Integrating Macro and Micro Level Crash Frequency Models Considering Spatial Heterogeneity and Random Effects

Shahrior Pervaz
Graduate Research Assistant
Department of Civil, Environmental & Construction Engineering
University of Central Florida
Tel: 1-407-561-0298; Fax: 1-407-823-3315
Email: shahrior.pervaz@knights.ucf.edu
ORCiD number: 0000-0001-7966-7083

Tanmoy Bhowmik
Postdoctoral Scholar
Department of Civil, Environmental & Construction Engineering
University of Central Florida
Tel: 1-407-927-6574; Fax: 1-407-823-3315
Email: tanmoy78@knights.ucf.edu
ORCiD number: 0000-0002-0258-1692

Naveen Eluru
Professor
Department of Civil, Environmental & Construction Engineering
University of Central Florida
Tel: 407-823-4815, Fax: 407-823-3315
Email: naveen.eluru@ucf.edu
ORCiD number: 0000-0003-1221-4113

*Corresponding author

ABSTRACT

Safety literature has traditionally developed independent model systems for macroscopic and microscopic level analysis. The current research effort contributes to the literature on crash frequency by building a bridge between these two divergent streams of crash frequency research. The study proposes an integrated micro-macro level model for crash frequency estimation. Specifically, the study develops an integrated model system that allows for the influence of independent variables at the microscopic level to be incorporated within the macroscopic propensity estimation. The empirical analysis is based on the data drawn from 300 traffic analysis zones, 1,818 roadway segments, and 4,184 intersections from the City of Orlando, Florida for the years 2018 and 2019. The study considers a host of exogenous variables including roadway and traffic factors, land-use, built environment, and sociodemographic characteristics for the model estimation. The proposed model system can also accommodate for hierarchical correlations such as correlation between all segments or intersections in a zone. The study findings highlight the presence of common spatial unobserved factors influencing crash frequency across segment level and intersection level as well as presence of significant parameter variability across both micro and macro level in the crash frequency. The empirical analysis is further augmented by employing several goodness of fit and predictive measures. The results clearly demonstrate the improved performance offered by the proposed integrated micro-macro model relative to the non-integrated macro model. The overall model fit measures and interpretations encourage the application of the proposed model for crash frequency analysis.

Keywords: *Crash frequency; Integrated micro-macro model; Comparison exercise; Unobserved effects.*

1 BACKGROUND

Statistical and econometric models are important tools for understanding the factors affecting crash occurrence and their consequences. These models allow policy makers and transportation professionals to recommend strategies that reduce crash occurrence and/or alleviate the consequences of crashes. Crash occurrence is traditionally examined as crash frequency while crash consequences are analyzed as crash severity. Crash frequency models are developed for a spatial unit (such as traffic analysis zone (TAZ), county, and city) (Abdel-Aty et al., 2013; Bhowmik et al., 2021a, 2021b, 2019; Yasmin and Eluru, 2018, 2016; Zeng et al., 2019) or a facility (such as a segment or intersection) (Abdulhafedh, 2016; Dong et al., 2014; Mohammadnazar et al., 2021; Wali et al., 2018; Wang et al., 2020; Zeng et al., 2020). The aggregate spatial unit models are referred to as macroscopic models while the facility level models are referred to as microscopic models (Cai et al., 2019). Safety research has traditionally developed independent model systems for each of these levels. The current research contributes to growing literature on crash frequency models by building a bridge between the two traditionally divergent streams of crash frequency research.

The proposed research recognizes that crashes at a microscopic level (at segments or intersections) are aggregated for a spatial unit to generate crash frequency at the macroscopic level. When separate models are developed at the macroscopic and microscopic level, the embedded relationship within the data relating the microscopic level and macroscopic level crashes is neglected or lost. Earlier research efforts have attempted to address this issue by developing hierarchical model systems that recognize the multiple microscopic level dependent variables within a macroscopic unit (Alarifi et al., 2017, 2018; Huang and Abdel-Aty, 2010). However, these approaches accommodate for interactions across the different levels via common unobserved factors. While this approach is an enhancement over independent macro and micro level models, there are other drawbacks. In independent macroscopic models or hierarchical models, we can only incorporate the influence of observed macroscopic resolution variables on crash frequency. For example, segment level characteristics in a zone (such as speed limit or number of lanes) can only be considered at a zonal level. Given that crash frequency models often take the form of non-linear model systems, a simple aggregation of an independent variable (say average distance weighted speed limit in the zone) might not reflect the true impact of the variable at the macroscopic level. In fact, it would be appropriate to posit that macroscopic models that consider aggregate zonal variables are inadvertently introducing model parameter bias by not recognizing the inherently non-linear nature of variable impacts on crashes. A truly integrated approach allows us to test for the influence of factors affecting crashes at the microscopic level on the crash frequency at the macroscopic level.

In summary, the current research proposes an integrated model system that allows for the influence of independent variables at the microscopic level to be incorporated within the macroscopic propensity estimation. The approach would involve incorporating the sum of crash propensity by type of micro-facility within the macroscopic propensity computation. Within this proposed framework, two specific model structures are examined. In the first model structure, the propensity is computed as an expected value and an associated parameter for each facility type is estimated. In the second approach, the propensity of the microscopic models is incorporated while allowing for the microscopic model parameters to vary based on the macroscopic fit. The first approach fixes the microscopic models and accommodates for the cumulative propensity influence as any other independent variable. In the second approach, feedback between macroscopic and microscopic models is allowed (more details in the methodology section). In both approaches, the

model structure is quite flexible and truly builds on the individual microscopic and macroscopic components. The integrated model with simple constraints can be employed to estimate the traditional macroscopic and microscopic models.

The proposed model system is estimated using data drawn from the City of Orlando for the years 2018 and 2019. The study considers a total of 300 TAZs, 1,818 roadway segments, and 4,184 intersections for the analysis. An exhaustive set of independent variables including roadway and traffic factors, land-use attributes, built environment characteristics, and sociodemographic characteristics at both macroscopic and microscopic levels are considered. The model estimation procedure explicitly tested for potential temporal parameter instability by estimating separate models for 2018 and 2019 (as highlighted in multiple articles on temporal parameter stability, see Mannering, 2018; Marcoux et al., 2018; Behnood and Mannering, 2019; Islam and Mannering, 2020; Kabli et al., 2020; Tirtha et al., 2020). The model estimation results are further augmented by the evaluation of the predictive performance of the proposed framework.

2 EARLIER RESEARCH AND CURRENT STUDY IN CONTEXT

Crash frequency models at the macroscopic and microscopic levels have been explored extensively in safety literature. Macroscopic crash frequency models, with a focus on long-term safety planning, analyze crash occurrence for aggregated geographical regions such as state, county, city, traffic analysis districts (TADs), and traffic analysis zones (TAZs) (Abdel-Aty et al., 2013, 2011; Aguero-Valverde and Jovanis, 2006; Azimian et al., 2021; Bhowmik et al., 2022, 2021a, 2021b, 2019, 2018; Cheng et al., 2020; Cui and Xie, 2021; Huang et al., 2019, 2010; Li et al., 2019; Liu and Sharma, 2018; Noland and Oh, 2004; Soroori et al., 2019; Wang et al., 2020; Xie et al., 2019; Yasmin and Eluru, 2018, 2016; Zeng et al., 2019). On the other hand, microscopic crash frequency models examine factors affecting crash occurrence at a facility resolution such as a segment or an intersection to suggest facility specific treatments (Abdulhafedh, 2016; Aguero-Valverde and Jovanis, 2008; Alarifi et al., 2017, 2018; Alhomaidat et al., 2020; Dong et al., 2014; Gong et al., 2020; Kaaf and Abdel-Aty, 2015; Kim et al., 2007; Kim and Washington, 2006; Mohammadnazar et al., 2021; Mousavi et al., 2021; Oh et al., 2004; Saha et al., 2020; Satria et al., 2020; Shirani-bidabadi et al., 2020; Veeramisti et al., 2021; Wali et al., 2018; Wang et al., 2020; Wen et al., 2019; Ye et al., 2009; Yu et al., 2019; Zeng et al., 2020).

s

2.1 Independent Microscopic and Macroscopic Models

An exhaustive review of studies from the two levels is beyond the scope of the paper (for recent reviews, see Bhowmik et al., 2019; Lord and Mannering, 2010; Mannering and Bhat, 2014; Wang et al., 2021). We provide a summary of earlier work on independent macro and micro level models along two dimensions: (a) methodologies employed, and (b) important factors that affect crash frequency. On the methodological front, as the dependent variable is crash frequency in the two systems, there is a lot of commonalities. Most commonly used methods include Poisson regression model (Abdulhafedh, 2016; Oh et al., 2004), Poisson lognormal model (Oh et al., 2004), multi-level Poisson lognormal model (Alarifi et al., 2018, 2017), negative binomial (NB)/Poisson-Gamma model (Abdel-Aty et al., 2011; Aguero-Valverde and Jovanis, 2006; Alhomaidat et al., 2020; Gong et al., 2020; Noland and Oh, 2004; Wali et al., 2018; Wang et al., 2020), NB-ordered logit fractional split model (Yasmin and Eluru, 2018), latent segmentation Poisson/negative binomial model (Yasmin and Eluru, 2016), panel mixed NB model (Bhowmik et al., 2019), spatial Durbin model (Wang et al., 2019), multivariate Tobit model (Zeng et al., 2019), copula-based crash frequency model (Bhowmik et al., 2021a; Yasmin et al., 2018), Bayesian Poisson-

lognormal/hierarchical models (Abdel-Aty et al., 2013; Azimian et al., 2021; Cui and Xie, 2021; Huang et al., 2019, 2010; Li et al., 2019; Satria et al., 2020; Wang and Huang, 2016; Zeng et al., 2020), geographically weighted regression model (Li et al., 2013; Liu et al., 2017; Mohammadnazar et al., 2021), Poisson-Tweedie model (Saha et al., 2020), and Conway-Maxwell Poisson model (Shirani-bidabadi et al., 2020). The reader would recognize that microscopic crash frequency models are likely to be affected in most cases with excessive zeros. Thus, microscopic models might see increased application of zero-inflated models and/or hurdle models (Dong et al., 2014; Yu et al., 2019).

At the macroscopic level, the factors identified to be of significance include roadway and traffic factors, land-use, built environment, and sociodemographic factors. For instance, county level models identified factors include roadway density, intersection density, urban land-use, school density, population, proportion of males, population aged above 65, and median household income (Azimian et al., 2021; Cheng et al., 2020; Li et al., 2019). In addition to these factors, TAZ, TAD and census tract level models identified the effect of number of lanes, divided road length, sidewalk and shoulder width, speed limits, bikeway length, intersection types, signal intensity, Annual Average Daily Traffic (AADT), percentage of truck traffic, vehicle miles travelled, commercial, residential, office area, number of residential units, shopping centers, number of household with no vehicles, unemployment rate, commute mode share for drive alone, public transit, and non-motorized means of transport on crash counts (Bhowmik et al., 2019; Cai et al., 2019; Cui and Xie, 2021; Huang et al., 2019, 2016; Park et al., 2020; Pljakić et al., 2019; Soroori et al., 2019; Wang et al., 2019; Xie et al., 2019; Zeng et al., 2019; Zhai et al., 2019). At the microscopic level, for intersections, the factors identified include number of legs, number of exclusive left-turn and right-turn lanes, angle of intersection, presence of signs, bus stops, curvature and medians on approaches, major and minor road AADT, and percentage of truck traffic (Alarifi et al., 2018, 2017; Cai et al., 2019; Dong et al., 2014; Gong et al., 2020; Huang et al., 2016; Park et al., 2020; Saha et al., 2020; Veeramisti et al., 2021). For segment crash frequency models, factors identified include segment length, posted speed limit, median width, access control, number of lanes, curvature, gradient, AADT, and percentage of truck traffic (Alarifi et al., 2018, 2017; Alhomaidat et al., 2020; Huang et al., 2016; Mohammadnazar et al., 2021; Satria et al., 2020; Veeramisti et al., 2021; Wang et al., 2020; Wen et al., 2019; Yu et al., 2019; Zeng et al., 2020).

2.2 Integrated Microscopic and Macroscopic Models

The primary focus of our review is on studies that recognize the interconnectedness of macroscopic and microscopic crash models. Huang and Abdel-Aty (2010) presented the potential hierarchical relationships involved in safety literature such as crash frequency for county – corridor – intersection combination. They suggest the consideration of county and corridor level observed and unobserved variables in modeling intersection crash frequency. Huang et al. (2016) estimated separate models for microscopic and macroscopic levels. The microscopic model predictions were aggregated at the zonal level and were compared to the prediction from macroscopic model and observed counts. As expected, the microscopic model performed better given the larger amount of data that is incorporated in the micro model. Alarifi et al. (2017) proposed a multilevel joint model that examines segment and intersection level crash frequency simultaneously by relating them based on a corridor level identifier. The authors employed corridor level observed and unobserved variables to accommodate for the potential correlation across the two facility types (see Wang and Huang, 2016 for similar analysis at the zonal level). The research was further extended in Alarifi et al. (2018) by considering crash frequency by crash type. Wang et al. (2017) developed crash

frequency models by transportation mode at the microscopic level by considering macroscopic variables and concluded their consideration improved model fit. Park et al. (2020) built on earlier multilevel modeling research by considering segment/intersection membership as a weighted function to predict crashes. Overall, the summary of these research efforts clearly highlights how microscopic models were enhanced by considering macroscopic variables (observed and unobserved) in the analysis. It is important to note that none of these studies enhance the macroscopic model development process by embedding microscopic facility level attributes.

An exception to this is the Cai et al. (2019) that proposed a joint modeling approach. In this study, the authors recognize that sum of observed crashes at the microscopic level (segments and intersections) should add up to the macroscopic crashes in the zone. At the same time, to accommodate for potential error in microscopic model prediction, they employ an adjustment factor as a function of zonal variables. The adjustment factor acts as a calibration parameter for the macro level crashes. While this study is a substantial improvement on prior research, the authors restrict the macroscopic model estimation to macro level socioeconomic variables. Further, the model structure is such that the framework cannot be used to estimate traditional microscopic and macroscopic models within the same system.

2.3 Context of the Current Study

In our current study, we develop a representatively integrated model from the first principles. The model system proposed recognizes that crashes at the micro level facilities contribute to total macroscopic level crash counts. To allow for this influence, we add one component per micro level facility type (such as intersection or segment) in the form of an additional variable in the propensity of the macroscopic model system. The component for a facility type is evaluated as the sum of crash propensity for all facilities of that type in the spatial unit. A scalar parameter for each facility type component can be estimated in the model system. The introduction of component by micro level facility type provides two possible model frameworks. In the first framework, we focus on the optimization of the macroscopic model data fit by only estimating the scalar parameter per component. The parameters embedded with the micro level models are assumed fixed for this purpose. In the second approach, a joint log-likelihood function of macroscopic and microscopic models is considered, and the microscopic parameters are estimated based on their contribution to microscopic model directly and the macroscopic model via the microscopic propensity component in the macroscopic model. The proposed model system can also accommodate for hierarchical correlations such as correlation between all segments or intersections in a zone. These correlations also implicitly account for spatial variations for facilities in the zone.

The overall model development process is achieved using a negative binomial regression framework at the micro and macro level. The approach can be extended readily to any possible mathematical model (such as Poisson lognormal). In terms of empirical analysis, this study incorporates both micro and macro level factors for crash frequency analysis of both zonal level, and segment and intersection level. The empirical analysis is based on the traffic analysis zone (TAZ) level crash count data, and segments and intersections level crash count data from Orlando city of Florida. The model estimation results are further augmented by the evaluation of the predictive performance of the proposed model. The proposed approach will offer significant advantages methodologically and empirically. Methodologically, a single integrated model will allow us to employ a single model code for estimation of macroscopic and microscopic models. The integrated model developed can be employed to estimate the traditional macroscopic and microscopic models by setting the appropriate parameters to zero. In addition, incorporating micro

level variables into the macro level propensity offers improved model performance. Thus, a better understanding of the crash occurrence can be achieved with this framework. On the empirical side, a better understanding of crash contributing factors will allow us to provide pragmatic and efficient crash countermeasures. Since the proposed model considers variables from both macro and micro levels, it is possible to select countermeasures for both levels from this framework. The proposed model can also be applied to identify crash hotspots, which is a top priority for safety treatment.

The rest of the paper is organized as follows: The methodology section will provide mathematical details of the proposed model system. The data section will describe data compilation procedures. The results section will present the findings of the model estimation and discuss the predictive performance of the model. The final section will conclude the paper and offer thoughts for future work.

3 METHODOLOGY

The econometric formulation of the proposed integrated micro-macro model is presented here (please see Bhowmik et al., 2021a; Wang et al., 2020 for details methodology of simple macro and micro model).

3.1 Integrated Micro-Macro Approach

For any spatial unit, the general form of the probability equation of the NB formulation can be written as follows:

$$P(y_t \,|v_t, \lambda_t') = \frac{\Gamma\left(y_t + \frac{1}{\lambda_t'}\right)}{\Gamma(y_t + 1)\Gamma\left(\frac{1}{\lambda_t'}\right)}\left(\frac{1}{1+\lambda_t' v_t}\right)^{\frac{1}{\lambda_t'}}\left(1-\frac{1}{1+\lambda_t' v_t}\right)^{y_t} \quad (1)$$

where, t represents the different spatial units including segments (s: 1,2 …. $S = 1{,}818$), intersections (i: 1,2 … $I = 4{,}184$) and TAZs (z: 1,2 …. $Z = 300$). y_t be the index for crash counts occurring over a period of time in the corresponding spatial unit t . $P(y_t)$ is the probability that unit t has y_t number of total crashes. λ_t' is NB over-dispersion parameter specific to the spatial unit t and v_t is the expected number of crashes occurring in t over a given time period. This, v_t can be expressed as a function of explanatory variables using a log-link function. In our analysis, v_t at the zonal level is labelled as v_z; v_t at the segment level is labelled as v_s and v_t at the intersection level is labelled as v_i. For the micro level specifications (segments and intersections), the formulation of v_s and v_i is collectively defined as follows:

$$v_{(s,i)} = E\left(y_{(s,i)}|\boldsymbol{x}_{(s,i)}\right) = exp\left((\boldsymbol{\beta}_{s,i} + \boldsymbol{\varrho}_{(s,i)})\boldsymbol{x}_{(s,i)} + \boldsymbol{\theta}_{z\,(s,i)} + \varepsilon_{(s,i)}\right) \quad (2)$$

where, $v_{(s,i)}$ is the expected number of crashes that correspond to each micro level spatial unit (segments (s) and intersections (i)). $\boldsymbol{x}_{(s,i)}$ is a vector of explanatory variables and $\boldsymbol{\beta}_{s,i}$ is a vector of mean coefficients to be estimated corresponds to the spatial unit(s, i). $\boldsymbol{\varrho}_{(s,i)}$ is a vector of unobserved factors moderating the influence of attributes in $\boldsymbol{x}_{(s,i)}$ on the total crash count propensity for analysis unit (s, i). $\boldsymbol{\theta}_{z(s,i)}$ is a vector of unobserved effects specific to the zone for either segments or intersections highlighting the spatial arrangement of the segments and intersections within the same zone. This $\boldsymbol{\theta}_{z(s,i)}$ will be same across the spatial unit (s, i) if they

correspond to same zone (TAZ) and thus the adjacency heterogeneity (dependency) will be captured through the proposed system. Finally, $\varepsilon_{(s,i)}$ term represents a gamma distributed error term with mean 1 and variance $\lambda_{(s,i)}{}'$.

The reader would note that, the spatial unobserved heterogeneity can vary across the spatial unit (s,i). Therefore, in the current study, we parameterized the correlation parameter $\boldsymbol{\theta}_z$ as a function of observed attributes as follows:

$$\boldsymbol{\theta}_{z(s,i)} = \boldsymbol{\gamma}_{z(s,i)}\boldsymbol{s}_{z(s,i)} \tag{3}$$

where, $\boldsymbol{s}_{z(s,i)}$ is a vector of exogenous variables at the zonal level z (including a constant) employed for segment *s* or intersection *i*, $\boldsymbol{\gamma}_{z(s,i)}$ is a vector of parameters to be estimated.

Once the micro level propensities are estimated, we adopt two alternative approaches to estimate the zonal level expected number of crashes (v_z) as presented in equation 4 and 5 respectively.

$$v_z = E\left(y_z|c_z\right) = exp\left(\left(\propto +\boldsymbol{\phi}_z\right)\boldsymbol{c}_z + \rho_s * \ln\left(\sum_{p=1}^{Sz}(v_s)\right) + \rho_i * \ln\left(\sum_{p=1}^{Iz}(v_i)\right) + \varepsilon_z\right) \tag{4}$$

$$\begin{aligned} v_z = E\left(y_z|c_z\right) = exp\Bigg(&\left(\propto +\boldsymbol{\phi}_z\right)\boldsymbol{c}_z + \rho_s * \ln\left(\sum_{p=1}^{Sz}\left(exp\left((\boldsymbol{\beta}_s + \boldsymbol{\varrho}_s)\boldsymbol{x}_s + \boldsymbol{\theta}_{zs}\right)\right)\right) \\ &+ \rho_i * \ln\left(\sum_{p=1}^{Iz}\left(exp\left((\boldsymbol{\beta}_i + \boldsymbol{\varrho}_i)\boldsymbol{x}_i + \boldsymbol{\theta}_{zi}\right)\right)\right) + \varepsilon_z\Bigg) \end{aligned} \tag{5}$$

where, $\boldsymbol{c}_z$ is a vector of exogenous variables at zonal level z, $\propto$ is a vector of mean parameters to be estimated. $\boldsymbol{\phi}_z$ is a vector of unobserved factors moderating the influence of attributes in $\boldsymbol{c}_z$ on the total crash count propensity for zone z. ρ_s and ρ_i is a scalar associated with the corresponding micro level spatial unit (segments and intersections) highlighting the share of each micro level propensities to be linked with the macro level propensities. p is a counter here ranging from 1 to S_z (I_z) where S_z (I_z) represents the segments (intersections) in zone z. For example, if 5 segments are present in the zone z_1, then we will sum the propensity for these 5 segments to obtain a value for z_1 . ε_z is a gamma distributed error term with mean 1 and variance $\lambda_z{}'$. The main difference between the two approaches is that the micro level propensities will remain fixed and only the scalar parameters will be estimated for approach 1. In the second approach, we allow the micro level parameters to be jointly influenced by microscopic and macroscopic fit.

In estimating the model, it is necessary to specify the structure for the unobserved vectors $\boldsymbol{\varrho}$, $\boldsymbol{\theta}_{(s,i)}$, $\boldsymbol{\phi}$ represented by Ψ. In this paper, it is assumed that these elements are drawn from

independent normal distribution: $\Psi \sim N(0, (\boldsymbol{\pi^2}, \boldsymbol{\sigma^2_{(s,i)}}, \boldsymbol{\zeta^2}))$. Thus, conditional on Ψ, the likelihood function for approach 1 (equation 6) and 2 (equation 7) across TAZ can be expressed as follows:

$$L_z = \int_{\Psi} P(y_z) f(\Psi) d\Psi \tag{6}$$

$$L_z = \int_{\Psi} P(y_z) * \prod_{p=1}^{Sz} P(y_s)^{w_s} * \prod_{p=1}^{Iz} P(y_i)^{w_i} f(\Psi) d\Psi \tag{7}$$

where w_s (w_i) is a dummy variable taking a value of 1 if the corresponding zone z has segments (intersections) in it or 0 otherwise. Finally, the log-likelihood function is:

$$LL = \sum_{z} Ln(L_z) \tag{8}$$

All the parameters in the model are estimated by maximizing the joint logarithmic function LL presented in equation 8. We apply Quasi-Monte Carlo simulation techniques based on the scrambled Halton sequence to approximate this integral in the likelihood function (with 150 draws) and maximize the logarithm of the resulting simulated likelihood function (See Bhat, 2001; Yasmin and Eluru, 2013 for more details). We tested the model with higher number of scrambled draws (200 and 250) and found the model estimation was stable. We use the GAUSS Matrix Programming software for estimating the models (Aptech, 2015).

4 DATA PREPARATION

The current research considers Orlando city region of Florida as the area of analysis, which is composed of 300 TAZs, 1,818 segments, and 4,184 intersections (see Figure 1). The study focuses on total crashes for the years 2018 and 2019. The crash frequency variables for the TAZs, segments and intersections were computed as a sum of crashes during 2018 and 2019. The data were compiled from the Signal Four Analytics databases. All the crash records were aggregated at TAZ level using the Geographic Information System (GIS). The crashes nearest to a TAZ were counted as the crashes of that TAZ. A 250 feet buffer around each intersection was created, and the spatial join tool was used to assign crashes into the intersections by intersecting crash map and intersection buffer map. The remaining crashes were assigned to road segments by using the proximity tool in GIS. In this process, a total of 42,086 TAZ crashes were classified into 30,886 intersection crashes and 11,200 segment crashes.

4.1 Variables Considered

A comprehensive set of independent variables including roadway, traffic, land-use, built environment, and sociodemographic characteristics are considered in our study. Information about these variables were collected from different data sources including Transportation Statistics Division of Florida Department of Transportation (FDOT), US Census Bureau, American Community Survey, and Florida Geographic Data Library databases. These explanatory variables

were aggregated at the zonal level using the GIS for macro level dataset. We have considered different spatial aggregations based on the independent variable specific data source. For example, sociodemographic data were extracted from US Census Bureau (American Community Survey) database at census tract level. In cases where the buffer area overlaps multiple census tracts, the data was aggregated by calculating the proportion of census tract area within that buffer area. Macro level analyses use roadway and traffic factors (such as proportion of roads by functional class, number of lanes, average speed limit, average shoulder width, average sidewalk width and median width, intersection density, traffic signal per intersection, AADT, and truck AADT), land-use attributes (such as proportion of residential, commercial, institutional, industrial and recreational area), built environment characteristics (such as number of restaurants, business centers, commercial centers, educational centers, and shopping centers), and sociodemographic characteristics (such as population density, proportion of males and females, household density, median household income, proportion of car, drive alone, non-motorized means of transport, different population group by age level, household with vehicle availability, and population with different races). The micro level (segments and intersections) explanatory variables also include similar roadway and traffic factors, land-use, built environment, and sociodemographic variables. For segment level variables, roadway and traffic attributes were assigned to the segments by using the proximity tool in GIS. Aggregation of land-use, built environment, and sociodemographic variables for segments and intersections, and roadway and traffic variables for intersections were performed by intersecting maps of the attributes and 0.5-mile buffer map of respective facility types (segments and intersections). For modeling intersection data, it is customary to consider AADT data for major and minor roads. However, in our case, the AADT data was not available labelled as major and minor AADT. Hence, in our analysis, we considered 0.5-mile buffer area of an intersection as influence zone and generated the sum of AADT of all approaches to compute intersection AADT.

Table 1 lists the independent variables with the appropriate definition considered for final model estimation along with the minimum, maximum, mean, and standard deviation (SD) values at both micro and macro level. For model estimation purposes, several functional forms of the variables were also considered. The variables that offered the best model fit were retained. The final specification includes the statistically significant variables at 90% confidence level. The reader would also note that the model estimation process carefully examined for potential collinearity/correlations across independent variables. The variable selection was finalized after ensuring the covariance across the independent variables is within an acceptable range post-model estimation.

5 EMPIRICAL ANALYSIS

5.1 Model Specification and Overall Measure of Fit[1]

The empirical analysis involves a series of model estimations. First, we estimate the simple non-integrated NB models at both micro (segment and intersection) and macro levels. Second, we

[1] The reader would note that an exhaustive estimation process was implemented to test for parameter stability across 2018 and 2019. In this estimation process, information for traffic, land-use and sociodemographic factors were considered separately for each year. However, the information on the roadway and built environment factors were considered from 2019 data only as this information was not available for 2018. The estimation steps considered included: (1) estimating two separate model systems for 2018 and 2019 and comparing their performance relative to the performance of the pooled model system (combining 2018 and 2019), (2) considering for the influence of

estimate an existing integrated micro-macro model, as proposed by Cai et al. (2019). Third, we developed our proposed integrated system following two approaches: a) integrated model 1: focusing on optimizing the macroscopic model data fit by only estimating the scalar parameter while fixing the micro level parameters; and b) integrated model 2: the microscopic parameters are estimated based on their contribution to microscopic model directly and the macroscopic model via the microscopic propensity component in the macroscopic model. Fourth, we identify the best model by comparing model performance based on Bayesian Information Criterion (BIC). The BIC for a given empirical model is equal to:

$$BIC = -2LL + K \ln(Q) \quad (9)$$

where LL is the log-likelihood value at convergence, K is the number of parameters and Q is the number of observations. The model with the lower BIC is the preferred model.

Finally, we incorporate unobserved heterogeneity in terms of spatial variations and random effects in the model selected in the fourth step and compare its performance with the models without unobserved heterogeneity.

The corresponding BIC (LL) values are: (1) non-integrated models (with 29 parameters): 34,798.048 (-17,316.319), (2) existing integrated micro-macro model (with 20 parameters): 34,782.556 (-17,334.240), (3a) integrated model 1 (with 24 parameters): 34,765.831 (-17,314.470), (3b) integrated model 2 (with 23 parameters): 34,787.607 (-17,328.210), and (4) integrated model 1 with unobserved heterogeneity (with 28 parameters): 34,297.126 (-17,068.710). Based on these BIC values, three specific observations could be drawn. First, all the integrated systems (our proposed two approaches and the existing integrated approach) provide improved data fit as evidenced by the lower BIC values in comparison to the non-integrated model. Second, within the integrated systems, our proposed model 1 provides the lowest BIC indicating the best data fit in comparison to the other models. The performance of the other two integrated models (our proposed model 2 and existing integrated model) are quite close to each other as highlighted by the marginal differences in BIC across these models. Finally, we accommodate unobserved heterogeneity in our integrated model 1 (the best model in terms of data fit) and find that the model accommodating unobserved heterogeneity provides further improved BIC (lower) compared to its independent counterpart, thus reinforcing the importance of incorporating the influence of common unobserved factors in crash frequency analysis.

5.2 Model Estimation Results

In this section, we will discuss the factors affecting crash counts across macro and micro levels. For the sake of brevity, we will only discuss the results of the best fit model identified above. Table 2 presents the model estimation results for the proposed integrated micro-macro model 1 with

unobserved heterogeneity specific to the year indicator variable in the pooled dataset and (3) various systematic interactions of the independent variables with the year indicator variable. The results from these efforts indicated that temporal instability was not detected for our dataset. While this was a surprising finding, it is possible that the time frame of 2 years might is too small to affect parameter stability in our context. The documentation of the temporal parameter stability is summarized in the Supplemental Documentation (Table S1). It would be interesting to explore the temporal stability finding using data from additional years as a specific direction of future research. Given this finding, for ease of model estimation, the models were considered as a single observation for each spatial unit across the two years for our analysis (while considering a year offset variable of 2).

correlation and random effects (please refer to the supplementary material for results of the simple non-integrated models, Table S2). It is to be noted that a positive (negative) sign for a variable in the crash counts of Table 2 indicates that an increase in the variable is likely to result in more (less) crashes.

5.3 Crash Specific Constants

The model constants do not have any substantive interpretation.

5.4 Segment Level Attributes

The segment level model shows that the parameter associated with the segment length has a positive impact on crash frequency. The result is consistent with the expectation because the longer road segment is correlated with higher exposure between drivers, and other road users. The result is similar to some other studies (Alhomaidat et al., 2020; Mohammadnazar et al., 2021; Yu et al., 2019; Zeng et al., 2020). The parameters for number of lanes and average inside shoulder width also show positive association with crash frequency. This is intuitive as the roads with higher number of lanes usually have higher traffic volume, higher lane change rates and conflict risk resulting in higher number of crashes. Inside shoulder may provide shelter for emergency stop and pull over. However, stopped vehicles may be hazardous and contribute to certain type of crashes (for example, rear end crashes), especially on freeways. Moreover, wider shoulders may encourage higher operating speeds which can also increase crash risk (Stamatiadis et al., 2009). We also found that average inside shoulder width has significant variability as indicated by the standard deviation parameter in Table 2. This distributional parameter indicates that the overall impact of the variable on crash count at segment level is likely to be positive (95.35%). The parameter associated with average sidewalk width shows negative effect on segment level crash count. The presence of wider sidewalks provides additional safety for non-motorists from colliding with a motorized vehicle and thus contributes to a lower risk (Bhowmik et al., 2019). An increase in traffic signal per unit road length increases the likelihood of crashes. This is intuitive as a higher number of traffic signals may lead to an increase in certain types of crashes (such as rear-end crashes) in dilemma zones (Abdel-Aty and Wang, 2006; Lee et al., 2017; Park et al., 2020). In addition, AADT is found to be positively associated with crash frequency. The results indicate that the segments with higher AADT have higher likelihood of crashes. This result is consistent with previous studies (Alarifi et al., 2017; Alhomaidat et al., 2020; Cai et al., 2019; Huang et al., 2016; Mohammadnazar et al., 2021; Satria et al., 2020; Veeramisti et al., 2021; Wang et al., 2020). The land-use mix variable indicates crash count is negatively associated with land-use mix. The result indicates that segments in the vicinity of mixed land-use developments (residential, industrial, institutional, commercial, and recreational areas) are likely to experience lower crash risk potentially because of reduced driving speeds.

5.5 Intersection Level Attributes

In the intersection level model, the parameter associated with the proportion of interstate-expressway roads at the intersection has a negative impact on intersection crashes. It is possible that these roadways are well maintained in terms of pavement quality, lighting and enforcement improving overall safety. As the length of bike lane in the vicinity of the intersection increases, the number of crashes is likely to reduce. The result is quite interesting and might be important for encouraging bike infrastructure additions. It is possible that there might be self-selection at play as well. Specifically, it is possible that bike lanes are added on safer roadways.

As expected, the parameter associated with AADT has a positive impact on intersection crash frequency. This finding is in line with previous studies (Alarifi et al., 2018, 2017; Cai et al., 2019; Huang et al., 2016; Park et al., 2020; Saha et al., 2020; Veeramisti et al., 2021). The AADT parameter also exhibits significant variation across intersections as evidenced by the significant random parameter. The overall impact of this variable on crash count at an intersection level is likely to be overwhelmingly positive (99.99%).

5.6 Macro Level Attributes

Several variables influenced the macroscopic model including the micro level propensity component for segments and intersections. The coefficients for the scalar parameters for segments and intersections are positive as expected indicating that an increase in propensity of micro level crashes contributes to an increase in macro level crash frequency.

Among other zonal variables considered, average speed limit in the zone is associated with increased crash risk. The findings indicate that higher average speeds are associated with higher number of crashes. The result while counter intuitive should be carefully considered. Usually, on freeways or larger facilities, higher speed limits are likely to reduce crashes. However, in a zone with mixed facility type, higher average speed might highlight increased speed transitions and potential conflicts. In addition, the variable truck AADT is associated with increased crash risk. An increased presence of trucks in the zone can affect traffic flow and increase speed variance leading to higher number of crashes. Trucks are also likely to reduce visibility for other vehicles and might increase crash risk (Cui and Xie, 2021; Xie et al., 2017).

An increase in zonal commercial area is associated with higher crash risk. The result might indicate that commerce related activities such as loading/unloading, movement of heavy vehicles and increased traffic conflicts might contribute to higher crash risk (Cui and Xie, 2021; Mohammadnazar et al., 2021; Soroori et al., 2019; Xie et al., 2019). In terms of household density, the model results indicate that increased density is associated with higher crash risk. The result is quite intuitive and indicates as density increases, traffic is likely to increase and contribute to additional crash risk. Interestingly, a random parameter associated with household density indicates significant variability in the risk associated with household density. To be sure, while the impact of the variable varies across zones, the net impact is predominantly positive (99.99%). Finally, the zones with higher portion of African-American minorities are associated with higher crash risk. The result is a potential manifestation of inadequate facilities in low-income and minority neighborhoods in the region.

5.7 Unobserved Heterogeneity

As discussed earlier, the proposed model system accommodates spatial variations for facilities in the zone through hierarchical correlations such as a correlation between all segments or intersections in a zone. The last two variables in Table 2 correspond to these correlations. The significant effect of these parameters clearly highlights the presence of common unobserved factors across facilities present in the same zone. The result further reinforces our hypothesis that incorporating such correlations in crash frequency analysis is important.

5.8 Predictive Performance of the Model

Along with the log-likelihood and BIC measures, we assess the predictive performance of the proposed integrated micro-macro model by comparing RMSE (Root mean square error) values with non-integrated macro model (the lower value represents better prediction result). The RMSE

values of the proposed integrated model and non-integrated macro model are 83.761 and 110.958, respectively. The aggregate measures clearly highlight the improved predictive performance offered by our proposed integrated approach. To further evaluate the predictive performance of the estimated models, we carried out a comparison exercise between the proposed integrated micro-macro model 1 with correlation and random effects and the non-integrated macro model by comparing RMSE values across different crash groups (see Figure 2). The exercise shows that our proposed model exhibits either lower or similar RMSE values compared to the non-integrated model across majority of the groups (13 out of 15). To be specific, the proposed integrated method provides significantly better performance across 9 crash groups while for other 4, the RMSE value for both models are quite close to each other. However, for the remaining two crash groups, the non-integrated model performs slightly better, as evidenced by the marginal differences in RMSE across the two systems.

6 CONCLUSIONS

Crashes at macroscopic level are usually generated by aggregating the crashes at microscopic level, still majority of the earlier research developed individual model systems for each of these levels while ignoring the interconnectedness between them. The current research contributes to the safety literature by integrating the two traditionally divergent streams of crash frequency research (micro and macro) into a unified framework while allowing the influence of independent variables at the microscopic level to be incorporated within the macroscopic propensity estimation. The single integrated model will allow us to employ a single model code for estimation of macroscopic and microscopic models. The model system proposed recognizes that crashes at the micro level facilities contribute to total macroscopic level crash counts. This influence is incorporated by adding the sum of crash propensity specific to each micro-facility (such as segments and intersections) in the form of an additional variable in the propensity of the macroscopic model system. Based on the introduction of the micro level component, two frameworks are proposed in the current research effort. In the first approach, we fix the microscopic models and accommodate for the cumulative propensity influence as any other independent variable. On the other hand, the second approach is developed while allowing the feedback between macroscopic and microscopic models through a joint log-likelihood function. For both approaches, a scale parameter is estimated for accommodating the share of each micro level propensities to be linked with the macro level propensities. The proposed model system also incorporates hierarchical correlations (like correlation between all segments or intersections in a zone) for considering the spatial variations for facilities in the zone. The overall model development process is achieved using a negative binomial regression framework at the micro and macro level.

The empirical analysis is conducted using zonal, segment and intersection level crash count data for the years 2018 and 2019 from Orlando city of Florida while considering a comprehensive set of exogenous variables from both micro and macro levels including roadway and traffic factors, land-use, built environment, and sociodemographic characteristics. The empirical analysis involves a series of model estimations including: 1) non-integrated models; 2) existing integrated micro-macro model (proposed by Cai et al., 2019); 3) our proposed integrated approach 1 and 4) our proposed integrated approach 2. The comparison exercise, based on the Bayesian Information Criterion (BIC) value clearly highlighted the improved performance of our proposed integrated approach 1 relative to the other models. Further, within the proposed approach 1, the model accommodating unobserved heterogeneity outperforms its independent counterparts, thus highlighting the importance of incorporating the influence of common unobserved factors in crash

frequency analysis. The model estimation results are further augmented by the evaluation of the predictive performance of the proposed model. The findings further reinforce the superiority of our proposed model over the non-integrated system.

The study is not without limitations. The proposed integrated approach requires compiling data collectively across all micro level facilities in the region. The compilation can be cumbersome for major urban regions as several small intersections and segments might lead to large datasets and substantial data processing resources. The crash frequency variable considered in our analysis combines all motorized crashes of various types and non-motorized crashes in a single variable. It would be interesting to examine crash frequency by crash type (and severity) as separate dependent variables to offer more reasonable interpretations of different variables on crash frequency. Further, as the methodology proposed is focused on crash frequency at a spatial aggregation (TAZ) or facility aggregation (segment or intersection), it is not possible to accommodate for driver behavior in our model system. It would be worthwhile to develop frameworks that can accommodate for such crash level variables in crash frequency models in future research efforts.

ACKNOWLEDGMENT

The authors would like to gratefully acknowledge Signal Four Analytics (S4A) and Florida Department of Transportation (FDOT) for providing access to Florida crash and geospatial data.

REFERENCES

Abdel-Aty, M., Lee, J., Siddiqui, C., Choi, K., 2013. Geographical unit based analysis in the context of transportation safety planning. Transportation Research Part A 49, 62-75.

Abdel-Aty, M., Siddiqui, C., Huang, H., Wang, X., 2011. Integrating trip and roadway characteristics to manage safety in traffic analysis zones. Transportation Research Record 2213, 20–28.

Abdel-Aty, M., Wang, X., 2006. Crash estimation at signalized intersections along corridors: Analyzing spatial effect and identifying significant factors. Transportation Research Record 1953, 98–111.

Abdulhafedh, A., 2016. Crash frequency analysis. Journal of Transportation Technologies 6, 169.

Aguero-Valverde, J., Jovanis, P.P., 2008. Analysis of road crash frequency with spatial models. Transportation Research Record 2061, 55–63.

Aguero-Valverde, J., Jovanis, P.P., 2006. Spatial analysis of fatal and injury crashes in Pennsylvania. Accident Analysis and Prevention 38, 618–625.

Alarifi, S.A., Abdel-Aty, M., Lee, J., 2018. A Bayesian multivariate hierarchical spatial joint model for predicting crash counts by crash type at intersections and segments along corridors. Accident Analysis and Prevention 119, 263–273.

Alarifi, S.A., Abdel-Aty, M.A., Lee, J., Park, J., 2017. Crash modeling for intersections and segments along corridors: A Bayesian multilevel joint model with random parameters. Analytic Methods in Accident Research 16, 48–59.

Alhomaidat, F., Kwigizile, V., Oh, J.S., Houten, R. Van, 2020. How does an increased freeway speed limit influence the frequency of crashes on adjacent roads? Accident Analysis and Prevention 136, 105433.

Aptech, 2015. Aptech Systems Inc.

Azimian, A., Dimitra Pyrialakou, V., Lavrenz, S., Wen, S., 2021. Exploring the effects of area-level factors on traffic crash frequency by severity using multivariate space-time models. Analytic Methods in Accident Research 31, 100163.

Behnood, A., Mannering, F., 2019. Time-of-day variations and temporal instability of factors affecting injury severities in large-truck crashes. Analytic Methods in Accident Research 23, 100102.

Bhat, C.R., 2001. Quasi-random maximum simulated likelihood estimation of the mixed multinomial logit model. Transportation Research Part B 35, 677–693.

Bhowmik, T., Yasmin, S. and Eluru, N., 2022. Accommodating for systematic and unobserved heterogeneity in panel data: Application to macro-level crash modeling. Analytic Methods in Accident Research, 33, 100202.

Bhowmik, T., Rahman, M., Yasmin, S., Eluru, N., 2021a. Exploring analytical, simulation-based, and hybrid model structures for multivariate crash frequency modeling. Analytic Methods in Accident Research 31, 100167.

Bhowmik, T., Yasmin, S. and Eluru, N., 2021b. A new econometric approach for modeling several count variables: a case study of crash frequency analysis by crash type and severity. Transportation research part B, 153,172-203.

Bhowmik, T., Yasmin, S., Eluru, N., 2019. Do we need multivariate modeling approaches to model crash frequency by crash types? A panel mixed approach to modeling crash frequency by crash types. Analytic Methods in Accident Research 24, 100107.

Bhowmik, T., Yasmin, S. and Eluru, N., 2018. A joint econometric approach for modeling crash counts by collision type. Analytic Methods in Accident Research, 19,16-32.

Cai, Q., Abdel-Aty, M., Lee, J., Huang, H., 2019. Integrating macro- and micro-level safety analyses: a Bayesian approach incorporating spatial interaction. Transportmetrica A 15, 285–306.

Cheng, W., Gill, G.S., Zhou, J., Ensch, J.L., Kwong, J., Jia, X., 2020. Alternative multivariate multimodal crash frequency models. Journal of Transportation Safety and Security 12, 628–652.

Cui, H., Xie, K., 2021. An accelerated hierarchical Bayesian crash frequency model with accommodation of spatiotemporal interactions. Accident Analysis and Prevention 153, 106018.

Dong, C., Clarke, D.B., Yan, X., Khattak, A., Huang, B., 2014. Multivariate random-parameters zero-inflated negative binomial regression model: An application to estimate crash frequencies at intersections. Accident Analysis and Prevention 70, 320–329.

Gong, H., Wang, F., Zhou, B. (Brenda), Dent, S., 2020. Application of random effects negative binomial model with clustered dataset for vehicle crash frequency analysis. Journal of Transportation Science and Technology 9, 183–194.

Huang, H., Abdel-Aty, M., 2010. Multilevel data and Bayesian analysis in traffic safety. Accident Analysis and Prevention 42, 1556–1565.

Huang, H., Abdel-Aty, M.A., Darwiche, A.L., 2010. County-level crash risk analysis in Florida: Bayesian spatial modeling. Transportation Research Record 2148, 27–37.

Huang, H., Chang, F., Zhou, H., Lee, J., 2019. Modeling unobserved heterogeneity for zonal crash frequencies: A Bayesian multivariate random-parameters model with mixture components for spatially correlated data. Analytic Methods in Accident Research 24, 100105.

Huang, H., Song, B., Xu, P., Zeng, Q., Lee, J., Abdel-Aty, M., 2016. Macro and micro models for zonal crash prediction with application in hot zones identification. Journal of Transport Geography 54, 248–256.

Islam, M., Mannering, F., 2020. A temporal analysis of driver-injury severities in crashes involving aggressive and non-aggressive driving. Analytic Methods in Accident Research 27,

100128.
Kaaf, K. Al, Abdel-Aty, M., 2015. Transferability and calibration of highway safety manual performance functions and development of new models for urban four-lane divided roads in Riyadh, Saudi Arabia. Transportation Research Record 2515, 70–77.
Kabli, A., Bhowmik, T., Eluru, N., 2020. A multivariate approach for modeling driver injury severity by body region. Analytic Methods in Accident Research 28, 100129.
Kim, D.G., Lee, Y., Washington, S., Choi, K., 2007. Modeling crash outcome probabilities at rural intersections: Application of hierarchical binomial logistic models. Accident Analysis and Prevention 39, 125–134.
Kim, D.G., Washington, S., 2006. The significance of endogeneity problems in crash models: An examination of left-turn lanes in intersection crash models. Accident Analysis and Prevention 38, 1094–1100.
Lee, J., Abdel-Aty, M., Cai, Q., 2017. Intersection crash prediction modeling with macro-level data from various geographic units. Accident Analysis and Prevention 102, 213–226.
Li, Z., Chen, X., Ci, Y., Chen, C., Zhang, G., 2019. A hierarchical Bayesian spatiotemporal random parameters approach for alcohol/drug impaired-driving crash frequency analysis. Analytic Methods in Accident Research 21, 44–61.
Li, Z., Wang, W., Liu, P., Bigham, J.M., Ragland, D.R., 2013. Using geographically weighted poisson regression for county-level crash modeling in California. Safety Science 58, 89–97.
Liu, C., Sharma, A., 2018. Using the multivariate spatio-temporal Bayesian model to analyze traffic crashes by severity. Analytic Methods in Accident Research 17, 14–31.
Liu, J., Khattak, A.J., Wali, B., 2017. Do safety performance functions used for predicting crash frequency vary across space? Applying geographically weighted regressions to account for spatial heterogeneity. Accident Analysis and Prevention 109, 132–142.
Lord, D., Mannering, F., 2010. The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives. Transportation Research Part A 44, 291–305.
Mannering, F., 2018. Temporal instability and the analysis of highway accident data. Analytic Methods in Accident Research 17, 1–13.
Mannering, F.L., Bhat, C.R., 2014. Analytic methods in accident research: Methodological frontier and future directions. Analytic Methods in Accident Research 1, 1–22.
Marcoux, R., Yasmin, S., Eluru, N., Rahman, M., 2018. Evaluating temporal variability of exogenous variable impacts over 25 years: An application of scaled generalized ordered logit model for driver injury severity. Analytic Methods in Accident Research 20, 15–29.
Mohammadnazar, A., Mahdinia, I., Ahmad, N., Khattak, A.J., Liu, J., 2021. Understanding how relationships between crash frequency and correlates vary for multilane rural highways: Estimating geographically and temporally weighted regression models. Accident Analysis and Prevention 157, 106146.
Mousavi, S.M., Osman, O.A., Lord, D., Dixon, K.K., Dadashova, B., 2021. Investigating the safety and operational benefits of mixed traffic environments with different automated vehicle market penetration rates in the proximity of a driveway on an urban arterial. Accident Analysis and Prevention 152, 105982.
Noland, R.B., Oh, L., 2004. The effect of infrastructure and demographic change on traffic-related fatalities and crashes: A case study of Illinois county-level data. Accident Analysis and Prevention 36, 525–532.
Oh, J., Washington, S., Choi, K., 2004. Development of accident prediction models for rural highway intersections. Transportation Research Record 1897, 18–27.

Park, H.C., Yang, S., Park, P.Y., Kim, D.K., 2020. Multiple membership multilevel model to estimate intersection crashes. Accident Analysis and Prevention 144, 105589.
Pljakić, M., Jovanović, D., Matović, B., Mićić, S., 2019. Macro-level accident modeling in Novi Sad: A spatial regression approach. Accident Analysis and Prevention 132, 105259.
Saha, D., Alluri, P., Dumbaugh, E., Gan, A., 2020. Application of the Poisson-Tweedie distribution in analyzing crash frequency data. Accident Analysis and Prevention 137, 105456.
Satria, R., Aguero-Valverde, J., Castro, M., 2020. Spatial analysis of road crash frequency using Bayesian models with Integrated Nested Laplace Approximation (INLA). Journal of Transportation Safety and Security, 1–23.
Shirani-bidabadi, N., Mallipaddi, N., Haleem, K., Anderson, M., 2020. Developing bicycle-vehicle crash-specific safety performance functions in Alabama using different techniques. Accident Analysis and Prevention 146, 105735.
Soroori, E., Mohammadzadeh Moghaddam, A., Salehi, M., 2019. Application of local conditional autoregressive models for development of zonal crash prediction models and identification of crash risk boundaries. Transportmetrica A 15, 1102–1123.
Stamatiadis, N., Pigman, J.G., Sacksteder, J., Ruff, W., Lord, D., 2009. Impact of shoulder width and median width on safety. NCHRP Report 633, Transportation Research Board, Washington, D.C.
Tirtha, S.D., Yasmin, S., Eluru, N., 2020. Modeling of incident type and incident duration using data from multiple years. Analytic Methods in Accident Research 28, 100132.
Veeramisti, N., Paz, A., Khadka, M., Arteaga, C., 2021. A clusterwise regression approach for the estimation of crash frequencies. Journal of Transportation Safety and Security 13, 247–277.
Wali, B., Khattak, A.J., Waters, J., Chimba, D., Li, X., 2018. Development of safety performance functions: Incorporating unobserved heterogeneity and functional form analysis. Transportation Research Record 2672, 9–20.
Wang, C., Chen, F., Cheng, J., Bo, W., Zhang, P., Hou, M., Xiao, F., 2020. Random-parameter multivariate negative binomial regression for modeling impacts of contributing factors on the crash frequency by crash types. Discrete Dynamics in Nature and Society, 1–13.
Wang, J., Huang, H., 2016. Road network safety evaluation using Bayesian hierarchical joint model. Accident Analysis and Prevention 90, 152–158.
Wang, J., Huang, H., Zeng, Q., 2017. The effect of zonal factors in estimating crash risks by transportation modes: Motor vehicle, bicycle and pedestrian. Accident Analysis and Prevention 98, 223–231.
Wang, K., Bhowmik, T., Zhao, S., Eluru, N., Jackson, E., 2021. Highway safety assessment and improvement through crash prediction by injury severity and vehicle damage using multivariate Poisson-lognormal model and joint negative binomial-generalized ordered probit fractional split model. Journal of Safety Research 76, 44–55.
Wang, S., Chen, Y., Huang, J., Chen, N., Lu, Y., 2019. Macrolevel traffic crash analysis: A spatial econometric model approach. Mathematical Problems in Engineering 2019, 1–10.
Wen, H., Zhang, X., Zeng, Q., Sze, N.N., 2019. Bayesian spatial-temporal model for the main and interaction effects of roadway and weather characteristics on freeway crash incidence. Accident Analysis and Prevention 132, 105249.
Xie, K., Ozbay, K., Kurkcu, A., Yang, H., 2017. Analysis of traffic crashes involving pedestrians using big data: Investigation of contributing factors and identification of hotspots. Risk Analysis 37, 1459–1476.

Xie, K., Ozbay, K., Yang, H., 2019. A multivariate spatial approach to model crash counts by injury severity. Accident Analysis and Prevention 122, 189–198.
Yasmin, S., Eluru, N., 2018. A joint econometric framework for modeling crash counts by severity. Transportmetrica A 14, 230–255.
Yasmin, S., Eluru, N., 2016. Latent segmentation based count models: Analysis of bicycle safety in Montreal and Toronto. Accident Analysis and Prevention 95, 157–171.
Yasmin, S., Eluru, N., 2013. Evaluating alternate discrete outcome frameworks for modeling crash injury severity. Accident Analysis and Prevention 59, 506–521.
Yasmin, S., Momtaz, S.U., Nashad, T., Eluru, N., 2018. A Multivariate copula-based macro-level crash count model. Transportation Research Record 2672, 64–75.
Ye, X., Pendyala, R.M., Washington, S.P., Konduri, K., Oh, J., 2009. A simultaneous equations model of crash frequency by collision type for rural intersections. Safety Science 47, 443–452.
Yu, R., Wang, Y., Quddus, M., Li, J., 2019. A marginalized random effects hurdle negative binomial model for analyzing refined-scale crash frequency data. Analytic Methods in Accident Research 22, 100092.
Zeng, Q., Guo, Q., Wong, S.C., Wen, H., Huang, H., Pei, X., 2019. Jointly modeling area-level crash rates by severity: A Bayesian multivariate random-parameters spatio-temporal Tobit regression. Transportmetrica A 15, 1867–1884.
Zeng, Q., Wen, H., Huang, H., Wang, J., Lee, J., 2020. Analysis of crash frequency using a Bayesian underreporting count model with spatial correlation. Physica A 545, 123754.
Zhai, X., Huang, H., Xu, P., Sze, N.N., 2019. The influence of zonal configurations on macro-level crash modeling. Transportmetrica A 15, 417–434.

LIST OF FIGURES

LIST OF TABLES

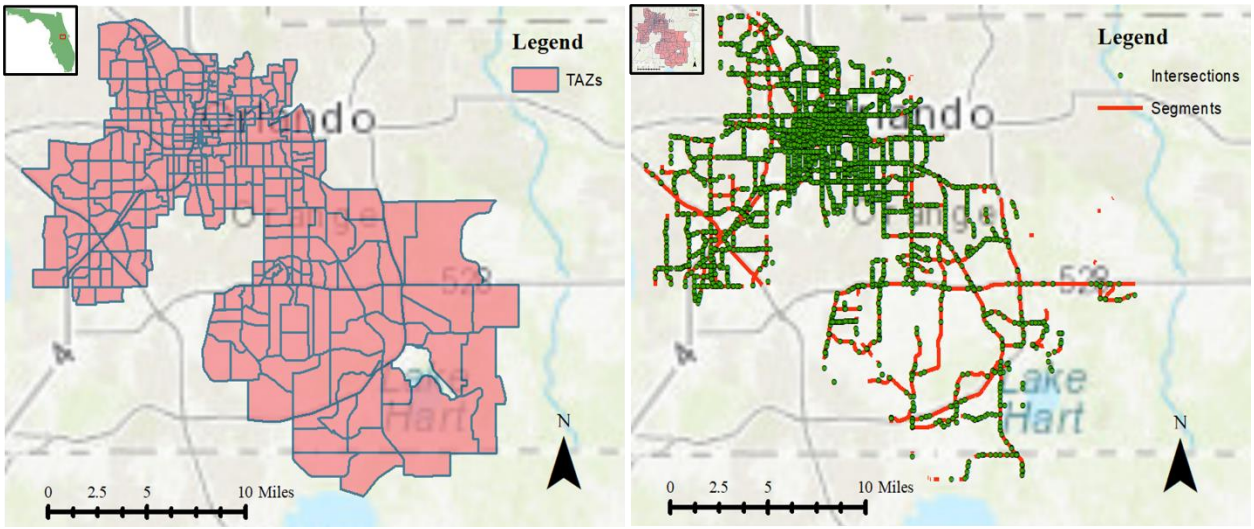

FIGURE 1: Illustration of a) Traffic Analysis Zones (left), and b) Segments and Intersections (right)

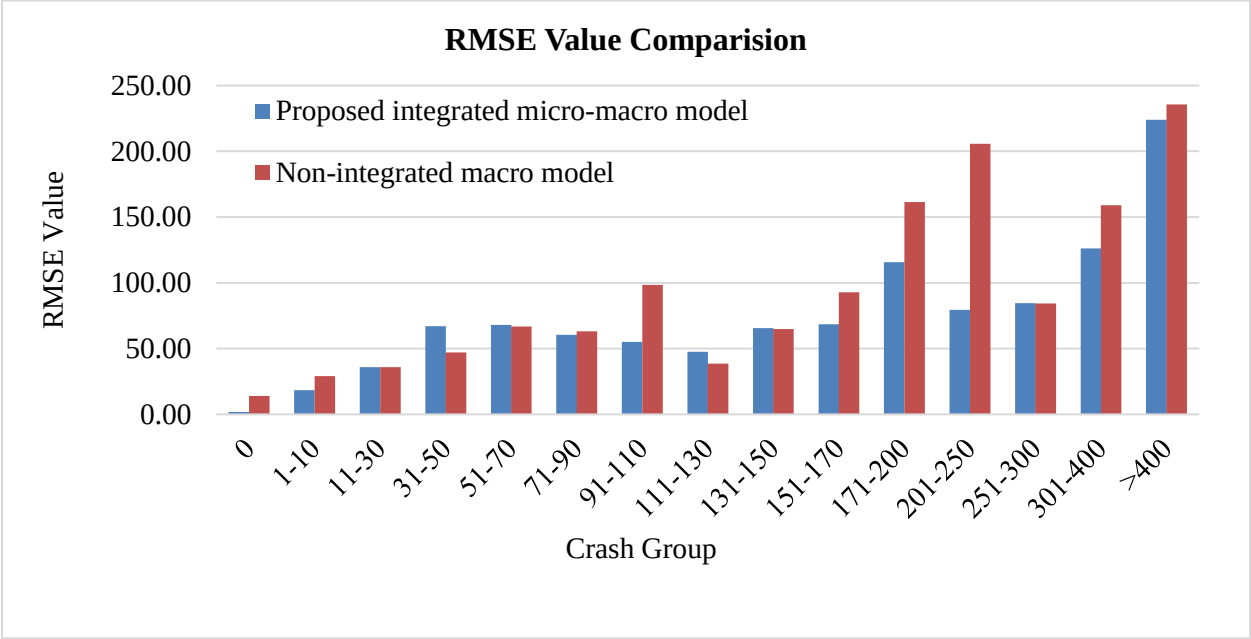

FIGURE 2: RMSE Values of Proposed Integrated Micro-Macro Model and Non-Integrated Macro Model across Different Crash Groups

TABLE 1 Summary Statistics of the Exogenous Variables at Micro and Macro Level

Variables	Definition	Min	Max	Mean	SD
Segment level (micro)					
Crash	Segment crash	0.000	481.000	6.161	19.241
Segment length	Ln (Segment length in mile)	-12.852	1.144	-2.758	1.770
No of lanes	No of lanes of the segment	1.000	5.000	2.128	0.720
Average inside shoulder width	Average inside shoulder width in feet	0.000	36.000	3.028	4.596
Average sidewalk width	Ln (Average sidewalk width in feet + 1)	0.000	2.708	1.789	0.491
No of traffic signals per unit segment length	No of traffic signals/ segment length in meter	0.000	474.048	0.862	13.620
AADT	Ln (AADT of the segment + 1)	6.686	12.095	9.987	0.891
Land-use mix	Land-use mix = $[-\frac{\sum(p_k(\ln p_k))}{\ln N}]$, where k is the category of land-use, p is the proportion of the developed land area for specific land-use, N is the number of land-use categories; here N=5	0.000	0.997	0.545	0.226
Intersection level (micro)					
Crash	Crash within 250 feet buffer intersection area	0.000	202.000	7.382	15.152
Proportion of interstate and expressways	Interstate and expressways length/total road length	0.000	1.000	0.092	0.139
Average bike lane length	Average bike lane length in mile	0.000	0.472	0.014	0.024
AADT	Ln (AADT of the intersection + 1)	0.000	20.554	15.217	2.147
TAZ level (macro)					
Crash	TAZ crash	0.000	765.000	140.287	130.587
Proportion of interstate and expressways	Interstate and expressways length/total road length in TAZ area	0.000	1.000	0.095	0.200
Proportion of divided road	Divided roads length/total road length in TAZ area	0.000	1.000	0.610	0.357
Average speed limit	Average speed limit in mph	0.000	70.000	36.422	10.725
Average sidewalk width	Ln (Average sidewalk width in feet + 1)	0.000	2.646	1.775	0.506
Intersection density	No of intersections/ area of TAZ in acre	0.000	0.770	0.085	0.115
Traffic signal per intersection	No of traffic signal/total no of intersections in the TAZ	0.000	1.000	0.058	0.106
AADT	Ln (AADT of TAZ + 1)	0.000	13.507	11.189	1.864
Truck AADT	Ln (Truck AADT of TAZ + 1)	0.000	11.302	8.326	1.613
Proportion of heavy vehicle	Total truck AADT/total AADT	0.000	0.170	0.056	0.024
Proportion of residential area	Residential area/total land-use area	0.000	0.989	0.412	0.326
Proportion of commercial area	Commercial area/ total land-use area	0.000	1.000	0.203	0.248
No of restaurants	Z score*: No of restaurants	-0.597	6.690	0.000	1.000
No of educational centers	Z score: No of educational centers	-0.649	3.879	0.000	1.000
Household density	No of households per acre TAZ area	0.084	8.621	2.016	1.574
Non-motorized means of transport	Ln (Non-motorized means of transport + 1)	0.000	5.366	2.152	1.166
Proportion of African American population	No of African American population/total population	0.000	0.978	0.222	0.246

**Z-score represents the standardized form of the actual variable*

TABLE 2 Results of Proposed Integrated Micro-Macro Model

Variables	Proposed integrated macro-micro model	
	Estimate	**t-stat**
Segment level (micro)		
Constant	-0.813	-1.08
Segment length	0.515	5.753
No of lanes	0.179	2.389
Average inside shoulder width	0.037	2.475
Standard deviation*	0.022	1.746
Average sidewalk width	-0.210	-1.808
No of traffic signals per unit segment length	0.025	4.24
AADT	0.306	4.838
Land-use mix	-0.818	-3.191
Over-dispersion parameter	2.189	11.727
Intersection level (micro)		
Constant	0.922	3.964
Proportion of interstate and expressways	-0.400	-1.725
Average bike lane length	-3.263	-2.362
AADT	0.030	1.992
Standard deviation	0.007	1.986
Over-dispersion parameter	3.000	39.23
TAZ level (macro)		
Constant	-2.730	-9.085
Average speed limit	0.022	4.178
Truck AADT	0.316	7.774
Proportion of commercial area	0.439	3.623
Household density	0.092	4.053
Standard deviation	0.002	2.458
Proportion of African American population	0.253	2.381
Over-dispersion parameter	2.933	14.653
Parameter for segment level propensity sum	0.226	7.809
Parameter for intersection level propensity sum	0.497	10.007
Unobserved heterogeneity		
Segment correlation	0.682	8.875
Intersection correlation	0.739	17.948
For proposed integrated micro-macro model: Log-likelihood: -17,068.710; BIC: 34,297.126		

** The standard deviation denotes random parameter associated with respective variable which indicates significant variation of the impact of that variable across the study level.*

Supplementary Material

TABLE S1: Results of Integrated Micro-Macro Models with 2018, 2019, and Pooled Records

Variables	Integrated macro-micro model (2018 records)		Integrated macro-micro model (2019 records)		Integrated macro-micro model (pooled records)	
	Estimate	**t-stat**	**Estimate**	**t-stat**	**Estimate**	**t-stat**
Segment level (micro)						
Constant	-0.412	-0.675	-0.550	-0.773	-0.456	-0.988
Segment length	0.602	7.465	0.564	6.559	0.583	9.904
No of lanes	0.133	1.890	0.229	2.879	0.183	3.398
Average inside shoulder width	0.033	2.257	0.034	2.374	0.034	3.299
Average sidewalk width	-0.218	-1.828	-0.230	-1.917	-0.224	-2.650
No of traffic signals per unit segment length	0.029	4.924	0.027	4.752	0.028	6.857
AADT	0.290	5.717	0.276	4.502	0.280	7.098
Land-use mix	-0.678	-2.710	-0.662	-2.497	-0.671	-3.688
Over-dispersion parameter	1.997	10.607	2.155	10.366	2.078	14.789
Intersection level (micro)						
Constant	1.302	22.623	1.286	22.905	1.669	10.062
Average inside shoulder width	0.017	1.805	0.018	1.953	0.031	3.903
Average no of lanes	--	--	--	--	-0.167	-2.086
Proportion of local road	-0.481	-1.668	--	--	-0.496	-2.234
Proportion of institutional area	-0.790	-2.558	-0.754	-2.379	-0.754	-3.404
No of finance centers	--	--	-0.067	-1.850	-0.058	-2.156
Proportion of household with no vehicle	--	--	--	--	-0.746	-1.964
Over-dispersion parameter	3.898	33.837	3.808	34.049	3.843	47.857
TAZ level (macro)						
Constant	-3.035	-7.401	-2.513	-6.018	-2.794	-9.372
Average speed limit	0.016	2.325	0.012	1.732	0.015	3.164
Truck AADT	0.439	7.851	0.378	6.447	0.412	10.081
Proportion of commercial area	0.688	3.863	0.608	3.991	0.677	5.706
Household density	0.119	4.373	0.106	3.701	0.112	5.611
Proportion of African American population	0.536	3.095	0.593	3.716	0.565	4.834
Over-dispersion parameter	1.126	7.943	0.989	7.299	1.058	10.737
Parameter for segment level propensity sum	0.234	4.697	0.295	4.413	0.259	6.087
Parameter for intersection level propensity sum	0.432	7.885	0.441	7.318	0.430	10.816
Number of parameters	23		23		26	
Log-likelihood	-13,548.720		-13,645.050		-27,189.480	
Bayesian Information Criterion (BIC) value	54, 649.914				54,545.280	
Sample size	300				600	

TABLE S2 Results of Non-Integrated NB Micro and Macro Models

Variables	Non-integrated NB model results	
	Estimate	**t-stat**
Non-Integrated Segment level (micro)		
Constant	-0.813	-1.08
Segment length	0.515	5.753
No of lanes	0.179	2.389
Average inside shoulder width	0.037	2.475
Average sidewalk width	-0.210	-1.808
No of traffic signals per unit segment length	0.025	4.24
AADT	0.306	4.838
Land-use mix	-0.818	-3.191
Over-dispersion parameter	2.189	11.727
Log-likelihood: -4,238.812; BIC: 8,545.174		
Non-Integrated Intersection level (micro)		
Constant	0.922	3.964
Proportion of interstate and expressways	-0.400	-1.725
Average bike lane length	-3.263	-2.362
AADT	0.030	1.992
Over-dispersion parameter	3.000	39.23
Log-likelihood: -11,617.002; BIC: 23,275.698		
Non-Integrated TAZ level (macro)		
Constant	-6.422	-9.476
Proportion of interstate and expressways	-0.524	-2.716
Proportion of divided road	0.457	3.286
Average speed limit	0.026	2.575
Average sidewalk width	0.256	1.758
Intersection density	1.709	3.347
Traffic signal per intersection	0.623	2.213
AADT	0.696	12.138
Proportion of heavy vehicle	2.655	1.743
Proportion of residential area	0.248	1.954
No of restaurants	0.273	5.788
No of educational centers	0.123	3.301
Non-motorized means of transport	0.085	2.758
Proportion of African American population	0.447	3.498
Over-dispersion parameter	3.019	15.261
Log-likelihood: -1,460.505; BIC: 3,006.567		
Non-integrated (combining micro and macro) models: Log-likelihood: -17,316.319; BIC: 34,798.048		