AER')
library(corrplot)
corrplot(cor(dat[,c(4:13)]))

library(AER)
nb_model<-function(data,train_percent){
  set.seed(3)
data$POP <- scale(data$POP)
data$RET <- scale(data$RET)
data$NRET <- scale(data$NRET)
data$VEH <- scale(data$VEH)
data$HHINC <- scale(data$HHINC)
data$HH <- scale(data$HH)

boxplot(data[, c('POP', 'RET', 'NRET', 'HHINC')],
        main = 'Boxplot after normalization')

ntrain <- round(train_percent * nrow(data))
index <- sample(nrow(data), ntrain)
train <- data[index, ]
test <- data[-index, ]

modl <- glm.nb(Sum_IC ~ POP + RET + NRET + HHINC +
               LOG_SUM_INT, data = train)
print(summary(modl))

plot(train$Sum_IC, modl$fitted.values,
      main = 'Showing training performance')
abline(a = 0, b = 1)
pred <- predict(modl, test, type = 'response')

plot(test$Sum_IC, pred,
      main = 'Showing test performance')
abline(a = 0, b = 1)

pois <- glm(Sum_IC ~ POP + RET + NRET + HHINC +
            LOG_SUM_INT, data = train, family = 'poisson')

ovd <- dispersiontest(pois, trafo = 1)
```r
out <- list()
out$model <- modl
out$test <- test
out$pred <- pred
out$train_fit <- modl$fitted.values
out$test_response <- test$Sum_IC
out$train_prediction <- pred
out$dispersion <- ovd$estimate
out$dispersion_pval <- ovd$p.value
out$MAD <- mean(abs(test$Sum_IC - pred))
out$MSPE <- mean((test$Sum_IC - pred)^2)
return(out)
}

run_model <- nb_model(data = dat, train_percent = 0.9)
run_model$dispersion
run_model$dispersion_pval
run_model$MAD
run_model$MSPE

MAD_E <- mean(abs(run_model$model$residuals))
MSPE_E <- mean(run_model$model$residuals^2)

summary(run_model$model)
summary(run_model$model)$coefficients
```
write.csv(summary(run_model$model)$coefficients,'c:/Data Excel File/Output/NB/NB_IC_STATE.csv',row.names = T,quote = F)

**R Programming Codes for Segment Models:**

library(MASS)

dat<-read.csv(file="c:/Data Excel File/Input4NB.csv", header=TRUE, sep=",")
dat$LOG_SUM_RS<-log(dat$Sum_SegmentLength)
dat<-dat[dat$Sum_RC>0,]

#install.packages('corrplot')
#install.packages('AER')

library(corrplot)
corrplot(cor(dat[,c(4:13)]))

library(AER)
nb_model<-function(data,train_percent){
  set.seed(12345)
  data$POP<-scale(data$POP)
data$RET<-scale(data$RET)
data$NRET<-scale(data$NRET)
  data$VEH <- scale(data$VEH)
  data$HHINC <- scale(data$HHINC)
```r
data$HH <- scale(data$HH)

boxplot(data[,c('POP','RET','NRET','HHINC')],
    main='Boxplot after normalization')

ntrain<round(train_percent*nrow(data))
index<-sample(nrow(data),ntrain)

train<-data[index,]

train<data[-index,]

modl<-glm.nb(Sum_RC~POP+RET+NRET+HHINC+
    LOG_SUM_RS, data = train)

print(summary(modl))

plot(train$Sum_RC,modl$fitted.values,
    main = 'Showing training performance')

abline(a=0,b=1)

pred<-predict(modl,test,type = 'response')

plot(test$Sum_RC,pred,
    main = 'Showing test performance')

abline(a=0,b=1)

pois<-glm(Sum_RC~POP+RET+NRET+HHINC+
    LOG_SUM_RS, data = train, family = 'poisson')

ovd<-dispersiontest(pois,trafo = 1)

out<-list()

out$model<-modl

out$test<-test

out$pred<-pred

out$train_fit<-modl$fitted.values
```
```r
out$test_response <- test$Sum_RC
out$train_prediction <- pred
out$dispersion <- ovd$estimate
out$dispersion_pval <- ovd$p.value
out$MAD <- mean(abs(test$Sum_RC - pred))
out$MSPE <- mean((test$Sum_RC - pred)^2)
return(out)
}

run_model <- nb_model(data = dat, train_percent = 0.90)
run_model$dispersion
run_model$dispersion_pval
run_model$MAD
run_model$MSPE

MAD_E <- mean(abs(run_model$model$residuals))
MSPE_E <- mean(run_model$model$residuals^2)

summary(run_model$model)
summary(run_model$model)$coefficients

write.csv(summary(run_model$model)$coefficients, 'c:/Data Excel File/Output/NB/NB_RC_STATE.csv', row.names = T, quote = F)
```