Southeastern Transportation Center

Final Year 1 Report

MRI 4: Big Data for Safety Monitoring, Assessment, and Improvement

Principal Investigator
Mohamed Abdel-Aty
Essam Radwan

Researcher
Jaeyoung Lee
Qi Shi
Ling Wang
Qing Cai

University of Central Florida
Department of Civil, Environmental & Construction Engineering
Orlando, FL 32816-2450

August 2015

CONTENTS

1 MACROSCOPIC ANALYSIS

1.1 Background and Objectives

The objective of this research is to generate new ideas to collect Big Data for the macroscopic traffic crash analysis. There have been many efforts to assess traffic safety at various scopes such as zonal-level, segments, intersections or corridors. Among these scopes, macroscopic safety analysis focuses on traffic crash occurrence at the zonal-level and attempts to relate zonal socio-demographic features with crash counts. In recent years, efforts to incorporate traffic safety into transportation planning has been made, which is termed as transportation safety planning (TSP). The Safe, Affordable, Flexible Efficient, Transportation Equity Act – A Legacy for Users (SAFETEA-LU), which is compliant with the United States Code, compels the United States Department of Transportation to consider traffic safety in the long-term transportation planning process. Most of macroscopic studies have analyzed traffic safety using planning data based on traffic analysis zones (TAZs). Nevertheless, it is expected that there are much more meaningful contributing factors for the crash occurrence at the zonal-level, and it will be helpful to find out zones with higher traffic crash risks and their contributing factors. For the macroscopic safety analysis, planning data that are used for the travel demand forecasting, or other socio-demographic data have been used so far. Thus, it is required to try more variables from various sources for the specific crash types. To summarize, the key objectives are to:

- Produce new ideas for acquisition and use of Big Data to facilitate safety assessment at the macroscopic level.
- Macroscopic traffic safety analytics using Big Data
- Visualize and analyze Big Data for various crash types using GIS techniques

In this first year, we also present some ideas for microscopic safety analysis using big data.

However, the second year of the project will mostly address this issue.

1.2 Data Collection

In order to analyze traffic crashes at the macroscopic level, the research team has collected "Big Data" from multiple sources as shown in Figure 1.

- Layer 1: TAZs and TADs (Traffic Analysis Districts) maps were obtained from the Florida Department of Transportation (FDOT) and U.S. Census Bureau, respectively.
- Layer 2: The demographic data and socioeconomic data were obtained from the U.S. Census Bureau and the FDOT Central Office.
- Layer 3: The roadway and traffic data were collected from the FDOT Transportation Statistics Office (TRANSTAT).
- Layer 4: The crash data were obtained from the FDOT CAR (Crash Analysis Reporting) system database and Signal Four Analytics (S4A).

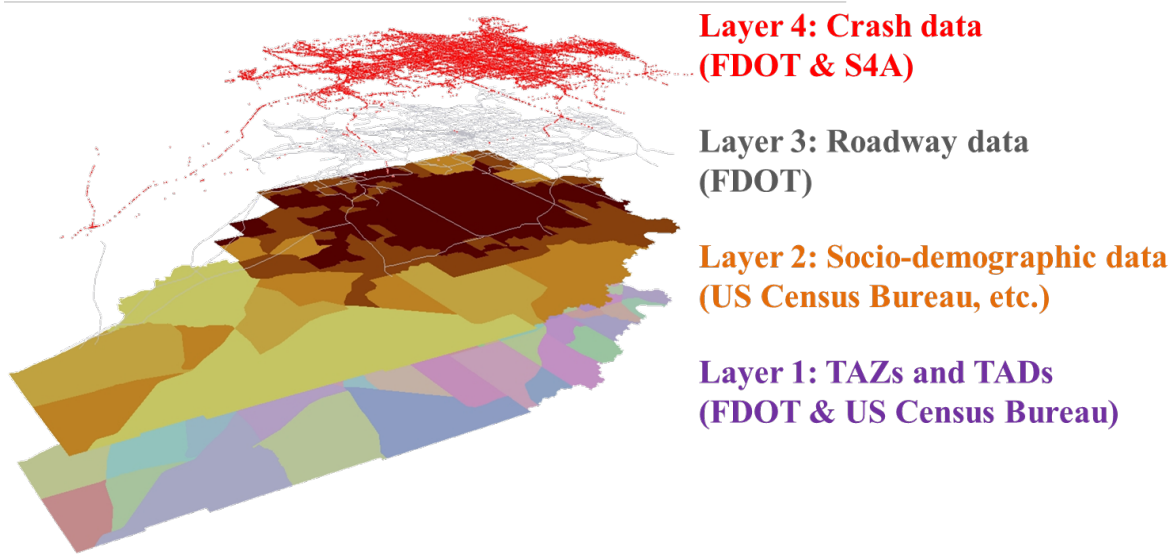

Figure 1 Multiple Data Sources

1) Layer 1: Geographic boundary maps

Three types of geographic boundary maps were collected: TAZ, TAD, and county maps. TAZs are used by the FDOT Central Office for statewide transportation plans. TAZs have been widely used for macroscopic traffic analysis since they are the only traffic/transportation related spatial boundaries. TADs are new and highly aggregated geographic unit for traffic analysiss. TAD may be useful if practitioners want to define crash pattern at a higher aggregate level. The key spatial characteristics of TAZs and TADs are summarized in Table 1. Considering the overall average TAZ area is 6.472mi^2, TADs are approximately 16 times larger (103.314mi^2) than TAZs. Figure 2 and Figure 3 depict TAZs and TADs in Florida, respectively.

Table 1 Summary of TAZ, TAD and county maps

Geographic Units	Area (mi^2)	No of TAZs	Avg. area/TAZ
TAZ	55,127	8,518	6.472
TAD	61,368	594	103.314

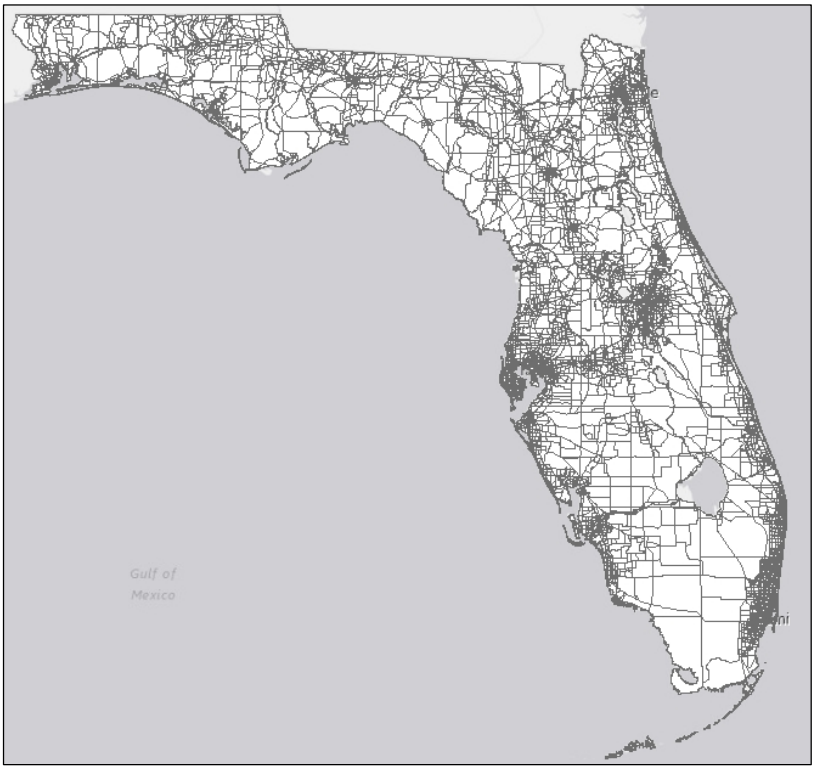

Figure 2 TAZ map (N=8,518)

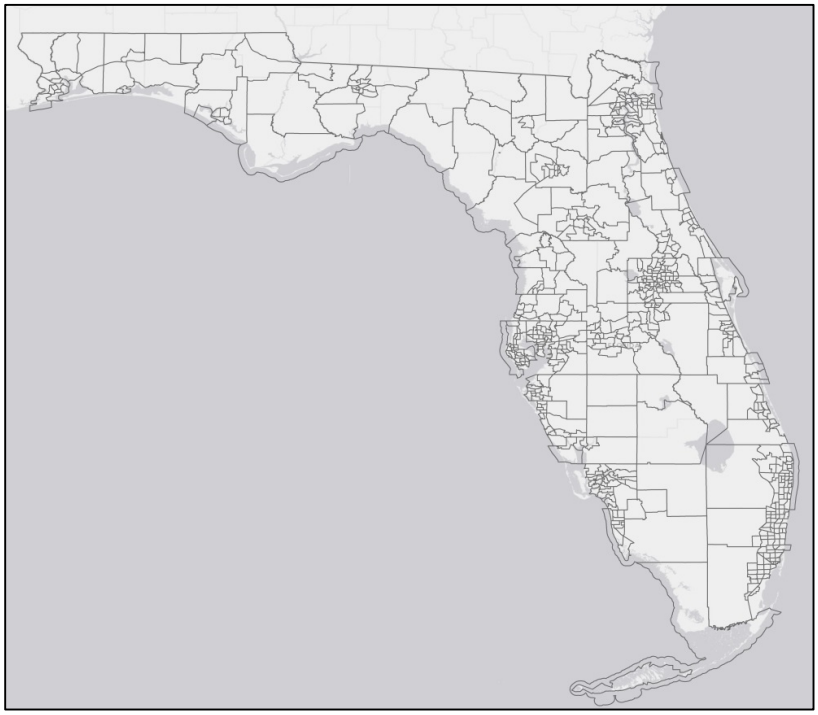

Figure 3 TAD map (N=594)

2) Layer 2: Demographic and socioeconomic data

Demographic and socioeconomic data, which can serve as surrogate for traffic volumes that affect crash occurrence, are collected (Table 2). The demographic data such as population, population by race/ethnicity, and population by age group based on the census block were acquired from the U.S. Census Bureau. TAZ-based data were provided by FDOT Central Office, which are called Zone Data (ZDATA). Single Family Units (SFU), MFU (Multi Family Units), and HMT (Hotel, Motel, and Timeshare) data were acquired, which are very closely related to trip generation. Furthermore, trip attraction factors such as employments and school enrollments are obtained based on TAZ maps. The urban boundary map was collected from the FGDL (Florida Geographic Data Library) and median household income data were obtained from the U.S. Census Bureau.

Table 2 Summary of demographic and socioeconomic data

Category	Variables	Base units	Sources
Demographic	Population Population by race/ethnicity Population by age group	Census block	U.S. Census Bureau
	Number of SFU Percentage of the nonpermanent vacant in SFU Percentage of the single family vacant Population of SFU in residential area Number of MFU Percentage of the nonpermanent vacant in MFU Percentage of the multiple family vacant Population of MFU in residential area	TAZ	FDOT Central Office
Socioeconomic	Percentage of SFU owns no vehicle Percentage of SFU owns one vehicle Percentage of SFU owns two or more vehicles Percentage of MFU owns no vehicle Percentage of MFU owns one vehicle Percentage of MFU owns two or more vehicles		
	Number of HMT rooms Percentage of HMT occupancy Number of HMT occupants		
	Industrial Employment Commercial Employment Service Employment Total Employment School Enrollment		
	Urban boundaries	Polygon	FGDL
	Median household income	Block Group	U.S. Census Bureau

3) Layer 3: Roadway and traffic data

Roadway/traffic data were collected from FDOT TRANSTAT and FDOT UBR (Table 3). The roadway data includes the location of intersections and traffic signals, total roadway length, and roadways by speed limits. Traffic data contain AADT (Annual Average Daily Traffic) and truck

traffic volume. Roadway and traffic data are expected to be important contributing factors for the crash occurrence.

Table 3 Summary of roadway and traffic data

Category	Variables	Base units	Sources
Roadway	Intersection Traffic signal locations	Point	FDOT TRANSTAT
	Total roadway length Roadway by speed limits	Polyline	FDOT UBR
Traffic	AADT Truck traffic volume	Polyline	FDOT TRANSTAT

4) Layer 4: Crash data

Figure 4 presents the overall process of the crash data collection from the two sources: FDOT CAR and S4A.Two forms of crash report are used in the State of Florida. They are short form and long form crash reports. Crashes reported on the long forms involve either higher injury severity level or criminal activities such as hit-and-run or DUI. Since only long form crashes have been coded and archived in FDOT's CAR database. The research team has collected short form crashes from S4A. Therefore, the research team is able to use more complete crash data in this research project.

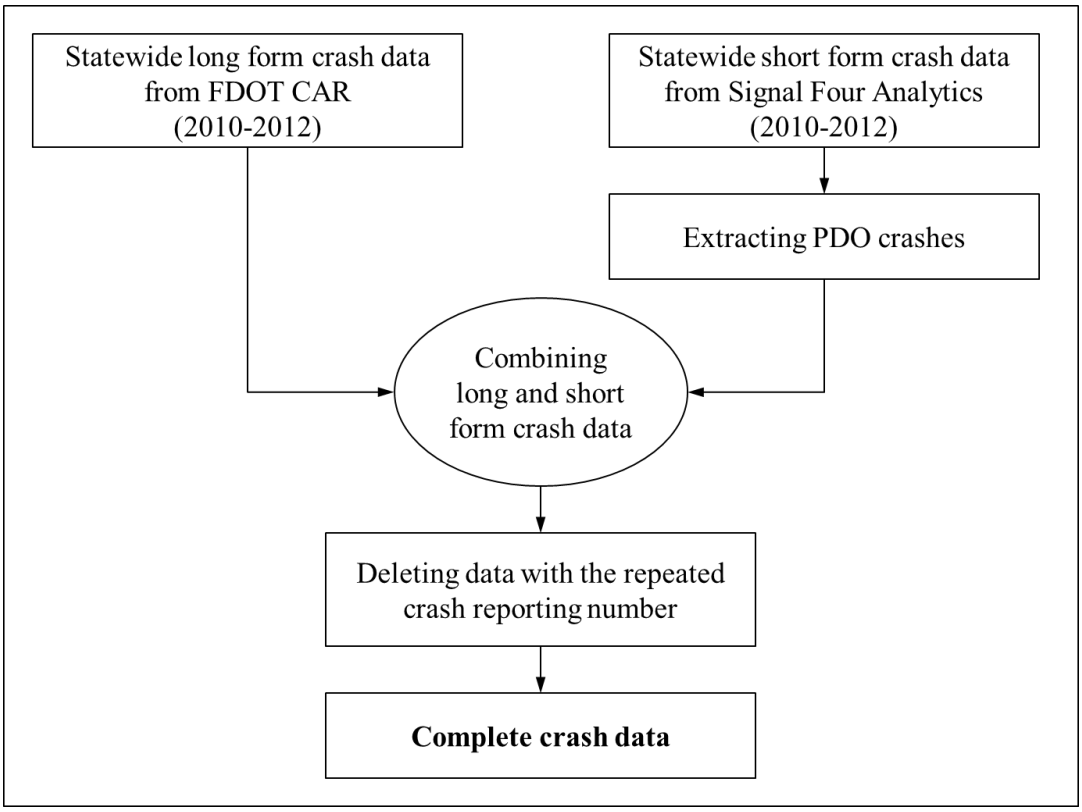

Figure 4 Crash Data Collection Process

The number of crashes by severity levels, form types, and years is shown in Table 4. The number of injury and fatal crashes are stable across 3 years. However, it is evident that many PDO (Property Damage Only) crashes in 2010-2011 are under-reported compared to the number of PDO crashes in 2012. The possible reasons for the underreporting of PDO crashes are as follows: First, S4A started to collected short form crash data from all counties in Florida from 2010 onward. However, very few short form crash data were collected in 2010 except for select counties. The number of reported short form crashes has significantly increased since 2011. Second, the crash report form has been changed in 2011, and thus it is thought that there was confusion in submitting crash reports. Third, The Florida Statutes regarding the crash reporting rules (F.S. 316.066) have been amended, and the number of reportable long form crashes has

increased since 2012. The amended Florida Statutes regulate that traffic crashes should be reported by long form if a crash: 1) resulted in death of, personal injury to, or any indication of complaints of pain or discomfort by any of the parties or passenger involved in the crash; 2) involves DUI (Driving Under the Influence or alcohol or drugs) or hit-and-run (F.S. 316.061(1) and 316.193); 3) rendered a vehicle inoperable to a degree that required a wrecker to remove it from the scene of the crash; or 4) involved a commercial motor vehicle. These possible reasons may increase the number of PDO reported long crashes in 2012. The State is moving in the right direction and the data appear to be more complete. More PDO crashes are captured by both long and short forms. There is an indication that the percent of PDO crashes reported on Long forms is increasing. In July 2010 agencies were no longer required to submit short forms, this led to some agencies to change to all long forms. We are trying to use as much complete crash database as possible, while maintaining consistency. This is difficult as it is apparent that the changes in 2010 and 2011 are impacting the number of reported crashes. Up to the time of writing this report, the 2013 geocoded crash data were not available from FDOT.

Table 4 The number of crashes by severity levels, form types, and years

Year	Severity levels			Source		Sum
	PDO	Injury	Fatal	S4A	CAR	
2010	147,872	122,288	2,183	15,370	256,973	272,343
2011	169,484	102,398	2,103	53,343	220,642	273,985
2012	241,321	111,450	2,136	99,885	255,022	354,907
Sum	558,677	336,136	6,422	168,598	732,637	901,235

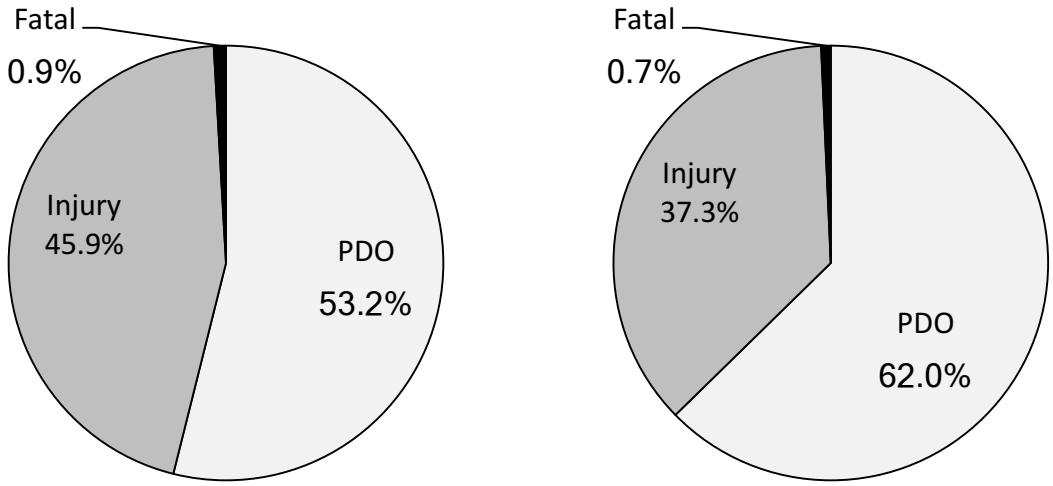

Figure 5 Comparison of the proportion of crashes by severity levels between long form only (left) and complete data (right)

As shown in Figure 5 Comparison of the proportion of crashes by severity levels between long form only (left) and complete data (right), crash data without short form reports (long form only data) have 45.9% of injury crashes and 53.2% PDO crashes. On the other hand, the percentage of injury crashes was dropped to 37.3% whereas PDO crashes were 62.0%, which is obviously more reasonable. Using data with many missing PDO crashes may result in biased model estimation, particularly for total and PDO SPFs (no effect for injury and fatal SPFs). Therefore, the complete data including both short and long form data were used in this research project.

Each yellow point in Figure 6 represents the location of a crash. Figure 7 shows the result of Kernel Density Estimation (KDE) of crashes, which defines the spread of risk as an area around a defined cluster in which there is an increased likelihood of a traffic crash to occur based on spatial dependency. As seen in Figure 4, the largest cluster is located in Miami-Dade County, and Hillsborough and Pinellas Counties and Orange County have the second and third largest

clusters, respectively. Also, Duval and Escambia Counties show the relatively high concentration of traffic crashes.

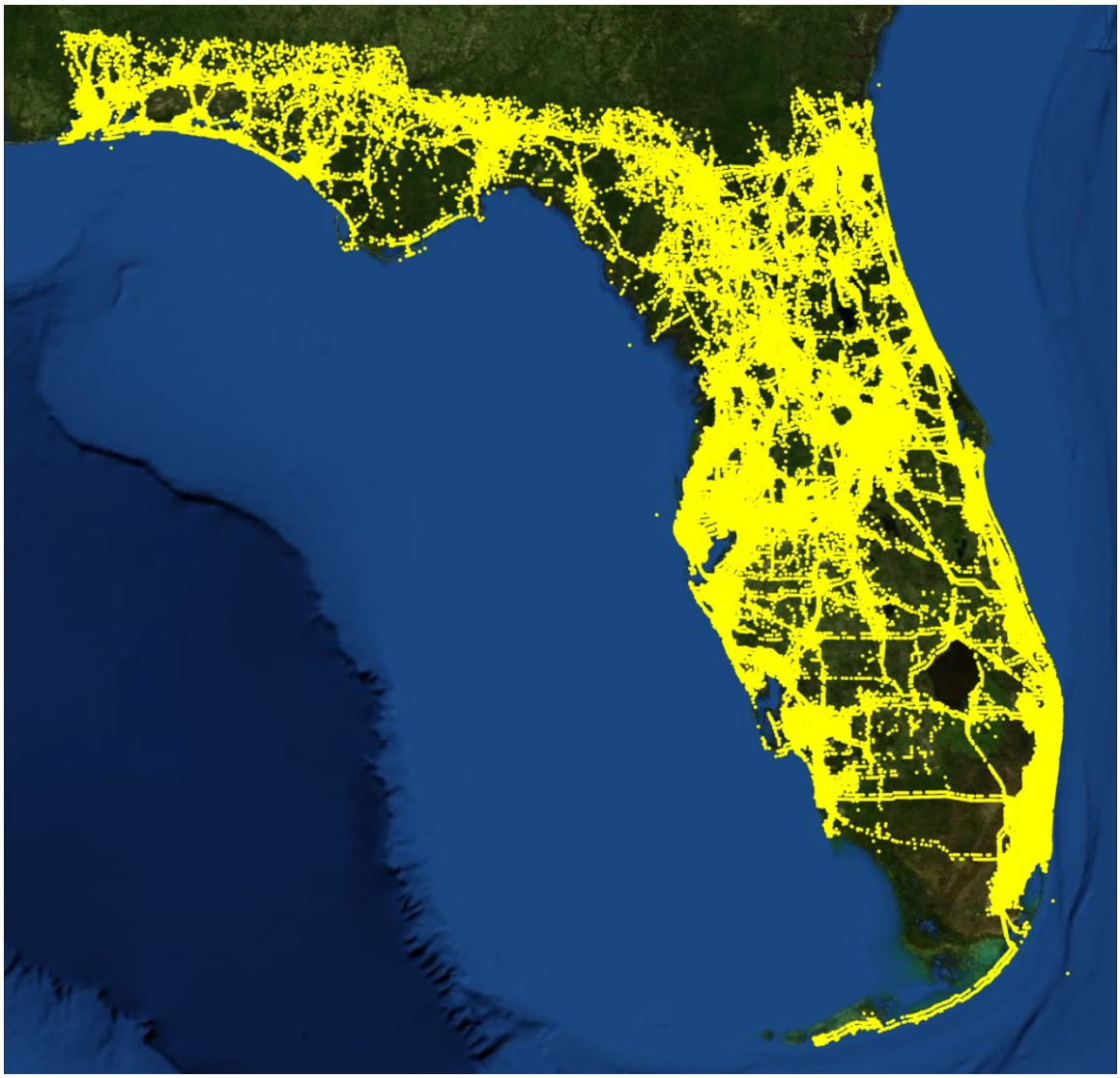

Figure 6 Spatial distribution of traffic crashes (2010-2012)

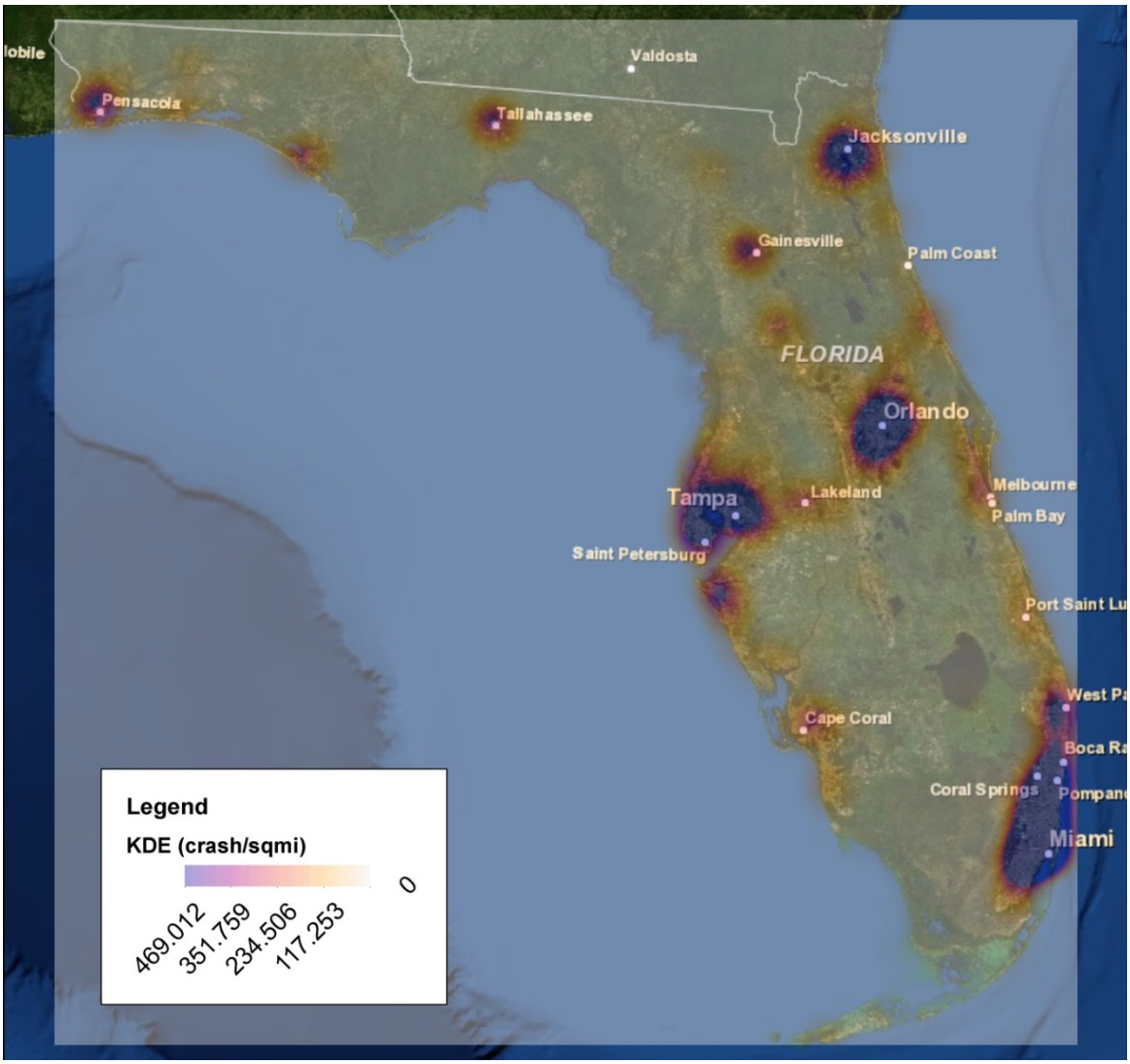

Figure 7 Kernel density estimation of traffic crashes

1.3 Exploratory Analysis of the Collected Data

The newly collected data has been processed for developing various SPFs (Safety Performance Functions) in this chapter. This chapter summarizes descriptive statistics and several spatial distributions of the collected data based on TAZs and TADs.

1.3.1 Traffic Analysis Zones

There are 8,518 TAZs in the State of Florida (Figure 2). TAZs cover the whole state and are used by FDOT Central Office for statewide long-term transportation plans. The collected data were processed based on TAZs and socio-demographic, roadway, and crash variables are summarized in Table 5,Table 6, and Table 7, respectively. Also, roadways by functional classifications and spatial distribution of total crashes are shown in Figure 8 and Figure 9, correspondingly. TAZs will be used for developing Traffic Safety Analysis Zones (TSAZs) in the following chapter.

Table 5 Descriptive statistics for socio-demographic variables in TAZs

Variables	Mean	Stdev	Min	Max
Total population	2172	3007	0	38980
Number of family unit	817	1147	0	18200
Proportion of the nonpermanent vacant	0.107	0.091	0	0.500
Proportion of the families vacant	0.071	0.068	0	0.500
Proportion of families have no vehicle	0.095	0.123	0	1.000
Proportion of families have 1 vehicle	0.372	0.146	0	1.000
Proportion of families have 2 or more vehicles	0.490	0.205	0	1.000
Number of HMT rooms per square mile	172.486	941.718	0	32610.839
Total employment	1140	1722	0	31931
Proportion of industry employment	0.176	0.232	0	1.000
Proportion of commercial employment	0.299	0.235	0	1.000
Proportion of service employment	0.492	0.259	0	1.000
School enrollments per square mile	775.020	5983.006	0	255140.358

Table 6 Descriptive statistics for roadway variables in TAZs

Variables	Mean	Stdev	Min	Max
Area (mi^2)	6.47	24.80	0	885.32
Road density	9.396	28.397	0	2496.049
Proportion of freeway/expressway	0.016	0.084	0	1.000
Proportion of principle arterial	0.104	0.202	0	1.000
Proportion of minor arterial	0.117	0.211	0	1.000
Proportion of collector road	0.191	0.246	0	1.000
Proportion of local road	0.572	0.329	0	1.000
Proportion of roadway length with low speed limit 5-30 mph	0.747	0.277	0	1.000
Proportion of roadway length with medium speed limit 35-50 mph	0.170	0.218	0	1.000
Proportion of roadway length with high speed limit 55-70 mph	0.059	0.150	0	1.000
Number of intersection per mile	16.699	230.370	0	8614.967
Number of signal per mile	2.904	86.103	0	6347.763
Number of intersection per square mile	57.081	149.704	0	4857.521
Number of signal per square mile	8.257	47.040	0	1619.174
Daily vehicle miles travel	31381.035	41852.293	0	684758.350
Proportion of daily heavy vehicle miles travel	0.067	0.052	0	0.519
Proportion of urban area	0.722	0.430	0	1.000

Table 7 Descriptive statistics for crashes in TAZs

Crash variables	Mean	Stdev	Min	Max	Sum	%
Total	105.80	142.25	0	1507	901235	100.0
Incapacitating injury	5.12	7.21	0	110	43631	4.8
Fatal	0.75	1.24	0	14	6408	0.7
Pedestrian	1.91	3.31	0	39	16240	1.8
Bicycle	1.80	3.31	0	88	15307	1.7
DUI	3.82	5.06	0	86	32545	3.6

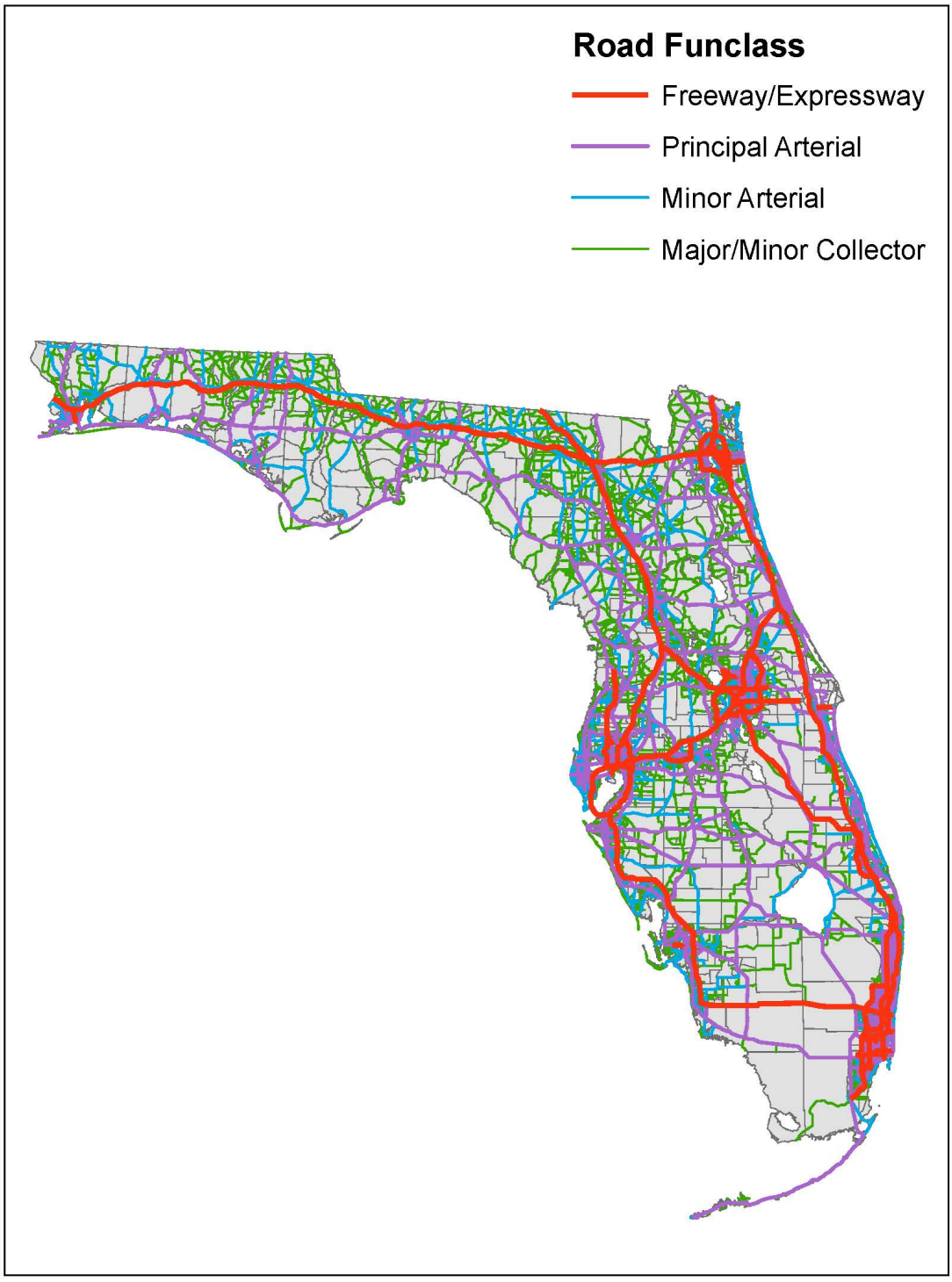

Figure 8 Roadways by functional classifications in TAZs

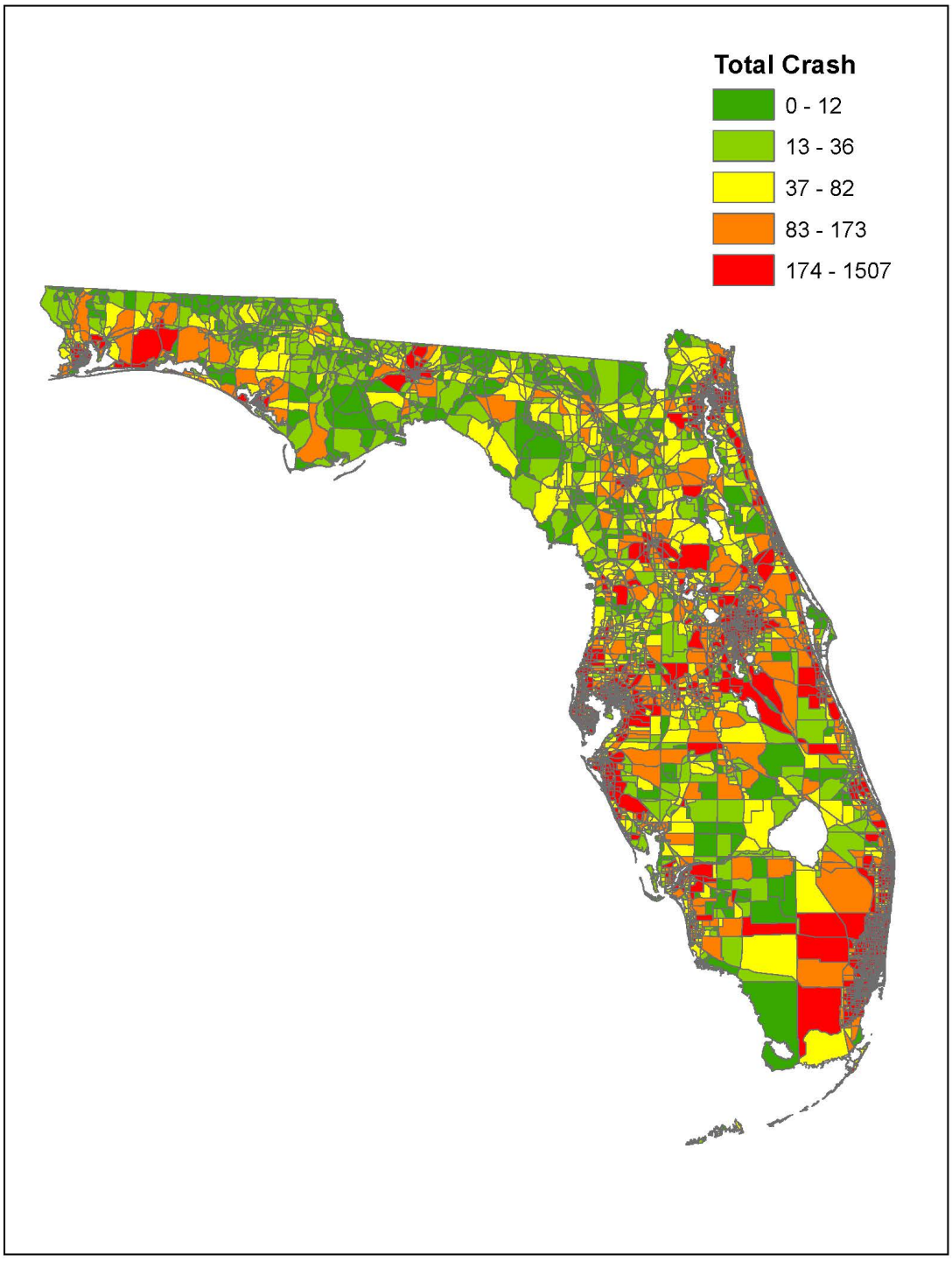

Figure 9 Spatial distributions of total crashes in TAZs

1.3.2 Traffic Analysis Districts

Similar to TAZs, TADs (Traffic Analysis Districts) cover the whole state (Figure 3). However, TAD is much more highly aggregated geographic unit compared to TSAZ. TAD may be useful if practitioners want to define crash pattern at a higher aggregate level. The collected data were prepared based on TAD and processed socio-demographic, roadway, and crash variables are summarized in Table 8, Table 9, and Table 10, correspondingly. Moreover, population density, roadways by functional classifications, and total crash maps are displayed in Figure 10, Figure 11, and Figure 12, respectively.

Table 8 Descriptive statistics for socio-demographic variables in TADs

Variables	Mean	Stdev	Min	Max
Total population	103.314	260.083	2.617	3095.520
Number of family unit	30718	35919	8	358901
Proportion of the nonpermanent vacant	11557	12454	2	108195
Proportion of the families vacant	0.102	0.045	0.000	0.310
Proportion of families have no vehicle	0.065	0.034	0.000	0.286
Proportion of families have 1 vehicle	0.077	0.065	0.004	0.544
Proportion of families have 2 or more vehicles	0.386	0.068	0.170	0.675
Number of hotel, motel, timeshare rooms per square mile	0.536	0.105	0.078	0.825
Total employment	38.145	96.745	0.000	766.641
Proportion of industry employment	16150	18159	7	157003
Proportion of commercial employment	0.177	0.136	0.000	0.819
Proportion of service employment	0.338	0.139	0.012	0.854
School enrollments per square mile	0.485	0.134	0.045	0.977

Table 9 Descriptive statistics for roadway variables in TADs

Variables	Mean	Stdev	Min	Max
Area (mi^2)	212.454	283.916	25.774	2685.062
Road density	7.613	5.311	0.074	24.561
Proportion of freeway/expressway	0.022	0.032	0.000	0.316
Proportion of principle arterial	0.053	0.045	0.000	0.314
Proportion of minor arterial	0.058	0.041	0.000	0.280
Proportion of collector road	0.112	0.066	0.000	0.603
Proportion of local road	0.755	0.108	0.077	0.935
Proportion of roadway length with low speed limit 5-30 mph	0.831	0.085	0.432	0.987
Proportion of roadway length with medium speed limit 35-50 mph	0.121	0.058	0.005	0.445
Proportion of roadway length with high speed limit 55-70 mph	0.048	0.057	0.000	0.425
Number of intersection per mile	1.995	1.115	0.217	7.881
Number of signal per mile	0.121	0.126	0.000	1.363
Number of intersection per square mile	17.895	19.765	0.130	126.392
Number of signal per square mile	1.171	1.728	0.000	13.376
Daily vehicle miles travel	599647	428747	38547	4632469
Proportion of heavy vehicle	0.071	0.038	0.015	0.290
Proportion of urban area	0.720	0.376	0.000	1.000

Table 10 Descriptive statistics for crashes in TADs

Crash variables	Mean	Stdev	Min	Max	Sum	%
Total	1517.23	1603.29	188	15094	901235	100.0
Incapacitating injury	73.45	54.57	4	457	43631	4.8
Fatal	10.79	8.13	0	77	6408	0.7
Pedestrian	27.34	33.39	1	344	16240	1.8
Bicycle	25.77	29.59	0	312	15307	1.7
DUI	54.79	36.31	6	345	32545	3.6

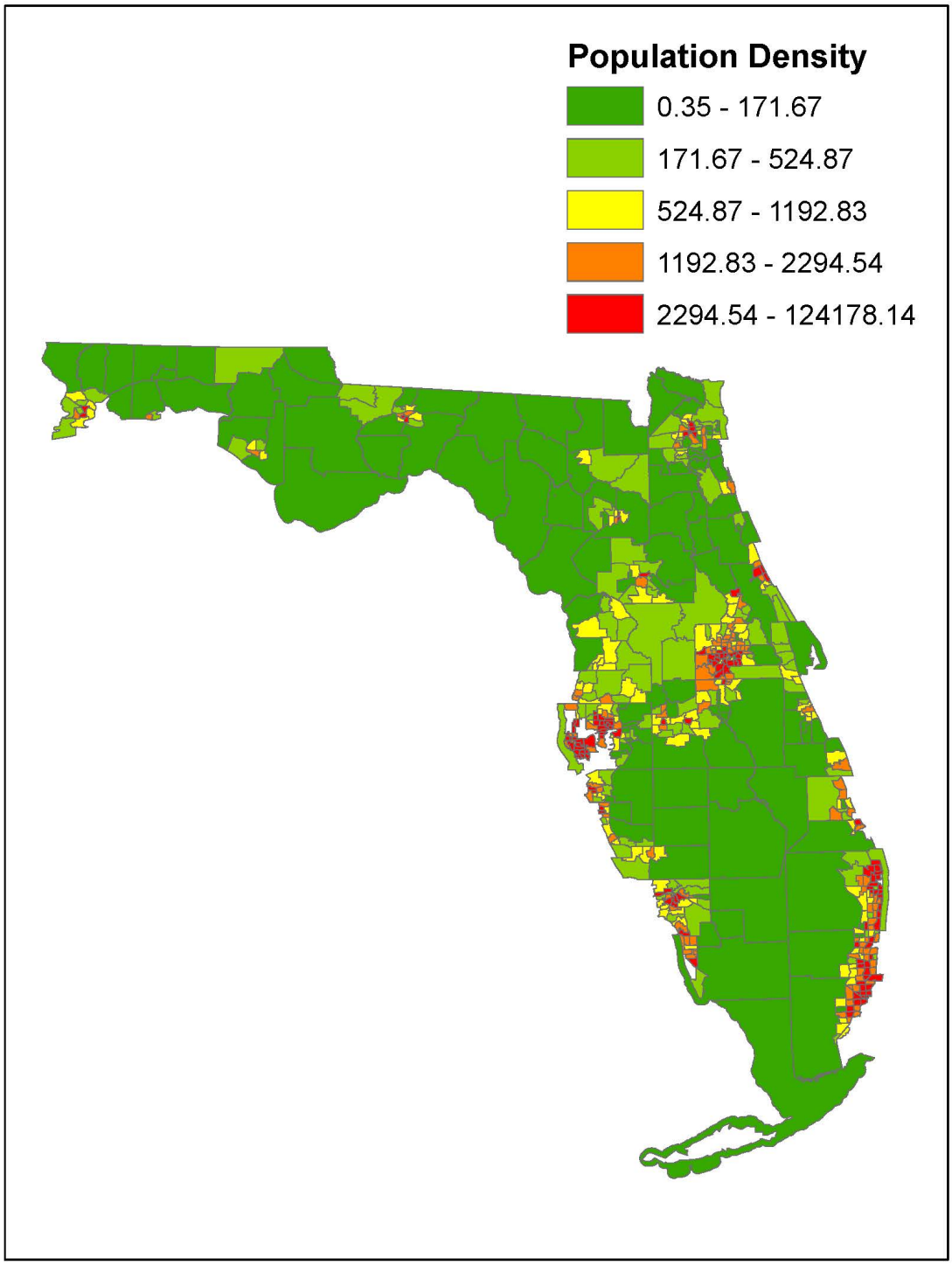

Figure 10 Population density based on TADs

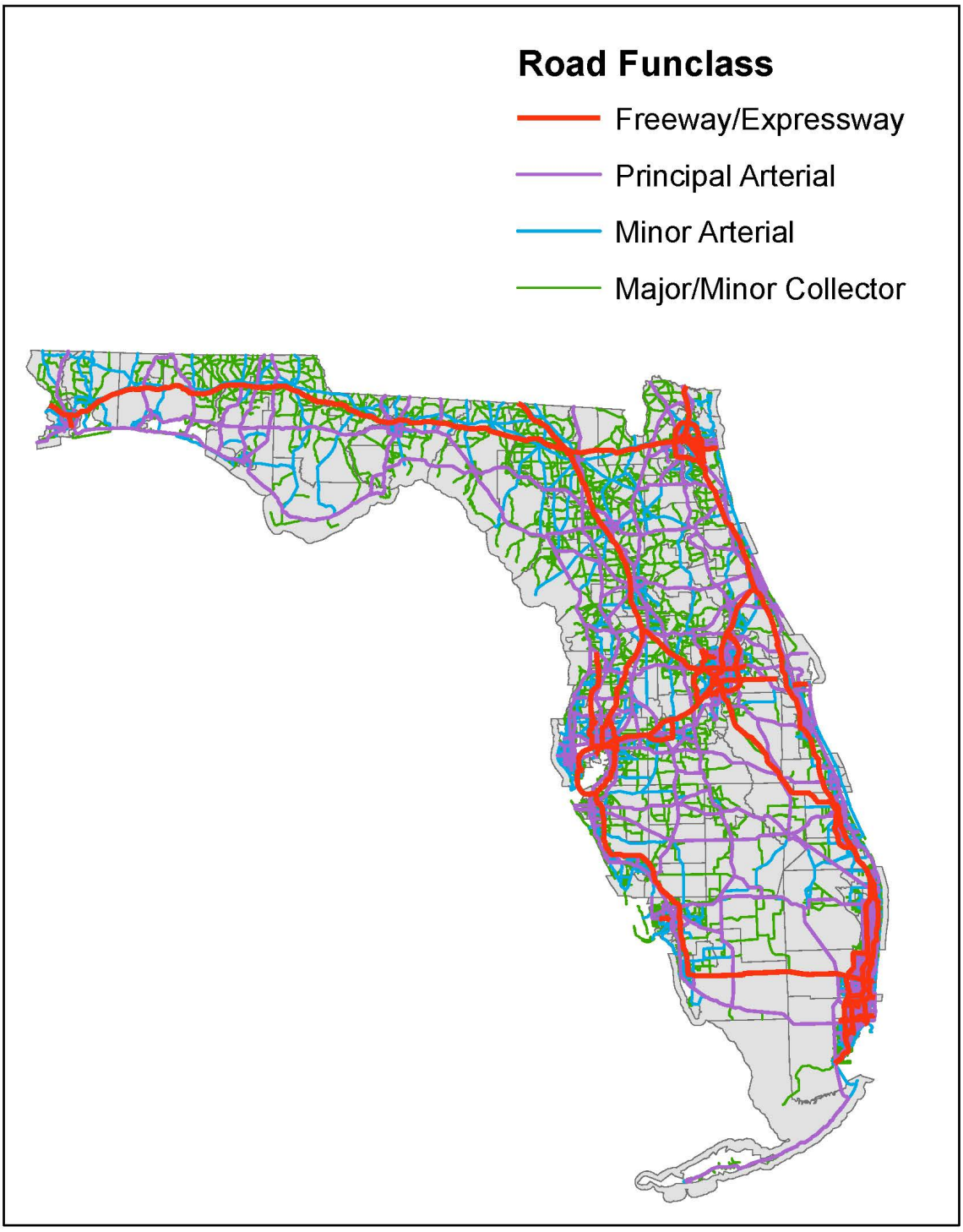

Figure 11 Roadways by functional classification in TADs

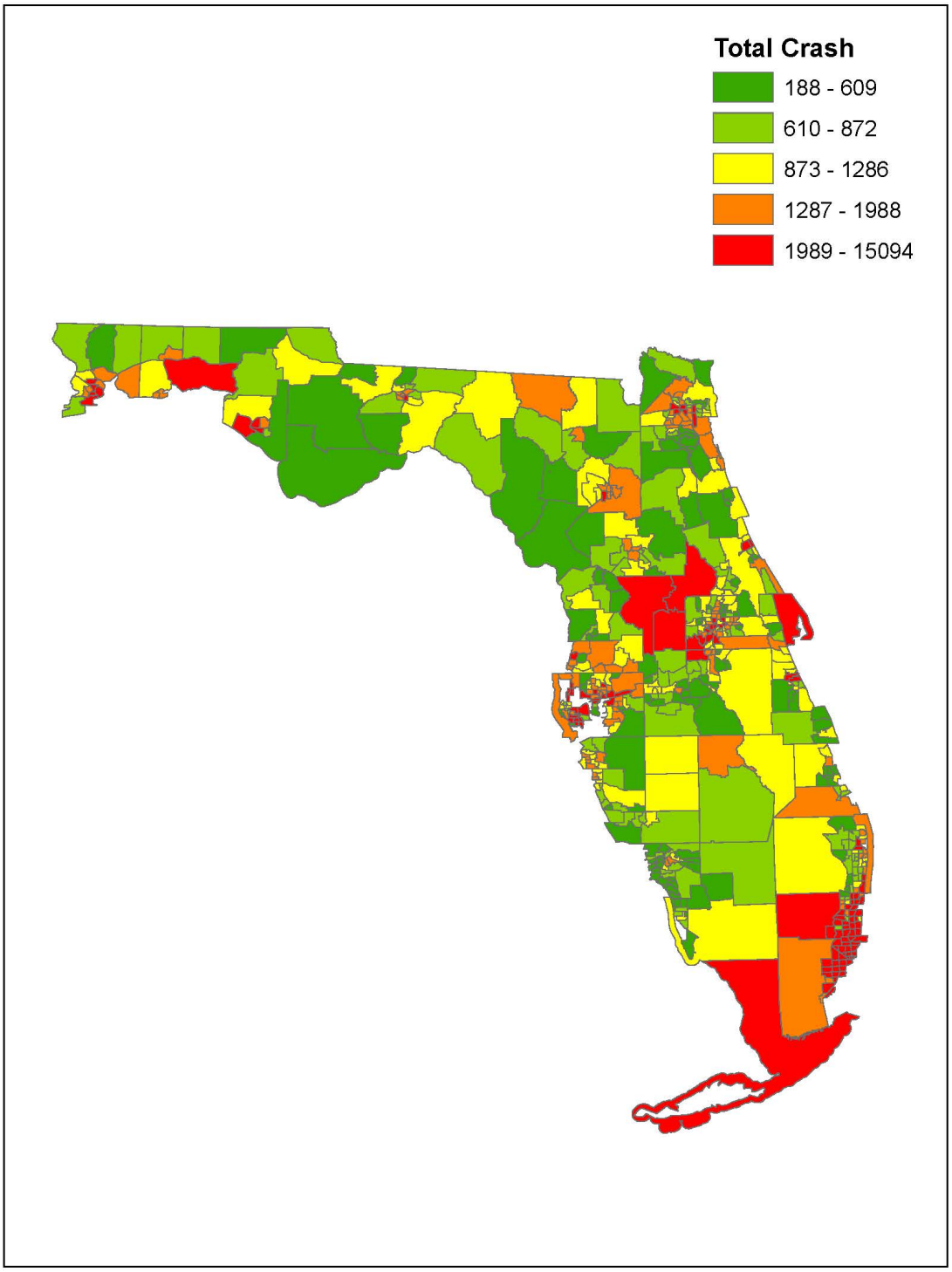

Figure 12 Spatial distributions of total crashes in TADs

1.4 Development of Traffic Safety Analysis Zones

Basically TAZs are designed to analyze origin-destination pairs of trips generated from each zone. Thus, transportation planners need to minimize trips which start and end in the same zone. It is inferred that minimizing intra-zonal trips end up with the small size of TAZs. On the other hand, traffic safety analysts need to consider traffic crashes that occur inside zones. So they can be related to zonal characteristics with traffic safety of the zones. Therefore, it is possible that TAZs are too small to analyze traffic safety at macroscopic level. Moreover, the small size of zones makes many zones with zero crash counts, especially for rarely occurring crashes such as fatal crashes. The second criterion abovementioned indicates that TAZs are usually divided by physical boundaries, mostly arterial roadways. Considering that many crashes occur on arterial roads, between zones, inaccurate results will be made from relating traffic crashes on the boundary of the zone to only the characteristics of that zone. A simple way to overcome these two issues while using TAZs for safety analysis is to aggregate TAZs into sufficiently large and homogenous traffic crash patterns. The existing TAZs were aggregated if they meet the following conditions:

- Zones are spatially contiguous; and
- Zones have same crash rate levels

All TAZs were classified into several categories based on their crash rates (crashes per square mile) as shown in Table 11. Subsequently, the neighboring zones with same categories are combined and new five zone systems were created (TSAZ1-5). The optimal zone scale for TSAZs (Traffic Safety Analysis zones) was determined using Brown-Forsythe (FB_F) test. FB_F test evaluates whether the variance of variables of interests (i.e. crash rates) is equal when the scales of zone systems change. The underlying assumption of FB_F is that there is greater

variance in crash rates among smaller zones and a lower variance among larger zones. A high variance value means that the crash risks are local, whereas a low variance means that more global crash patterns can be captured. The optimal zone scale ensures that the variance of crash rates is somewhere in between.

Table 11 Classification of TAZs by crash rates

No	Number of classifications	Classifications (based on percentile crash rate)	Range (crash per mile)
1	2	Class 1: 50-100%	20000-8.122
		Class 2: 0-50%	8.120-0.000
2	3	Class 1: 66-100%	20000-15.614
		Class 2: 33-66%	15.609-3.751
		Class 3: 0-33%	3.744-0.000
3	5	Class 1: 80-100%	20000-30.249
		Class 2: 60-80%	30.229-11.978
		Class 3: 40-60%	11.975-5.260
		Class 4: 20-40%	5.258-1.616
		Class 5: 0-20%	1.615-0.000
4	7	Class 1: 86-100%	20000-44.702
		Class 2: 71-86%	44.690-19.305
		Class 3: 57-71%	19.296-10.660
		Class 4: 43-57%	10.658-6.058
		Class 5: 29-43%	6.056-2.879
		Class 6: 14-29%	2.878-0.952
		Class 7: 0-14%	0.951-0.000
5	10	Class 1: 90-100%	20000-66.773
		Class 2: 80-90%	66.681-30.249
		Class 3: 70-80%	30.229-18.126
		Class 4: 60-70%	18.102-11.978
		Class 5: 50-60%	11.975-8.122
		Class 6: 40-50%	8.120-5.260
		Class 7: 30-40%	5.258-3.118
		Class 8: 20-30%	3.116-1.616
		Class 9: 10-20%	1.615-0.548
		Class 10: 0-10%	0.546-0.000

F_{BF} statistics are calculated using the following formula:

$$F_{BF} = \left[\sum_{i=1}^{t}\left(\overline{D}_i-\overline{D}\right)^2/(t-1)\right] \Big/ \left[\sum_{i=1}^{t}\sum_{j=1}^{n_1}\left(\overline{D}_{ij}-\overline{D}_1\right)^2/(N-t)\right] \quad (1)$$

where, n_i is the number of samples in the ith zone system, N is the total number of samples for all zone systems, t is the number of neighborhood groups y_{ij} is the crash rates of the jth sample

from the *i*th zone system, $\bar{y}_i$ is the median of crash rate from the *i*th zone system, and $D_{ij}=\left|y_{ij}-\bar{y}_i\right|$ is the absolute deviation of the *j*th observation from the *i*th zone system median, $\overline{D}_i$ is the mean of D_{ij} for zone system *i*, and $\overline{D}$ is the mean of all D_{ij}. The test assumes that the variances of different zones are equal under the null hypothesis. The calculated value was obtained using an *F* distribution and α=0.1 was used to test for statistical significance.

There are two steps involved in the F_{BF} test. First, the variance between each zone system from TSAZ5 (N=4,907) to TSAZ1 (N=1,064) (Table 12). The largest zone system (TSAZ1) is compared for a total of 4 separate calculations of F_{BF}, as shown in the F_{BF1} column of Table 12. Second, the variance between each neighborhood group from TSAZ1 to TSAZ4 and the smallest zone system (TSAZ5) is compared (F_{BF2}). TSAZ5 was used as the smallest zone system instead of SAZ (N=8,518) since the variance of crash rates based on TAZs is very large (Standard deviation=3035.39), which shows the crash rates are not relevant to TAZs. A significant value of F_{BF1} implies that the zone system does not reflect the global pattern of crash data; in essence each zone is so small that it only captures local crash patterns. On the contrary, the significant value of F_{BF2} indicates that the zone data are not local; they are so large that local level crash patterns are undetectable. The zone systems between lower and upper limits identify a spatial scale at which local level variation is still detectable but also captures larger zonal level crash characteristics.

Table 12 Brown-Forsythe test for determining optimal zone scale

Zone system	No of zones	Crashes per miles		Brown-Forsythe test			
		Mean	Stdev	F_{BF1}	p-value	F_{BF2}	p-value
TAZ	8,518	144.588	3035.390	-	-	-	-
TSAZ5	4,907	14.614	53.510	3.630	0.0567	-	-
TSAZ4	3,920	14.678	59.152	2.810	0.0936	0.010	0.9436
TSAZ3	**3,041**	**14.947**	**66.557**	**1.960**	**0.1617**	**0.060**	**0.8134**
TSAZ2	**1,754**	**15.634**	**86.843**	**0.440**	**0.5081**	**1.070**	**0.3002**
TSAZ1	1,064	18.036	110.703	-	-	3.630	0.0567

The F_{BF} test results for homogeneity of variance for crash rates under various zone scales are presented in Table 12. The F_{BF1} test statistics shows that zone systems smaller than TSAZ3 (i.e., TSAZ4 and TSAZ5) have significantly different variance from that of TSAZ1. Thus, zone systems smaller than TSAZ3 are too small to capture global crash patterns. On the other hand, F_{BF2} test statistics indicates that the zone system larger than TSAZ2 (i.e., TSAZ1) is so large that it cannot capture local crash characteristics. Given the result, systems with TSAZ2 and TSAZ3 are considered optimal for macro-level crash analysis. In conclusion, TSAZ2 was chosen as the final TSAZ since it can minimize boundary crashes and zones without certain types of crashes.

Table 13 contrasts the areas in TSAZ and TAZ. As shown in the table, the number of zones in TSAZ (N=1,754) is one-fifth of TSAZ (N=8,518), and the average area in TSAZ is 18.061 mi^2 whereas that in TAZ is 6.472 mi^2.

Table 13 Areas in TAZ and TSAZ

Zone system	No of zones	Average (mi^2)	Stdev	Min	Max
TAZ	8,518	6.472	24.803	0.000	885.322
TSAZ	1,754	18.061	226.645	0.000001	9395.0400

Table 14 compares the crash rates in TAZ and TSAZ. The mean crash rate in TAZ is 144.588 crashes per mile while that in TSAZ is almost one-tenth, 15.634 crashes per mile. Moreover, the standard deviation of crash rate in TAZ is very large, it is 3035.390. After the regionalization, the standard deviation of crash rate in TSAZ became 86.843. It may imply that the new zone system, TSAZ have more homogenous crash rates compare to TAZ.

Table 14 Crash rates in TAZ and TSAZ

Zone system	Average (crash per mi)	Stdev	Min	Max
TAZ	144.588	3035.390	0.000	2517.986
TSAZ	15.634	86.843	0.000	20000

Table 15 contrasts the numbers and percentages of zones without crashes in TAZ and TSAZ. There is no big difference in the percentage of zones without total crashes before and after the regionalization. However, when it comes to fatal crashes, the percentage of zones without fatal crashes in TAZ is 63.0% while it is smaller in TSAZ (54.1%).

Table 15 Zones without crashes in TAZ and TSAZ

Zone system	Zones without total crashes		Zones without fatal crashes	
	Zones	Percentage	Zones	Percentage
TAZ	291	3.4%	5363	63.0%
TSAZ	90	3.0%	1664	54.1%

Table 16 compares the numbers and percentages of boundary crashes in TAZ and TSAZ. There are 68.2% boundary crashes in TAZ whereas there are 47.0% boundary crashes in TSAZ. In other words, more than 20% of boundary crashes has been reduced after the regionalization.

Table 16 Boundary crashes in TAZ and TSAZ

Zone system	Boundary crashes	Total crashes	Percentage
TAZ	614,671	901,235	68.2%
TSAZ	423,275		47.0%

Figures 1-4 compare TAZ and TSAZ maps in, Districts 2 (Jacksonville area), 5 (Orlando area), 6 (Miami-Dade and Broward area), 7 (Tampa and St. Petersburg area), respectively. As presented in the Figures, the zones, especially in the urban area, are highly aggregated.

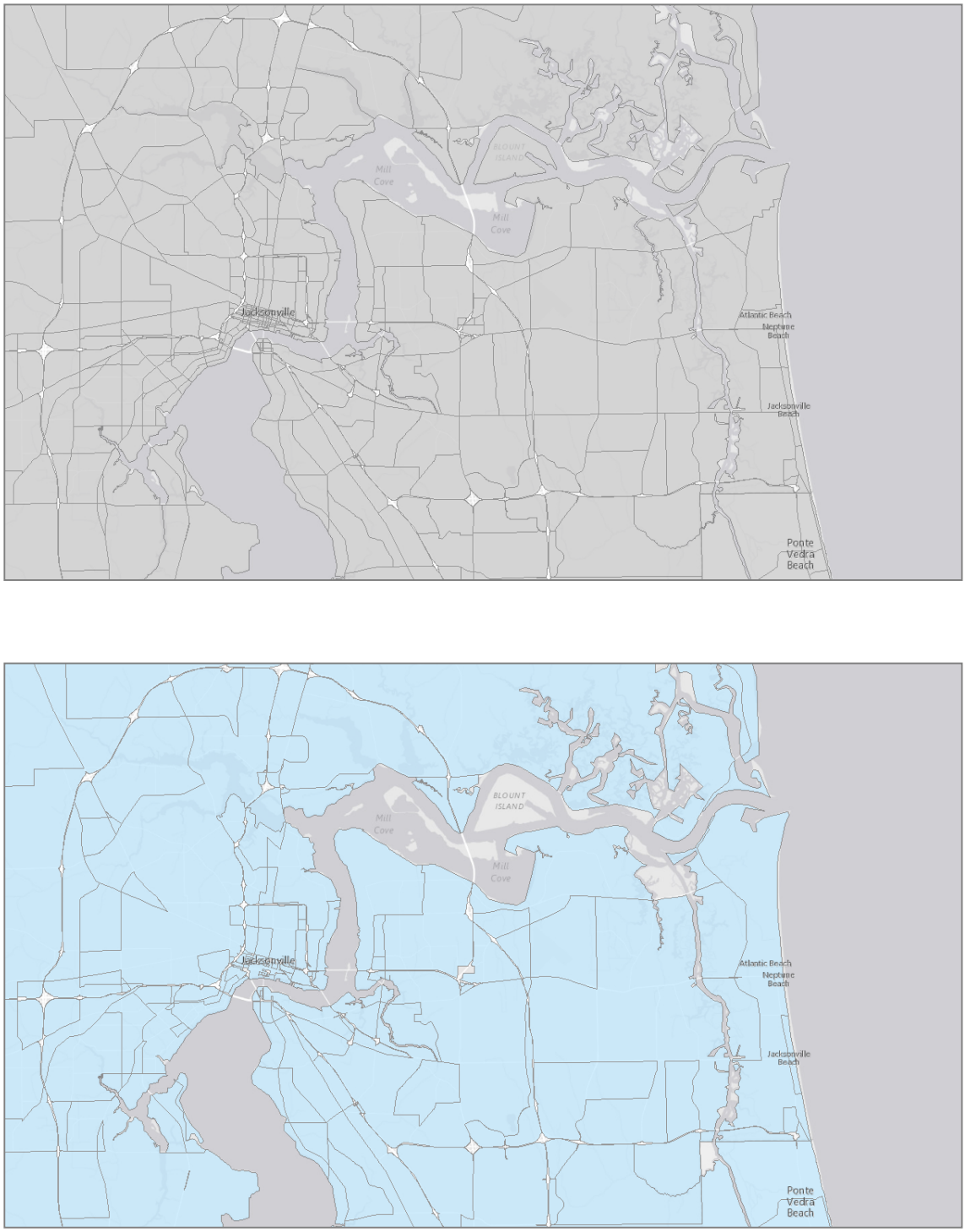

Figure 13 TAZ (upper) and TSAZ (lower) in District 2 – Jacksonville Area

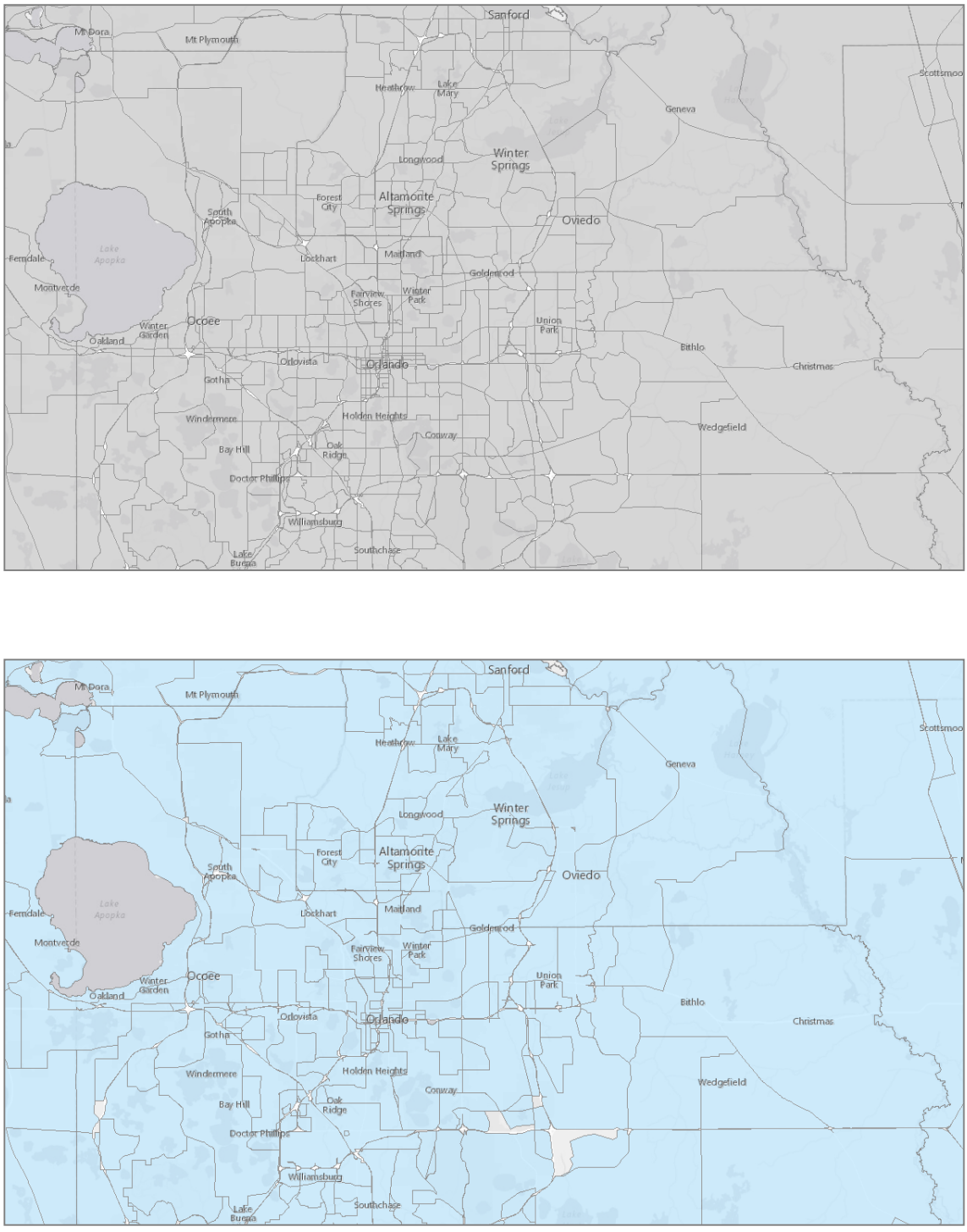

Figure 14 TAZ (upper) and TSAZ (lower) in District 5 – Orlando Metropolitan Area

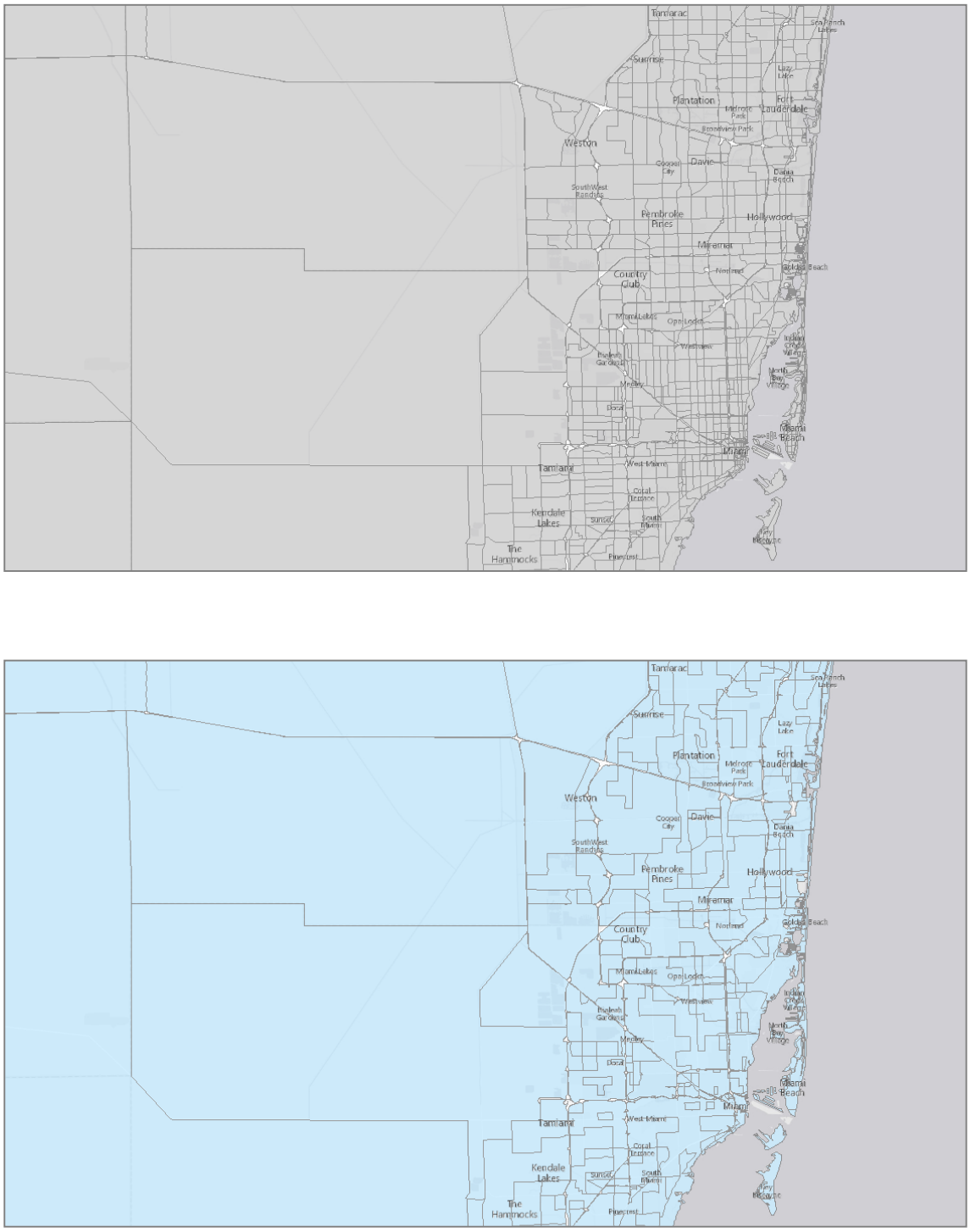

Figure 15 TAZ (upper) and TSAZ (lower) in District 6 – Miami-Dade and Broward Area

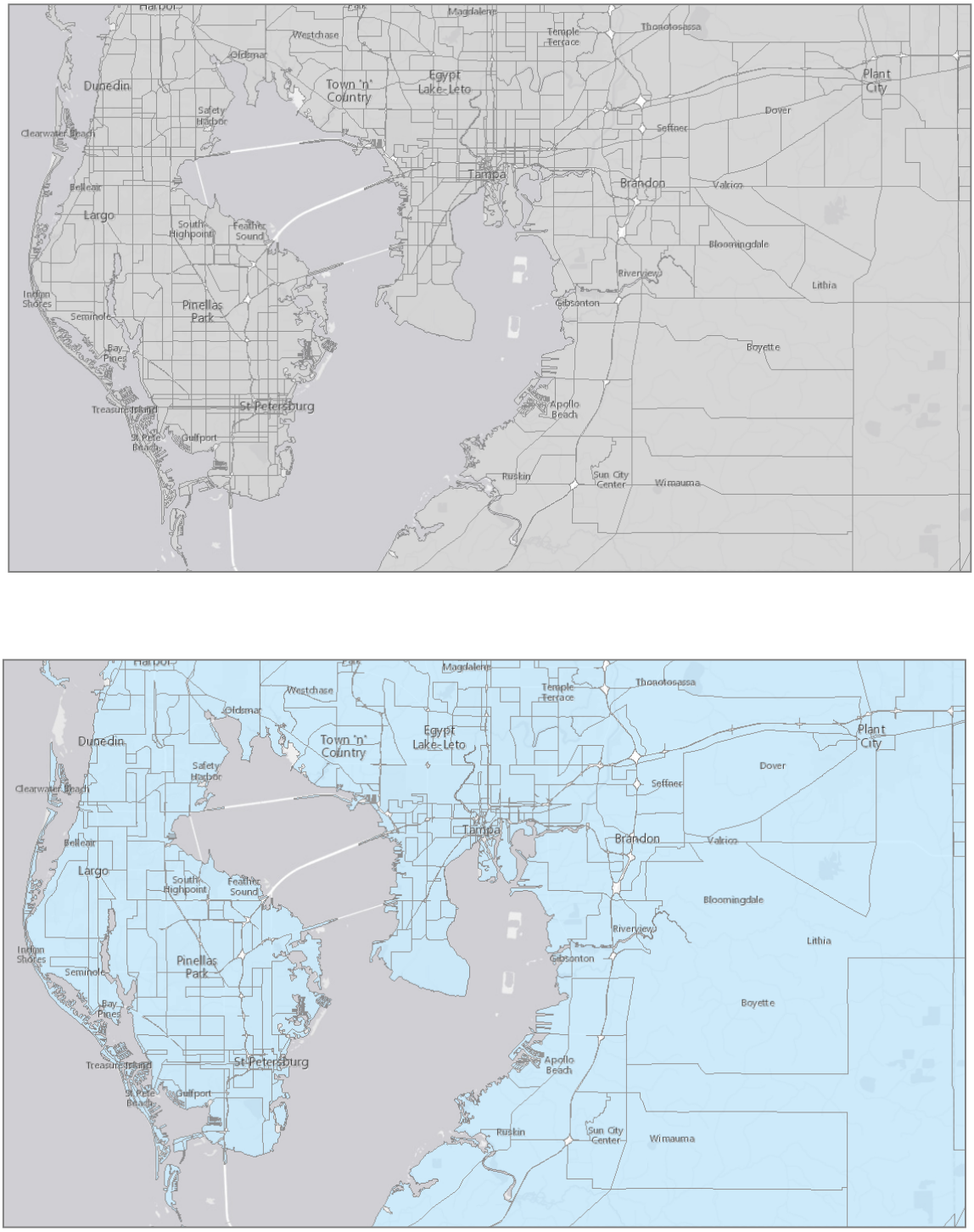

Figure 16 TAZ (upper) and TSAZ (lower) in District 7- Tampa and St. Petersburg Area

1.5 Estimation of Safety Performance Functions for TAZs

TAZs have been widely adopted for macroscopic traffic safety analysis since they are the only spatial unit related transportation. One of the advantages of using TAZs is that we can directly use transportation planning data based on TAZs for traffic safety analysis. Table 17 presents the SPFs for total, severe, pedestrian, bicycle, and DUI crashes based on TAZs.

Table 17 Safety Performance Functions (SPFs) based on TAZs

Parameters	Total	Severe	Pedestrian	Bicycle	DUI
Intercept	0.5239 (<.0001)	-3.1497 (<.0001)	-5.3065 (<.0001)	-5.2053 (<.0001)	-3.7878 (<.0001)
Log of hotel, motel, timeshare rooms per square mile	0.0184 (<.0001)	-0.0281 (<.0001)	0.0250 (<.0001)	0.0184 (0.0022)	0.1932 (<.0001)
Log of total employments	0.2513 (<.0001)	0.2175 (<.0001)	0.2888 (<.0001)	0.2508 (<.0001)	
Log of school enrollments per square mile	0.0483 (<.0001)	0.0118 (0.0013)	0.0434 (<.0001)	0.0439 (<.0001)	0.0101 (0.0053)
Proportion of collectors	-0.6579 (<.0001)	-0.3388 (<.0001)	0.9312 (<.0001)		-0.1478 (0.0120)
Proportion of local roads		0.2688 (<.0001)		0.9430 (<.0001)	0.5341 (<.0001)
Log of signals per mile				0.1524 (<.0001)	
Log of vehicle-miles-traveled	0.1228 (<.0001)	0.2294 (<.0001)	0.1788 (<.0001)	0.2124 (<.0001)	0.2504 (<.0001)
Proportion of heavy vehicles	-1.5487 (<.0001)		-2.3843 (<.0001)	-5.5383 (<.0001)	-2.4019 (<.0001)
Log of bike lane length		0.0888 (0.0008)	-0.1039 (0.0015)		0.1021 (<.0001)
Log of sidewalk length	0.2633 (<.0001)	0.1002 (<.0001)	0.4166 (<.0001)	0.4169 (<.0001)	0.1772 (<.0001)
Proportion of commuters using public transportation	3.6087 (<.0001)	0.5743 (0.0315)	5.9113 (<.0001)	3.1999 (<.0001)	-0.8465 (0.0017)
Proportion of commuters using bicycle	-1.1734 (0.0050)	-1.0187 (0.0319)	1.3750 (0.0124)	5.7402 (<.0001)	1.4123 (0.0008)
Proportion of commuters by walking	-1.1944 (<.0001)	-1.6823 (<.0001)	1.4496 (<.0001)	1.1972 (0.0012)	
Log of distance to the nearest urban area	-0.0462 (<.0001)	-0.0124 (<.0001)	-0.0589 (<.0001)	-0.1409 (<.0001)	-0.0181 (<.0001)
Over-dispersion	0.7844	0.5993	0.5762	0.6208	0.4603
LL	-44338.2	-21595.1	-12783.8	-12145.9	-18145.0
AIC	88702.4	43218.2	25597.5	24319.9	36315.9
BIC	88794.1	43316.9	25703.3	24418.6	36407.6
MAD	60.74	3.66	1.42	1.36	2.36
Adj_R2	0.435	0.368	0.418	0.398	0.432

1.6 Estimation of Safety Performance Functions for TSAZs

As explained in Section 1.3, TSAZs were developed to overcome the existing TAZs. The estimated TSAZ SPFs for total, severe, pedestrian, bicycle and DUI crashes are summarized in Table 8.

Table 18 Safety Performance Functions (SPFs) based on TSAZs

Parameters	Total	Severe	Pedestrian	Bicycle	DUI
Intercept	0.4533 (<.0001)	-3.3923 (<.0001)	-6.8921 (<.0001)	-7.6585 (<.0001)	-4.3574 (<.0001)
Log of hotel, motel, timeshare rooms per square mile	0.1197 (<.0001)				
Log of school enrollments per square mile			0.1136 (<.0001)	0.0925 (<.0001)	0.0498 (<.0001)
Proportion of arterials				-0.4590 (<.0001)	
Proportion of local roads	1.0929 (<.0001)	1.2773 (<.0001)	1.5750 (<.0001)	1.3080 (0.0215)	1.2475 (<.0001)
Log of signals per mile			0.4700 (<.0001)	0.4601 (<.0001)	0.1239 (0.0449)
Log of vehicle-miles-traveled	0.2676 (<.0001)	0.3709 (<.0001)	0.5112 (<.0001)	0.5722 (<.0001)	0.4385 (<.0001)
Proportion of heavy vehicles	-4.2794 (<.0001)		-5.3685 (<.0001)	-6.1078 (<.0001)	-5.3321 (<.0001)
Proportion of urban areas		-0.1972 (0.0072)		0.5582 (<.0001)	-0.1766 (0.0194)
Log of bike lane length	0.5317 (<.0001)	0.5235 (<.0001)	0.3627 (<.0001)	0.3351 (<.0001)	0.4256 (<.0001)
Proportion of commuters using public transportation	7.9379 (<.0001)	4.7546 (<.0001)	10.3840 (<.0001)	5.8575 (<.0001)	2.1931 (0.0039)
Proportion of commuters using bicycle	-4.2149 (0.0029)			8.6063 (<.0001)	
Proportion of commuters by walking	-4.0968 (<.0001)				
Over-dispersion	1.2057	0.8927	0.6988	0.6865	0.5783
LL	-10570.0	-5727.2	-3586.6	-3466.3	-4924.2
AIC	21159.9	11468.4	7191.2	6956.7	9868.3
BIC	21214.6	11506.7	7240.5	7022.3	9923.0
MAD	326.61	16.00	5.24	4.59	9.21
Adj_R2	0.472	0.698	0.648	0.725	0.714

1.7 Estimation of Safety Performance Functions for TADs

The research team built various SPFs for Traffic Analysis Districts (TADs). TADs are new, higher-level geographic entities for traffic analysis. Compared to TAZs and TSAZs, TADs are a much more highly aggregated geographic unit. TADs can be useful if practitioners want to analyze crash pattern at a higher aggregate level than TAZs. Table 19 shows SPFs for total, severe, pedestrian, bicycle, and DUI crashes, respectively.

Table 19 Safety Performance Functions (SPFs) based on TADs

Parameters	Total	Severe	Pedestrian	Bicycle	DUI
Intercept	-5.7374 (<.0001)	0.0628 (<.0001)	-0.8715 (0.0534)	-1.1564 (0.0294)	-3.5760 (<.0001)
Natural log of hotel, motel, timeshare room density	0.0359 (0.0023)		0.0468 (0.0031)	0.0443 (0.0126)	
Proportion of families with no vehicle					0.5708 (0.0426)
Proportion of roadway length with Posted Speed Limit higher than 55 mph					-1.2143 (0.0055)
Natural log of intersections per mile	0.3141 (<.0001)		0.3588 (<.0001)	0.3891 (<.0001)	0.1715 (<.0001)
Natural Log of VMT	0.4093 (<.0001)	0.3579 (<.0001)	0.0615 (0.0948)	0.0820 (0.0428)	0.2992 (<.0001)
Proportion of urban area	0.3020 (<.0001)	-0.2571 (<.0001)		0.3980 (<.0001)	-0.1803 (0.0098)
Natural log of bike lane length				0.0708 (0.0027)	
Natural log of sidewalk length	0.0786 (<.0001)	0.1105 (<.0001)	0.1803 (<.0001)		0.1976 (<.0001)
Natural log of number of total commuters	0.3020 (<.0001)	0.3739 (<.0001)			0.2218 (<.0001)
Proportion of commuters using public transportation			0.1849 (<.0001)	0.1339 (<.0001)	
Proportion of commuters using bicycle				0.1958 (<.0001)	
Proportion of commuters by walking			0.1017 (<.0001)		
Over-dispersion	0.1400	0.2177	0.2103	0.2528	0.1579
LL	-4445.9	-2940.0	-2210.2	-2204.6	-2625.7
AIC	8907.7	5894.0	4436.4	4427.3	5269.4
BIC	8942.8	5924.7	4471.5	4466.7	5308.9
MAD	447.82	30.74	10.75	10.56	18.22
Adj_R2	0.755	0.492	0.636	0.584	0.498

1.8 Hot Zone Identification

The PSI (Potential for Safety Improvement), or excess crash frequency, is a measure of how many crashes can be effectively reduced for a particular site and is suggested in the Highway Safety Manual (AASHTO, 2010). The PSI for each zone is the difference between the expected crash count and the predicted crash count. A hot zone in this study is defined as a zone with top 10% highest PSI.

The identified hot zones for total crashes are displayed in Figure 17. As shown in the figure, most of the hot zones are located in urban area with some exceptions. Figure 18 depicts the hot zones for severe crashes. Different from total crash hot zones, severe crash hot zones are more in rural areas rather than urban areas. Figure 19 shows the location of pedestrian crash hot zones. Including some areas in Miami-Dade County, they are mostly placed in urban or suburban areas. It is thought that these locations are related to residential land-use. Figure 20 presents the bicycle crash hot zones. It is interesting that the identified bicycle hot zones are quite similar to those in the pedestrian hot zones. Lastly, Figure 21 displays the DUI crash hot zones. It seems they are more concentrated in rural or suburban areas. However, it was shown that Tampa and St. Petersburg areas are dangerous for DUI crashes. It is worthy to note that the far south area, Everglades and Key West areas were classified as hot zones for all crash types. Possibly it is because there are many tourists in these areas, who are more likely to be vulnerable to traffic crashes.

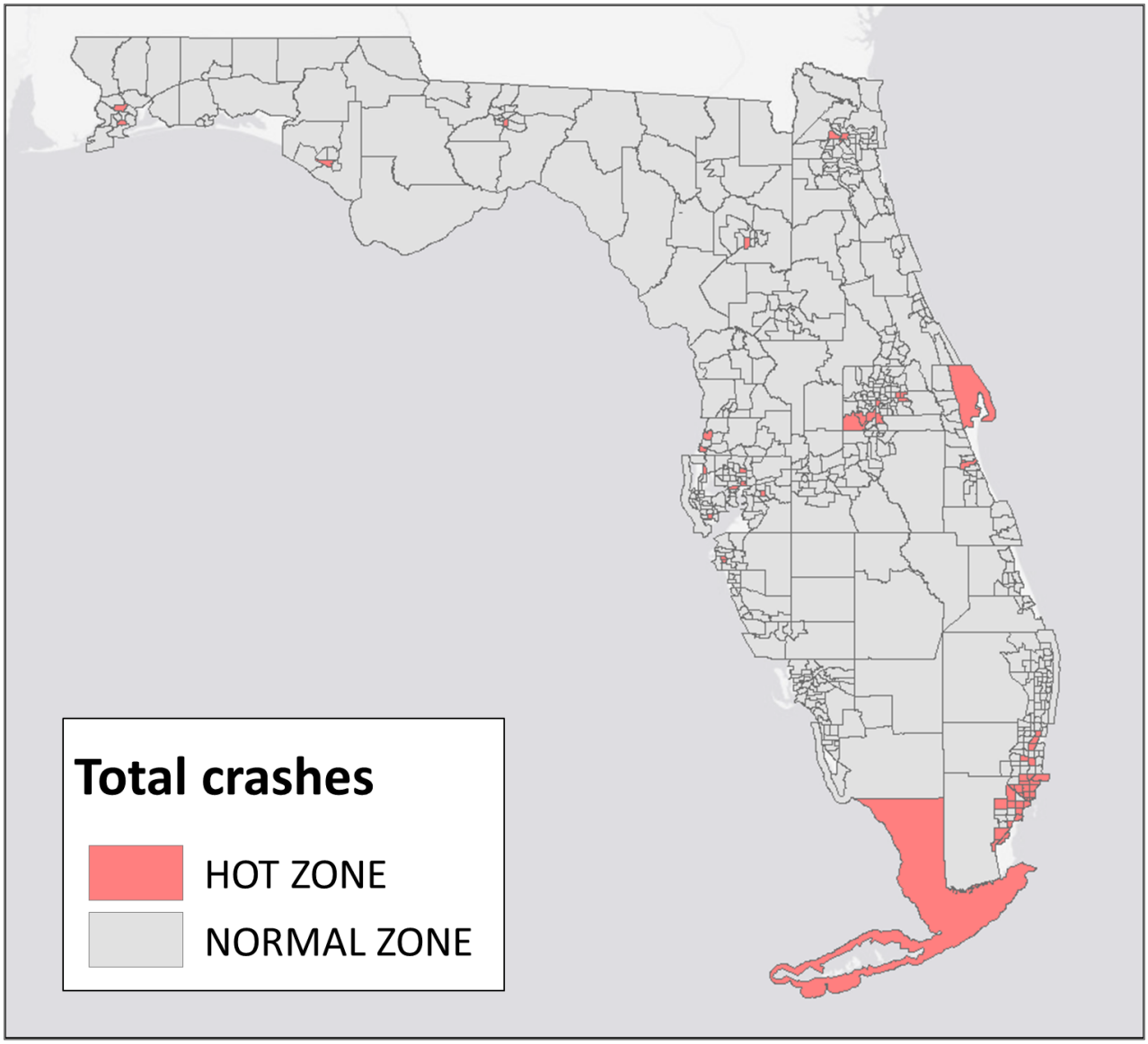

Figure 17 Hot zones for total crashes

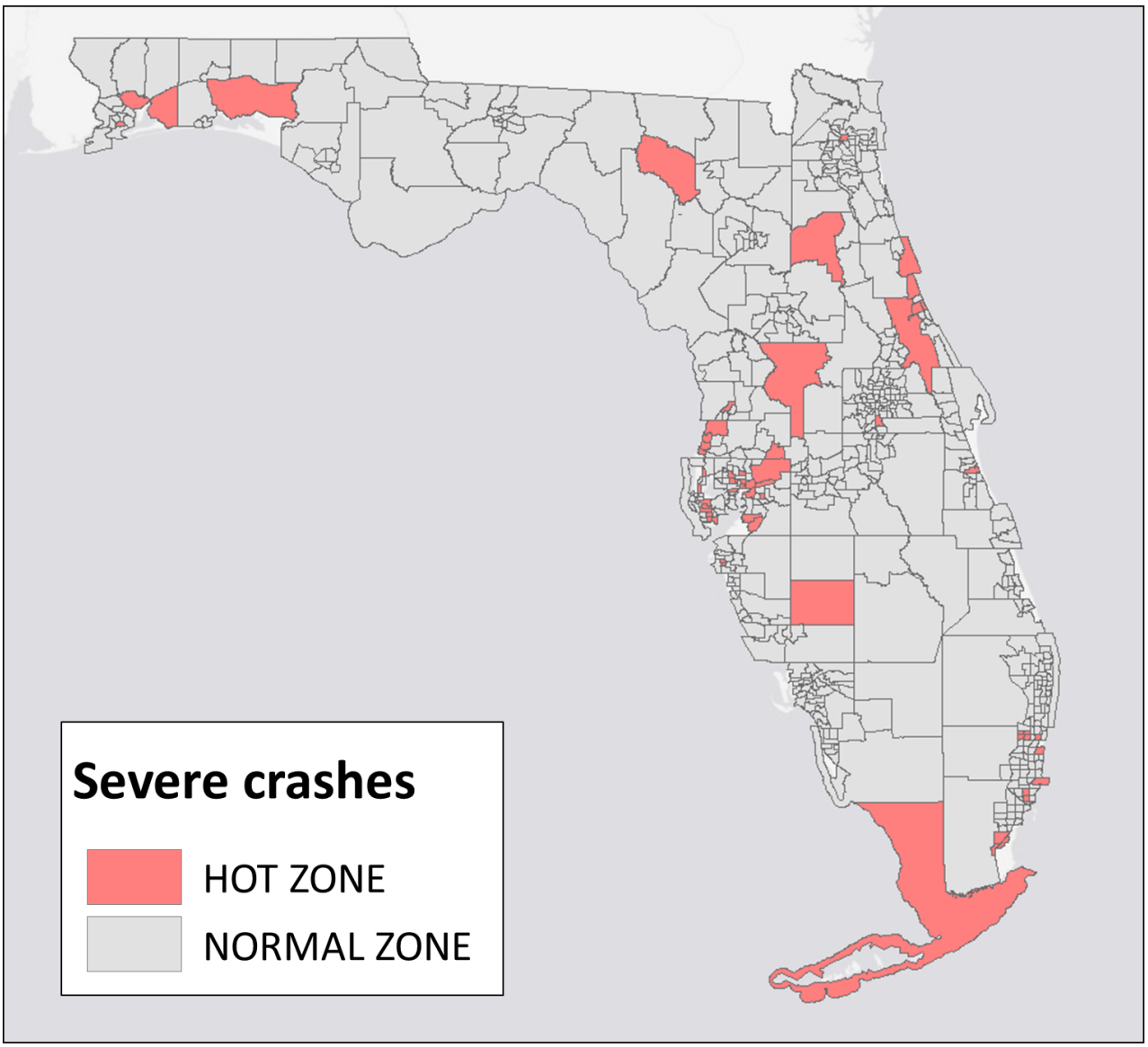

Figure 18 Hot zones for severe crashes

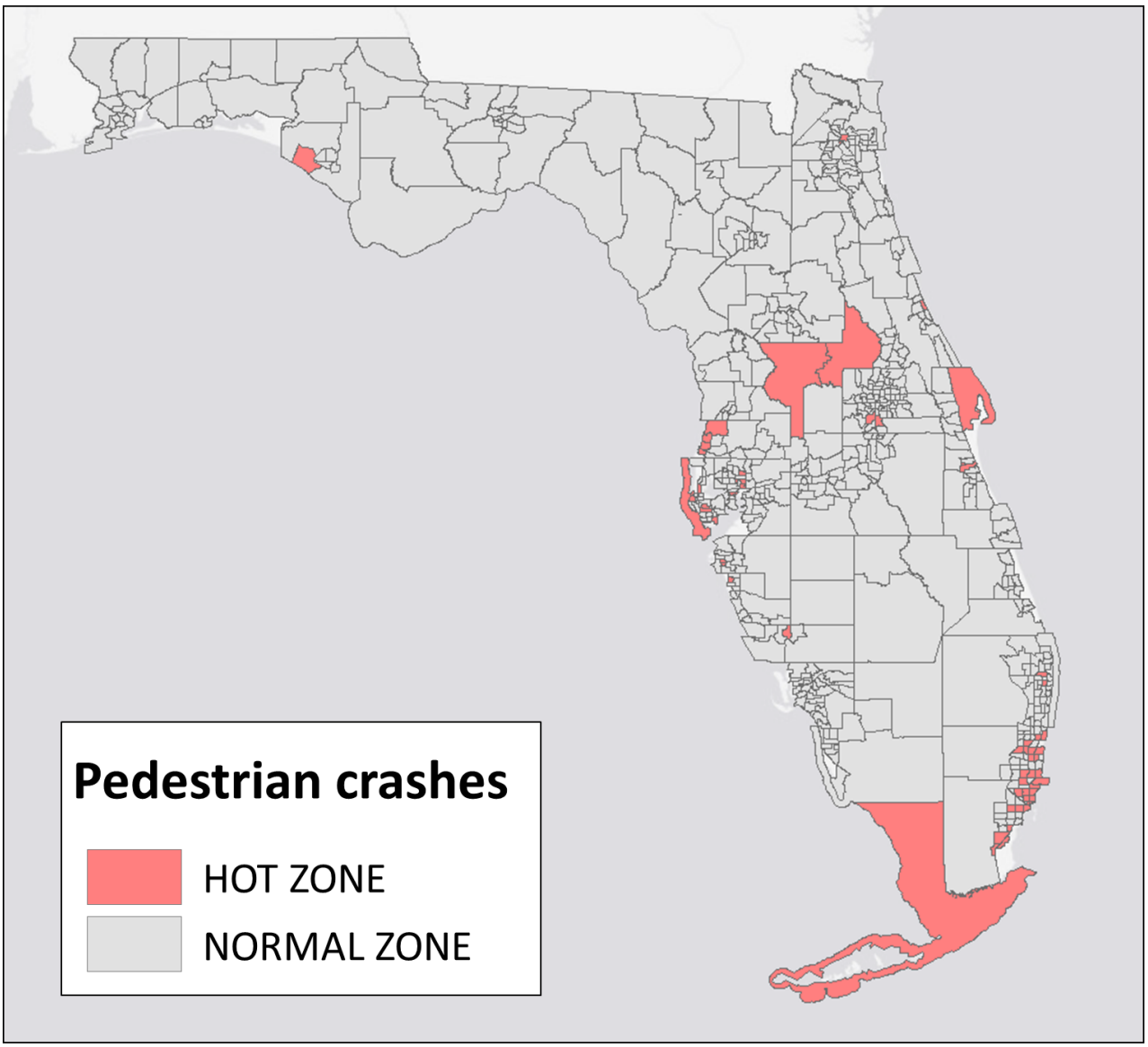

Figure 19 Hot zones for pedestrian crashes

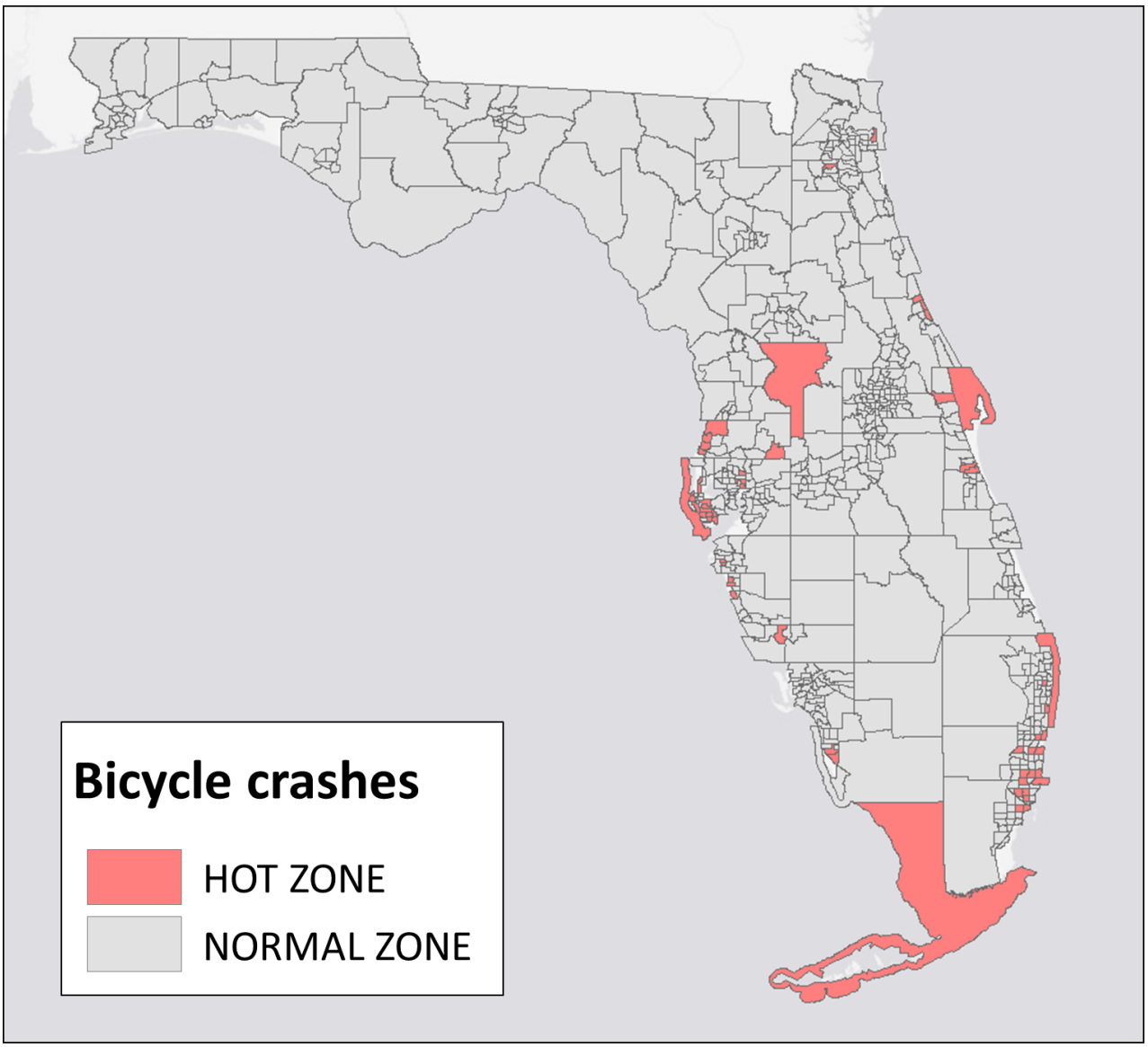

Figure 20 Hot zones for bicycle crashes

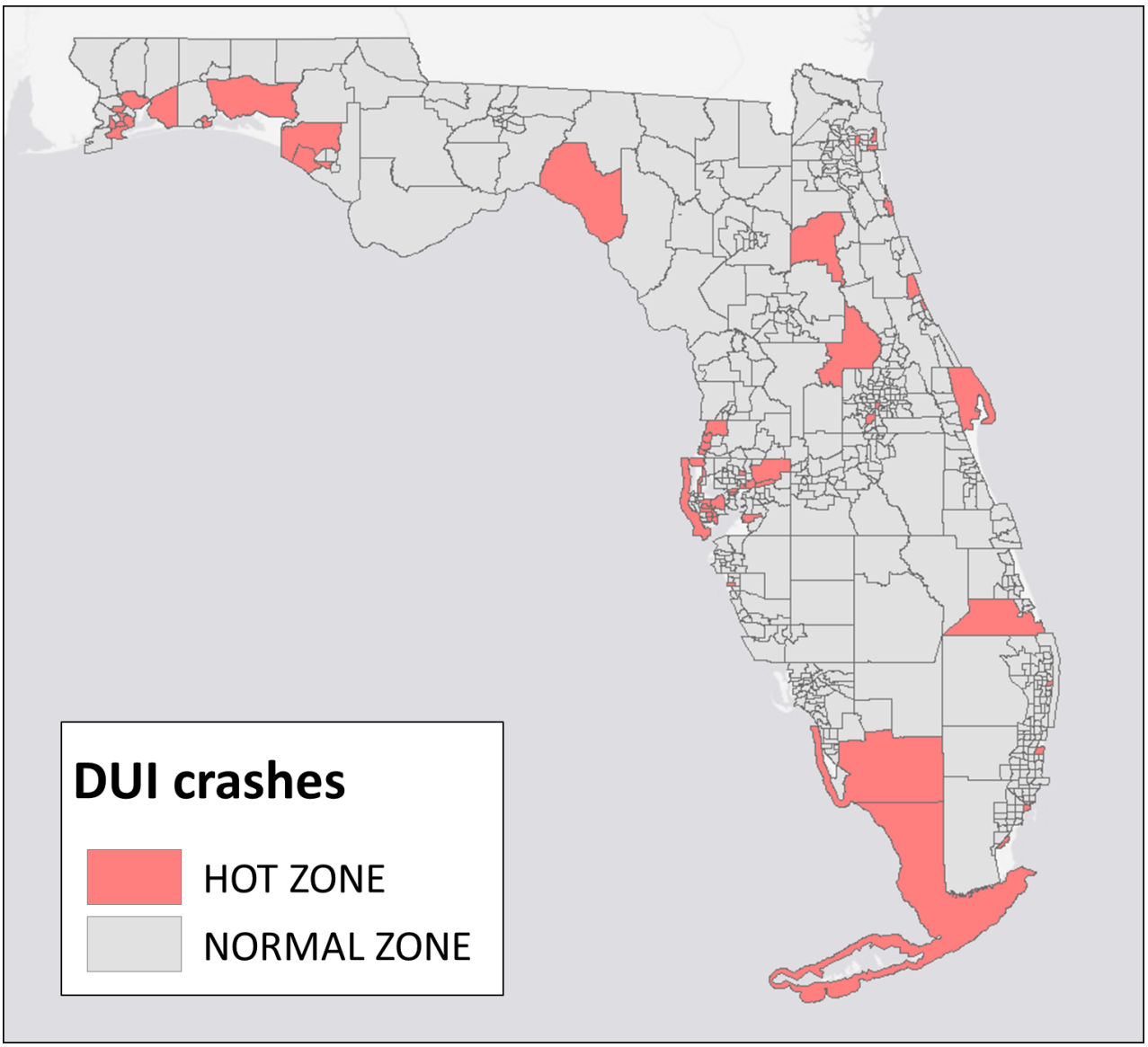

Figure 21 Hot zones for DUI crashes

1.9 Summary and Conclusion

This chapter aims to produce new ideas for collecting and utilizing Big Data to explore traffic crashes at the macroscopic level. In order to achieve the objective, the research team tried to acquire all available various data as follows:

- Layer 1: boundary maps (TAZs and TADs)
- Layer 2: socio-economic data
- Layer 3: roadway and traffic data
- Layer 4: crash data

The collected data were processed for developing safety performance functions based on TAZs and TADs. Several contributing factor were found to be statistically significant for traffic crashes. It was found that significant variable sets were quite different by crash types. Subsequently, a series of screening analyses were conducted for the five crash types based on TADs. The PSIs (Potential for Safety Improvements) were calculated by subtracting the predicted crash counts from the expected crash counts. Thus, the PSI stands for the number of crashes can be effectively reduced. A hot zone in this study is defined as a zone with top 10% highest PSI. This chapter suggests what types of data can be collected and how we can utilize these "Big Data" for macroscopic traffic safety analysis. It is expected that more various and larger Big Data will contribute more reliable and meaningful results.

2 MICROSCOPIC ANALYSIS

Compared with the "Big Data" used for macroscopic analysis, "Big Data" for microscopic analysis is "Big" because of the continuous detection manner in which the data are collected and the level of details that the collected data contains. The continuously collected data enlarge the data sample size quickly. Meanwhile, different data sources drastically increase the data dimension. The datasets are so large and complex that traditional data processing tools cannot handle them efficiently. Data mining is an analytic process which is applied to explore big data, with the target to find the relationship between variables, predicting and classifying events, etc. Thus, several types of data mining methods were used in the microscopic analysis.

2.1 Data Collection

In total, four types of data were collected, namely traffic data, Dynamic Message Signs (DMS) data, roadway geometric characteristics data and crash data. The detailed information of these four datasets is as follows,

1) Traffic data

Traffic data are provided by Central Florida Expressway Authority (CFX), formerly known as Orlando-Orange County Expressway Authority (OOCEA). On their system, five expressways are under its operation and maintenance. Currently two traffic detection systems are deployed on the five expressways. The Automatic Vehicle Identification (AVI) system is installed for both Electronic Toll Collection (ETC) and travel time estimation. In 2013, the expressway network is covered by a newly introduced Microwave Vehicle Detection System (MVDS). Both of these two detection systems archive the traffic information continuously at one-minute interval basis.

Thus they serve as the major source of "Big Data" in the microscopic analysis. The data are collected and processed each month.

2) DMS data

There are 37 DMS installed on CFX's expressway network. These DMS can display real-time traveling information to motorists. They have significant potential for congestion mitigation and safety warning. The DMS data contain information about DMS identification, messages displayed and the timestamps for the messages. The DMS data are provided by CFX as well. Currently, one-year data from September, 2012 to September, 2013 has been collected.

3) Roadway geometric characteristics data

Roadway geometric characteristics data are collected from FDOT Roadway Characteristics Inventory (RCI) database. The RCI data contain comprehensive geometric information. As the research object for microscopic safety analysis is the CFX expressways, pertinent geometric information is collected and processed for the five expressways.

4) Crash data

Some issues related with the long form and short form crashes in the macroscopic analysis does not exist in microscopic analysis. The traffic data have been collected after July, 2012 from which time the S4A archives the complete long and short form crashes. Consequently, the crash data used for microscopic analysis are collected from S4A database. Nevertheless, in microscopic analysis the crash information is required to be much more precise compared to the macroscopic safety analysis. The S4A crash data are only coded with longitude and latitude.

Great effort has been made to locate these crashes on the expressways. First of all, the crashes whose locations contain key words of expressways are selected. Then they are assigned with roadway ID and milepost using a GIS network specifically developed by the research team to select the crashes on the expressways. The results of this selection classify the crashes to the mainline of expressways, ramps and toll plazas. Further analysis on each type of these lanes then is possible. Figure 22 illustrates the final result of the crash data selection. The dots in the Figure 22 mean crashes. Crash data are also updated each month to keep up with the traffic data collection.

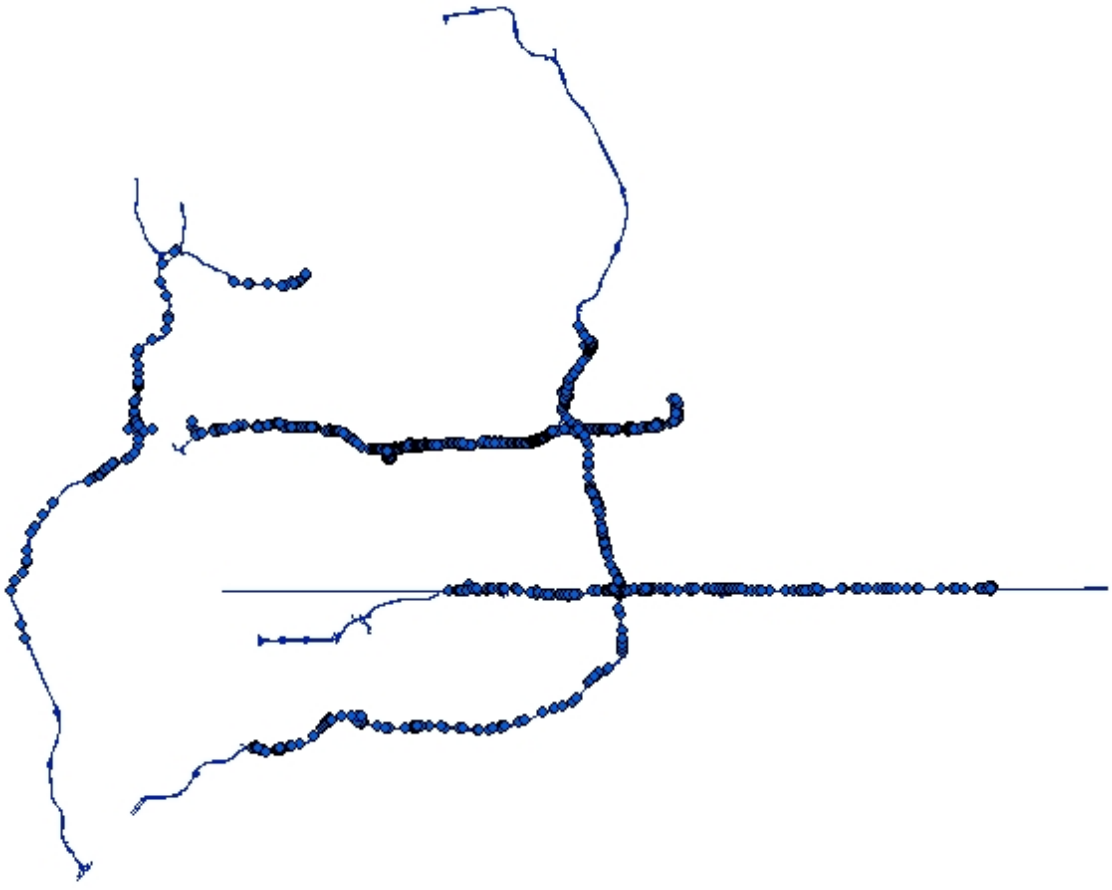

Figure 22 Crash locations on expressways in Central Florida

2.2 Data Mining Techniques

Data mining is one of the most widely used tools to explore data which are large and complex. It involves artificial intelligence, machine learning, statistics, etc. The overall goal of the data mining process in this study is the identification of traffic patterns which lead to high crash risk from a large amount of data. Four data mining methods are used: Support Vector Machine (SVM), Artificial Neural Network (ANN), Classification and Regression Tree (CART) and Logistic Regression. A brief discussion of the data mining methods is shown below.

SVM is used for classification analysis by constructing a hyperplane in a high dimensional space (Suykens and Vandewalle 1999). The hyperplane is chosen when it has the largest distance to the nearest training-data point, which means it represents the largest separation between two types of event. There are two types of SVM: linear and nonlinear. The choosing of SVM is based on the data type, e.g., linear SVM is better if data are linearly separated. The nonlinear SVM is achieved by applying kernel. By introducing kernel, SVM is flexible in the choice of the separation form and can handle the nonlinear issue (Deng et al. 2012).

ANN belongs to non-parametric models that can handle nonlinear relationships between predictors and target variables. A network may contain several units, and the units are grouped into layers: multiple input layers, a hidden layer, and multiple output layers (Stapelberg 2009). ANN may conduct nonlinear or linear transformations for input units in the hidden layer. The hidden units combine the input values, then the values calculated by hidden units are combined at the output units. In the output units, predicted values are computed and compared with the target value to obtain the error function, which the ANN intends to minimize. In addition to the

ability of detecting complex nonlinear relationships between target and explanatory variables, the ANN can also detect all possible interactions between explanatory variables (Tu 1996).

CART uses a tree-like graph or model of decisions and their possible consequences. It targets to classify objects by constructing a set of if-then rules (Markey et al. 2003). CART models with categorical target variable are called classification trees, and those with continuous target variable are called regression trees. Comparing to other data mining methods, CART can clearly perform variable screening or feature selection which makes the analytics easy to be interpreted. Meanwhile, it is flexible in handling nonlinear and noise relationship between parameters (Friedl and Brodley 1997).

Logistic Regression has been widely used in analysis of data whose target variable is categorical (Washington et al. 2010). It measures the relationship between target variables and explanatory variables by estimating probabilities based on a logistic function. Logistic Regression is easy for interpretation since the model result provides the coefficient value for each significant variable. The impact of each variable on odds ratio can be measured.

2.3 Real-time Safety Analysis

Accessibility of real-time traffic data from Intelligent Transportation Systems has triggered substantial efforts to explore their applications for better transportation system. Safety and mobility are often regarded as two important indicators of the system performance. Improvement of both safety and mobility depends on clear interpretation of their interrelationship. To achieve this objective, real-time traffic data show great potentials since they could reflect traffic states at microscopic level. The research team conducted a comprehensive investigation to evaluate the

relationship between mobility and safety using real-time ITS traffic data on urban expressways in central Florida. The expressways are heavily instrumented with traffic detection facilities which provide continuous monitoring of traffic mobility and precise traffic information near crash locations before their occurrence.

In this chapter, Congestion Index was introduced to represent the traffic mobility on the expressways. Congestion Index is defined as the reduction in speed caused by traffic congestion. The higher the value, the more mobility is reduced. Congestion Index is calculated as:

$$CI = \begin{cases} \frac{Free\ flow\ speed - actual\ speed}{Free\ flow\ speed} & if\ CI > 0 \\ 0 & Otherwise \end{cases} \quad (2)$$

Due to the SVM and ANN models lack of the capability of selecting significant variables, a binary logistic regression was applied to select variables. Based on the result, it was found that eight explanatory variables had significant impact on crash occurrence. The significant variables are peak hour, traffic volume by lane and speed variation at the nearest upstream station, Congestion Index and truck percentage at downstream station, lanes, and median and shoulder width. The description of these variables is shown in Table 20.

Table 20 Variable description

Parameter	Description
Peak	Peak hour indicator: Peak = 1: weekday 7:00 - 9:00 & 17:00 – 19:00 Peak = 0: otherwise
U1_lanevol	Average traffic volume by lane at U1 station
U1_spddiff	Speed difference between the inner and outer lanes at U1 station
D1_trkpct	Truck percentage at D1 station
D1_ci	Congestion Index at D1 station
Lane45	Number of lanes on cross section per direction: Lane45 = 1: 4 or 5 lanes at detection location Lane45 = 0: otherwise Base level: 2 lanes
Median	Median width (ft)
Shoulder	Shoulder width: Shoulder = 1: shoulder width ≥10ft Shoulder = 0: shoulder with < 10ft

Though machine learning and artificial intelligence models may outperform statistical models, SVM, ANN and CART have a limitation on providing the effects of explanatory variables in the target variable. However, with the sensitivity analysis, the relationship between crash risk and the chosen eight explanatory variables could be analyzed. Since the main objective of the study is to reduce the crash risk, the research team focused on the sensitivity analysis of crash events. For continuous variables, sensitivity analysis was conducted by changing each explanatory variable by a user-defined value while the other variables maintain their original value; for categorical variables, it was conducted by setting all their values to 0 and then 1. Then the trained data mining models were used to estimate the new dataset and to provide the mean predicted crash occurrence probability. This analysis could be used to detect the positive or

negative relationship between explanatory variables and crash risk for crash events. The sensitivity analysis of these 8 variables in four data mining models is shown in Table 21.

Table 21 Sensitivity analysis of the explanatory variables

Variable	Changing unit	Difference in Mean Probability			
		SVM	ANN	CART	Logistic Regression
Peak	0 to 1	0.025	0.000	0.000	0.020
U1_lanevol	Increase by 10	0.003	0.047	0.008	0.014
U1_spddiff	Increase by 5	-0.006	0.000	0.001	0.012
D1_trkpct	Increase by 0.1	-0.005	0.000	0.000	0.014
D1_ci	Increase by 0.1	0.007	0.001	0.040	0.056
Lane45	0 to 1	0.015	0.003	0.000	0.040
Median	Increase by 10	-0.007	-0.052	0.000	-0.003
Shoulder	Increase by 2	-0.032	0.000	0.000	-0.029

Table 21 indicates that the impact of explanatory variables on target variable varies in different models. SVM and Logistic Regression model show that all these eight variables have impacts on crash risk, among which six variables' positive or negative relationships with the crash risk are identical. However, two variables' effects on crash risk are not consistent between SVM and Logistic Regression model. The increasing of speed variation at nearest upstream station and truck percentage at the nearest downstream station result in lower crash risk in SVM, but lead to higher crash probability in Logistic Regression. In ANN model, four variables were found to have no impact on crash. They are peak hour, speed variation at the nearest upstream station, truck percentage at the nearest downstream station and shoulder width. The other four variables' influences on the crash are consistent with SVM and Logistic Regression. It is noteworthy that the two inconsistent variables are not significant in ANN model. The CART model does not

include truck percentage at the nearest downstream station in the model, and shows the negligible impact of speed variation at the nearest upstream station. The different results may be due to different model structure and mechanism.

Model performances were compared by using ROC (Receiver Operating Characteristics) Curve. It is one of the most useful indexes for evaluating and comparing the performance of models where the response variable is binary. The higher ROC value, the better fitting the model generates. The ROC values of the four models are shown in Table 22.

Table 22 ROC of models

Index	SVM	ANN	CART	Logistic Regression
ROC	0.736	0.734	0.657	0.717

Table 22 shows that the model performances of SVM and ANN are almost the same and are the best. In contrast, the CART model has the lowest ROC. The results are expected, as CART is only based on a series of simple rules while SVM and ANN allow interaction between variables and also can account for the non-linear relationship between target variable and explanatory variables.

2.4 Summary and Conclusion

This chapter focuses on explore the impact of traffic and geometric parameters on traffic safety from the microscopic aspect. Four datasets were collected. They are traffic data, DMS data, roadway geometric characteristics data and crash data.

Several contributing factors were found to be significant for crash occurrence. In order to deeply understand their impacts on crash risk, this study applied four data mining methods, i.e., SVM, ANN, CART, and Logistic Regression. Since these four models' structure and mechanism differ, the sensitivity analysis showed that the impact of some explanatory variables on target variable varies. However, the majority of variables' positive or negative relationships with the crash risk are consistent in different models. Model performances were measured by ROC. The higher ROC indicates better model performance. The results showed that ANN and SVM provided the highest ROC value, and CART was the worst model. ANN and SVM are much more complicated than CART, but CART is easier to be understood and calculated. When choosing the data mining method, both model complexity and performance should be taken into consideration. In future, it is expected that more microscopic data can be collected and implemented in microscopic safety analysis.

.

Reference

AASHTO, 2010, Highway Safety Manual, American Association of State Highway and Transportation Officials.

Deng, N., Tian, Y., and Zhang, C., 2012. Support vector machines: Optimization based theory, algorithms, and extensions. CRC Press.

Friedl, M.A., and Brodley, C.E., 1997. Decision tree classification of land cover from remotely sensed data. Remote sensing of environment 61 (3), 399-409.

Markey, M.K., Tourassi, G.D., and Floyd, C.E., 2003. Decision tree classification of proteins identified by mass spectrometry of blood serum samples from people with and without lung cancer. Proteomics 3 (9), 1678-1679.

Stapelberg, R.F., 2009. Handbook of reliability, availability, maintainability and safety in engineering design Springer Science & Business Media.

Suykens, J.A., and Vandewalle, J., 1999. Least squares support vector machine classifiers. Neural processing letters 9 (3), 293-300.

Tu, J.V., 1996. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. Journal of clinical epidemiology 49 (11), 1225-1231.

Washington, S.P., Karlaftis, M.G., and Mannering, F.L., 2010. Statistical and econometric methods for transportation data analysis CRC press.