Assessing the Variation of Curbside Safety at the City Block Level

Aditya Medury[1], Dimitris Vlachogiannis[1] and Offer Grembek[1]

[1] Safe Transportation Research and Education Center, University of California Berkeley

Berkeley UNIVERSITY OF CALIFORNIA

Introduction

The abundance of data in today's world generates opportunities for deeper comprehension of the various parameters affecting crash frequency. This study incorporates data from many different sources (summary statistics are provided in Table 1) including geocoded police-reported crash data, curbside infrastructure data and socio-demographic data for the city of San Francisco, CA. To handle over-dispersion, negative binomial (NB) models were developed, and to capture additional unobserved heterogeneity, two-component finite mixture NB models were formulated, one with fixed priors (FMNB) and another with varying priors (GFMNB).

Table 1: Summary statistics of variables considered

Variable	Min	q_1	$\tilde{x}$	$\bar{x}$	q_3	Max
No. of collisions (2013-2017)	0.00	0.00	0.00	0.42	0.00	17.00
Length	0.17	57.20	86.02	99.57	134.93	1242.03
AADT	0.00	240.60	1592.56	4433.23	5465.72	62287.93
Peak hour bicycle volumes	0.00	3.82	27.04	189.02	129.51	7392.41
Median presence (0/1)	0.00	0.00	0.00	0.10	0.00	1.00
No. of lanes	0.00	2.00	2.00	2.20	2.00	6.00
On-street parking spaces	0.00	0.00	0.00	3.09	0.00	110.00
Fraction of local streets	0.00	0.00	1.00	0.71	1.00	1.00
Presence of bus lines	0.00	0.00	0.00	0.35	1.00	1.00
Ave. daily TNC activity	3.68	5.51	6.07	6.03	6.59	7.79
Percentage of zero-vehicle households	0.00	0.12	0.23	0.29	0.41	0.98

Negative Binomial Model

Given the non-negative integer nature of crashes, negative binomial (NB) models are extensively used for crash frequency modeling. The negative binomial regression arises from a two-stage model for the distribution of number of crashes, n_i, which follows a Poisson distribution with the mean λ_i,and,λ_i follows the Gamma distribution with shape parameter ϕ and scale parameter, μ_i/ϕ:

$$n_i|\lambda_i \sim \text{Poisson}(\lambda_i) \quad \text{and} \quad \lambda_i \sim \Gamma(\phi, \mu_i/\phi) \tag{1}$$

$$p(n_i) = \int_0^\infty p(n_i|\lambda_i) f(\lambda_i) d\lambda_i = \frac{\Gamma(\phi+n_i)}{\Gamma(\phi)\,\Gamma(n_i+1)}\left(\frac{\phi}{\phi+\mu_i}\right)^{\phi}\left(\frac{\mu_i}{\phi+\mu_i}\right)^{n_i} \tag{2}$$

Herein, $\Gamma()$ is the Gamma function and $f(n_i)$ is the marginal probability of observing n_i crashes. Given a vector of explanatory variables, $\mathbf{X_i}$, the mean, $E(n_i)$ for a given segment, i, can be expressed as:

$$E_{NB}(n_i) = \mu_i = exp(\beta \mathbf{X_i}) \tag{3}$$

To correct for regression-to-the-mean, an Empirical Bayes (EB) adjustment is applied, combining a site's historical crash data with the expected number of crashes estimated based on the site characteristics:

$$E_{NB}(\lambda_i|n_i) = \left(\frac{\mu_i}{\mu_i+\phi}\right) n_i + \left(\frac{\phi}{\mu_i+\phi}\right)\mu_i. \tag{4}$$

Potential for Safety Improvement

Another metric considered for prioritization is the potential for safety improvement (PSI), defined as excess expected average crash frequency with EB adjustment:

$$\text{PSI}_{NB,i} = E_{NB}(\lambda_i|n_i) - \mu_i \tag{5}$$

Generalized Finite Mixture NB Model

A finite mixture negative binomial model utilizes a finite number (K) of unobserved categories/latent classes to capture the unobserved heterogeneity in crash data. The crash count at a location, n_i, follows the Poisson distribution with the mean crash rate λ_i, and λ_i in turn follows a K-component finite mixture of gamma distribution:

$$n_i|\lambda_i \sim \text{Poisson}(\lambda_i) \quad \text{and} \quad p(\lambda_i) = \sum_{k=1}^{K} \pi_{ik} p_k(\lambda_{ik}), \tag{6}$$

where, $\pi_{ik} = \pi_k(\gamma, z_i)$ is the prior probability of component k, with $\sum_{k=1}^{K} \pi_{ik} = 1$ with $\pi_{ik} > 0, \forall k$.

The marginal distribution of n_i follows a mixture of NB distributions with probability density function and mean defined as follows:

$$p(n_i|\mu_i, \mathbf{\Theta}) = \sum_{k=1}^{K} \pi_{ik}\left(\frac{\Gamma(\phi_k+n_i)}{\Gamma(\phi_k)\Gamma(n_i+1)}\left(\frac{\phi_k}{\phi_k+\mu_{ik}}\right)^{\phi_k}\left(\frac{\mu_{ik}}{\phi_k+\mu_{ik}}\right)^{n_i}\right) \tag{7}$$

$$E(n_i|\mathbf{X_i}, \mathbf{\Theta}) = \sum_{k=1}^{K} \pi_{ik}\mu_{ik} \tag{8}$$

The prior probability is modeled using multinomial/binary logit framework using explanatory variables z and coefficients γ. $\mathbf{\Theta} = \{(\beta_1, \dots, \beta_K), (\phi_1, \dots, \phi_K), \gamma\}$ is the vector of all parameters. The posterior probability of an observation, i belong to component, k is given by:

$$w_{ik} = \pi_{ik} p_k(n_i)/p(n_i), \tag{9}$$

where, $p_k(n_i)$ represents marginal probability of observing n_i crashes conditional on the observation belonging to component k.

The EB estimate for the GFMNB model is given by:

$$E_{GFMNB}(\lambda_i|n_i) = \sum_{k=1}^{K} w_{ik}\left\{\left(\frac{\mu_{ik}}{\mu_{ik}+\phi_k}\right) n_i + \left(\frac{\phi_k}{\mu_{ik}+\phi_k}\right)\mu_{ik}\right\} \tag{10}$$

Two different functions for the derivation of the priors are implemented in the context of this research. One for constant component weights without a class membership model (FMNB),i.e., $\pi_{ik} = \pi_k$ and one for a finite mixture model with a class membership model estimated using multinomial logit models (GFMNB).

Methodology

Three nested NB models where tested on three model selection criteria (Log-Likelihood (LL), Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC)). The first model considered only segment-level variables, while the second incorporated adjoining intersection information. The third NB model alternative incorporated TAZ-level variables on TNC pick-ups/drop-offs and demographic information.

Results

Based on all the three likelihood-based model selection criteria, we conclude that the third model best explains the segment-level crashes among the NB model specifications. Once the final NB model was identified, two-component FMNB and GFMNB models were estimated using the same variables as in the NB model.

Variables	NB	FMNB		GFMNB	
		Comp. 1	Comp. 2	Comp. 1	Comp. 2
Intercept	-7.94^{***}	-8.68^{***}	-7.83^{***}	-9.09^{***}	-4.38^{***}
Log (Segment length)	0.87^{***}	0.90^{***}	0.88^{***}	0.91^{***}	0.79^{***}
Log (AADT)	0.07^{***}	0.09^{***}	$0.06^{\cdot}$	0.07^{***}	-0.01
Log (Peak-hour bike vol.)	0.09^{***}	-0.04^{*}	0.37^{***}	-0.00	$0.08^{\cdot}$
Presence of median	-0.26^{***}	-0.37^{*}	$-0.31^{\cdot}$	-0.41^{*}	-0.28^{*}
Number of lanes	0.16^{***}	0.14^{**}	0.17^{**}	0.33^{***}	-0.01
Local street (YES/NO)	-0.38^{***}	-0.39^{**}	-0.32^{*}	-0.31^{**}	-0.30^{**}
Within bicycle network	0.24^{***}	0.06	0.14	0.09	0.01
Number of bus stops	0.09^{**}	0.00	0.15^{*}	-0.02	0.17^{***}
Presence of bus lines	0.17^{**}	0.28^{*}	0.07	0.40^{***}	-0.02
Transit station (YES/NO)	0.37^{***}	0.45^{***}	0.24^{*}	0.45^{***}	0.27^{**}
Number of on-street parking meters	0.03^{***}	0.02^{*}	0.08^{***}	0.03^{***}	0.02^{**}
Presence of off-street parking	0.18^{***}	0.22^{*}	0.12	0.33^{***}	0.06
Mean PMT duration (multi-space)	-0.08^{***}	-0.05	-0.28^{***}	-0.31^{**}	-0.06^{*}
Mean PMT duration (single-space)	-0.23^{***}	-0.02	-1.34^{***}	-0.26^{*}	-0.24^{*}
Speed limit ($\geq$ 35 mph)	0.42^{***}	1.08^{***}	-0.36	0.95^{***}	-0.10
One-way street	-0.20^{***}	-0.30^{*}	-0.03	-0.05	-0.25^{*}
Adjoining intersections, signalized	0.50^{***}	0.60^{***}	0.34^{*}	0.51^{***}	0.23^{*}
Adjoining intersections, 4-legged	-0.10^{*}	0.21^{*}	-0.39^{***}	0.11	-0.31^{***}
Log (Daily average TNC activity)	0.21^{***}	0.29^{***}	0.10	0.27^{***}	0.03
Percentage of zero vehicle households	0.67^{***}	1.23^{***}	-0.16	0.61^{**}	0.82^{***}
ϕ	2.46	9.40	4.72	10.88	3.62
Class Membership	–	52.0%	48.0%	52.0%	48.0%
AIC	10988.92	10898.42		**10876.93**	
BIC	**11143.22**	11214.03		11234.62	
Log Likelihood	–5472.46	-5404.21		**-5387.47**	

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\cdot}p < 0.1$

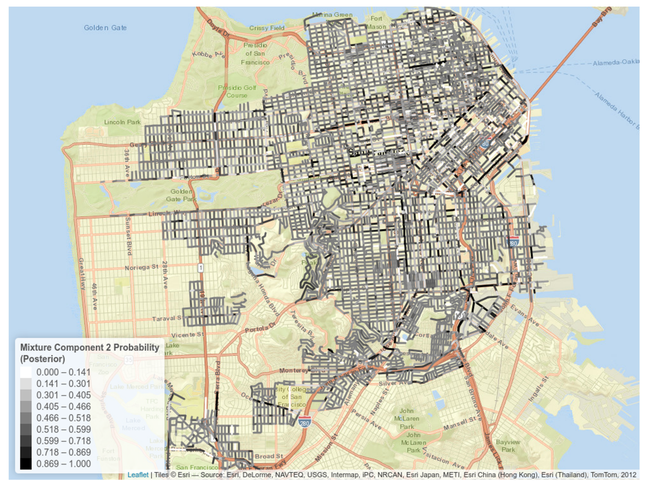
The summary statistics of component characteristics for the FMNB and GFMNB models are provided below:

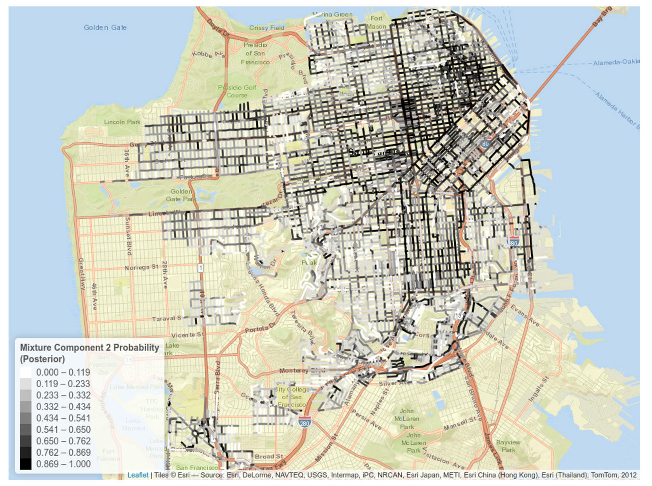
Variables	FMNB		GFMNB	
	Comp. 1	Comp. 2	Comp. 1	Comp. 2
Total collisions (mean)	0.3	0.7	0.2	0.6
Total collisions (variance)	0.6	1.8	0.6	1.4
% freeways	1.3%	1.2%	0.7%	0.2%
% major streets	8.3%	13.4%	7.1%	12.8%
% secondary streets	16.1%	22.0%	10.9%	25.9%
% local streets	74.4%	63.4%	81.3%	59.3%
Peak hour bicycle volumes	167.8	246.8	65.5	340.3
Presence of bus line	33.0%	40.0%	13.0%	61.0%
% one-way	16.3%	22.0%	11.7%	25.3%
TNC activity (mean)	513.7	647.7	450.0	671.8
AADT (mean)	4205.4	5053.2	3695.0	5337.7
% of zero vehicle households	27.2%	33.9%	20.8%	39.0%
Mean segment length	92.5	118.9	112.7	83.5

Results

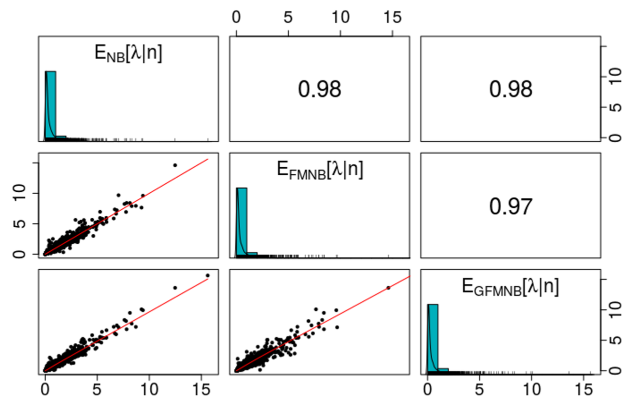
Indicatively, the coefficients for the class membership model for GFMNB component 2 which are estimated using a binary logit model are presented, as well as the posterior probabilities of component 2 for both FMNB (left) and GFMNB (right) for the segments in scope.

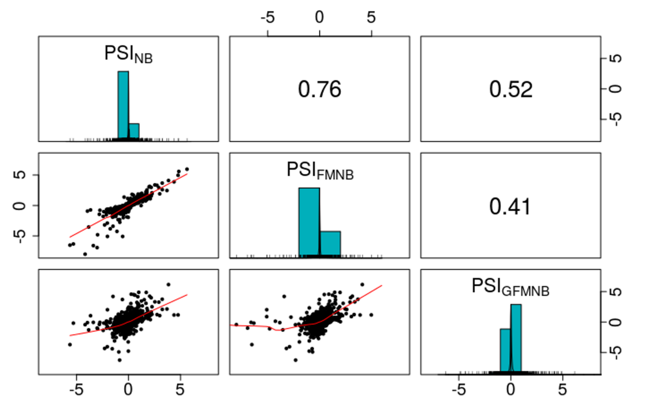
Variables	Intercept	Log (Segment length)	Log (AADT)	Log (Peak-hour bike vol.)
Coefficients	-15.33***	0.84*	0.45**	1.19***
Variables	Presence of bus lines	Log (Daily average TNC activity)	Percentage of zero vehicle households	
Coefficients	-0.2	0.27	-1.11	





The figure below, evaluates the correlation between models. The EB estimates (left) show high degree of correlation, implying that the different models may produce high overlaps when prioritizing sites for investigation. In comparison, the PSI estimates (right) demonstrate much lower correlations, especially between GFMNB and other models.





Discussion and Future Research

- The differences observed in the component-specific models illustrate the possibility that finite-mixture models may capture different safety regimes which can collectively explain the overall crash data.
- The use of random parameter models and spatial correlation structures will be explored in future research.
- Future studies must also consider more detailed safety assessments of curbside infrastructure such as temporary pick-up/drop-off loading zones, no-parking zones to assure causality in the variables considered.
- Definitive assessment of the impact of TNC pick-ups/drop-offs on safety will require more detailed data collection efforts and collaboration from TNC companies to account for the temporal variation in TNC activity.