

**IMPROVEMENTS TO ROAD SAFETY IMPROVEMENT
SELECTION PROCEDURES FOR CONNECTICUT**

FINAL REPORT

June 2016

John Ivan, Amy Burnicki, Kai Wang and Sha Mamun

JHR 16-328

Project 14-01

This research was sponsored by the Joint Highway Research Advisory Council (JHRAC) of the University of Connecticut and the Connecticut Department of Transportation and was performed through the Connecticut Transportation Institute of the University of Connecticut.

The contents of this report reflect the views of the authors who are responsible for the facts and accuracy of the data presented herein. The contents do not necessarily reflect the official views or policies of the University of Connecticut or the Connecticut Department of Transportation. This report does not constitute a standard, specification, or regulation.

TECHNICAL REPORT DOCUMENTATION PAGE

1. Report No. JHR 16-328	2. Government Accession No. N/A	3. Recipient's Catalog No.	
4. Title and Subtitle Improvements to Road Safety Improvement Selection Procedures for Connecticut		5. Report Date June 2016	
		6. Performing Organization Code N/A	
7. Author(s) John Ivan, Amy Burnicki, Kai Wang and Sha Mamun		8. Performing Organization Report No. JHR 16-328	
9. Performing Organization Name and Address University of Connecticut Connecticut Transportation Institute Storrs, CT 06269-5202		10. Work Unit No. (TR AIS) N/A	
		11. Contract or Grant No. N/A	
12. Sponsoring Agency Name and Address Connecticut Department of Transportation 2800 Berlin Turnpike Newington, CT 06111		13. Type of Report and Period Covered FINAL	
		14. Sponsoring Agency Code CCTRP 14-01	
15. Supplementary Notes This study was conducted under the Connecticut Cooperative Transportation Research Program (CCTRP), http://www.cti.uconn.edu/cctrp/ .			
16. 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 report 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. The SPFs are applied in two Connecticut towns to illustrate the usefulness of the SPFs as a network screening tool.			
17. Key Words Safety performance function, crash count, local road, cluster analysis		18. Distribution Statement No restrictions. This document is available to the public through the National Technical Information Service Springfield, Virginia 22161	
19. Security Classif. (of this report) Unclassified	20. Security Classif. (of this page) Unclassified	21. No. of Pages 77	22. Price N/A

ACKNOWLEDGMENTS

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.

SI* (MODERN METRIC) CONVERSION FACTORS

APPROXIMATE CONVERSIONS TO SI UNITS

SYMBOL	WHEN YOU KNOW	MULTIPLY BY	TO FIND	SYMBOL
LENGTH				
in	inches	25.4	millimeters	mm
ft	feet	0.305	meters	m
yd	yards	0.914	meters	m
mi	miles	1.61	kilometers	km
AREA				
in ²	square inches	645.2	square millimeters	mm ²
ft ²	square feet	0.093	square meters	m ²
yd ²	square yard	0.836	square meters	m ²
ac	acres	0.405	hectares	ha
mi ²	square miles	2.59	square kilometers	km ²
VOLUME				
fl oz	fluid ounces	29.57	milliliters	mL
gal	gallons	3.785	liters	L
ft ³	cubic feet	0.028	cubic meters	m ³
yd ³	cubic yards	0.765	cubic meters	m ³
NOTE: volumes greater than 1000 L shall be shown in m ³				
MASS				
oz	ounces	28.35	grams	g
lb	pounds	0.454	kilograms	kg
T	short tons (2000 lb)	0.907	megagrams (or "metric ton")	Mg (or "t")
TEMPERATURE (exact degrees)				
°F	Fahrenheit	5 (F-32)/9 or (F-32)/1.8	Celsius	°C
ILLUMINATION				
fc	foot-candles	10.76	lux	lx
fl	foot-Lamberts	3.426	candela/m ²	cd/m ²
FORCE and PRESSURE or STRESS				
lbf	poundforce	4.45	newtons	N
lbf/in ²	poundforce per square inch	6.89	kilopascals	kPa

APPROXIMATE CONVERSIONS FROM SI UNITS

SYMBOL	WHEN YOU KNOW	MULTIPLY BY	TO FIND	SYMBOL
LENGTH				
mm	millimeters	0.039	inches	in
m	meters	3.28	feet	ft
m	meters	1.09	yards	yd
km	kilometers	0.621	miles	mi
AREA				
mm ²	square millimeters	0.0016	square inches	in ²
m ²	square meters	10.764	square feet	ft ²
m ²	square meters	1.195	square yards	yd ²
ha	hectares	2.47	acres	ac
km ²	square kilometers	0.386	square miles	mi ²
VOLUME				
mL	milliliters	0.034	fluid ounces	fl oz
L	liters	0.264	gallons	gal
m ³	cubic meters	35.314	cubic feet	ft ³
m ³	cubic meters	1.307	cubic yards	yd ³
MASS				
g	grams	0.035	ounces	oz
kg	kilograms	2.202	pounds	lb
Mg (or "t")	megagrams (or "metric ton")	1.103	short tons (2000 lb)	T
TEMPERATURE (exact degrees)				
°C	Celsius	1.8C+32	Fahrenheit	°F
ILLUMINATION				
lx	lux	0.0929	foot-candles	fc
cd/m ²	candela/m ²	0.2919	foot-Lamberts	fl
FORCE and PRESSURE or STRESS				
N	newtons	0.225	poundforce	lbf
kPa	kilopascals	0.145	poundforce per square inch	lbf/in ²

*SI is the symbol for the International System of Units. Appropriate rounding should be made to comply with Section 4 of ASTM E380. (Revised March 2003)

TABLE OF CONTENTS

LIST OF FIGURES	vi
LIST OF TABLES.....	vi
INTRODUCTION	1
LITERATURE REVIEW	3
METHODOLOGY AND DATA PREPARATION	5
Roadway Network Shape Features	5
TAZ Level Demographic Records.....	5
TAZ Level Geographic/Land Cover Features	5
Crash Records and Integration of Crash to TAZ	5
Clustering of TAZs	6
Statistical Methodology	9
VARIABLE SELECTION AND SPF RESULTS	13
EXAMPLE NETWORK SCREENING APPLICATIONS.....	17
CONCLUSIONS AND FUTURE RESEARCH	21
REFERENCES	23
APPENDICES	25

LIST OF FIGURES

Figure 1 Clustering Results and Cluster Distribution	6
Figure 2 Distributions of KAB Crashes by Cluster	7
Figure 3 Distributions of Independent Variables by Cluster	8
Figure 4 Distributions of Independent Variables by Cluster (Continued).....	9
Figure 5 Example Network Screening Application – Groton.....	18
Figure 6 Example Network Screening Application – Stamford.....	19

LIST OF TABLES

Table 1 Goodness-of-fit of the Cluster Based SPF	14
Table 2 SPF Prediction Performance.....	14
Table 3 Coefficient Estimates for KAB Intersection Crashes	15
Table 4 Coefficient Estimates for KAB Segment Crashes	16

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 safety performance functions (SPFs) is estimating the expected number of crashes on traffic facilities to identify road locations with higher crash counts, and implement cost-effective countermeasures to reduce 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 and implementing effective countermeasures to reduce 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 USA States, 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 km²). A different SPF was estimated for each cluster, and the similarities and differences among these functions are discussed. We also include an example application of the functions as a network screening tool in two Connecticut towns.

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) TAZ level SPFs were initially introduced by Levine *et al.* (11). Their study uses a set of socioeconomic and network variables to predict the number of crashes by TAZ. Similarly, Pulugurtha *et al.* (12) 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.* (13), Lovegrove and Sayed (14), Lovegrove (15) and Hadayeghi *et al.* (16) 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.* (17) found that TAZ level SPFs can safely be transferred both temporally and spatially. Noland and Quddus (18) showed that TAZs with high employment density had more traffic crashes, whereas in urbanized areas with more densely populated TAZs fewer crashes were observed.

Recently, an analysis tool (PLANSAFE) was developed on a National Cooperative Highway Research Program (NCHRP) project (19) 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.

Although these TAZ 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 injury and fatal 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. In the final section, the SPFs are applied to the City of Stamford and the Town of Groton to illustrate the usefulness of the functions 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 (20). 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 A.

TAZ Level Demographic Records

TAZ level demographic records were collected from the Census Transportation Planning Package Database (CTPP, 2010) (21). 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) (22). 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

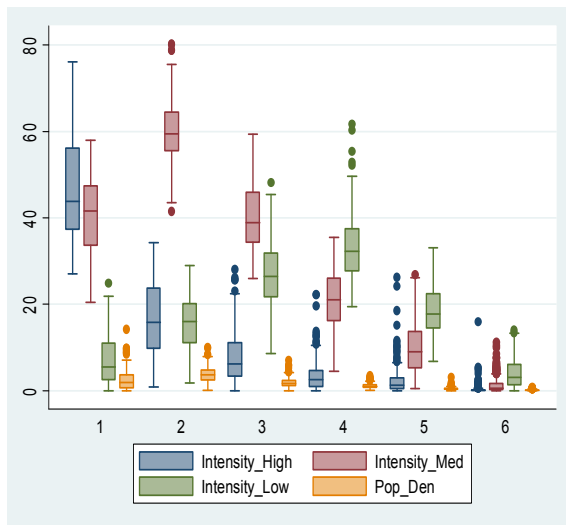
Intersection and segment crash records were collected from the Connecticut Crash Data Repository (CTCDR) (23). As more severe crashes lead to more serious consequence and generate more interest (particularly among the members of the steering committee for this project), only K (fatal injury), A (incapacitating 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 5347 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 A.

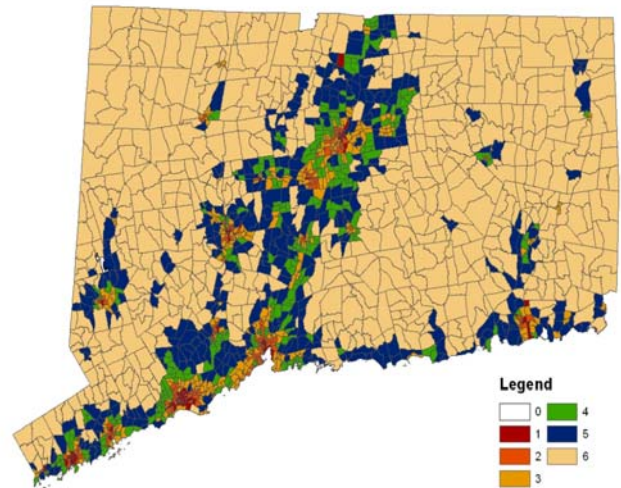
Clustering of TAZs

K-means clustering analysis (24) 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 (25, 26). Different numbers of clusters were respectively tested, and the Calinski and Harabase pseudo-F index (27) 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 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.



(a) land-use intensities and population density distributions



(b) cluster distribution over Connecticut

Figure 1 Clustering Results and Cluster Distribution

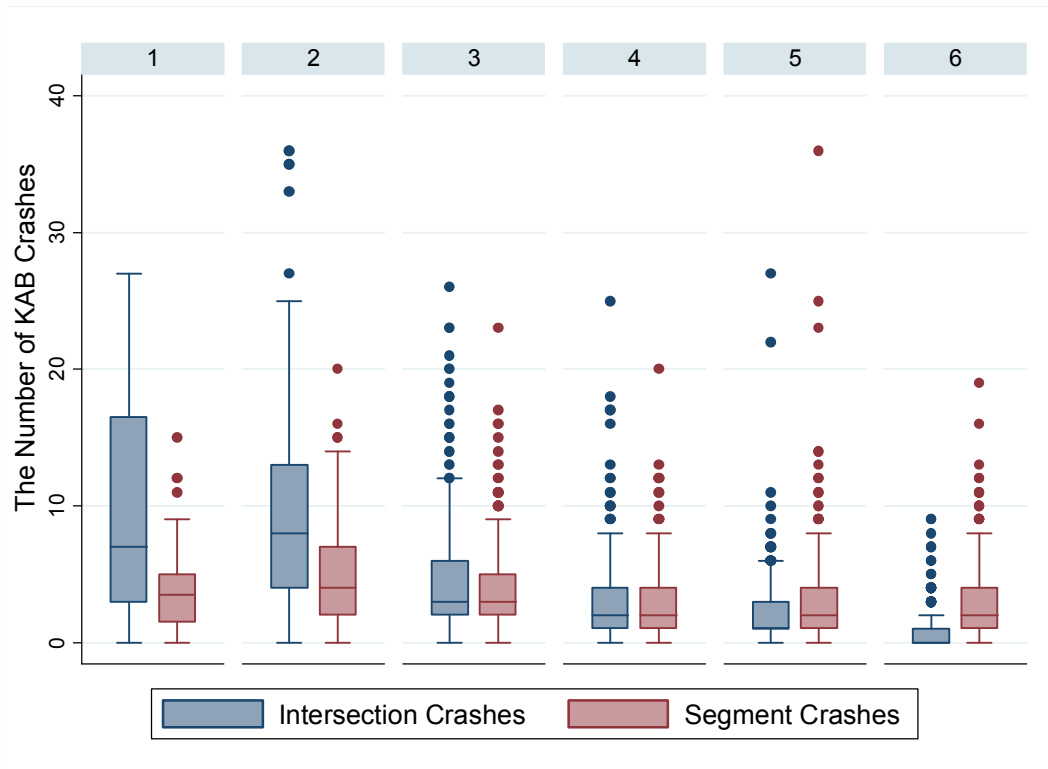


Figure 2 Distributions of KAB Crashes by Cluster

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.

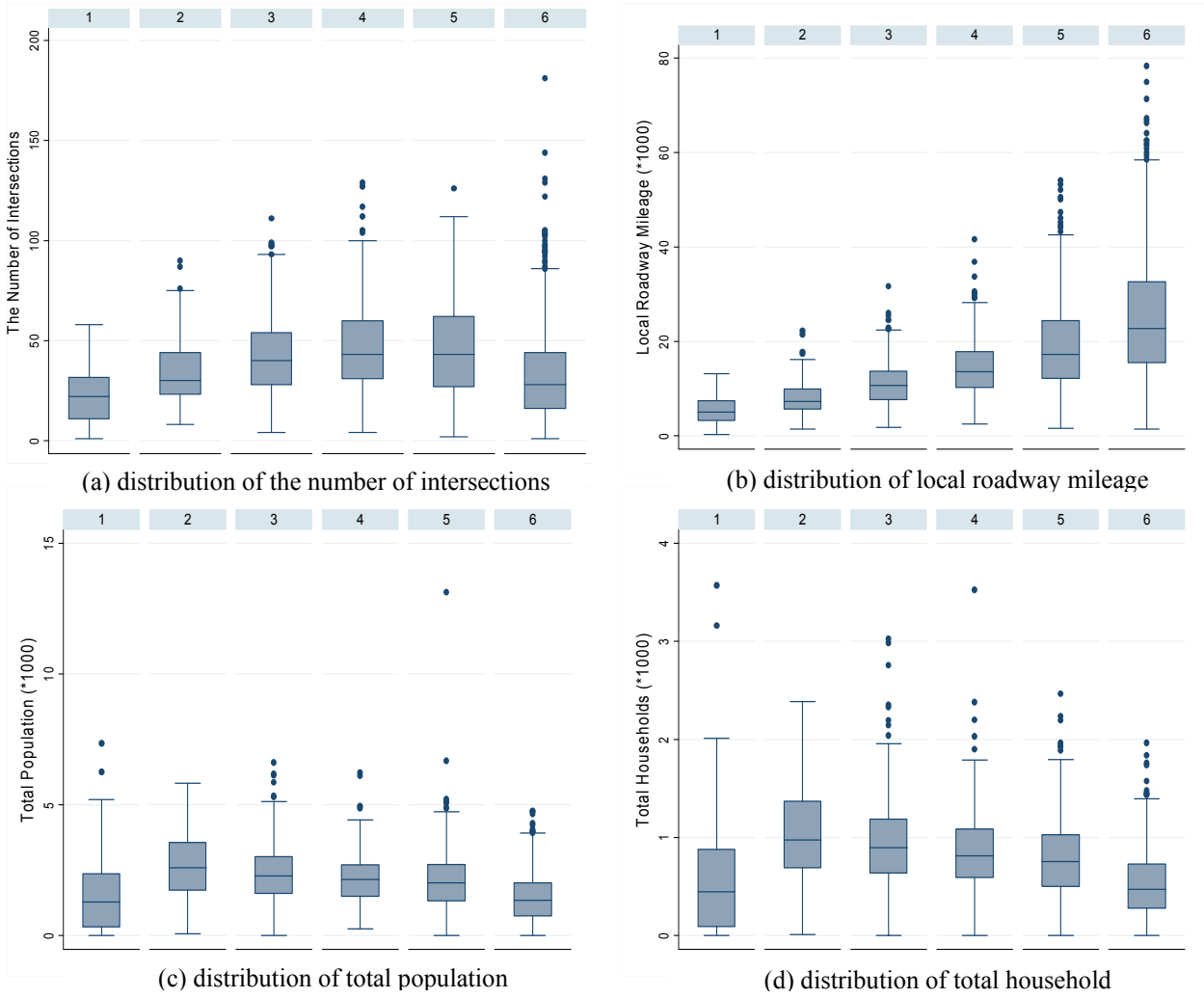


Figure 3 Distributions of Independent Variables by Cluster

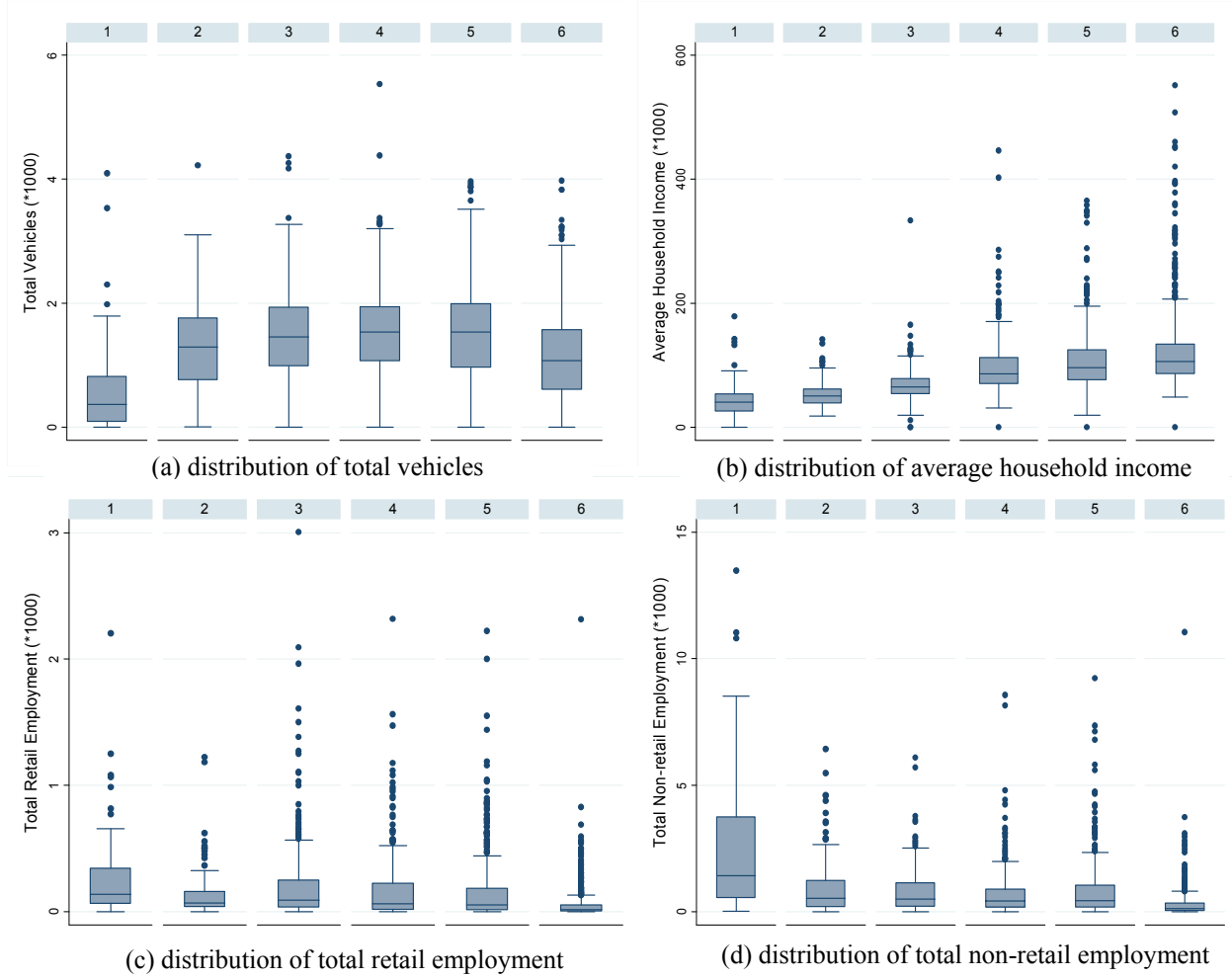


Figure 4 Distributions of Independent Variables by Cluster (Continued)

Statistical Methodology

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 (28):

$$Prob[y_i|\mu_i] = \frac{\exp(-\mu_i)\mu_i^{y_i}}{y_i!} \quad (1)$$

where $Prob[y_i|\mu_i]$ is the probability of y crashes occurring at TAZ i and μ_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 β , the μ_i can be estimated by the equation:

$$\ln(\mu_i) = \beta X_i \quad (2)$$

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

$$Var(y_i) = E(y_i) = \mu_i \quad (3)$$

This constraint might be questionable as the variance of crash data is usually greater than the mean, which is known as over-dispersion (28). The negative binomial regression model addresses this issue, which is derived by rewriting Equation 2 such that:

$$\mu_i = exp(\beta X_i + \varepsilon_i) \quad (4)$$

where $exp(\varepsilon_i)$ is an error term assumed to follow a gamma distribution with mean 1 and variance σ^2 . The distribution of the negative binomial model has the form (28):

$$Prob[y_i|\mu_i] = \frac{\Gamma\left[\left(\frac{1}{\sigma}\right) + y_i\right]}{\Gamma\left(\frac{1}{\sigma}\right) y_i!} \left[\frac{1}{\sigma}\right]^{\frac{1}{\sigma}} \left[\frac{\mu_i}{\left(\frac{1}{\sigma}\right) + \mu_i}\right]^{\mu_i} \quad (5)$$

where Γ is a gamma function and the variance of negative binomial model can be written as follows:

$$Var(y_i) = \mu_i(1 + \sigma\mu_i) = \mu_i + \sigma\mu_i^2 \quad (6)$$

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

$$\mu_{int,i} = Y I_i^{\beta_I} exp(\beta_0 + \beta_P P_i + \beta_R R_i + \beta_N N_i + \beta_V V_i + \beta_C C_i + \beta_H H_i) \quad (7)$$

Where

$\mu_{int,i}$	=	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
β_s	=	the estimated parameters

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

$$\mu_{seg,i} = Y L_i^{\beta_L} exp(\beta_0 + \beta_P P_i + \beta_R R_i + \beta_N N_i + \beta_V V_i + \beta_C C_i + \beta_H H_i) \quad (8)$$

Where

$\mu_{seg,i}$ = **predicted segment crashes in TAZ i**

L_i = 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.* (29). These criteria are calculated as:

$$AIC = 2K - 2 \ln(L) \quad (9)$$

$$BIC = K * \ln(N) - 2\ln(L) \quad (10)$$

$$\text{Mean Absolute Deviation (MAD)} = \frac{1}{N} \sum_{i=1}^N |\hat{y}_i - y_i| \quad (11)$$

$$\text{Mean Squared Predictor Error (MSPE)} = \frac{1}{N} \sum_{i=1}^N (\hat{y}_i - y_i)^2 \quad (12)$$

Where

- K = **the number of estimated parameters**
- L = 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

Cluster SPF	Intersection SPF				Segment SPF			
	Population		Households		Population		Households	
	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC
1	432	448	428	444	330	346	334	350
2	887	908	896	917	692	713	718	739
3	1,231	1,256	1,246	1,271	1,081	1,105	1,109	1,134
4	1,110	1,135	1,120	1,145	1,051	1,075	1,063	1,088
5	1,220	1,247	1,219	1,246	1,475	1,502	1,489	1,516
6	1,247	1,278	1,246	1,277	2,120	2,151	2,125	2,155
Statewide SPF	6,935	6,972	6,977	7,015	6,826	6,863	6,970	7,008

Table 2 displays the SPF performance for the statewide and cluster-based functions using both estimation data and prediction data. Based on the MOEs, the cluster-based SPFs using either population or households are proven to 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

MOEs	Statewide SPF (Population)	Statewide SPF (Households)	Cluster-based SPF (Population)	Cluster-based SPF (Households)
Intersection SPF				
MAD Estimation	2.65	2.72	1.95	1.95
MAD Prediction	2.65	2.74	1.62	1.75
MSPE Estimation	18.25	20.72	11.14	11.29
MSPE Prediction	13.29	14.95	6.41	7.50
Segment SPF				
MAD Estimation	2.00	2.01	1.77	1.87
MAD Prediction	1.52	1.58	1.30	1.47
MSPE Estimation	8.28	9.13	7.55	7.62
MSPE Prediction	4.00	4.48	3.51	3.74

Table 3 shows the coefficient estimates for the intersection SPFs using population as a predictor. Coefficients for all models are provided in the Appendix I. 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 cluster 4, 5 and 6 functions. The amount of non-retail employment is positively associated with the number of intersection crashes in cluster 1, 2 and 6 functions. 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

Variables	Coefficient Estimates by Cluster					
	1	2	3	4	5	6
Intercept	-1.275	0.270	-0.150	-0.984	-2.688	-4.908
	(0.001)	(0.487)	(0.717)	(0.044)	(0.000)	(0.000)
Log (number of intersections)	0.682	0.170	0.078	0.040	0.606	0.844
	(0.000)	(0.225)	(0.587)	(0.810)	(0.000)	(0.000)
Population (*1000)	0.161	0.282	0.360	0.372	0.054	0.129
	(0.014)	(0.000)	(0.000)	(0.000)	(0.368)	(0.145)
Retail employment (*1000)	0.196	-0.295	-0.221	0.462	0.845	0.992
	(0.530)	(0.451)	(0.261)	(0.045)	(0.000)	(0.000)
Non-retail employment (*1000)	0.090	0.182	0.121	-0.003	-0.064	0.174
	(0.003)	(0.000)	(0.072)	(0.966)	(0.195)	(0.008)
Average household income (*1000)	-0.005	-0.013	-0.010	-0.002	-0.003	0.002
	(0.067)	(0.000)	(0.000)	(0.240)	(0.009)	(0.001)
Over dispersion	0.258	0.280	0.422	0.616	0.357	0.227
	(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.

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

Variables	Coefficient Estimates by Cluster					
	1	2	3	4	5	6
Intercept	-3.648	-1.769	-1.300	-1.621	-5.429	-5.946
	(0.008)	(0.213)	(0.305)	(0.265)	(0.000)	(0.000)
Log (roadway length in miles)	0.403	0.248	0.160	0.100	0.539	0.504
	(0.020)	(0.161)	(0.297)	(0.552)	(0.000)	(0.000)
Population (*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)
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)
Non-retail employment (*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)
Average household income (*1000)	-0.003	-0.012	-0.012	-0.003	-0.002	0.001
	(0.327)	(0.001)	(0.000)	(0.027)	(0.009)	(0.015)
Over dispersion	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.

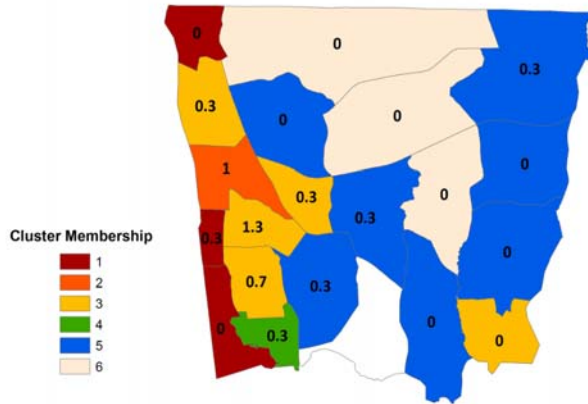
EXAMPLE NETWORK SCREENING APPLICATIONS

As an application exercise, we carried out two screenings using the cluster-based SPFs to predict expected annual crashes and analyze safety in the towns of Stamford and Groton. These towns were chosen because each includes TAZs representing all six clusters, and thus permit application of all six cluster-based SPFs. Here 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). 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 (30). Details about our procedures for applying the EB method and developing the network screening application tool are provided in the Appendix A and Appendix D.

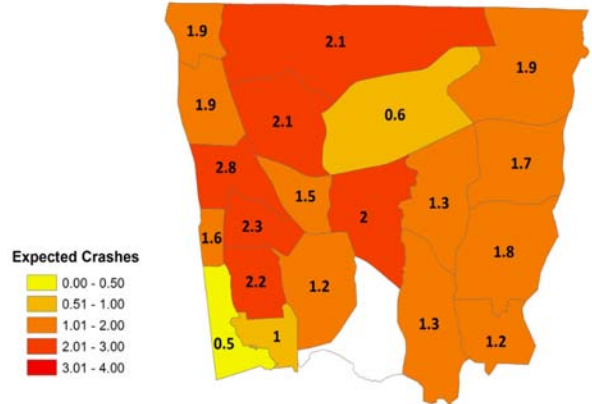
Figure 5 shows the screening analysis results for Groton. Figure 5(a) and 5(c) show the cluster type for each TAZ and the number of observed intersection and segment crashes in each TAZ, respectively. Note there are some decimal observed crashes shown in Figure 5(a) and 5(c). This is because when we allocated crashes, if the crash occurred at the boundary of more than one TAZ, it was evenly allocated among the TAZs. The areas with higher land-use intensities and high population density are mainly located in the western parts of the town where the US submarine base and CBD are located. The areas with lower land-use intensities and lower population density are primarily located in the north central parts of the town. The number of observed crashes for all TAZs in Groton is very low. Figure 5(b) and 5(d) respectively show the expected number of intersection and segment crashes. It is clear that the expected crash distribution is quite different from the observed distribution, which indicates the importance of using the EB method to avoid making decisions on the basis of spurious crash count observations.

Figure 6 shows the results of the network screening for Stamford. The areas with higher land-use intensities and higher population density are mainly located in the southern parts of the town. The TAZs with higher number of observed intersection and segment crashes are mainly located in the southern and middle parts, and the TAZs with higher number of expected intersection and segment crashes are mainly located in the middle and northern parts of the town. Again, the expected crash distribution is quite different from the observed distribution.

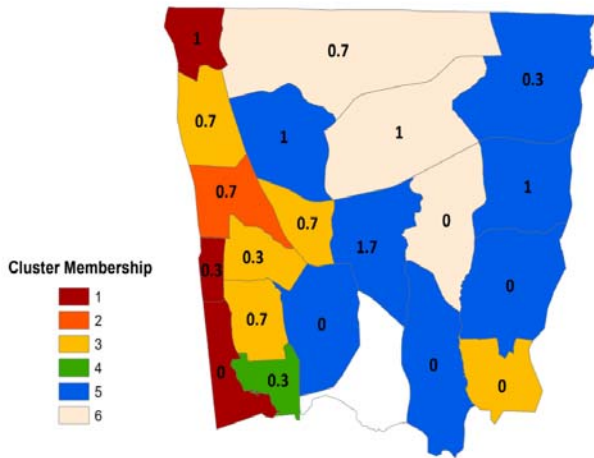
(a) Number of Observed Intersection Crashes by TAZ Groton, CT



(b) Number of Expected Intersection Crashes by TAZ Groton, CT



(c) Number of Observed Segment Crashes by TAZ Groton, CT



(d) Number of Expected Segment Crashes by TAZ Groton, CT

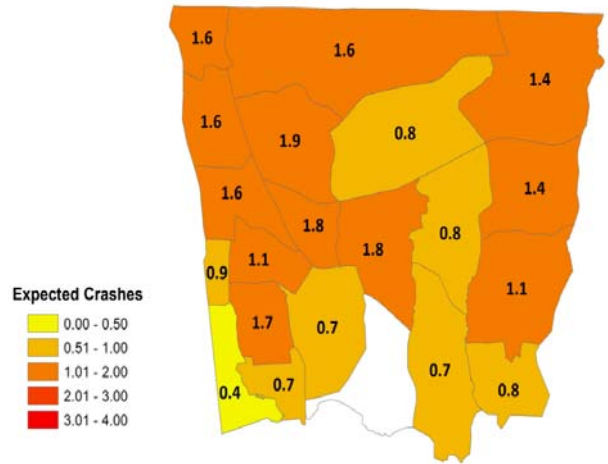
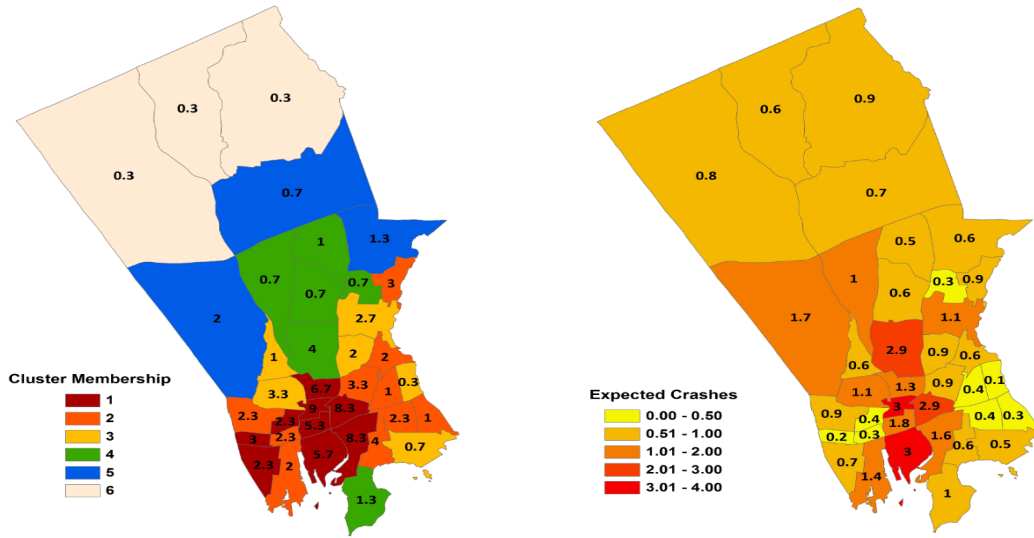


Figure 5 Example Network Screening Application – Groton

(a) Number of Observed Intersection Crashes by TAZ Stamford, CT (b) Number of Expected Intersection Crashes by TAZ Stamford, CT



(c) Number of Observed Segment Crashes by TAZ Stamford, CT (d) Number of Expected Segment Crashes by TAZ Stamford, CT

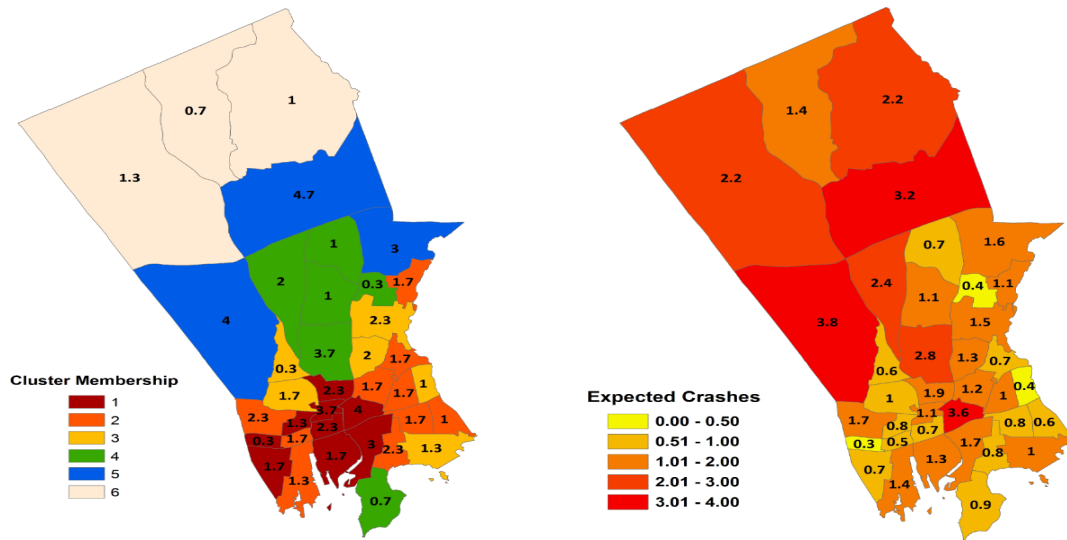


Figure 6 Example Network Screening Application – Stamford

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 to the towns of Stamford and Groton as a network screening tool. In Groton, the TAZs with higher number of expected intersection and segment crashes are mainly located in the middle and northern parts of the town. In Stamford, the TAZs with higher number of expected intersection and segment crashes are mainly located in the middle and northern parts of the town. It is anticipated that the example applications can help local agencies develop cost-effective countermeasures to improve safety for local roads by identifying the areas of town in which to focus safety improvement projects.

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 roadway network and economic development. 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 following the procedure documented in the Appendix A.

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 estimates the new SPFs to improve the prediction accuracy. Additionally, crash counts might vary from TAZs with small geographical size to large ones. Some analysts might want to investigate crash rates by TAZ area for a normalized comparison. The process to calculate crash rates is provided in Appendix E.

REFERENCES

1. Highway Safety Manual, 1st Edition, American Association of State Highway and Transportation Officials, Washington D.C., 2010.
2. Jonsson, T., Ivan, J., and Zhang, C. Crash prediction models for intersections on rural multilane highways: differences by collision type. In *Transportation Research Record: Journal of the Transportation Research Board*, No. 2019, Transportation Research Board of the National Academies, Washington, D.C., 2007, pp. 91-98.
3. Ceifetz, A., Bagdade, J., Nabors, D., Sawyer, M., and Eccles, K. *Developing safety plans: A manual for local rural road owner*. Project Report, Project 12-017, Federal Highway Administration (FHWA), March 2012.
4. Souleyrette, R., Caputcu, M., Cook, D., McDonald, T., Sperry, R., and Hans, Z. *Safety Analysis of Low-Volume Rural Roads in Iowa*, Final Report, Project 07-309, Institute for Transportation, Iowa State University, Dec. 2010.
5. Ivan, J. New approach for including traffic volumes in crash rate analysis and forecasting. In *Transportation Research Record: Journal of the Transportation Research Board*, No. 1897, Transportation Research Board of the National Academies, Washington, D.C., 2004, pp. 134-141.
6. Vogt, A. Crash Models For Rural Intersections: Four-Lane by Two-Lane Stop-Controlled and Two-Lane by Two-Lane Signalized US Department of Transportation, Federal Highway Administration Report, FHWA-RD-99-128, 1999.
7. Vogt, A., and Bared, J. Accident Models for Two-Lane Rural Segments and Intersections. In *Transportation Research Record: Journal of the Transportation Research Board*, No. 1635, Transportation Research Board of the National Academies, Washington, D.C., 1998, pp. 18-29.
8. Oh, J., S. Washington, and K. Choi. Development of Accident Prediction Models for Rural Highway Intersections. In *Transportation Research Record: Journal of the Transportation Research Board*, No. 1897, Transportation Research Board of the National Academies, Washington, D.C., 2004, pp. 18-27.
9. S. Bindra, J. Ivan and T. Jonsson, *Predicting Segment-Intersection Crashes with Land Development Data*. In *Transportation Research Record: Journal of the Transportation Research Board*, No. 2102, Transportation Research Board of the National Academies, Washington, D.C., 2009, pp. 9-17.
10. Pirdavani, A., Brijs, T., Bellemans, T., Kochan, B., and Wets, G. Application of Different Exposure Measures in Development of Planning-level Zonal Crash Prediction Models. In *Transportation Research Record: Journal of the Transportation Research Board*, No. 2280, Transportation Research Board of the National Academies, Washington, D.C., 2012, pp. 145-153.
11. Levin, N., K. Kim and L. Nitz. Spatial Analysis of Honolulu Motor Vehicle Crashes. II. Zonal Generators. *Accident Analysis and Prevention*. Vol. 50. pp. 678-687. 2013.
12. Pulugurtha S., V. R, Duddu and Y. Kotagiri. Traffic Analysis Zone Level Crash Estimation Models Based on Land Use Characteristics. *Accident Analysis and Prevention*, Vol. 36, No. 6, 2004, pp. 973-984.

13. Ladron de Guevara, F., S. P. Washington, and J. Oh. Forecasting Crashes at the Planning Level: Simultaneous Negative Binomial Crash Model Applied in Tucson, Arizona. In *Transportation Research Record: Journal of the Transportation Research Board*, No. 1897, Transportation Research Board of the National Academies, Washington, D.C., 2004, pp. 191–199.
14. Lovegrove, G. R., and T. Sayed. *Macro-level Collision Prediction Models for Evaluating Neighborhood Traffic Safety*. Canadian Journal of Civil Engineering. Vol. 33. No. 5. pp 609-621. 2006.
15. Lovegrove, G. *Road Safety Planning, New Tools for Sustainable Road Safety and Community Development*. AV Akademikerverlag. Berlin. 2012.
16. Hadayeghi, A., A. Shalaby and B. Persaud. *Macro-level Accident Prediction Models for Evaluating the Safety of Urban Transportation System*. Transportation Research Board. National Research Council. Washington, D.C..2003.
17. Khondakar, B. T. Sayed and G. Lovegrove. *Transferability of Community-based Collision Prediction Models for Use in Road Safety Planning Applications*. Journal of Transportation Engineering. Vol. 136. No. 10. pp. 871-880. 2010.
18. Noland, R. B., and M. A. Quddus. A Spatially Disaggregate Analysis of Road Casualties in England. *Accident Analysis and Prevention*, Vol. 36, No. 6, 2004, pp. 973–984.
19. Washington, S., Van Schalkwyk, I., Mitra, S., Meyer, M., Dumbaugh, E., Zoll, M. Incorporating Safety into Long-Range Transportation Planning. NCHRP Report 546, National Cooperative Highway Research Program, Transportation Research Board, Washington DC, 2006.
20. United States Census Bureau 2010.
<https://www.census.gov/geo/maps-data/data/tiger-line.html>
21. CTPP 2010, Census Transportation Planning Package Database.
<http://ctpp.transportation.org/Pages/5-Year-Data.aspx>
22. National Land Cover Database 2011 (NLCD 2011). http://www.mrlc.gov/nlcd11_data.php
23. Connecticut Crash Data Repository. <http://www.ctcrash.uconn.edu/>
24. STATA. Clustering Kmeans and Kmedians. Release 12. A Stata Press, StataCorp LP. College Station, Texas, 2011.
<http://www.stata.com/manuals13/myclusterkmeansandkmedians.pdf>.
25. Hair L., R. Anderson, R. Tatham and W. Black. *Multivariate Data Analysis*. Prentive Hall. 1998.
26. Mohamed M., N. Saunier, L. Miranda-Moreno and S. Ukkusuri. *A Clustering Regression Approach: A Comprehensive Injury Severity Analysis of Pedestrian-Vehicle Crashes in New York, US and Montreal, Canada*. Safety Science. Vol. 54, pp. 27-37. 2013.
27. Calinski T. and J. Harabasz. *A Dendriter Method for Cluster Analysis*. Communications in Statistics. Vol. 3, pp. 1-27. 1974.
28. Washington, S., M. Karlaftis, F. L. Mannering. *Statistical and Econometric Methods for Transportation Data Analysis, 2nd ed*. Chapman and Hall/CRC, Boca Raton, FL. 2011.
29. Oh, J., C. Lyon, S. Washington, B. Persaud and J. Bared. *Validation of FHWA Crash Models for Rural Intersections: Lessons Learned*. In *Transportation Research Record: Journal of the Transportation Research Board*. No. 1840, Transportation Research Board of the National Academies, Washington, D.C., pp. 41-49. 2003.
30. Hauer, E., D. W. Harwood, F. M. Council and M. S. Griffith. *The Empirical Bayes Method for Estimating Safety: A Tutorial*. In *Transportation Research Record*. No. 1784. pp.126-131. 2002.

APPENDICES

Appendix A – Data Collection, Compilation and Analysis Procedures

Appendix B – Comprehensive SPF Estimation Results

Appendix C – List of Data Sets Used for Analysis

Appendix D – Instructions for Use of Visualization Tool

Appendix E – Instructions for Computation of Crash Rates by TAZ Area

Appendix A Data Collection, Compilation and Analysis Procedures

Figure A-1 presents a flow chart of the process of collecting, compiling and analyzing the data for the project. The Roman numerals and capital letters indicate the sections of this Appendix where each section is covered in detail.

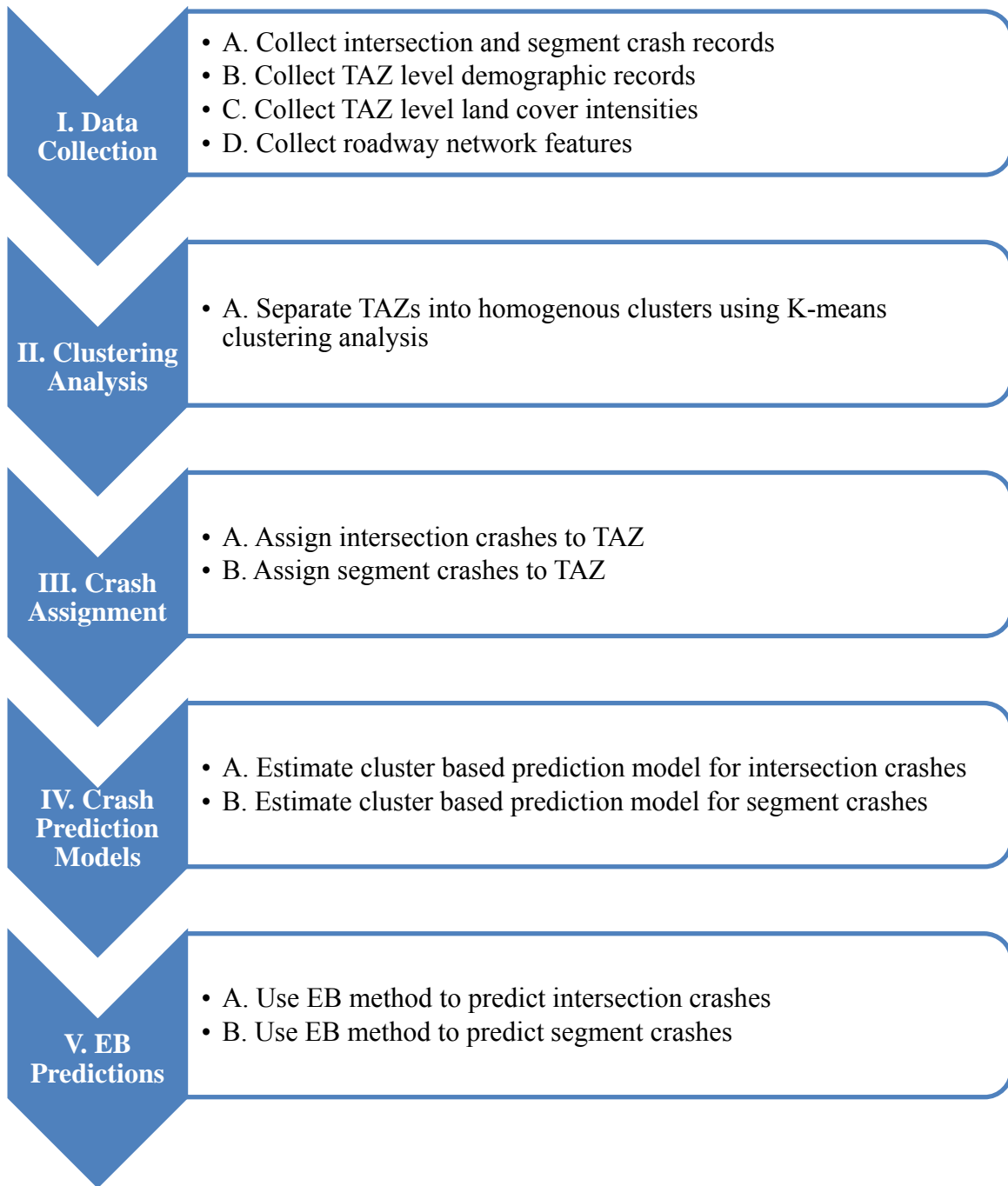


Figure A-1. Data Collection, Compilation and Analysis Flow Chart

I DATA COLLECTION

I.A Collect Intersection and Segment Crash Records

Intersection and segment crash data were collected from the Connecticut Crash Data Repository (CTCDR) <http://www.ctcrash.uconn.edu/>. As more severe crashes lead to more serious consequence, and generate more interest (particularly among the members of the steering committee for this project), only the type K (fatal injury), A (incapacitating injury) and B (non-incapacitating injury) crashes occurring on any roads under local jurisdiction in Connecticut from 2010 to 2012 were considered. In total, 5403 intersection crashes and 5347 segment crashes were extracted and used in estimating the crash prediction models. The intersection crash data includes road name and intersection name; the segment crash data includes road name, milepost and the nearest intersection name.

I.B Collect TAZ Level Demographic Records

TAZ level demographic data includes population, retail and non-retail employment, households, vehicles and household income. All of these variables are summarized by TAZ, and are used as the independent variables in crash prediction models. The demographic variables were collected from the Census Transportation Planning Package Database (CTPP 2010). 1806 TAZs were defined for the state of Connecticut in 2010. Two of these TAZs were apparently defined to represent special generators, and have no population or employment, so they were eliminated from the analysis and the remaining 1804 TAZs were used to estimate the crash prediction models.

I.C Collect TAZ Level Land Cover Intensities

Land cover data was acquired from the 2011 National Land Cover Database (Jin *et al.* 2013), which classifies each pixel in a Landsat image acquired at a spatial resolution of 30 meters into one of eighteen land-cover categories. Three land-cover classes were of interest: a) low intensity developed – single family housing, less than 50% impervious surface; b) medium intensity developed – single-family housing, between 50-80% impervious surface; and c) high intensity developed – apartment complexes, commercial and industrial areas, greater than 80% impervious surface. Land cover intensities were determined for all three land-cover classes by calculating their areal percentages within each TAZ.

I.D Collect Roadway Network Features

The 2010 TIGER/Line shapefiles for Connecticut were extracted from the United States Census Bureau (U.S. Census Bureau 2012). All roads not under local jurisdiction were removed to produce a new file consisting of city streets, neighborhood roads, and rural roads (MTFCC code = S1400). TIGER/Line shapefiles contain both spatial and attribute errors – for example, incorrect or missing roadway names, inaccurate spatial location of roadway features, missing roadways. Major errors were manually identified and edited for each town using ArcGIS 10 (ESRI 2010). Corrections included re-aligning roadway features (*e.g.*, extending roadways that should intersect but did not), adding missing roadway names, adjusting mislabeled roadway names (*i.e.*, roadways with incorrect names), and editing roadway names to ensure consistency (*e.g.*, 7th Street and Seventh

Street → 7th Street). The final editing step merged roadway line segments that shared coincident endpoints and roadway name to produce a single roadway feature for each roadway under local jurisdiction.

Two roadway attributes were determined using the resulting shapefile: number of intersections and total length of roadways under local jurisdiction for each TAZ. Only named roadways were considered when calculating total length, as unnamed roadways included private driveways and private roads for which crash records would not be available. First, the roadway and TAZ shapefiles were overlaid using an intersect operation to split roadway features at TAZ boundaries. This ensured only the length of the roadway segment falling within the TAZ boundary, and not the entire length of the roadway feature, was included in the summation. Unfortunately, the intersection operation also created duplicate line segments for all roadways that fell along the border of two TAZs (*e.g.*, roadway A intersected with TAZ 1 and roadway A intersected with TAZ 2). These roadways were identified using a spatial selection operation (*i.e.*, select all roadway features that share a line segment with TAZ features). A new field was added to the attribute table such that all selected roadway features were assigned a value of 2. This field was used to divide the length of the shared roadways in half in order to proportionally allocate the roadway feature's length among the two TAZs. The final step entailed summarizing the total length of roadways by TAZ.

Two intersection attributes were calculated for each TAZ: number of intersections when considering named roadways only and number of intersections when considering both named and unnamed roadways. The processing steps were the same in both calculations – only the selected roadway features differed (named roadways *vs.* all roadways). First, the roadway shape file was intersected with itself. This created a point feature at each location where one roadway feature intersected a second roadway feature. Unfortunately, this created multiple points at each intersection (*e.g.*, A intersected with B and B intersected with A). To remove the duplicate intersections, the x y coordinates of each point were added to the intersection attribute table. A dissolve operation was then run to remove all duplicate points (*i.e.*, points that shared the same coordinate pair). As noted above, roadways often fell along the boundary of TAZs, which meant that multiple intersections were also located on the boundary of two or more TAZs. An additional processing step was needed to proportionally allocate intersections among TAZs. A spatial join operation was used to join TAZs to each intersection. For intersections falling along the border of two or more TAZs, the spatial join operation records the number of TAZs connected to each intersection. The resulting field was used to proportionally allocate each intersection (*e.g.*, if spatial join returned a value of 3, the intersection was assigned a value of $\frac{1}{3}$). The final step entailed summarizing the total number of intersections found within each TAZ.

II CLUSTERING ANALYSIS

II.A Separate TAZs into Homogenous Clusters

K-means clustering analysis, sometimes referred to as portioning-based or objective function-based clustering approach, defines an objective distance function (e.g. Euclidean distance or Canberra distance), and categorizes the data by optimizing this objective function (STATA 2011). To select the optimum number of clusters in K-means clustering analysis, different numbers of clusters should be respectively tested, and the Calinski and Harabase pseudo-F index (Calinski and Harabase 1974) are used to determine the final number of clusters. The larger the Calinski and Harabase pseudo-F index, the more accurate is the clustering analysis. Figure A-2 shows that the optimum number of clusters was found to be six as this number achieved the highest value (2464) of the Calinski and Harabase pseudo-F index. Table II.1 describes the distribution interval of four variables - low intensity, moderate intensity, high intensity and population density in each cluster type.

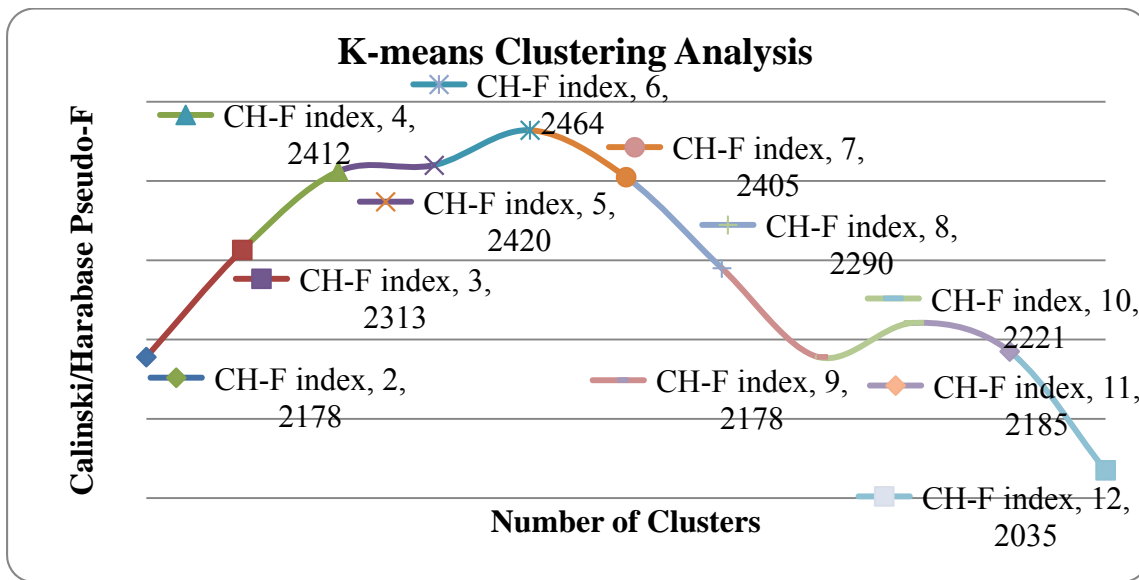


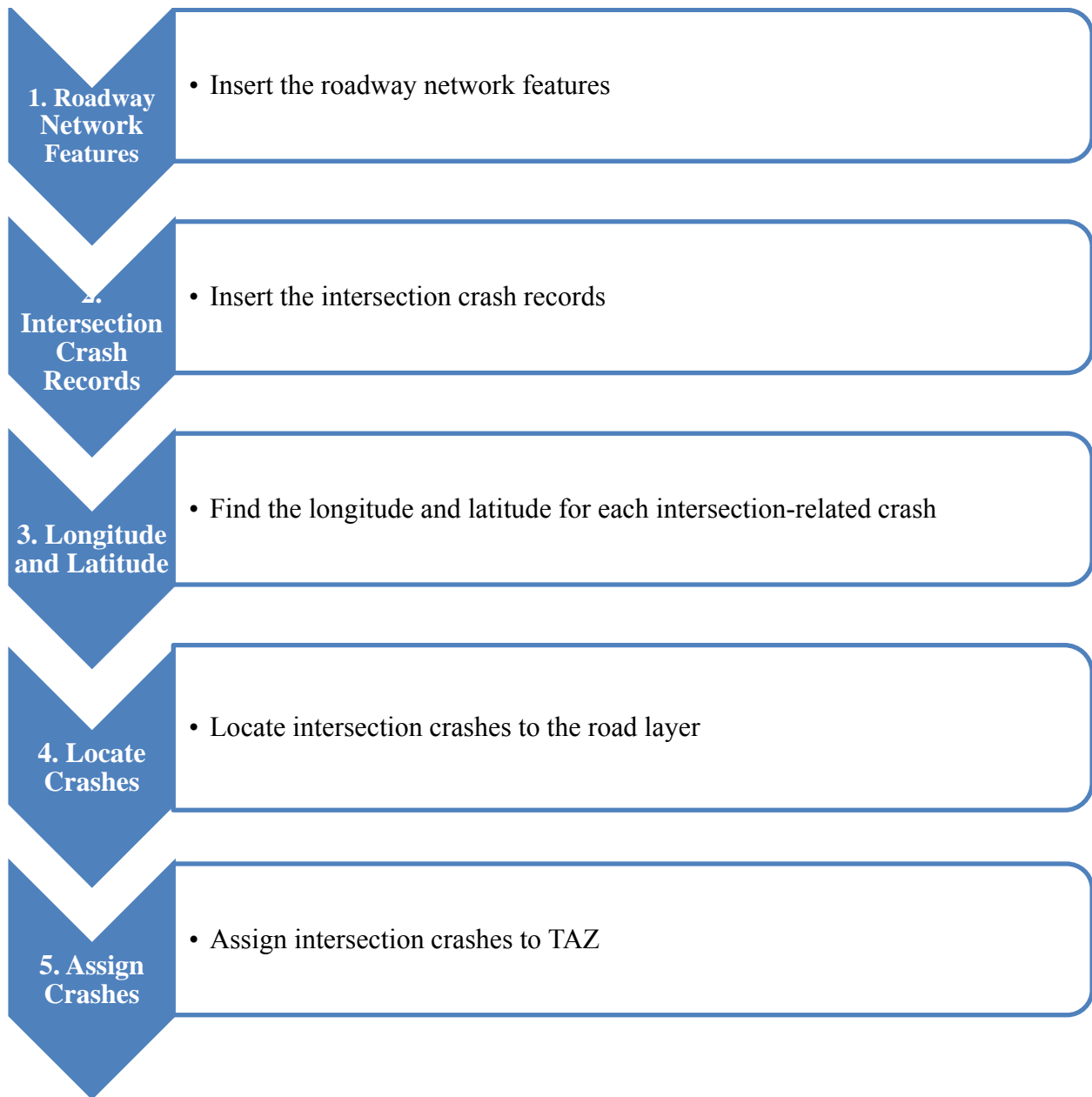
Figure A-2 K-mean Clustering Analysis Evaluation

Table A.1 Interval Values of Each Clustering Variable by Cluster

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
Low Intensity (%)	[0, 24.8]	[1.7, 29.0]	[8.6, 48.1]	[19.4, 61.7]	[6.8, 33.1]	[0, 14.0]
Moderate Intensity (%)	[20.4, 57.9]	[41.5, 80.2]	[25.9, 59.4]	[4.5, 35.4]	[0.5, 26.8]	[0, 11.2]
High Intensity (%)	[27.1, 76.1]	[0.9, 34.3]	[0, 28.1]	[0, 22.3]	[0, 26.3]	[0, 15.9]
Population Density (per km ²)	[0, 14.2]	[0.1, 10.0]	[0, 7.1]	[0.1, 3.5]	[0, 3.1]	[0, 0.8]

III CRASH ASSIGNMENT

III.A Assign Intersection Crashes to TAZ



III.B Insert the Roadway Network Features

See Section I.D

III.C Insert the Intersection Crash Records

See Section I.A

III.D Find the Longitude and Latitude for each Intersection-Related Crash

For the intersection crashes, an approximation of the longitude and latitude was achieved by inputting the road name, intersection name, town name and State for those crashes (*e.g.* North Eagleville Road and Bone Mill Road, Mansfield, CT) in Google Map API <http://www.gpsvisualizer.com/geocoder/>. Crashes whose longitude and latitude could not be automatically identified from the Google Map API were manually identified using Google Earth. Google Earth reports longitude and latitude based on the World Geodetic System of 1984 (WGS84) datum.

III.E Locate Intersection Crashes to the Road Layer

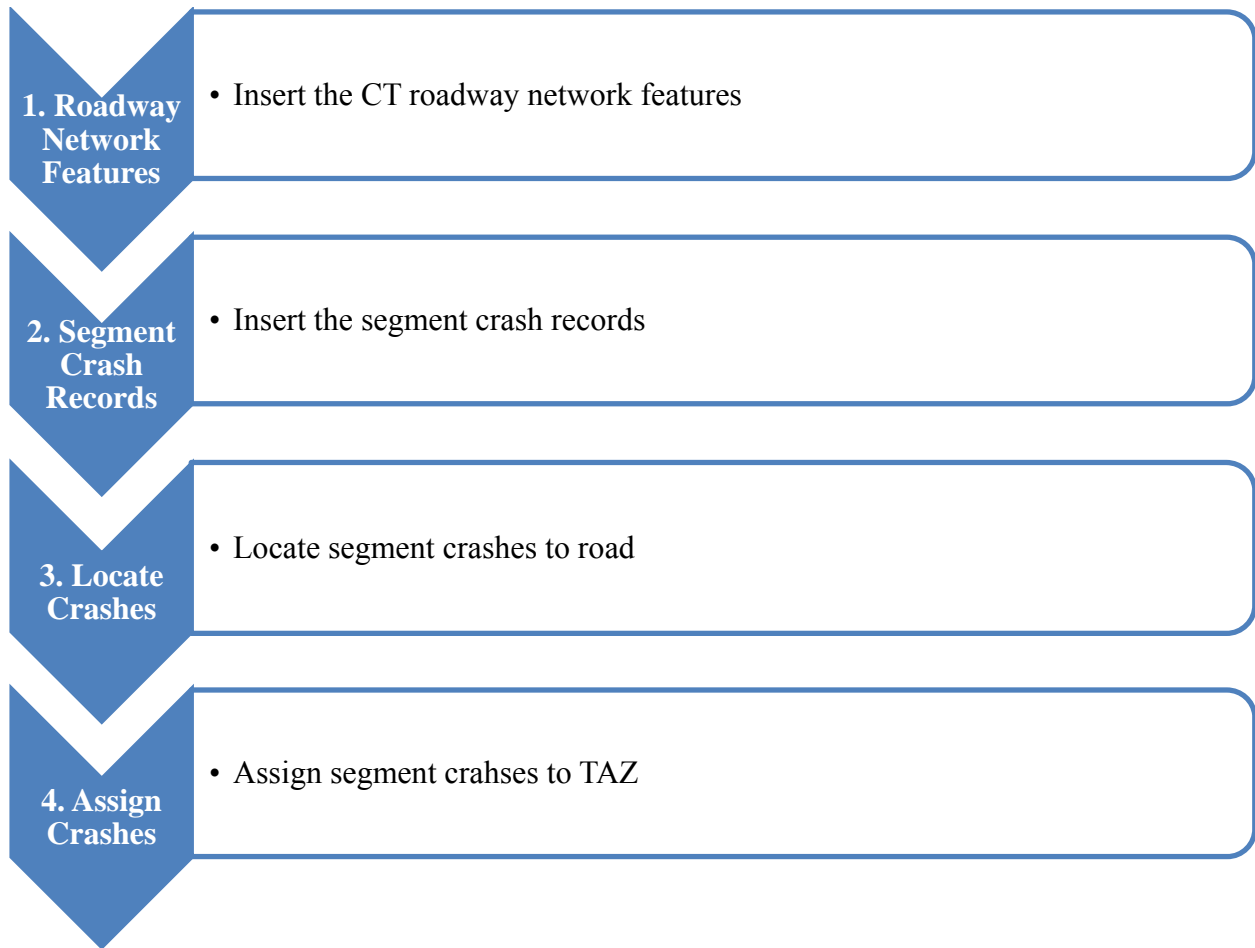
The resulting table of intersection crashes was added to ArcGIS using the longitude and latitude coordinates. This produced a new point feature shapefile, where each point represented a unique intersection crash. The intersection crash shapefile was re-projected to match the spatial reference system associated with the roadway and TAZ shapefiles (State Plane Coordinate System using the North American Datum of 1983). All geospatial data have some degree of positional inaccuracy (location of the geographic feature in a database compared to its true location on the surface of the earth). This is true for both the roadway features in the TIGER/Line shapefile and the roadway features in Google Maps. This meant that not all intersection crashes geocoded using the Google Map API occurred at the intersection of two roadway features when the intersection crash shapefile was overlaid on the TIGER/Line shapefile. To ensure all intersection crashes were located at the intersection of two roadway features, the intersection crash shapefile was edited such that all intersection crash points were moved to the nearest roadway feature endpoint using a snap editing tool. Most intersection crash points that required editing were moved less than 100 feet. Finally, we randomly selected a few intersection crashes, and check their locations on the TIGER/Line shapefile with the locations shown in crash data to verify the process of locating crashes is accurate.

III.E.1 Assign Intersection Crashes to TAZ

Similar to the calculation of the number of intersections per TAZ, the number of intersection crashes per TAZ necessitated an additional processing step to proportionally allocate intersection crashes that were located on the boundary of two or more TAZs. A spatial join operation was used to join TAZs to each intersection crash, which resulted in a count of the number of TAZs associated with each crash point. If an intersection crash occurred completely within a TAZ, the spatial join

returned a value of 1. If an intersection crash occurred on boundary of two or more TAZs, the spatial join returned a value of 2 (or 3 or 4). The resulting spatial join count field was used to proportionally allocate each intersection crash (*e.g.*, a value of 2 means the intersection crash was assigned a value of ½ and the intersection crash was divided between the two associated TAZs). The final step entailed summarizing the total number of intersection crashes within each TAZ.

III.E.2 Assign Segment-related Crashes to TAZ



III.E.3 Insert the Roadway Network Features

See Section to I.D.

III.E.4 Insert the Segment Crash Records

See Section I.A.

III.E.5 Locate Segment Crashes to Road

III.E.5.1 Associating Roadway Features with TAZs

In order to assign each segment crash to an individual TAZ, the individual roadway features (*i.e.*, segments along which the crash occurred) must be associated with TAZ(s). The shapefile created by intersecting roadway features and TAZ features (see section I.D.) served as the base file for this process. Within this file, all roadway features that intersect multiple TAZs are split into two line segments at the boundary of the TAZs, and multiple line segments are created for roadway features that fall along the boundary of two TAZs (*i.e.*, one line segment for each TAZ). The line segments within the intersected shapefile contain both roadway attribute information and TAZ attribute information. Two merge operations were performed using the roadway feature identification (FID) number, TAZ feature identification (FID) number, and roadway name. First, a merge was performed to combine all roadway features that shared the same roadway FID and TAZ FID. This created a single line feature for all roadways that had a segment completely within a TAZ and a segment along the border of the TAZ. Second, a merge was performed to combine all roadway features that shared coincident endpoints, roadway name, and TAZ FID. This created a single line feature representing a unique roadway within each TAZ. The attribute table of resulting merged shapefile was exported for use in Microsoft Excel.

As previously noted, several roadways under local jurisdiction occur along the boundary of two or more TAZs. For crashes occurring on these segments, proportional allocation was used to assign crashes to each TAZ (*e.g.*, if the crash occurred on a roadway that bordered two TAZs, each TAZ was assigned $\frac{1}{2}$ of the crash). To identify roadways occurring along multiple TAZs, a consolidation procedure was performed. First, a new column was created in the database by concatenating the roadway FID number with the roadway name. This was done to ensure that roadways that shared the same name but had different FID numbers were treated as individual roadways (*e.g.*, there were several roadway features named Main Street across the state). The consolidation procedure was performed to count the number of TAZs associated with each unique roadway name and FID combination. Approximately 85% of all roadways under local jurisdiction were associated with a single TAZ, while the remaining 15% were associated with two or more TAZs.

III.E.5.2 Assigning Segment Crashes to Roads

For segment crashes occurring on roadways associated with a single TAZ, the crash was assigned to that TAZ. For segment crashes occurring on roadways associated with multiple TAZs, the procedure described below was used to create a point feature shapefile that correctly located each segment crash along the length of the roadway in order to identify the TAZs associated with the crash for proportional allocation.

First, we built a route layer for the Connecticut roadway network features, using the command 'Create Routes' under the 'Linear Referencing Tools' of ArcGIS. Then we inserted the crash data, and located all crashes that occurred in the roads associated with more than one TAZ to the route layer using the command 'Make Route Event Layer' under the 'Linear Reference Tools' of ArcGIS, based on the information of roadway name and milepost.

Next, it was necessary to translate the crash location from route and milepost to a geolocation on a segment. Because we did not know which end of the road corresponded to milepost “0”, we used the following steps to do this:

1. Add a buffer with 0.3-mile radius to each crash that occurred on a road associated with more than one TAZ, using the command ‘Buffer’ under the ‘Analysis Tools’ of ArcGIS.
2. Use the command ‘Spatial Join’ under the ‘Analysis Tools’ to find all roads that the buffer intersects with.
3. Compare the road name of the nearest intersection for each crash with all the roads that the buffer of the crash intersects with.
4. Select all observations where the name of the nearest intersection matches none of the roads with which the buffer intersects.
5. In the crash records, recalculate the new mileposts of all selected crashes in last step. The new milepost of each crash was calculated as the total length of the roadway where the crash occurred minus the original milepost of the crash.
6. Locate all segment crashes to the route layer with the updated segment-related crash records, using the command ‘Make Route Event Layer’ under the ‘Linear Reference Tools’ of ArcGIS, based on the information of roadway name and milepost.
7. We randomly selected a few segment crashes, and check their locations on the TIGER/Line shapefile with the locations shown in crash records to verify the process of locating crashes is accurate.

III.E.6 Assign Segment Crashes to TAZ

See Section III.A.5.

IV CRASH PREDICTION MODELS

IV.A Estimate Cluster Based Prediction Model for Intersection Crashes

Details are provided in the Final Report.

IV.B Estimate Cluster Based Prediction Model for Segment Crashes

Details are provided in the Final Report.

V EB PREDICTIONS

In order to properly evaluate the safety of any roadway location, it is necessary to estimate the long run expected crash count before comparing it with other locations. This is because any crash count is just a single observation, and is not necessarily the average or expected count at the location. For example, a location with a high crash count could in the next year observe a much smaller crash count, or vice versa, just due to random fluctuations in crash counts from year to year. This phenomenon is known as “Regression to the mean (RTM)”, and failing to account for it could lead to serious bias in the estimates and corresponding analysis errors.

To avoid RTM bias, the Empirical Bayesian (EB) prediction method was used, because instead just predicting the crash counts for a location using the crash prediction models, or only referring to the observed crash counts for the location, the EB method uses Bayesian statistics to estimate the long run expected crash counts by combining the predicted crashes from crash prediction models with the observed crash counts. It significantly increases the precision of predictions for the future when only limited historical crash data is available, and it corrects the RTM bias (Hauer *et al.* 2002). To apply the EB method, we calculated the predicted number of crashes using the cluster-based models, and then estimated the expected number of crashes using the Empirical Bayesian (EB) method as prescribed in the HSM (HSM, 2010), as follows:

V.A Use EB Method to Predict Expected Intersection Crashes

Equation V.1 and V.2 are used directly to estimate the expected intersection crash frequency for a TAZ by combining the predicted crash counts with the observed crash counts.

$$N_{i,expected,int} = w_{i,int} \times N_{i,predicted,int} + (1 - w_{i,int}) \times N_{i,observed,int} \quad (V.1)$$

$$w_{i,int} = \frac{1}{1 + k_{i,int} \times (\sum_{all\ study\ years} N_{i,predicted,int})} \quad (V.2)$$

Where

$N_{i,expected,int}$ = estimate of expected intersection crash frequency for the study period in TAZ i

$N_{i,predicted,int}$ = estimate of predicted intersection crash frequency for the study period in TAZ i

$N_{i,observed,int}$ = observed intersection crash frequency for the study period in TAZ i

$w_{i,int}$ = weighted adjustment for the EB intersection prediction in TAZ i

$k_{i,int}$ = over-dispersion parameter in the intersection crash prediction model for TAZ i

V.B Use EB Method to Predict Expected Segment Crashes

Equation V.3 and V.4 are used directly to estimate the expected segment crash frequency for a TAZ by combining the predicted crash counts with the observed crash counts.

$$N_{i,expected,seg} = w_{i,seg} \times N_{i,predicted,seg} + (1 - w_{i,seg}) \times N_{i,observed,seg} \quad (V.3)$$

$$w_{i,seg} = \frac{1}{1+k_{i,seg} \times (\sum_{all\ study\ years} N_{i,predicted,seg})} \quad (V.4)$$

Where

- $N_{i,expected,seg}$ = estimate of expected segment crash frequency for the study period in TAZ i
- $N_{i,predicted,seg}$ = estimate of predicted segment crash frequency for the study period in TAZ i
- $N_{i,observed,seg}$ = observed segment crash frequency for the study period in TAZ i
- $w_{i,seg}$ = weighted adjustment for the EB segment prediction in TAZ i
- $k_{i,seg}$ = over-dispersion parameter in the segment crash prediction model for TAZ i

References

Calinski T. and J. Harabasz. A Dendriter Method for Cluster Analysis. Communications in Statistics. Vol. 3, pp. 1-27. 1974.

ESRI (Environmental Systems Resource Institute). 2010. ArcGIS 10. Redlands, California.

Jin, S., Yang, L., Danielson, P., Homer, C., Fry, J. and G. Xian. A comprehensive change detection method for updating the National Land cover Database to circa 2011. Remote Sensing of Environment. Vol. 132, pp. 159-175. 2013.

STATA. Clustering Kmeans and Kmedians. Release 12. A Stata Press, StataCorp LP. College Station, Texas, 2011. <http://www.stata.com/manuals13/mvclusterkmeansandkmedians.pdf>.

U.S. Census Bureau., 2010 TIGER/Line Shapefiles. Accessed in June 2014.

Appendix B Comprehensive SPF Estimation Results

This appendix presents the estimation results for the SPFs that were not selected for prediction.

Table B-2 Coefficient Estimates for KAB Intersection Crashes (Statewide SPFs)

Variables	Coefficient Estimates		
	Population	Households	Vehicles
Intercept	-0.960 (0.000)	-1.096 (0.000)	-2.119 (0.000)
Log (number of intersections)	0.151 (0.014)	0.229 (0.000)	0.720 (0.000)
Population (*1000)	0.448 (0.000)	NA (NA)	NA (NA)
Households (*1000)	NA (NA)	0.986 (0.000)	NA (NA)
Vehicles (*1000)	NA (NA)	NA (NA)	0.015 (0.799)
Retail employment (*1000)	0.097 (0.365)	0.017 (0.874)	0.212 (0.101)
Non-retail employment (*1000)	-0.010 (0.302)	-0.002 (0.832)	0.251 (0.000)
Average household income (*1000)	-0.006 (0.000)	-0.006 (0.000)	-0.008 (0.000)
Over dispersion	0.918 (0.000)	0.964 (0.000)	1.016 (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. NA means the variable is not applicable in the model.

Table B-3 Coefficient Estimates for KAB Intersection Crashes (Cluster-based SPFs Using Households)

Variables	Coefficient Estimates by Cluster					
	1	2	3	4	5	6
Intercept	-1.169	0.085	-0.239	-1.273	-2.532	-4.882
	(0.001)	(0.829)	(0.579)	(0.010)	(0.000)	(0.000)
Log (number of intersections)	0.623	0.298	0.209	0.211	0.521	0.820
	(0.000)	(0.031)	(0.154)	(0.207)	(0.000)	(0.000)
Households (*1000)	0.628	0.569	0.634	0.539	0.287	0.398
	(0.001)	(0.000)	(0.000)	(0.016)	(0.140)	(0.089)
Retail employment (*1000)	0.004	-0.502	-0.255	0.363	0.796	0.975
	(0.991)	(0.208)	(0.207)	(0.117)	(0.000)	(0.000)
Non-retail employment (*1000)	0.092	0.176	0.110	0.033	-0.058	0.173
	(0.001)	(0.000)	(0.112)	(0.685)	(0.234)	(0.008)
Average household income (*1000)	-0.005	-0.014	-0.011	-0.002	-0.002	0.002
	(0.058)	(0.000)	(0.000)	(0.312)	(0.037)	(0.000)
Over dispersion	0.231	0.306	0.460	0.657	0.355	0.222
	(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.

Table B-4 Coefficient Estimates for KAB Intersection Crashes (Cluster-based SPFs Using Vehicles)

Variables	Coefficient Estimates by Cluster					
	1	2	3	4	5	6
Intercept	-1.310	0.388	-0.426	-1.590	-2.487	-4.873
	(0.001)	(0.419)	(0.394)	(0.003)	(0.000)	(0.000)
Log (number of intersections)	0.736	0.321	0.404	0.437	0.517	0.818
	(0.000)	(0.057)	(0.015)	(0.017)	(0.000)	(0.000)
Vehicles (*1000)	0.338	0.325	0.193	0.047	0.148	0.193
	(0.056)	(0.014)	(0.052)	(0.729)	(0.147)	(0.081)
Retail employment (*1000)	0.173	-0.631	-0.152	0.295	0.842	0.995
	(0.596)	(0.128)	(0.457)	(0.209)	(0.000)	(0.000)
Non-retail employment (*1000)	0.097	0.202	0.116	0.069	-0.058	0.179
	(0.003)	(0.000)	(0.104)	(0.407)	(0.233)	(0.006)
Average household income (*1000)	-0.007	-0.018	-0.014	-0.003	-0.003	0.002
	(0.023)	(0.000)	(0.000)	(0.053)	(0.010)	(0.000)
Over dispersion	0.275	0.338	0.510	0.677	0.356	0.224
	(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.

Table B-5 Coefficient Estimates for KAB Segment Crashes (Statewide SPFs)

Variables	Coefficient Estimates		
	Population	Households	Vehicles
Intercept	-1.521 (0.000)	-1.705 (0.000)	-0.552 (0.139)
Log (roadway length in miles)	0.082 (0.022)	0.109 (0.004)	0.002 (0.960)
Population (*1000)	0.329 (0.000)	NA (NA)	NA (NA)
Households (*1000)	NA (NA)	0.811 (0.000)	NA (NA)
Vehicles (*1000)	NA (NA)	NA (NA)	0.424 (0.001)
Retail employment (*1000)	0.178 (0.033)	0.114 (0.195)	0.197 (0.016)
Non-retail employment (*1000)	0.083 (0.000)	0.084 (0.000)	0.098 (0.000)
Average household income (*1000)	-0.001 (0.025)	-0.001 (0.144)	-0.002 (0.000)
Over dispersion	0.319 (0.000)	0.374 (0.000)	0.412 (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. NA means the variable is not applicable in the model.

Table B-6 Coefficient Estimates for KAB Segment Crashes (Cluster-based SPFs Using Households)

Variables	Coefficient Estimates by Cluster					
	1	2	3	4	5	6
Intercept	-3.199	-3.038	-2.041	-2.658	-4.379	-6.516
	(0.021)	(0.036)	(0.123)	(0.062)	(0.000)	(0.000)
Log (roadway length in miles)	0.354	0.424	0.261	0.237	0.413	0.561
	(0.041)	(0.017)	(0.101)	(0.151)	(0.004)	(0.000)
Households (*1000)	0.550	0.226	0.489	0.478	0.632	0.738
	(0.012)	(0.119)	(0.000)	(0.007)	(0.000)	(0.000)
Retail employment (*1000)	0.267	-0.813	0.227	0.437	0.357	0.404
	(0.455)	(0.042)	(0.073)	(0.028)	(0.046)	(0.070)
Non-retail employment (*1000)	0.071	0.114	0.109	0.008	-0.035	0.051
	(0.021)	(0.024)	(0.082)	(0.907)	(0.428)	(0.316)
Average household income (*1000)	-0.004	-0.011	-0.011	-0.002	-0.002	0.001
	(0.185)	(0.000)	(0.000)	(0.093)	(0.041)	(0.004)
Over dispersion	0.213	0.226	0.302	0.367	0.395	0.196
	(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.

Table B-7 Coefficient Estimates for KAB Segment Crashes (Cluster-based SPFs Using Vehicles)

Variables	Coefficient Estimates by Cluster					
	1	2	3	4	5	6
Intercept	-3.710	-4.292	-2.229	-4.361	-5.273	-6.372
	(0.007)	(0.013)	(0.122)	(0.011)	(0.000)	(0.000)
Log (roadway length in miles)	0.451	0.607	0.310	0.435	0.521	0.549
	(0.008)	(0.004)	(0.070)	(0.027)	(0.002)	(0.000)
Vehicles (*1000)	0.310	-0.028	0.214	0.170	0.241	0.345
	(0.021)	(0.831)	(0.036)	(0.139)	(0.021)	(0.000)
Retail employment (*1000)	0.359	-0.735	0.253	0.583	0.521	0.381
	(0.264)	(0.064)	(0.051)	(0.004)	(0.003)	(0.084)
Non-retail employment (*1000)	0.056	0.089	0.108	-0.027	-0.026	0.045
	(0.100)	(0.089)	(0.082)	(0.664)	(0.538)	(0.391)
Average household income (*1000)	-0.008	-0.012	-0.014	-0.003	-0.003	0.001
	(0.066)	(0.000)	(0.000)	(0.018)	(0.007)	(0.007)
Over dispersion	0.289	0.205	0.313	0.395	0.406	0.192
	(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.

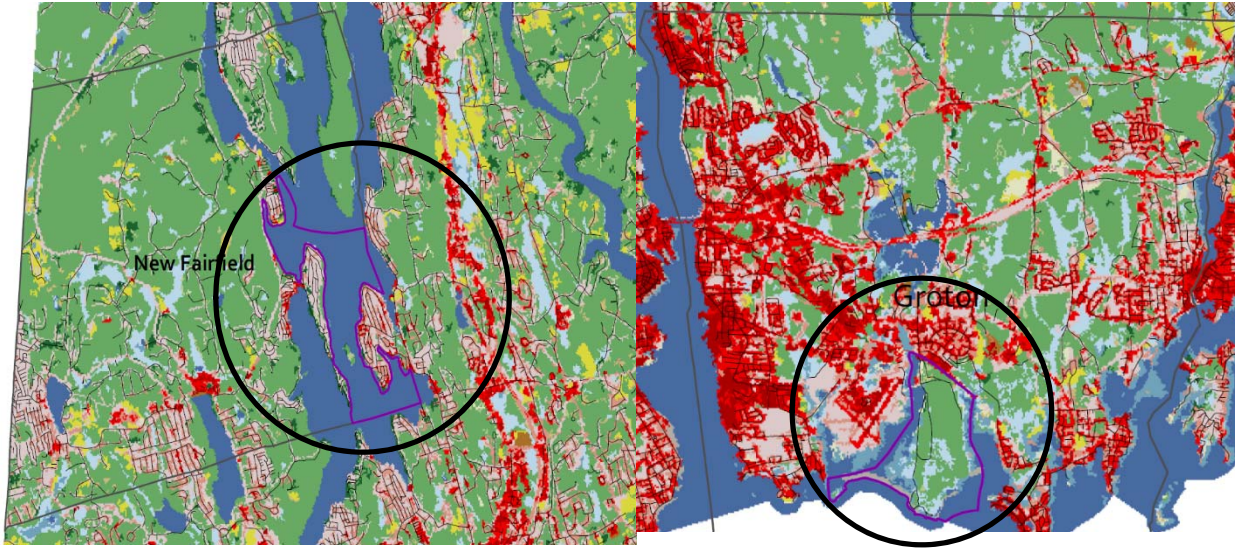


Figure B-7 The Two TAZs Without Population Eliminated from SPF Estimations

Appendix C List of Data Sets Used for Analysis

The following data that were used for the project are archived and available on request from the project team.

1. Roadway Network Shape Features

- 1.1. 2010 TIGER/Line Roadway Shapefile
- 1.2. Names of Roadway Under Local Jurisdiction
- 1.3. 2010 TAZ Boundary Shapefile
- 1.4. 2010 TAZ Size

2. TAZ Level Demographic Records

- 2.1. Raw Data
 - 2.1.1. *Household Income Data*
 - 2.1.2. *Retail and Non-Retail Employment Data*
 - 2.1.3. *Population Projection Data*
 - 2.1.4. *Vehicle Ownership Data*
- 2.2. Processed Data

3. TAZ Level Land Cover Features

- 3.1. 2011 National Land Cover Features for CT
- 3.2. Land Cover Intensities

4. Crash Records

- 4.1. Intersection Crash Records from Connecticut Crash Data Repository
- 4.2. Segment Crash Records from Connecticut Crash Data Repository

5. Crash Location Shapefile

- 5.1. Intersection Crash Location Shapefile
- 5.2. Segment Crash Location Shapefile

6. Assembled TAZ Level Data for Model Estimation and Expected Crash Data for Application Tool

Appendix D Instructions for Use of Visualization Tool

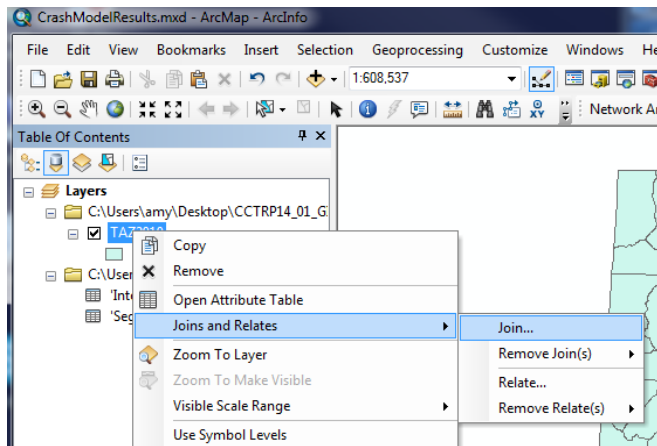
This Appendix outlines how to visualize the results of the crash prediction model by TAZ and lists the steps needed to update variables utilized by the crash prediction models – particularly crash counts, roadway features, land cover intensities, and demographic data. The Appendix assumes the user has previously used and has a basic understanding of the ArcGIS software package.

Requirements: ArcGIS 10 or higher, including a license for the Spatial Analyst extension

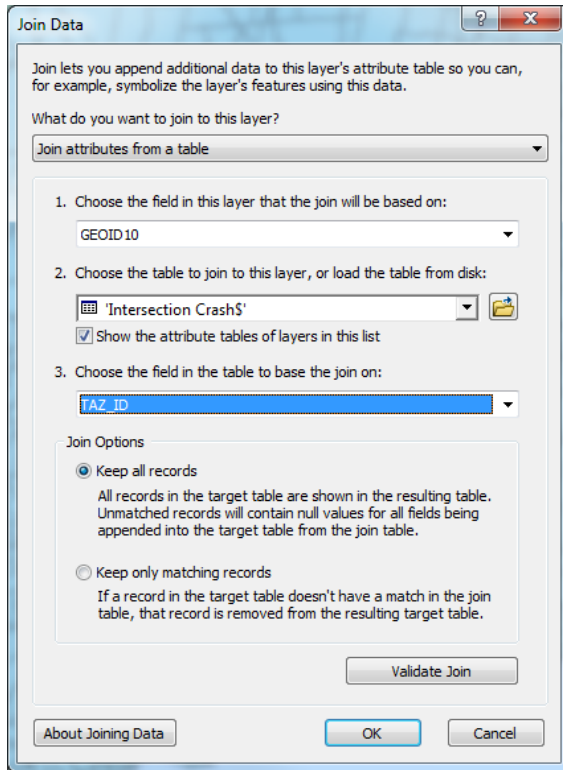
Datasets: CCTRP14_01_GIS.zip

Part 1. Visualizing model results.

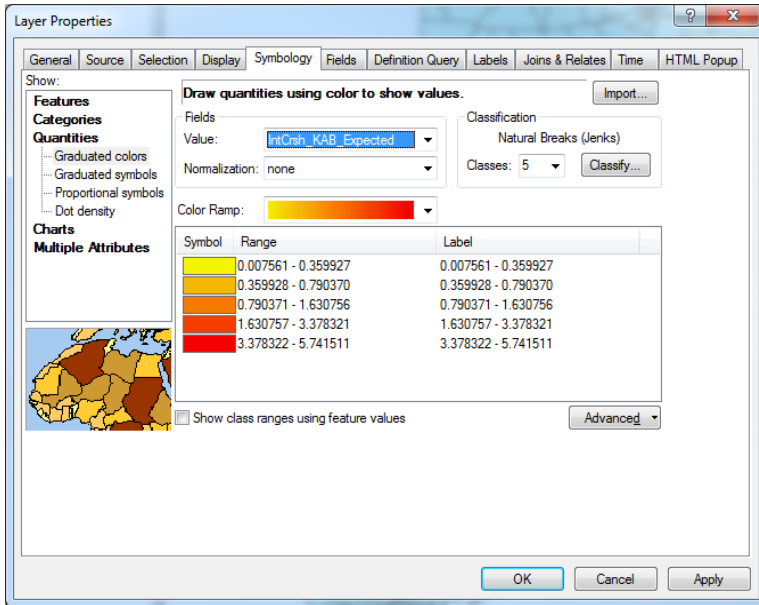
- a. Add the map document CrashModelResults.mxd to ArcMap (File → Open). The map document contains:
 - TAZ2010.shp: shapefile detailing the 2010 Traffic Analysis Zones for Connecticut
 - Intersection Crash.xls: Excel file containing the data for the prediction model for intersection crashes
 - Segment Crash.xls: Excel file containing the data for the prediction model for segment crashes
- b. To view model results for the prediction model for intersection crashes, first join the Intersection Crash table to the TAZ2010 attribute table. Right-click on TAZ2010 in the Table of Contents window and select Joins and Relates → Join.



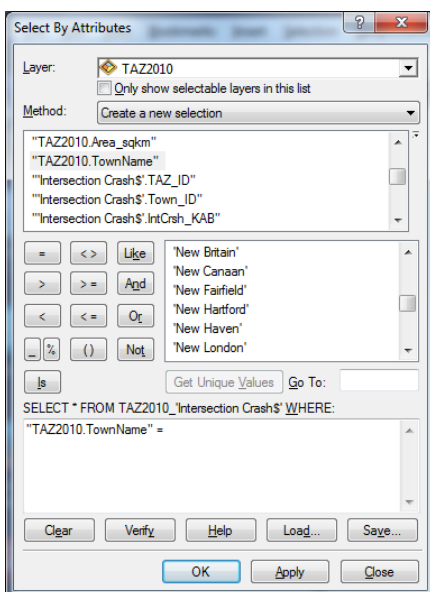
Specify GEOID10 as the join field for the TAZ2010 shapefile, select the Intersection Crash Excel file as the table to join, and specify TAZ_ID as the join field for the Intersection Crash Excel file. Click OK.



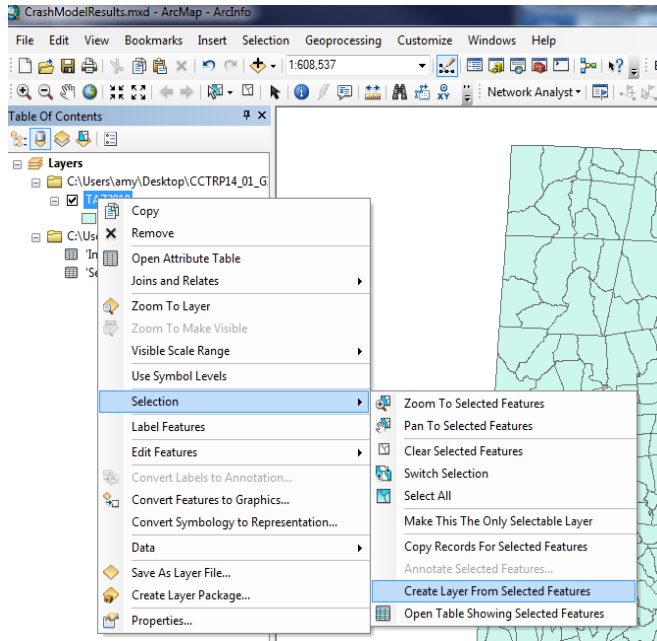
- c. The annual number of expected KAB intersection crashes as estimated using the empirical Bayes method corresponds to the field `IntCrsh_KAB_Expected`. To visualize this variable by TAZ, right-click on the TAZ2010 shapefile name in the Table of Contents window and select Properties to open the Layer Properties window. Under the Symbology tab, specify `IntCrsh_KAB_Expected` as the field to map (select Quantities within the Show window and `IntCrsh_KAB_Expected` as the Value field).
- You can control the classification scheme (*i.e.*, number of classes, classification break values) by clicking on the Classify button.
 - You can control the color scheme by selecting a pre-defined color ramp from the drop-down menu or by double-clicking on each individual symbol box to manually set the color corresponding to each class.



- d. To view model results for the prediction model for segment crashes, follow the steps above replacing the Intersection Crash Excel file with the Segment Crash Excel file.
- Note: Prior to joining the Segment Crash table to the TAZ2010 attribute table, it is recommended that you first remove the join to the Intersection Crash Excel file.
 - Right-click on TAZ2010 in the Table of Contents window, select Joins and Relates → Remove Join(s), and select 'Intersection Crash\$'.
- e. The TAZ2010 attribute table contains the field TownName, which identifies the town corresponding to each TAZ. If you are interested in visualizing modeling results for a single town or a subset of towns, use the Select by Attributes tool (Selection → Select by Attributes) to select your town(s) of interest using the field TownName.



After the TAZs corresponding to the town(s) of interest are selected, right-click on TAZ2010 in the Table of Contents window and select Selection → Create Layer from Selected Features. This will create a new layer that contains only the TAZs corresponding to the town(s) of interest.



- Note: To permanently save this new layer, right-click on the layer in the Table of Contents window and select Data → Export Data. You can then save the layer as a shapefile.

Part 2. Updating model variables.

This section details the processing steps required to update subsets of model variables. Specifically, it provides the steps needed to update:

1. Crash data – update the number of observed KAB intersection or segment crashes for each TAZ
2. Roadway data – update the number of intersections and total length of named roadways under local jurisdiction for each TAZ
3. TAZ cluster memberships – update cluster membership numbers for each TAZ based on an updated land cover map and/or updated population density demographic data
4. Demographic data – update TAZ-level demographic data

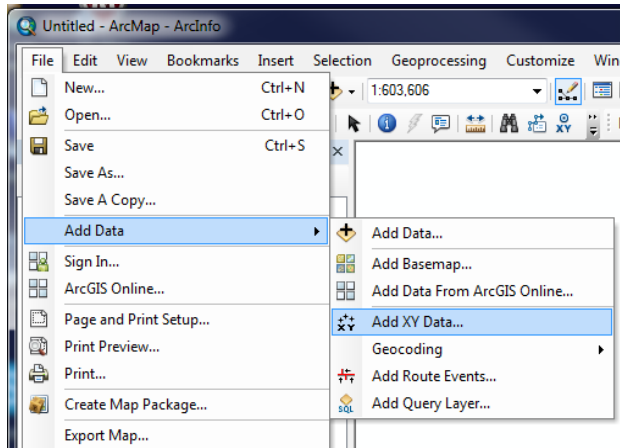
Note 1: Land-cover intensities were calculated using the 2011 National Land Cover Database. The next update will correspond to the release of the 2016 National Land Cover Database. The 2011 National Land Cover Database was made available to the public in December 2013.

Note 2: TAZ-level demographic data were collected from the 2010 Census. The next update will correspond to the 2020 Census.

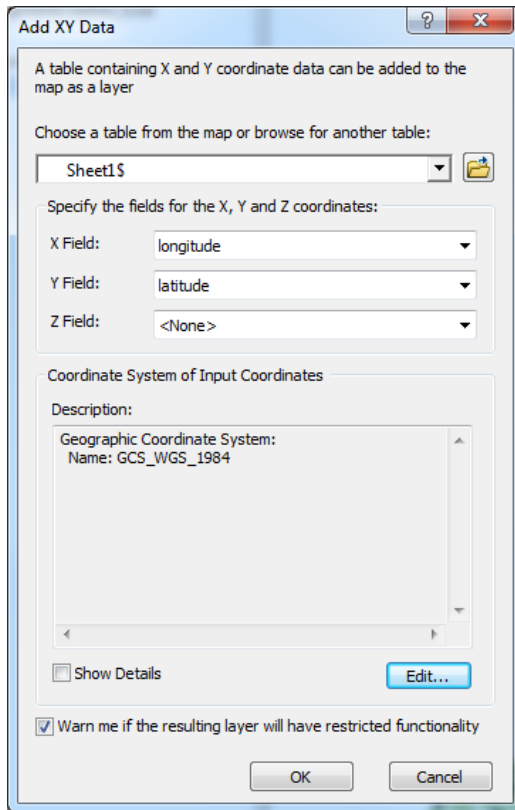
I. Updating crash data

Note: The steps below assume that all crashes have longitude and latitude coordinates stored in columns in a worksheet or database file (e.g., .xlsx, .dbf, or .txt).

- a. Add the updated crash database, stored as a .xlsx, .dbf or .txt file, to ArcMap by selecting File → Add Data → Add XY data.

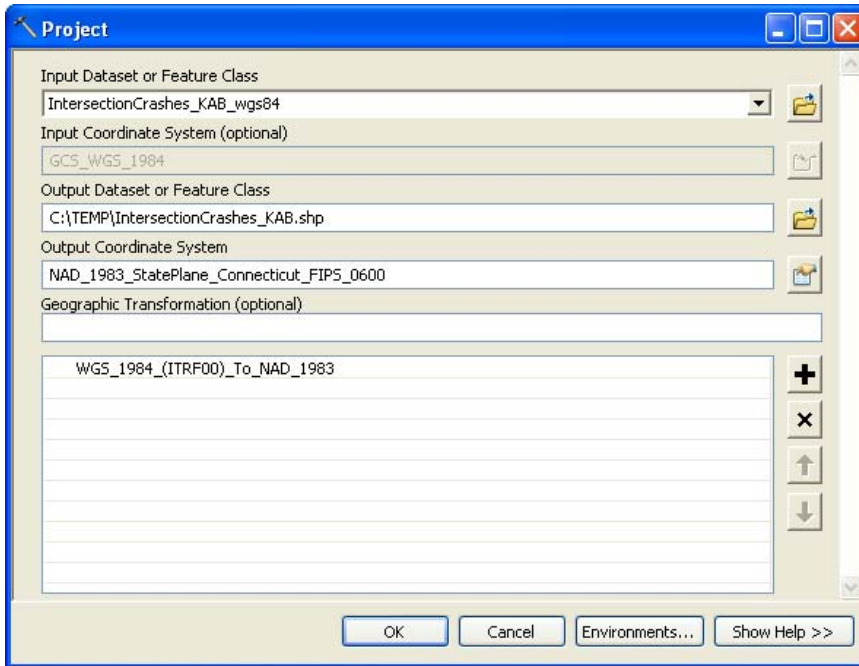



Select your updated crash database, set the X field to the column corresponding to the longitude of each crash point, and set the Y field to the column corresponding to the latitude of each crash point. Use the Edit button under the Coordinate System of Input Coordinates to specify the correct geographic coordinate system corresponding to your longitude and latitude values. For example, if your crash points were collected using the World Geodetic System of 1984 (WGS84) datum, this should be displayed under Description. Click OK to proceed.

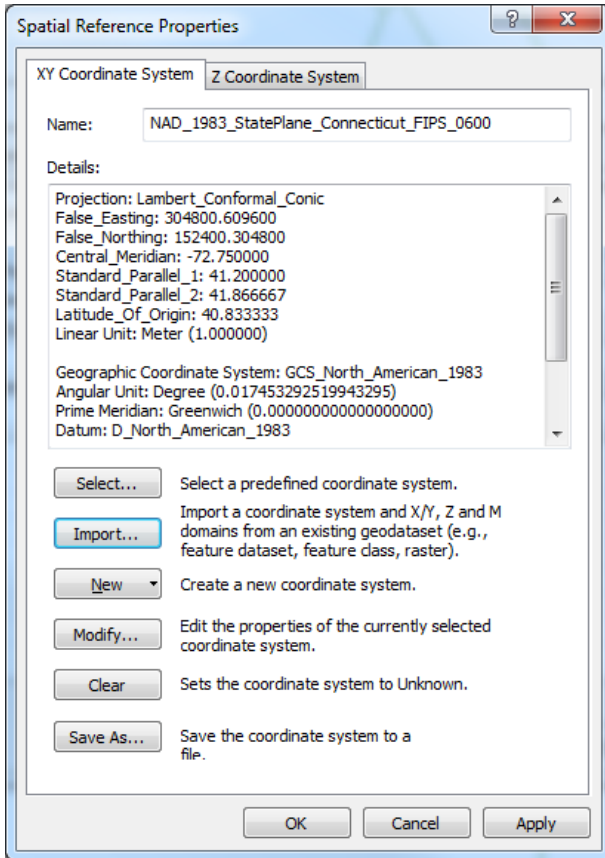


A point event layer will be added to ArcMap. To permanently save your updated crash points in a new shapefile, right-click on the layer name in the Table of Contents window and select Data → Export Data.

- b. All data layers must have the same projected coordinate system prior to data analysis. If the coordinate system associated with the newly created crash shapefile does not match the coordinate system associated with the TAZ and roadway shapefiles (State Plane Coordinate System for CT based on the North American Datum of 1983) or uses a geographic coordinate system solely, the crash point shapefile must be projected. Open ArcToolbox and select Data Management → Projection and Transformations → Feature → Project.

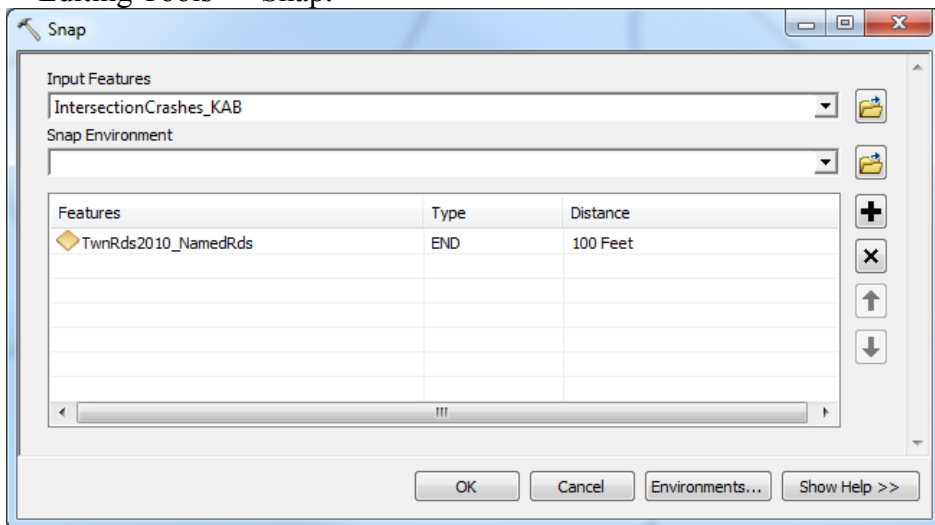


Select the updated crash point shapefile as the Input Dataset. Click the button () adjacent to the Output Coordinate System field to open the Spatial Reference Properties window. Click the Import button to import the projected coordinate system associated with the TAZ shapefile. This operation will ensure that the projected coordinate system associated with the crash point shapefile will exactly match the projected coordinate system associated with the TAZ shapefile. Click OK twice to proceed.

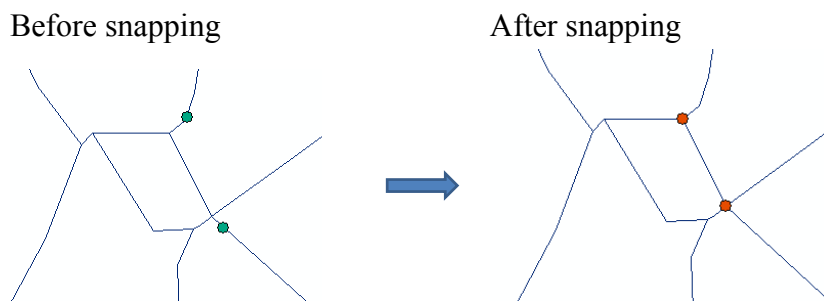


A new projected crash point shapefile is created.

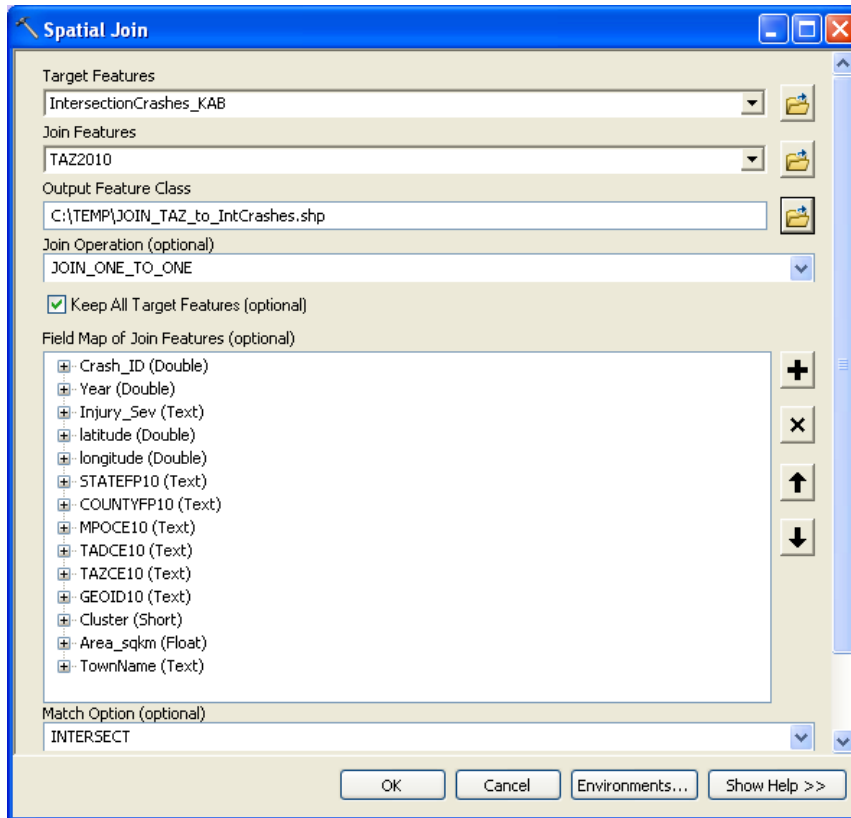
- c. In most cases, the crash points will not overlay exactly onto the roadway line features due to positional error in both datasets. This can be corrected using the Snap tool: ArcToolbox → Editing Tools → Snap.



Specify the projected crash point shapefile as the Input Features and the roadway shapefile as the snap environment (*i.e.*, features to which crash points will be snapped). The feature type setting controls the location of the snap. If the crash points correspond to intersection crashes, specify End as the feature type to snap to the nearest intersection (*i.e.*, snap to the nearest feature's endpoint). If the crash points correspond to segment crashes, specify Edge as the feature type to snap to the nearest roadway segment (*i.e.*, snap to the nearest feature's edge). Finally, specify a snapping distance – *i.e.*, maximum distance over which a point will be moved to the nearest roadway feature. For example, setting this value to 100 feet means that no point will be moved to a roadway or intersection if the roadway or intersection is more than 100 feet away from the current location of the crash point. Click OK.

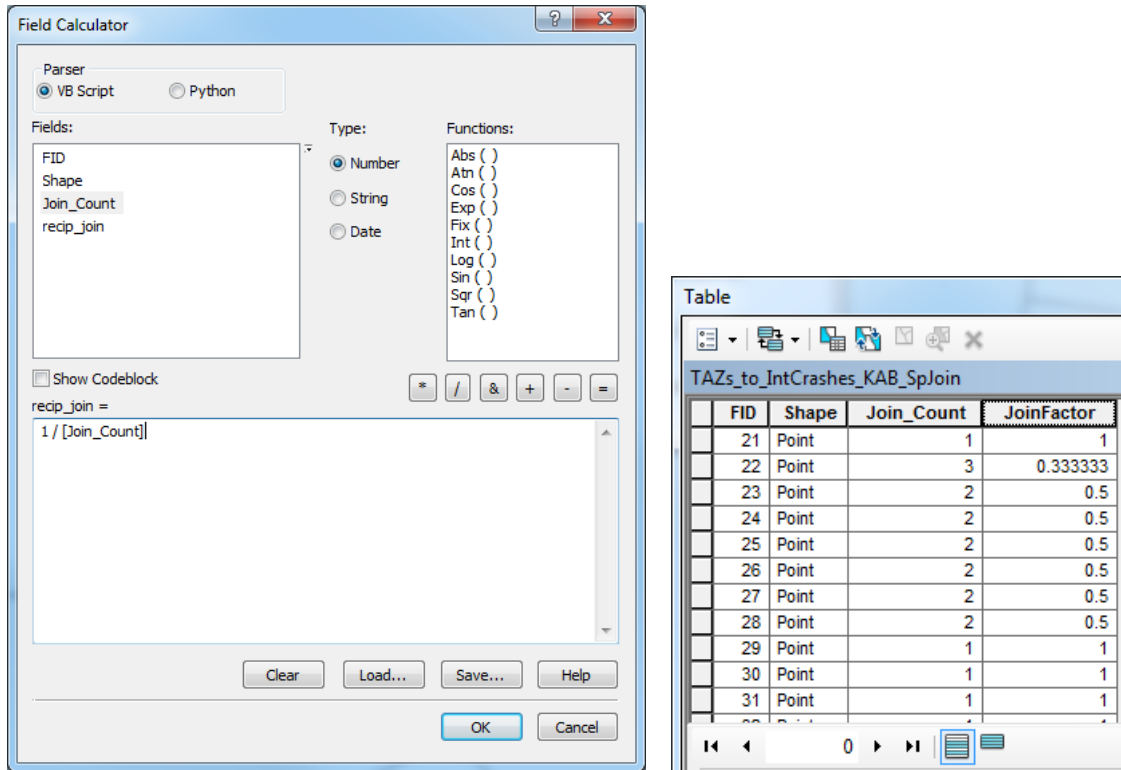


- d. Allocating crashes to each TAZ requires a two-step process because roadways define the boundaries of TAZs. For crashes located on a roadway that serves as the boundary for two TAZs (or located at an intersection that falls on the border of two or more TAZs), an additional processing step is needed to proportionally allocate crashes.
 - i. Use the Spatial Join tool to identify crashes that are located on the boundary of multiple TAZs: ArcToolbox → Analysis Tools → Overlay → Spatial Join. Specify the crash point shapefile as the Target Features and the TAZ shapefile as the Join Features. This spatial join will append to each crash point the attribute data corresponding to the TAZ(s) that intersect that crash point. The Join Operation should be set to One_to_One. As such, if a single crash point intersects two TAZs, then the join count statistic corresponding to this point will be 2. Set the Match Option to Intersect. Click OK.



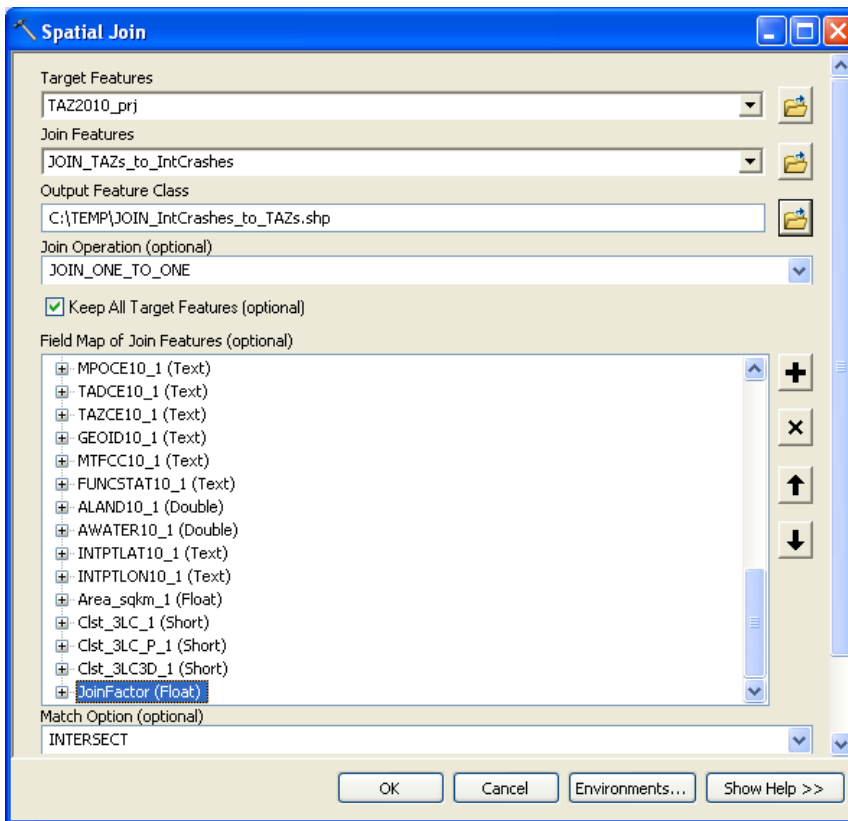
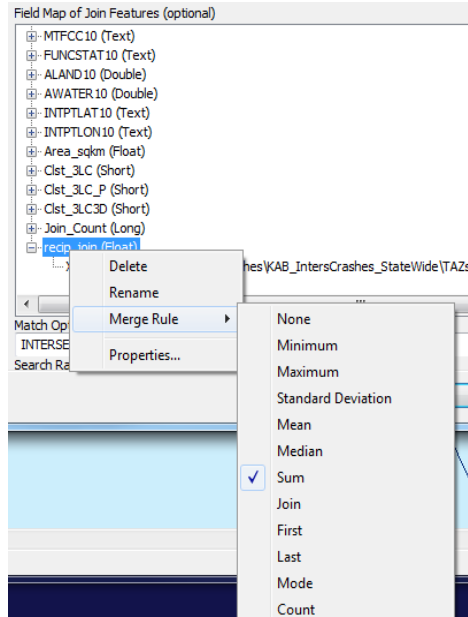
A new crash point shapefile is created that includes the field Join_Count. Join_Count records the number of TAZs associated with each crash point.

- ii. Add a new field to the attribute table associated with the newly created crash point shapefile to store the reciprocal of the Join_Count field. This field will be used to perform the proportional allocation. For example, if the Join_Count value for a crash point equals 2, then the reciprocal value will equal $\frac{1}{2}$ and the crash will be equally divided between the two associated TAZs. A new field can be added using the Add Field tool: ArcToolbox → Data Management Tools → Fields → Add Field. Specify the Field Name as JoinFactor and the Field Type as float.
- iii. Open the attribute table, right click on the column heading corresponding to the new field JoinFactor, and select Field Calculator. Enter the expression: $1 / [\text{Join_Count}]$. Click OK.



The JoinFactor field should correspond to the reciprocal of the Join_Count field.


- iv. Use the Spatial Join tool a second time to calculate the total number of crashes occurring within each TAZ accounting for proportional allocation. Specify the TAZ shapefile as the Target Features and the newly created crash point shapefile (*i.e.*, crash point shapefile containing the field JoinFactor) as the Join Features. The Join Operation should be set to One_to_One. As such, if several crash points are located in a single TAZ, then the join count statistic corresponding to this TAZ will represent the total number of crash points associated with the TAZ. Set the Match Option to Intersect. Under the Field Map of Join Features, right-click on the field JoinFactor and set the Merge Rule to Sum. This will sum the JoinFactor values for all crash points that intersect each TAZ and perform the proportional allocation. Click OK.



A new TAZ shapefile is created that includes the fields: Join_Count, which represents total number of crash points that intersect each TAZ, and JoinFactor, which represents the proportional allocation of crash points (*i.e.*, number of crash

points that intersect each TAZ accounting for crashes occurring along TAZ boundaries).

FID	Shape *	Join_Count	TADCE10	TAZCE10	GEOID10	Area_sqkm	JoinFactor
45	Polygon	23	00000010	00001194	0900300001194	0.796603	15
46	Polygon	0	00000010	00001116	0900300001116	3.27447	0
47	Polygon	29	00000010	00001156	0900300001156	0.727259	20
48	Polygon	23	00000010	00001193	0900300001193	0.428151	14.8333
49	Polygon	37	00000010	00001169	0900300001169	0.612059	25.8333
50	Polygon	4	00000010	00001174	0900300001174	0.601684	4
51	Polygon	16	00000010	00001179	0900300001179	0.402509	9
52	Polygon	9	00000010	00001154	0900300001154	0.563993	6.33333
53	Polygon	3	00000010	00001202	0900300001202	1.45821	1.5

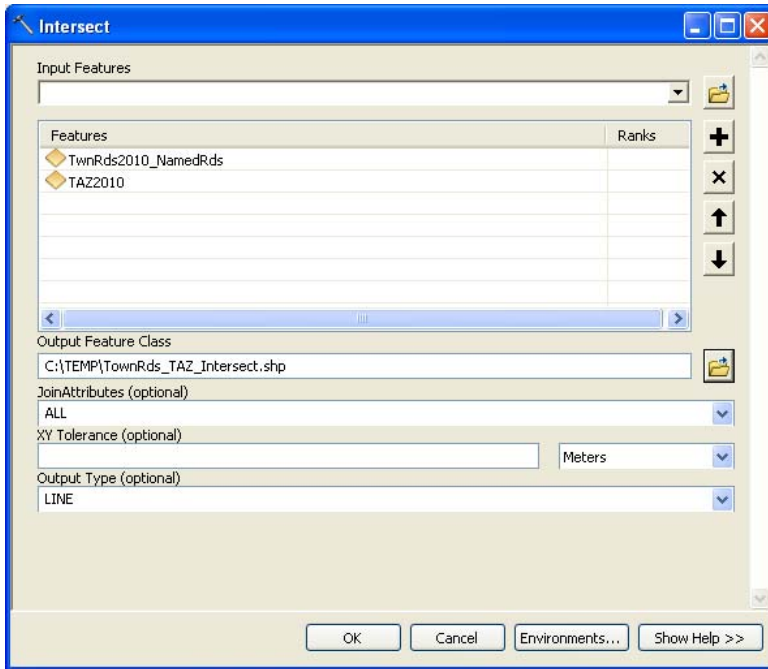
- v. The attribute table associated with the newly created TAZ shapefile can be exported as a .dbf or .txt file by selecting Export from the Table Options drop-down menu (). The JoinFactor field should be rounded up to the nearest whole number for modeling. The updated data replaces the IntCrsh_KAB or SegCrsh_KAB column in the model database, depending on type of crashes updated.

II. Updating roadway data

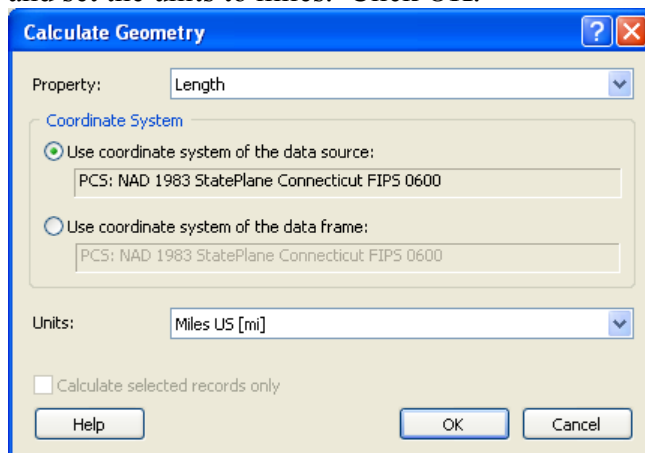
Note: The steps below assume that updated roadway features are stored as a shapefile. It is also assumed that the roadways features within the shapefile have been edited such that roadway line segments that shared coincident endpoints and the same roadway name were merged to produce a single roadway feature. In addition, it is assumed that roadways under state jurisdiction and unnamed roadways (*e.g.*, private driveways or private roads) have been removed from the roadway shapefile. Only named roadways under local jurisdiction were considered when calculating the number of intersections and total length of roadways within each TAZ.

- a. Add the updated roadway shapefile and the TAZ shapefile to ArcMap. Confirm that the projected coordinate system associated with the roadway shapefile matches the projected coordinate system associated with the TAZ shapefile. If not, follow the steps in section I.b. to project the roadway shapefile so that it matches the projected coordinate system associated with the TAZ shapefile.
- b. First, calculate the total length of named roadways under local jurisdiction associated with each TAZ. As previously mentioned, roadways define the boundaries of TAZs. Proportional allocation was used to divide the length of the shared roadway evenly between both TAZs.

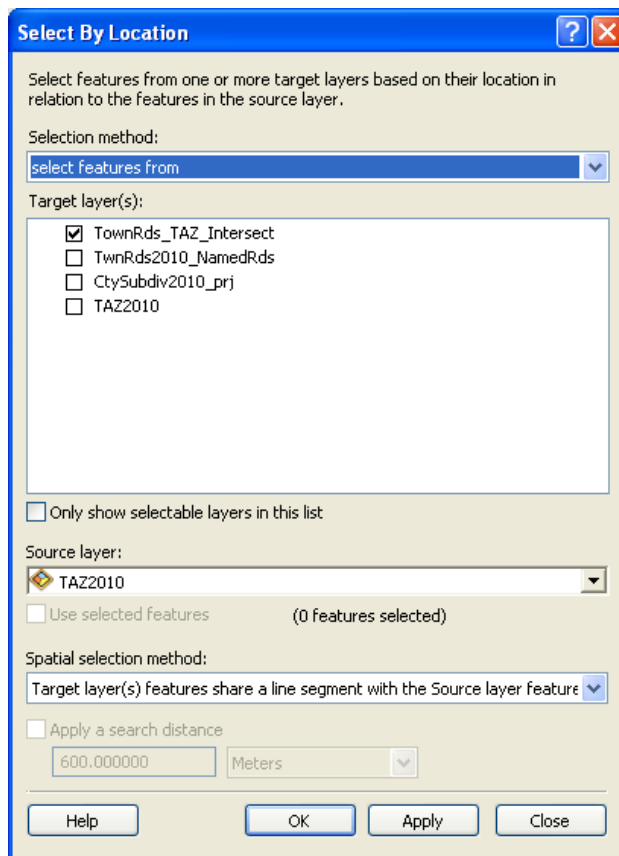
- i. Use the Intersect tool (ArcToolbox → Analysis Tools → Overlay → Intersect) to overlay TAZs onto the roadway features. This operation splits roadway features at TAZ boundaries, but also creates duplicate line segments for all roadways falling along the border of two TAZs (e.g., roadway A intersected with TAZ 1 and roadway A intersected with TAZ 2). Select the roadway and TAZ shapefiles under Input Features. Specify the JoinAttributes option as All and the Output type as Line. Click OK.



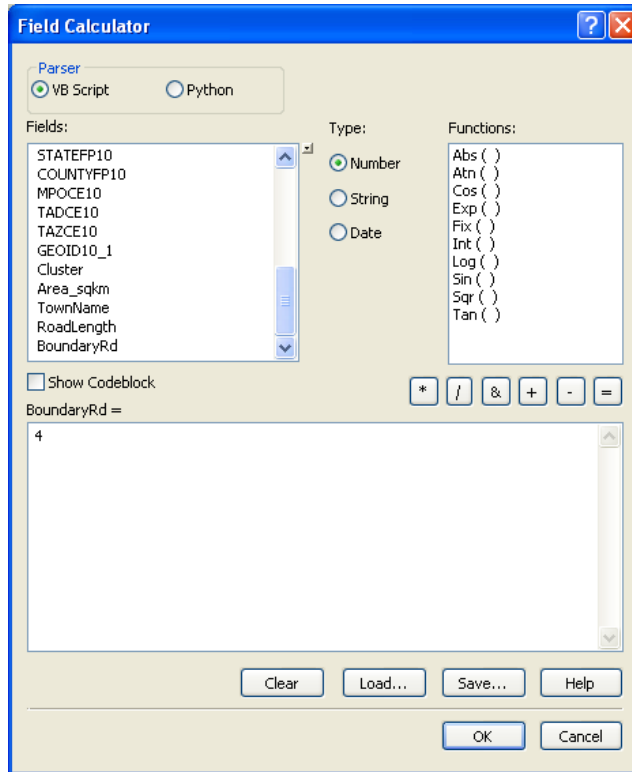
- ii. The length of the roadway line segments split at TAZ boundaries must be updated in the newly created shapefile. Add a new field to the attribute table using the Add Field tool (see section I.d.ii.). Specify the Field Name as RoadLength and the Field Type as float. Open the attribute table, right click on the column heading corresponding to the new field RoadLength and select Calculate Geometry. A warning box may appear – click Yes. Specify Length as the property to calculate and set the units to miles. Click OK.




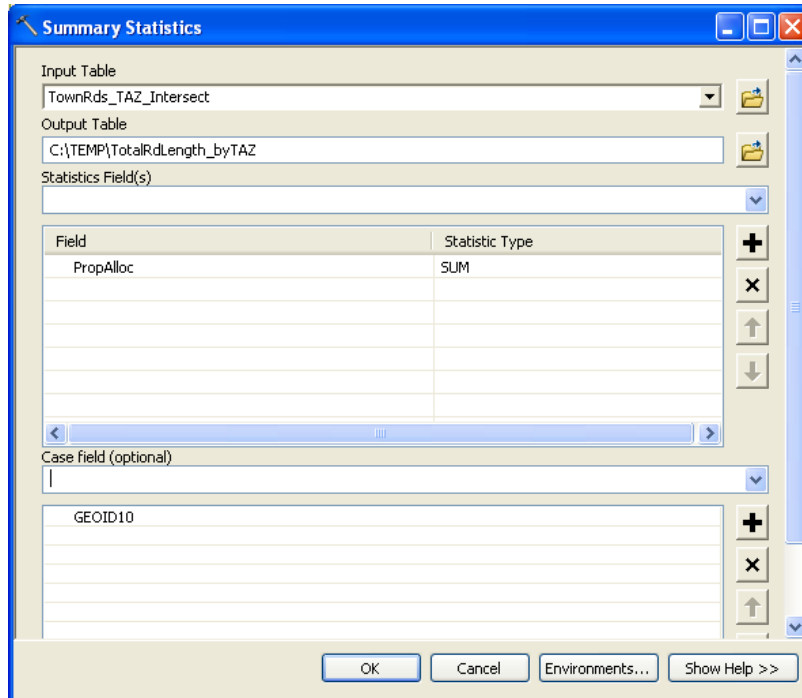
- iii. Add a second field named BoundaryRd to the attribute table and set the Field Type to short integer. This field will identify roadways that fall along the border of two TAZs and will be used to proportionally allocate one half of the total roadway length to each TAZ.
- iv. Use the Select by Location tool (Selection → Select by Location) to identify the roadways that fall along the border of two TAZs, open the. Specify the Target layer as the TAZ-roadway intersected shapefile, specify the Source layer as the TAZ shapefile and use the method to select “Target layer(s) features share a line segment with the Source layer features”. This will select all roadways that fall along the border of a TAZ. Click OK.



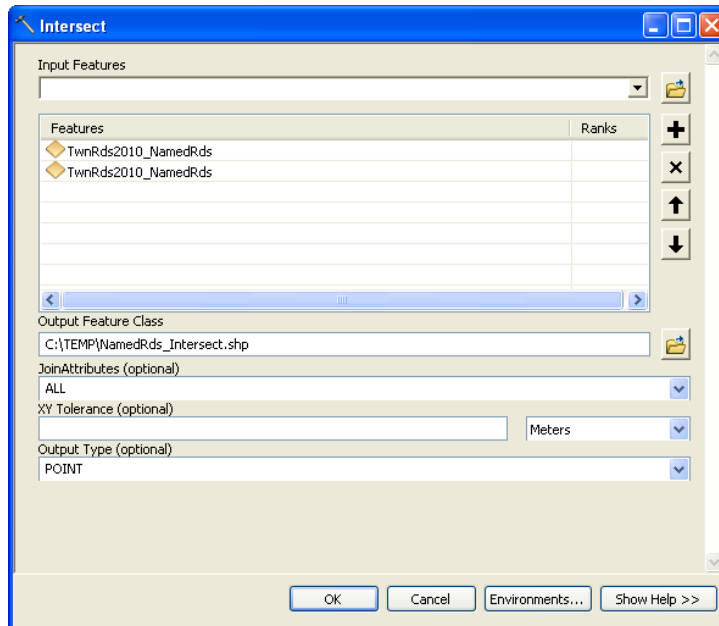
- v. With the border roadways selected, open the attribute table associated with the intersected shapefile. Right click on the column heading corresponding to the field BoundaryRd and select Field Calculator. Set the BoundaryRd field value to 4. Recall: The intersect operation created two duplicate roadway line segments for all roadways falling along the border of two TAZs. Each line segment must be divided by 4 to ensure one half of the roadway length is assigned to each TAZ. Only the roadway features currently selected will have their BoundaryRd field updated.



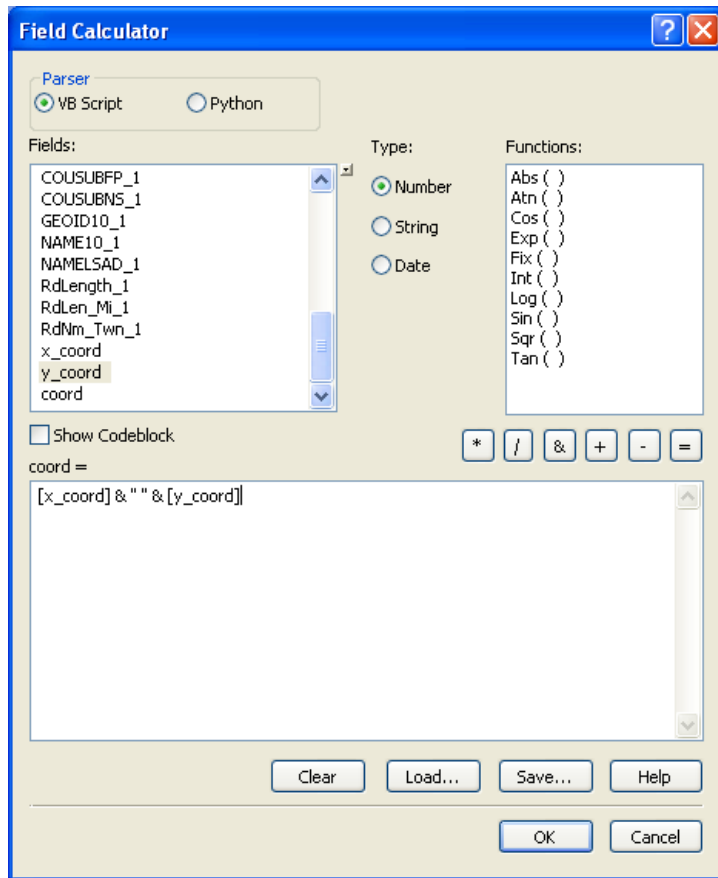
- vi. Reverse the selection using the Switch Selection button (). This will select all roadways that do not fall along a TAZ boundary. Follow the steps above to set the BoundaryRd field value to 0 for these roadways.
- vii. After the calculation is complete, clear all selected features (Selection → Clear Selected Features).
- viii. Add a third field named PropAlloc to the attribute table associated with the TAZ-roadway intersected shapefile. Set the Field Type to float. Use Field Calculator to enter the expression: $[RoadLength] / [BoundaryRd]$. Click OK.
- ix. Open the Summary Statistics tool to summarize the proportionally-allocated roadway lengths for each TAZ: ArcToolbox → Analysis Tools → Statistics → Summary Statistics. Specify the TAZ-roadway intersected shapefile as the Input Table, set the Statistics Field to PropAlloc, set the Statistic Type to sum, and specify the Case field as GEOID10 (*i.e.*, the field that corresponds to a unique identification number for each TAZ). Click OK.



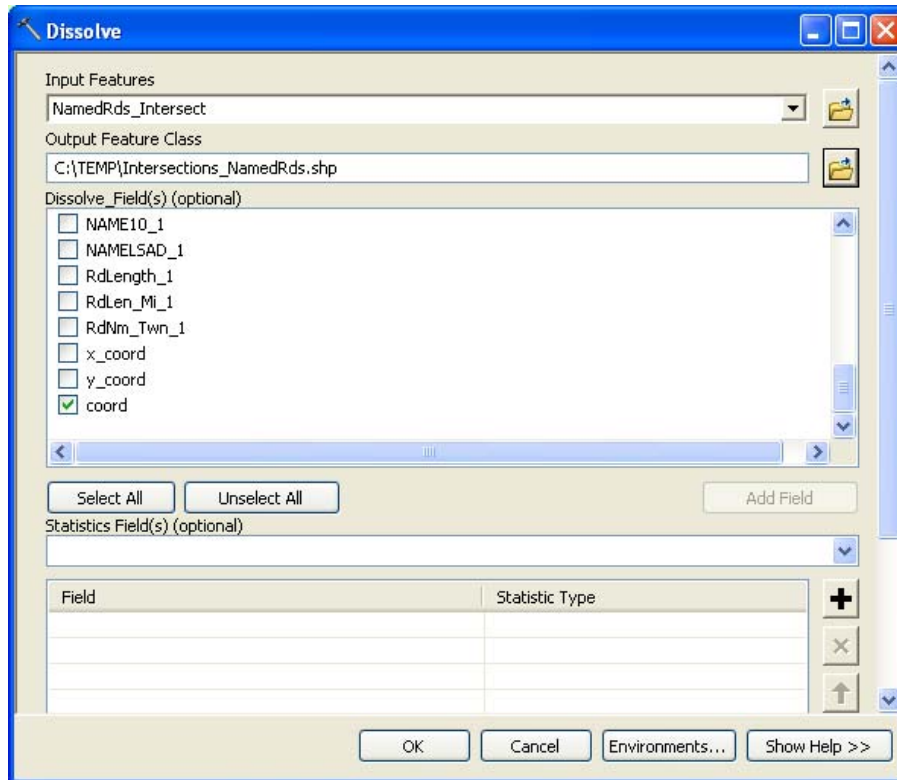
- x. The resulting table can be exported as a .dbf or .txt file. The field SUM_PropAlloc replaces the Length_Roadway column in the database corresponding to the prediction model for segment crashes.
- c. Next, calculate the total number of intersections involving named roadways under local jurisdiction associated with each TAZ.
 - i. Use the Intersect tool (ArcToolbox → Analysis Tools → Overlay → Intersect) to overlay the roadway shapefile onto itself; *i.e.*, roadway shapefile should appear twice under Input Features and the Output Type should be set to Point. This operation creates a point feature at the intersection of each pair of line segments, but also creates duplicate points at each intersection (*e.g.*, roadway A intersected with roadway B, roadway B intersected with roadway A).



- ii. To remove duplicate intersection points, create a new field that contains the x- and y-coordinates for each point then use this field to remove all points that share the same coordinate pair. Add a new field to the attribute table associated with the newly created intersection shapefile (see section I.d.ii.). Specify the Field Name as `x_coord` and the Field Type as float. Open the attribute table, right click on the column heading corresponding to the new field `x_coord` and select Calculate Geometry. A warning box may appear – click Yes. Specify as the property to calculate as X Coordinate of Centroid and set the units to meters. Click OK.
- iii. Repeat the above procedure to add a new field named `y_coord` and calculate the Y Coordinate of Centroid.
- iv. Create a new field named `coord` and set the Field Type to Text. This field will store the coordinate pair. Use Field Calculator to enter the expression: `[x_coord] & " " & [y_coord]`. This will concatenate the x- and y-coordinate fields to create a new field that contains the x-coordinate followed by a space followed by the y-coordinate. Note: Be sure to include a space between the set of quotations. Click OK.



- v. Open the Dissolve tool (ArcToolbox → Data Management → Generalization → Dissolve). You will use the field coord to remove all duplicate points. Dissolve works by aggregating all features that share the same attribute value, in this case the same coordinate pair. Click OK.



- vi. To calculate the number of intersections occurring within each TAZ based on proportional allocation, follow the procedure in section I.d.
- vii. The attribute table associated with the newly created TAZ shapefile can be exported as a .dbf or .txt file. The JoinFactor field replaces the Number_Intersection column in the database corresponding to the prediction model for intersection crashes.

III. Updating TAZ cluster membership

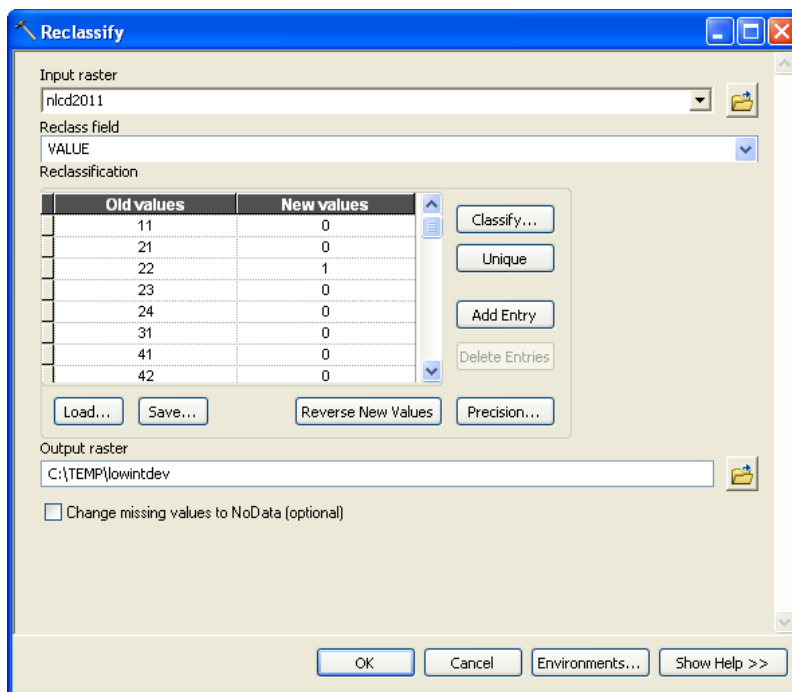
Note: The steps below assume that the 2016 National Land Cover Database for Connecticut has been downloaded in ArcGIS raster format – *i.e.*, GRID. A license for the Spatial Analyst extension is required for this procedure.

- a. Add the updated land cover grid and the TAZ shapefile to ArcMap.
- b. If the coordinate system associated with the new land cover grid does not match the coordinate system associated with the TAZ shapefile (State Plane Coordinate System for CT based on the North American Datum of 1983) or uses a geographic coordinate system solely, the land cover grid must be projected. Open ArcToolbox and select Data Management → Projection and Transformations → Raster → Project Raster, then follow the steps in section I.b.
- c. Land cover intensities were calculated based on the number of cells classified as developed within each TAZ. The National Land Cover Database classification system has three land-cover classes corresponding to developed land: a) low intensity developed – single family housing, less than 50% impervious surface [class code 22]; b) medium intensity developed – single-family housing, between 50-80% impervious

surface [class code 23]; and c) high intensity developed – apartment complexes, commercial and industrial areas, greater than 80% impervious surface [class code 24]. You will create three new raster layers, one for each developed class.

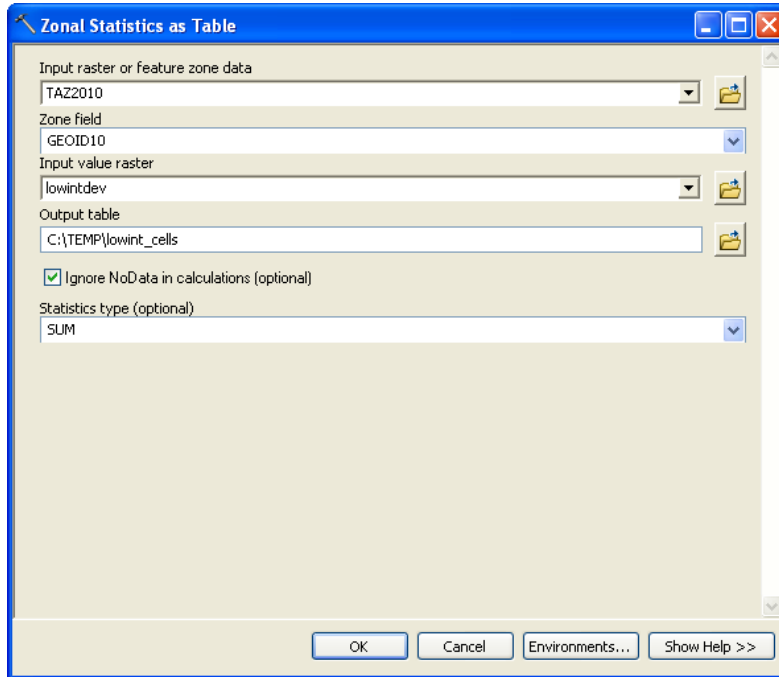
Note: Class codes may change values between 2011 and 2016; be sure to check the legend accompanying the land cover grid to confirm you have the correct class codes for low, medium and high intensity developed. The instructions below assume the class codes have not changed from 2011 to 2016.

- d. Use the Reclassify tool (ArcToolbox → Spatial Analyst Tools → Reclass → Reclassify) to create a new raster layer where all cells classified as low intensity developed are set to a value of 1 and all other cells are set to a value of 0. Specify the land cover grid as the Input raster, set the Reclass field to Value, and use the Reclassification table to set all cells with a current value of 22 to 1 and all other values to 0. Note: If the reclassification table displays a range of values in each row, hit Unique. In addition, be sure cells classified as NoData remain in the class NoData.



- e. Repeat the above procedure two additional times to create new grids for medium (23) and high (24) intensity developed.
- f. To determine the number of developed cells within each TAZ, you must sum all grid cells with a value of 1 that occur within the boundaries of each TAZ. This is accomplished using Zonal Statistics (ArcToolbox → Spatial Analyst Tools → Zonal → Zonal Statistics as Table). Specify the TAZ shapefile as the Feature Zone data and set the Zone field to GEOID10 (e.g., unique identification number for each TAZ). The Input value raster should be set to one of the developed grid (e.g., lowintdev). Set the Statistics type to Sum and confirm that the Ignore NoData in calculations option is

checked. The output of the procedure is a table with the field SUM, which records the number of developed cells within each TAZ.



- g. Repeat the above procedure two additional times to create tables for medium and high intensity developed.
- h. The resulting tables can be exported as .dbf or .txt files. Convert the updated cell counts to area in square kilometers using the equation: $(\text{number of cells} \times 900\text{m}^2) / 1,000,000$. Note: The cell size for NLCD raster data is 30m x 30m or 900m².
- i. Use the look-up table in Appendix A (Table A.1 Interval Values of Each Clustering Variable by Cluster) to adjust cluster membership values as needed.

IV. Updating demographic data

- a. Demographic data can be updated with the release of the 2020 U.S. Census by downloading the following variables from the Census Transportation Planning Package Database, specifying TAZ as the unit of analysis: population, retail and non-retail employment, and mean household income. The updated values replace the variables Population, Employment_Retail, Employment_Non-Retail, and Income_Mean in the model databases for intersection and segment crashes.
- b. To update the population density values used in the cluster analysis, open the TAZ shapefile in ArcMap and export the table to .dbf or .txt format. The TAZ shapefile includes the field Area_sqkm (*i.e.*, area of each TAZ in square kilometers). This field should be used to update the population density values for each TAZ (*i.e.*, number of people per square kilometer). Updated population density values can be combined with updated land cover intensities to adjust cluster membership values as needed in accordance with Table A.1 in Appendix A.

Appendix E Instructions for Computation of Crash Rates by TAZ Area

In this project, we intentionally didn't use crash rates by TAZ size to develop SPFs, as current practice in traffic safety analysis doesn't recommend making decisions on the basis of crash rates, but rather crash counts. However, if researchers are interested in comparing crash experience among TAZs using crash rates, the following procedures can be followed.

1. In the assembled TAZ level data for model estimation and expected crashes provided along with this document, calculate crash rates (crashes per km²) for each TAZ by dividing the observed and expected crash counts by the TAZ area for both the intersection and the segment files.
2. Insert the new crash rate variables into the visualization tool to conduct safety analysis, following the instructions in Appendix D.