Surrogate Safety and Network Screening:
Modelling Crash Frequency Using GPS Data and Latent Gaussian Models

Joshua Stipancic, PhD Candidate, McGill University
Luis Miranda-Moreno, Associate Professor, McGill University
Nicolas Saunier, Associate Professor, Polytechnique Montréal
Aurelie Labbe, Associate Professor, HEC Montréal

Paper prepared for presentation
at the Innovations in Road Safety Session

of the 2017 Conference of the
Transportation Association of Canada
St. John's, NL

Funding for this project was provided in part by the
Natural Sciences and Engineering Research Council of Canada

ABSTRACT

Improving road safety requires two steps of network screening and site diagnosis, which both require safety to be objectively quantified. In the screening phase, sites are identified and prioritized to maximize the efficiency of implemented countermeasures. Network screening methods commonly adopt regression techniques to estimate the expected number of crashes at sites across the network. Most existing techniques use crash-based ranking criteria which are subject to errors and omissions in collision databases, require long collection periods, and are reactive. GPS-enabled smartphones can collect reliable and spatio-temporally rich naturalistic driving data from regular drivers using an inexpensive, simple, and user-friendly tool that eliminates the need for external sensors. To date, very few studies have analyzed large volumes of smartphone GPS probe vehicle data or have considered advanced modelling techniques for screening in large road networks. The purpose of this paper is to develop a crash frequency model that incorporates surrogate safety measures (SSMs) extracted from the smartphones of regular drivers as predictive variables. After processing GPS data collected in Quebec City, Canada, several SSMs including vehicle manoeuvres (hard braking) and measures of traffic flow (congestion, average speed, and speed variation) were extracted. A Latent Gaussian Spatial Model was estimated using the INLA technique. Results showed that while negative binomial models outperformed Poisson models, the greatest improvement in model fit was achieved through a spatial model. In general, the relationships between SSMs and crash frequency established in previous studies were supported by the modelling results. Future work will include expanding the crash model to the entire Quebec City road network, comparing models estimated using INLA to those estimated using a traditional MCMC simulation, and incorporating collision severity estimation. The ability to screen the network based only on SSMs presents a substantial contribution to the field of road safety, and works towards the elimination of crash data in safety evaluation and monitoring.

INTRODUCTION

Improving road safety requires two principal steps of network screening and site diagnosis. Network screening is "the low-cost examination of all entities of a population" to identify a smaller subgroup for detailed investigation (diagnosis) (*1*). In road networks, this smaller subgroup (hotspots, blackspots, hazardous road locations, or high risk sites) includes locations where design or operation "create an increased risk of unforeseeable accidents" (*2*) with potential for crash reduction through engineering intervention (*3*). Considering parties involved in road safety management have limited budgets, sites should be identified and prioritized to maximize the efficiency of implemented countermeasures determined in the diagnosis phase. Network screening methods commonly use regression models, Bayesian statistics, or machine learning techniques to estimate the expected number of crashes at links or intersections throughout the road network (*4*) and uncover the risk factors contributing to crashes (*5*).

In both screening and diagnosis, safety must be objectively quantified. Most existing techniques use ranking criteria based on historical crash data (*6*) and establish relationships between attributes of traffic, geometry, environment, and driver (*7*) and collision risk (*8*). Though popular, crash-based methods are subject to errors in collision databases and are sensitive to underreporting (*9*). As traffic collisions are relatively rare events, long collection periods are required to accumulate the necessary volume of crash data for analysis (*10*). Therefore, traditional network screening cannot be carried out continuously, but is performed periodically (for example, every year) so crashes can accrue and databases can be updated. This highlights what is perhaps the most critical flaw of existing screening techniques. Crash-based methods are reactive, requiring collisions to occur before hazardous sites are identified and improvements are made (*2*). Lastly, crashes themselves are not perfect predictors of safety. For these reasons, alternative screening methods would be valuable for identifying high risk sites more quickly, more accurately, and with limited reliance on crash databases (*11*).

An alternative screening method requires an alternative data source from which risk can be constantly and systematically estimated across an entire road network. Instrumented, or probe, vehicles that act "as moving sensors, continuously feeding information about traffic conditions" (*12*) are among the only methods for collecting such data. Instrumented vehicles can assist in reducing dependence on crash data by supporting the development of screening methods based on surrogate safety measures (SSMs) rather than collision statistics. SSMs are non-crash measures that are physically and predictably related to crashes (*13*). With the proliferation of GPS-enabled smartphones, which themselves contain many of the same sensors used for instrumenting vehicles, large volumes of reliable and spatio-temporally rich naturalistic driving data can be collected from regular drivers (*14*) in crashes, near crashes, and under normal conditions (*15*, *16*). Smartphones are inexpensive and user-friendly, minimally impact behaviour, eliminate the need for external sensors (*17*, *18*), take advantage of widespread technology, and exploit existing communication infrastructure (*19*).

Despite advances to modelling techniques and data collection technologies, some areas of interest remain overlooked. Few studies have analyzed large volumes of GPS probe vehicle data collected from the smartphones of regular drivers. Even fewer have considered the link between GPS-derived SSMs and large volumes of historical crash data at the network level. No studies to date have considered advanced modelling techniques, including recent developments in Bayesian inference and spatial models, for screening in large road networks. The purpose of this paper is to develop a crash frequency model for an urban road network incorporating SSMs as predictive variables. Various SSMs, including vehicle manoeuvres and traffic flow characteristics, are first extracted. A full Bayesian Latent Gaussian Model is then developed using the R-INLA program. Finally, spatial autocorrelations are introduced in the model formulation.

LITERATURE REVIEW

Methods for modelling crash frequency have been varied in the existing literature, as demonstrated in a review completed by Lord and Mannering (*20*). The main approach has been to use statistical count models, primarily Poisson regression (*21*). Negative binomial (NB) or Poisson-Gamma models account for overdispersion (*22*) and

zero-inflated Poisson models address the overabundance of sites with zero observations (*21*). Lord, Washington, and Ivan (*22*) provide a comparison of these three primary model types. More recently, random effects, multivariate outcomes, and hierarchical structures have been incorporated in regression models (*20*). While regression techniques assume that coefficients take fixed values, Bayesian techniques assume that the coefficients are defined by a probability distribution (*23*). In Empirical Bayes (EB) models, the probability distribution is determined, in part, by using observed historical crash data (*24*). In Full Bayes (FB) techniques, the posterior distributions are determined by assuming a prior distribution and iteratively computing and updating the posterior marginal using a Monte Carlo Markov Chain (MCMC) simulation. Studies comparing EB and FB approaches (*6*, *25*) have shown their superiority to basic regression models. Miaou and Lord (*26*) compared EB and FB techniques, noting that EB estimates deviated from the FB estimates, and that those deviations could become significant for some data sets. FB techniques have been extended by incorporating temporal and spatial correlations in the model specification (*24*). Machine learning techniques have also been explored. Xie, Lord, and Zhang (*27*) showed that both a back-propagated neural network (BPNN) and Bayesian neural network (BNN) had better goodness-of-fit and prediction capabilities compared to a traditional NB model. Li et al. (*28*) focussed on assessing the predicting power of a Support Vector Machine (SVM), which was more accurate than an NB model.

Several studies have attempted to extract SSMs from probe vehicle data. Event-based techniques identify individual driver manoeuvres including steering, braking, or accelerating (*29*). Fazeen et al. (*30*) used smartphone accelerometer data to classify 'safe' accelerations and decelerations from 'unsafe' ones, though no evidence was provided demonstrating unsafe behaviour led to increased risk. Jun, Ogle, and Guensler (*14*) analyzed the relationship between spatio-temporal driving behavior and likelihood of crash involvement, finding that drivers involved in crashes tended to travel longer distances at higher speeds and "engaged in hard deceleration events" more frequently. Algerholm and Larhmann (*2*) correlated jerk and crash occurrence across drivers and sites (*2*). Using GPS, accelerometer, radar, and self-reported collision data, Bagdadi (*15*) proposed a jerk-based surrogate measure that correctly identified 86% of near misses. Traffic flow techniques use aggregate volume, speed, and density to measure risk (*31*). Though speed, flow, and variation in speed and flow have been suggested as potential SSMs in several studies (*7*), traffic flow SSMs have rarely been studied using GPS data. Speed profiles from GPS devices were considered by Moreno and Garcia (*32*) and Boonsiripant (*33*).

Despite previous work on crash frequency modelling and extracting SSMs from GPS data, several shortcomings remain. Although many methods for extracting and analyzing SSMs have been proposed, few studies have extracted SSMs from instrumented vehicles, and very few have extracted such measures from smartphone GPS data alone. In terms of crash modelling, although more complex models continue to improve estimates of road traffic crashes, very few studies to date have incorporated SSMs into statistical models of crash frequency. Finally, although Bayesian modelling is the most accurate and well-accepted approach for crash modelling, current estimation methods are computational expensive and time consuming. Despite recent advances in Bayesian inference, including the Integrated Nested Laplace Approximation (INLA) approach, no studies to date have applied Bayesian inference to the field of road safety. This paper aims to combine SSMs based on vehicle manoeuvres and traffic flow extracted from GPS smartphone data and network screening models to minimize the dependence on crash data in the network screening process.

METHODOLOGY

The methodology for this study consists of three steps. First, the process of collecting and processing the main data sources (GPS travel data, mapping data, and crash data) is discussed. Next, SSMs are extracted from the smartphone GPS data. Lastly, the crash frequency model is developed and calibrated based on a sample of data from an urban road network.

Data Collection and Processing

Smartphone GPS Data

Collecting and processing smartphone GPS data is covered in detail by Stipancic et al. (*34*) and illustrated in Figure 1. For each logged trip, i, GPS travel data is returned as a series of observations (O_{ij}), including the unique coordinate identifier (c_{ij}), datetime (t_{ij}), latitude (x_{ij}), longitude (y_{ij}), altitude (z_{ij}), and speed (v_{ij}). Time between consecutive observations is typically between 1 and 2 seconds. To reduce positional noise, the trip data is map matched to the OpenStreetMap (OSM) road network (*35*) using TrackMatching, a commercially available map matching service (*36*) that returns a new latitude and longitude, x'_{ij} and y'_{ij}, corresponding to a specific OSM link ID, l_{ij}, and the source, s_{ij}, and destination nodes, d_{ij}. Next, a Savitzky-Golay filter is applied to reduce noise in the GPS measured speeds. This filter is "a weighted moving average–based filter" suitable for time series' with fixed and uniform intervals and with limited discontinuities in the data (*37*). Optimal filter parameters were determined in a previous study (*38*). The filter is applied to speeds v_j to yield filtered speeds v'_{ij} and acceleration rate a_{ij} for every observation, yielding the final data structure below.

$$trip_i = \begin{Bmatrix} O_{i0} \\ O_{i1} \\ \vdots \\ O_{ij} \\ \vdots \\ O_{in_i} \end{Bmatrix} = \begin{Bmatrix} i,\ c_{i0},\ t_{i0},\ x_{i0},\ y_{i0},\ z_{i0},\ v_{i0} \\ i,\ c_{i1},\ t_{i1},\ x_{i1},\ y_{i1},\ z_{i1},\ v_{i1} \\ \vdots \\ i,\ c_{ij},\ t_{ij},\ x_{ij},\ y_{ij},\ z_{ij},\ v_{ij} \\ \vdots \\ i,\ c_{in_i},\ t_{in_i},\ x_{in_i},\ y_{in_i},\ z_{in_i},\ v_{in_i} \end{Bmatrix} \rightarrow \begin{Bmatrix} i,\ c_{i0},\ t_{i0},\ x'_{i1},\ y'_{i0},\ z_{i0},\ v'_{i0},\ a_{i0},\ l_{i0},\ s_{i0},\ d_{i0} \\ i,\ c_{i1},\ t_{i1},\ x'_{i1},\ y'_{i1},\ z_{i1},\ v'_{i1},\ a_{i1},\ l_{i1},\ s_{i1},\ d_{i1} \\ \vdots \\ i,\ c_{ij},\ t_{ij},\ x'_{ij},\ y'_{ij},\ z_{ij},\ v'_{ij},\ a_{ij},\ l_{ij},\ s_{ij},\ d_{ij} \\ \vdots \\ i,\ c_{in_i},\ t_{in_i},\ x'_{in_i},\ y'_{in_i},\ z_{in_i},\ v'_{in_i},\ a_{in_i},\ l_{in_i},\ s_{in_i},\ d_{in_i} \end{Bmatrix}$$

Network Map Data

Ideally, each network link should connect adjacent intersections (*39*). As the OSM road network is generated non-systematically by users, OSM links often connect multiple intersections, and should be redefined according to the following steps, also illustrated in Figure 2, which can be completed in any GIS software environment.

1. Identify all nodes that represent an intersection in the road network.
2. Split the road network at the identified nodes.
3. Rename each link according to its original ID and the nodes on either end of the link.
4. Remap the GPS observations to the new network.

Collision Data

Crash data must be assigned to links and intersections in the newly redefined road network to obtain crash counts for model calibration. Some collision reports may contain a latitude and longitude defining the location of the crash. However, for cases where coordinates are not provided, are provided inconsistently, or are inaccurate, a geocoding procedure may be used to determine the coordinates from text based fields, including address or intersection. This project used a geocoding procedure developed by Burns et al. (*40*). Even after geocoding, crashes may not fall directly onto the network. Buffers are used to assign crashes to individual links and intersections. Collision counts for links were generated by counting all crashes within a 50 m buffer around each link, while intersections utilized a 100 m buffer, based on results from previous studies (*38*).

Extraction of Surrogate Safety Measures

The extraction of SSMs from GPS smartphone data, and the strength of their relationships with crash frequency, is covered in detail in two previous studies (*38*, *41*). The SSMs incorporated into the network screening model are briefly defined below.

Decelerations and Accelerations

Deceleration is the most common evasive manoeuvre in urban areas (*2*). Though most studies have used accelerometers to calculate jerk as an SSM (*15*, *2*), and jerk cannot be calculated using GPS data alone, a simple deceleration threshold may be used to define hard breaking events (HBEs) (*30*), or conversely, hard acceleration events (HAEs) (*16*, *2*). Each a_{ij} is compared to a braking threshold, a_{min}, and acceleration threshold, a_{max}. The status of O_{ij} is determined using the following logic. For each series of consecutive negative (or conversely, positive) accelerations, l_{dec} (l_{acc}), the minimum (maximum) value is obtained. If this value is inferior (superior) to the threshold, $\min(l_{dec}) < a_{min}$ ($\max(l_{acc}) > a_{max}$), then consider that observation an HBE (HAE). This algorithm is illustrated in Figure 3. Based on previous work, a threshold of -2 m/s^2 was chosen for a_{min}. Total HBEs were divided by the number of trips on each link or intersection to obtain a rate of HBEs. HAEs were omitted from the model because they were highly correlated with HBEs.

Congestion Index

Though several congestion measures based on travel time have been proposed, link travel time is dependent on position which is altered during map matching. Therefore, a congestion measure based on speed is preferred. Dias et al. (*42*) proposed CI calculated as

$$\begin{aligned} CI &= \frac{\text{free flow speed} - \text{actual speed}}{\text{free flow speed}} && \text{if } CI > 0 \\ &= 0 && \text{if } CI \leq 0 \end{aligned} \tag{1}$$

This formulation yields CI values ranging from 0 (speed equal to the free flow speed) and 1 (speed is zero). As congestion is generally constrained to the peak periods, off-peak speeds can be used to estimate free flow speed. For each link, L_{ij}, the free flow speed, $FFS_{L_{ij}}$, is the average of all observed speeds falling on link L_{ij} outside of the peak periods (6:00 to 10:00 AM and 3:00 to 7:00 PM). CI for every observation is computed as

$$\begin{aligned} CI_{ij} &= \frac{FFS_{L_{ij}} - v'_{ij}}{FFS_{L_{ij}}} && \text{if } FFS_{L_{ij}} > v'_{ij} \\ &= 0 && \text{otherwise} \end{aligned} \tag{2}$$

where CI_{ij} is the congestion index for observation O_{ij}, $FFS_{L_{ij}}$ is the free flow speed on link L_{ij}, and v'_{ij} is the observed speed. Finally, CI was calculated for each link by taking the average of each observed CI_{ij}

$$CI_{L_{ij}} = \frac{\sum_i \sum_j CI_{ij}}{N} \tag{3}$$

where N is the count of all observations on link L_{ij}. As in previous work (*34*) filters were used to eliminate links with too few observations for a valid calculation (less than two trips with two observations per trip). Preliminary results indicated that CI during the PM peak had the strongest relationship with actual risk.

Average Speed

Travel speed has long been believed to be an indicator of crash risk at the link level. Average speed is calculated for each link as the average of the all smoothed speeds falling on that link, or

$$\bar{V}_{L_{ij}} = \frac{\sum_i \sum_j v'_{ij}}{N} \tag{4}$$

where N is the count of those observations on link L_{ij}. The off-peak period was chosen for calculation to avoid collinearity with CI (in this case, $\bar{V}_{L_{ij}}$ is exactly equal to $FFS_{L_{ij}}$ defined above).

Speed Uniformity

Although the magnitude of speed is widely believed to contribute to crash occurrence, much of the existing literature supports that variation in speed may be a better predictor of risk. The coefficient of variation of speed (CVS) was used in this study. For each link, L_{ij}, CVS is computed as

$$CVS_{L_{ij}} = \frac{\sigma(v'_{ij})}{\bar{V}_{L_{ij}}} \tag{5}$$

where $\sigma(v'_{ij})$ is the standard deviation of all speeds on link L_{ij} during the considered time period. CVS was found to be most strongly related to crash frequency during the off-peak period for this study.

Crash Frequency Model

Latent Gaussian Models

Latent Gaussian Models (LGM) are a subclass of structured additive models, in which the response variable y_i for each subject i is assumed to belong to an exponential family (Normal, Poisson, or Binomial distribution). In this case, y_i represents the crash frequency (number of crashes at link or intersection i) and the mean of y_i, noted μ_i , is related to the predictors through a link function $g()$ such that

$$g(\mu_i) = \eta_i = \beta_0 + \sum_{k=1}^{n_\beta} \beta_k z_{ki} + \sum_{j=1}^{n_f} f^{(j)}(u_{ji}) + \epsilon_i \tag{6}$$

where β_0 is the intercept, β_k are the coefficients representing the linear effect of covariates z_{ki}, $f^{(j)}$ are functions used to relax these linear relationships or introduce random effects, and ϵ_i is the unstructured error component (*43*). LGM are an extremely flexible family of models because of the forms that $f^{(j)}$ can take, such as introducing temporal or spatial dependence leading to dynamic or spatial models (*23*). The predictive structure, η_i, is called a structured additive predictor. LGM are a subset of all Bayesian structured additive models in which the prior distributions of all parameters β_0, β_k, $f^{(j)}$, and ϵ_i are assumed to be Gaussian (normally distributed). This set of the parameters define the latent Gaussian field.

From a likelihood perspective, LGM can be represented as a three-stage hierarchical structure, beginning with the conditionally independent likelihood function

$$\pi(y|x,\theta) = \prod_{i=1}^{n} \pi(y_i|\eta_i(x),\theta) \tag{7}$$

where y is the response vector, x is the latent field described above, θ is the vector of hyperparameters, and $\eta_i(x)$ is the ith additive predictor. Next, the latent Gaussian field is formally defined with a mean $\mu(\theta)$ and precision matrix $Q^{-1}(\theta)$ conditioned on hyperparameters θ:

$$x|\theta \sim N(\mu(\theta), Q^{-1}(\theta)) \tag{8}$$

Finally, a prior distribution is assigned to the hyperparameters:

$$\theta \sim \pi(\theta) \tag{9}$$

Although the prior distributions of the latent field must be Gaussian by definition, the prior distributions of the hyperparameters are not subject to this constraint. For more details on LGMs, readers are referred to Rue, Martino, and Chopin (*43*).

Integrated Nested Laplace Approximation

The greatest challenge with FB models is estimating the posterior distributions of the latent field and hyperparameters. Traditionally, MCMC simulations have been used to iteratively compute and update the posterior marginal of the latent field. However, this process is time and resource intensive (*43*). The INLA approach was proposed by Rue, Martino, and Chopin (*43*) to perform Bayesian approximations on LGMs. The INLA approach uses a combination of Laplace approximations and numerical integration to estimate the posterior marginal of the latent field, and offers accurate approximations with a significant reduction in computation time (*44*). The authors demonstrated that INLA provides precise estimates in seconds or minutes where MCMC computations would take hours or days. Furthermore, the authors note that the approximation error by INLA is nearly equal to the estimation error in typical MCMC methods (*43*). A package for programming language R has been developed to deploy INLA (R-INLA) in a useable and stable format (*44*).

Model Calibration

Calibration of the crash prediction model was completed in several steps, based on a subset of data from the larger road network. Using R-INLA, Poisson and NB Bayesian models were estimated at both the link and intersection levels. These models use observed crash counts as the outcome y, and the extracted SSMs, number of GPS trips, and the functional classification as independent variables. In the link level model, the link length is specified as an offset. Previous research highlights the need to account for spatial correlations, incorporated in the second stage of the hierarchical LGM structure, to account for similarities of adjacent links (*23*) due to the "effect of unknown confounders" (*45*). Failure to do so may lead to model biases (*45*, *23*). Due to the added complexity of the spatial component, the INLA approach "is particularly suitable in this context" (*23*). Spatial correlation was incorporated in the best performing Poisson or NB model based on Deviance Information Criteria (DIC) to create a third model for both links and intersections. The spatial component was accounted for using a Besag–York–Mollié (BYM) model, which combines an independent and identically distributed (IID) random term and an intrinsic conditional autoregressive (iCAR) (*23*). Such a model is typically referred to as an ecological regression model (*23*). A graph generated using Python scripts describes the network topology. Neighbours are defined as links or intersections which are immediately adjacent to one another. This third model is calibrated and compared to the non-spatial models to demonstrate the benefits of the spatial component. For more detail on the BYM model, readers are referred to Besag, York, and Mollié (*46*).

RESULTS

Data Description

The GPS data utilized for this study was collected using the Mon Trajet application (*47*), originally developed by Brisk Synergies (*48*) for the City of Quebec, Canada. The application was installed voluntarily by drivers who anonymously logged their trips using a simple interface, shown in Figure 4. This study made use of a sample of open data, which contained over 4000 drivers and nearly 22,000 individual trips recorded between April 28 and May 18, 2014. 11 years of crash data were obtained from the Ministry of Transportation Quebec (MTQ) for the period between 2000 and 2010. In total, 14,278 collisions involving at least one vehicle were identified.

Data Exploration

A subset of the road network near Laval University, shown in Figure 5, was used for model calibration. This area was selected for its road density, diversity of roadway functional class, and average trip volumes. At the link level, the calibration data set contained 83 motorway links, 225 arterial and collector links (grouped from the primary, secondary, and tertiary classifications within OSM), and 121 residential links, for 429 links in total. In terms of intersections, 114 were classified as motorway, 167 as arterial/collector, and 253 as residential, for 534

total intersections. Descriptive statistics for the model variables is provided in Table 1. From this table an average link experienced nearly 1 crash per year, though the maximum value was 10 crashes per year. 0.6 crashes per year occurred at an average intersection, up to 7 crashes at the most extreme site. The number of trips on each facility varied greatly, from a minimum of 4 to a maximum of several thousand. About one in eight trips along a link experienced an HBE, while for intersections, the number was one in five. Congestion tended to be higher along links, while CVS was higher at intersections. Average travel speed was approximately 43 km/h.

Model Calibration

Three models were estimated for both the link and intersections levels, resulting in six total models. First, the Poisson and NB Bayesian models were estimated using INLA and compared. Next, the BYM spatial model was added to the best performing model, and compared to determine superiority.

Poisson and Negative Binomial Models

Link level results for the Poisson and NB models are presented in Table 2. The NB model outperformed the Poisson by DIC, but had a worse goodness-of-fit by mean-square-error (MSE). Goodness-of-fit is further illustrated in Figure 6a and Figure 6c, where fitted values are plotted against observed crashes. It was observed that the NB model improves fitted values for sites at either extreme (highest and lowest crash counts) by accounting for overdispersion in the model formulation. Considering the posterior means of the covariates, results generally supported the relationships between SSMs and crash frequency established in previous studies (*38*, *41*). The posterior mean for number trips was positive, representing the effect of increasing exposure (volume) on increasing crash risk. The mean for the braking variable was also positive, and HBEs are generally related to increased crash frequency. For the traffic flow SSMs, the posterior means for both CI and CVS were positive (increasing congestion and speed variation generally increases crash frequency), while for speed, the mean was negative (facilities with higher average speeds tend to have fewer crashes), in agreement with previous results (*41*). For functional classification, all else being equal, motorways are less likely to have crashes than the reference category of residential streets (negative posterior mean), while arterials and collectors are more likely (positive posterior mean). Most variables are significant at 95 % confidence in both models.

Table 3 contains the results for the intersection level Poisson and NB models. The fit of these models is also illustrated in Figure 6b and Figure 6d. As with the link level model, the NB was superior by DIC though it had a higher MSE. Results for the intersection model were similar to the link model, except for the traffic flow SSMs. The posterior means for CI and CVS were negative in both intersection models, in contrast to earlier results. However, these variables were not significant at 95% confidence. This is likely because traffic flow measures actually occur along the link and must be aggregated to the intersection level for analysis. Therefore, traffic flow SSMs are better suited for modelling crashes at links.

Spatial Models

The NB model was the best performing for both links and intersections. Therefore, the spatial component was incorporated into the NB model formulation to yield NB BYM models. Results of the spatial NB BYM models are summarized in Table 4. By DIC, the spatial models were the best performing of all considered model types, and improved the goodness-of-fit compared to the non-spatial NB models (for the links level model, the MSE is improved drastically, as shown Figure 6e and Figure 6f). This result supports previous work (*24*) indicating the importance of including spatial correlations in crash frequency models.

Although these models were superior and improved goodness-of-fit, many more variables were observed to be non-significant in the BYM models. Much of the observed variation in the crash count is explained by the spatial autocorrelation present between adjacent links and intersections. The posterior means for variables trips and

HBEs/Trip remained positive, but are not consistently significant at 95 % confidence. The mean for CI remained positive in the link-level model and negative in the intersection-level model, but was no longer significant. CVS also became insignificant at 95 % confidence. The posterior mean for average speed remained approximately the same, but became insignificant in the intersection-level model. Although the insignificance of these variables is discouraging, they were not discarded for two reasons. First, models were estimated on a relatively small sample of the road network. Expanding the model to the entire network may reveal the true impact of these variables. Also, the 95 % confidence level is quite restrictive considering these are the first models of their kind.

From plots of fitted values against observed number of crashes in Figure 6, the effect of the spatial model on model fit was clearly observed. Although the NB models provided a small improvement in fit compared to the Poisson models, the NB BYM model provided a dramatic improvement, especially for the link-level model, where the observations fall almost directly on the ideal diagonal line. The reason for this relatively higher performance at the link level may be that adjacent links truly are adjacent (they are directly connected at the intersections) while adjacent intersections are separated by 150 m on average. Therefore, the spatial correlations may be much stronger for links than for intersections.

CONCLUSIONS

The purpose of this paper was to propose a method for modelling crash frequency using a Latent Gaussian Spatial Model estimated using the INLA technique, which uses GPS data, GPS-derived SSMs, and roadway functional class as predictive variables. The data set was an open sample of GPS data collected in Quebec City and was processed to reduce signal noise. Extracted SSMs were related to vehicle manoeuvres (hard breaking) or measures of traffic flow (congestion, average speed, and speed variation). A subset of the road network near Laval University was used to calibrate the statistical models, which were estimated on 11 years of crash data.

NB models outperformed Poisson models at both the link and intersection level, by improving fitted values for sites at both extremes of crash frequency. In general, the relationships between SSMs and crash frequency established in previous studies were supported by the modelling results. The positive posterior mean for number of trips and HBEs indicates that crashes are more likely at links and intersections with more trips (higher exposure) and more cases of hard braking. For the considered traffic flow SSMs, the direction of the effect at the link level supported previous research, while results at the intersection level either did not support previous work or were not statistically significant. This is likely because traffic flow measures occur along the link and must be aggregated to the intersection level. All else being equal, motorways experience fewer crashes than residential streets, while arterials and collectors experience a greater number of collisions.

The greatest improvement in model fit was achieved through a spatial model. By DIC, the NB BYM spatial models were the best performing of all considered model types, and improved the goodness-of-fit, particularly for the link level model. The reason for this relatively higher performance at the link level may be that adjacent links truly are adjacent and may have much stronger spatial correlations than intersections. This result supports previous work indicating the importance of including spatial correlations in crash frequency models. Many variables were observed to be non-significant in the NB BYM models. Much of the observed variation in the crash count is explained by the spatial autocorrelation present between adjacent links and intersections.

Future work will involve expanding the crash model to the entire Quebec City road network, which will reveal the true effect and significance of the proposed covariates. Running times and model estimates will be compared between the INLA approach and a traditional MCMC simulation. Collision frequency and severity are two independent dimensions of risk, and severity must also be considered. As INLA is currently incapable of estimating multivariate models, severity will be incorporated in a two-step model, combining an LGM with a discrete choice component. Additional model complexity, including interaction variables to increase flexibility, could be considered for future implementation. The ability to screen the network accurately based only on SSMs presents a substantial contribution to the field of road safety, and importantly works towards the elimination of crash data as a necessity in safety evaluation and monitoring.

REFERENCES

1. Hauer, E., J. Kononov, B. Allery, and M. S. Griffith. Screening the Road Network for Sites with Promise. *Transportation Research Record*, no. 1784, 2002, pp. 27-32.
2. Algerholm, N., and H. Lahrmann. Identification of Hazardous Road Locations on the basis of Floating Car Dat. *Road safety in a globalised and more sustainable world*, 2012.
3. Aguero-Valverde, J., and P. P. Jovanis. Bayesian Multivariate Poisson Lognormal Models for Crash Severity Modeling and Site Ranking. *Transportation Research Record*, no. 2136, 2009, pp. 82-91.
4. Park, P. Y., and R. Sahaji. Safety network screening for municipalities with incomplete traffic volume data. *Accident Analysis and Prevention*, no. 50, 2013, pp. 1062-1072.
5. Chang, L.-Y., and H.-W. Wang. Analysis of traffic injury severity: An application of non-parametric classification tree techniques. *Accident Analysis and Prevention*, no. 38, 2006, pp. 1019-1027.
6. Huang, H., H. C. Chin, and M.M. Haque. Empirical Evaluation of Alternative Approaches in Identifying Crash Hot Spots: Naive Ranking, Empirical Bayes, and Full Bayes Methods. *Transportation Research Record*, no. 2103, 2009, pp. 32-41.
7. Abdel-Aty, M., and A. Pande. Identifying crash propensity using specific traffic speed conditions. *Journal of Safety Research*, no. 36, 2005, pp. 97-108.
8. Lu, M. Modelling the effects of road traffic safety measures. *Accident Analysis and Prevention*, no. 38, 2007, pp. 507-517.
9. Kockelman, K. M., and Y.-J. Kweon. Driver injury severity: an application of ordered probit models. *Accident Analysis and Prevention*, Vol. 34, 2002, pp. 313-321.
10. Lee, C., B. Hellinga, and K. Ozbay. Quantifying effects of ramp metering on freeway safety. *Accident Anaysis and Prevention*, no. 38, 2006, pp. 279-288.
11. Cafiso, S., and A. Di Graziano. Surrogate Safety Measures for Optimizing Investments in Local Rural Road Networks. *Transportation Research Record*, no. 2237, 2011, pp. 20-30.
12. El Faouzi, N.-E., H. Leung, and A. Kurian. Data fusion in intelligent transportation systems: Progress and challenges – A survey. *Information Fusion*, no. 12, 2011, pp. 4-19.
13. Tarko, A., G. Davis, N. Saunier, T. Sayed, and S. Washington. Surrogate Measures of Safety. Transportation Research Board, 2009.
14. Jun, J., J. Ogle, and R. Guensler. Relationships between Crash Involvement and Temporal-Spatial Driving Behavior Activity Patterns Using GPS Instrumented Vehicle Data. in *Transportation Research Board Annual Meeting*, Washington, DC, 2007.
15. Bagdadi, O. Assessing safety critical braking events in naturalistic driving studies. *Transportation Research Part F*, no. 16, pp. 117-126.
16. Wu, K.-F., and P. P. Jovanis. Defining and screening crash surrogate events using naturalistic driving data. *Accident Analysis and Prevention*, no. 61, 2013, pp. 10-22.
17. Eren, H., S. Makinist, E. Akin, and a. Yilmaz. Estimating Driving Behavior by a Smartphone. in *2012 Intelligent Vehicles Symposium*, Alcalá de Henares, Spain, 2012, pp. 234-239.
18. Johnson, D. A., and M. M. Trivedi. Driving Style Recognition Using a Smartphone as a Sensor Platform. in *2011 14th International IEEE Conference on Intelligent Transportation Systems*, Washington, DC, 2011, pp. 1609-1615.
19. Herrera, J. C., D. B. Work, R. Herring, X. Ban, Q. Jacobson, and A. M. Bayen. Evaluation of traffic data obtained via GPS-enabled mobile phones: The Mobile Century field experiment. *Transportation Research Part C*, no. 18, 2010, pp. 568-583.
20. Lord, D., and F. Mannering. The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives. *Transportation Research Part A*, Vol. 44, no. 5, 2010, pp. 291-305.

21. Mannering, F. L., and C. R. Bhat. Analytic methods in accident research: Methodological frontier and future directions. *Analytic Methods in Accident Research*, 2014, pp. 1-22.
22. Lord, D., S. P. Washington, and J. H. Ivan. Poisson, Poisson-gamma and zero-inflated regression models of motor vehicle crashes: balancing statistical fit and theory. *Accident Analysis and Prevention*, no. 37, 2005, pp. 35-46.
23. Biangiardo, M., and M. Cameletti. *Spatial and Spatio-temporal Bayesian Models with R-INLA*. John Wiley & Sons, Ltd, 2015.
24. Jiang, X., M. Abdel-Aty, and S. Alamili. Application of Poisson random effect models for highway network screening. *Accident Analysis and Prevention*, no. 63, 2014, pp. 74-82.
25. Persuad, B., B. Lan, C. Lyon, and R. Bhim. Comparison of empirical Bayes and full Bayes approaches for before–after road safety evaluations. *Accident Analysis and Prevention*, no. 32, 2010, pp. 38-43.
26. Miaou, S.-P., and D. Lord. Modeling Traffic Crash–Flow Relationships for Intersections: Dispersion Parameter, Functional Form, and Bayes Versus Empirical Bayes Methods. *Transportation Research Record: Journal of the Transportation Research Board*, no. 1840, 2003, pp. 31-40.
27. Xie, Y., D. Lord, and Y. Zhang. Predicting motor vehicle collisions using Bayesian neural network models: An empirical analysis. *Accident Analysis and Prevention*, no. 39, 2007, pp. 922-933.
28. Li, X., D. Lord, Y. Zhang, and Y. Xie. Predicting motor vehicle crashes using Support Vector Machine models. *Accident Analysis and Prevention*, no. 40, 2008, pp. 1611-1618.
29. Dingus, T. A., S. G. Klauer, V. L. Neale, A. Petersen, S. E. Lee, J. Sudweeks, M. A. Perez, J. Hankey, D. Ramsey, S. Gupta, C. Bucher, Z. R. Doerzaph, J. Jermeland, and R. R. Knipling. The 100-Car Naturalistic Driving Study, Phase II – Results of the 100-Car Field Experiment. NHTSA, Washington, DC, DOT HS 810 593, 2006.
30. Fazeen, M., B. Gozick, R. Dantu, M. Bhukhiya, and M. C. Gonzalez. Safe Driving Using Mobile Phones. *IEEE Transactions on Intelligent Transportation Systems*, Vol. 13, no. 3, 2012, pp. 1462-1468.
31. Yan, X., M. Abdel-Aty, E. Radwan, X. Wang, and P. Chilakapati. Validating a driving simuator using surrogate safety measures. *Accident Analysis and Prevention*, no. 40, 2008, pp. 274-288.
32. Moreno, A.T., and A. Garcia. Use of speed profile as surrogate measure: Effect of traffic calming devices oncrosstown road safety performance. *Accident Analysis and Prevention*, no. 61, 2013, pp. 23-32.
33. Boonsiripant, S. Speed profile variation as a surrogate measure of road safety based on GPS-equipped vehicle data. Georgia Institute of Technology, PhD Thesis 2009.
34. Stipancic, J., L. Miranda-Moreno, A. Labbe, and N. Saunier. Measuring and Visualizing Space-Time Congestion Patters in an Urban Road Network Using Large-Scale Smartphone-Collected GPS Data. *Transportation Letters*, in press.
35. OpenStreetMap. About. *OpenStreetMap*, 2015. http://www.openstreetmap.org/about. Accessed May 11, 2015.
36. Marchal, F. TrackMatching. 2015. https://mapmatching.3scale.net/. Accessed May 1, 2015.
37. Zaki, M. H., T. Sayed, and K. Shaaban. Use of Drivers' Jerk Profiles in Computer Vision–Based Traffic Safety Evaluations. *Transportation Research Record: Journal of the Transportation Research Board*, no. 2434, 2014, pp. 103-112.
38. Stipancic, J., L. Miranda-Moreno, and N. Saunier. Vehicle Manoeuvres as Surrogate Safety Measures: Extracting Data From the GPS-Enables Smartphones of Regular Drivers. *Accident Analysis and Prevention*, in press.
39. Sioui, L., and C. Morency. Building congestion indexes from GPS data : Demonstration. in *13th WCTR*, Rio de Janeiro, 2013.
40. Burns, S., L. Miranda-Moreno, J. Stipancic, N. Saunier, and K. Ismail. Accessible and Practical Geocoding Method for Traffic Collision Record Mapping. *Transportation Research Record*, no. 2460, 2014, pp. 39-46.

41. Stipancic, J., L. Miranda-Moreno, and N. Saunier. The Impact of Congestion and Traffic Flow on Crash Frequency and Severity: An Application of Smartphone-Collected GPS Travel Data. *Transportation Research Record: Journal of the Transportation Research Board*, in press.
42. Dias, C., M. Miska, M. Kuwahara, and H. Warita. Relationship between congestion and traffic accidents on expressways: an investigation with Bayesian belief networks. in *Proceedings of 40th Annual Meeting of Infrastructure Planning (JSCE)*, Japan, 2009.
43. Rue, H., S. Martino, and N. Chopin. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society*, no. 71, 2009, pp. 319-392.
44. Schrodle, B., and L. Held. A primer on disease mapping and ecological regression using INLA. *Computation Statistics*, no. 26, 2011, pp. 241-258.
45. Latouche, A., C. Guihenneuc-Jouyaux, C. Girard, and D. Hémon. Robustness of the BYM model in absence of spatial variation in the residuals. *International Journal of Health Geographics*, Vol. 6, no. 39, 2007.
46. Besag, J., J. York, and A. Mollie. Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, no. 43, 1991, pp. 1-59.
47. City of Quebec. Mon Trajet. *City of Quebec*, http://www.ville.quebec.qc.ca/citoyens/deplacements/mon_trajet.aspx. Accessed May 13, 2015.
48. Brisk Synergies. *Brisk Synergies*, http://www.brisksynergies.com/. Accessed July 22, 2015.

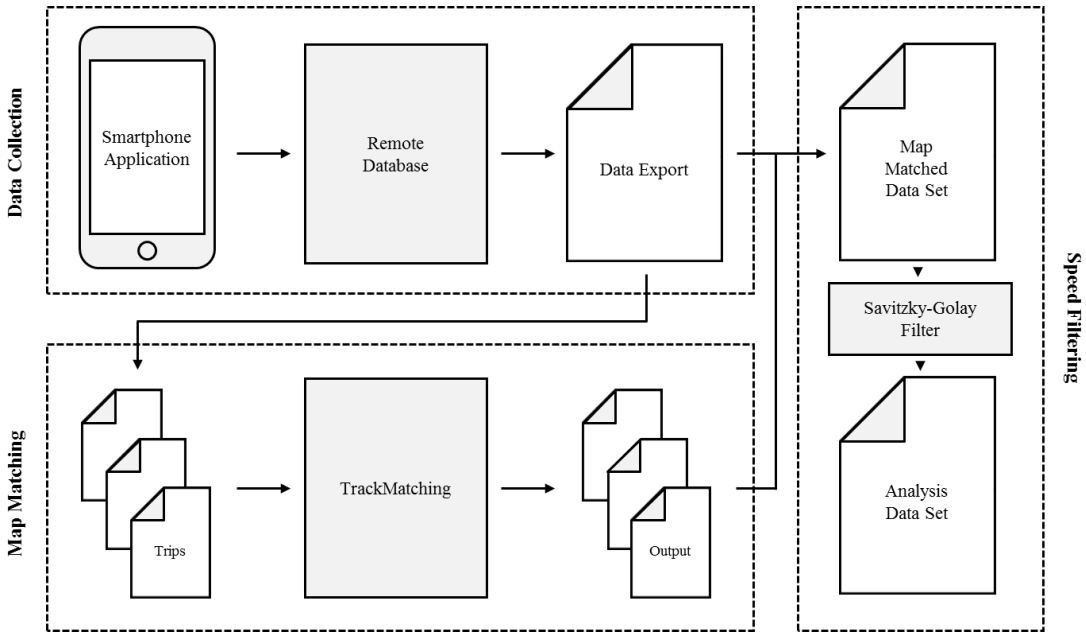

FIGURE 1 Collection and processing of smartphone-collected GPS data

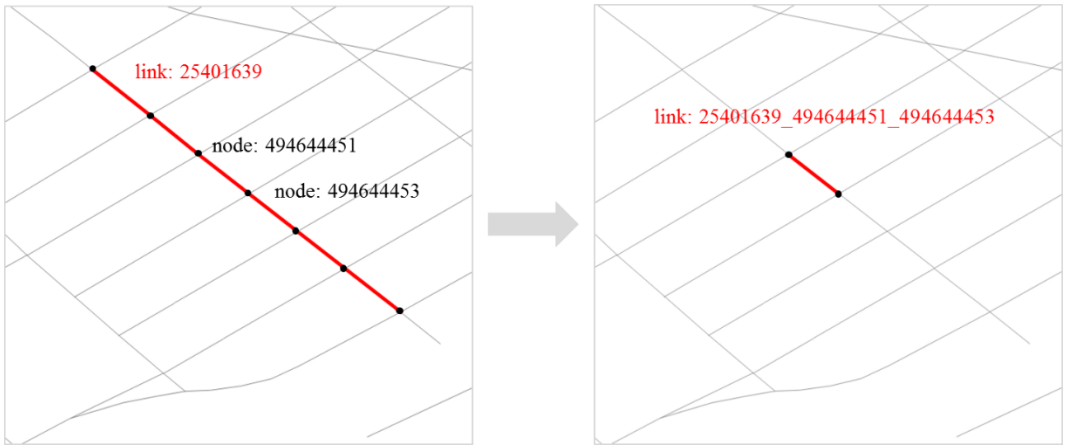

FIGURE 2 Redefinition of OSM Links

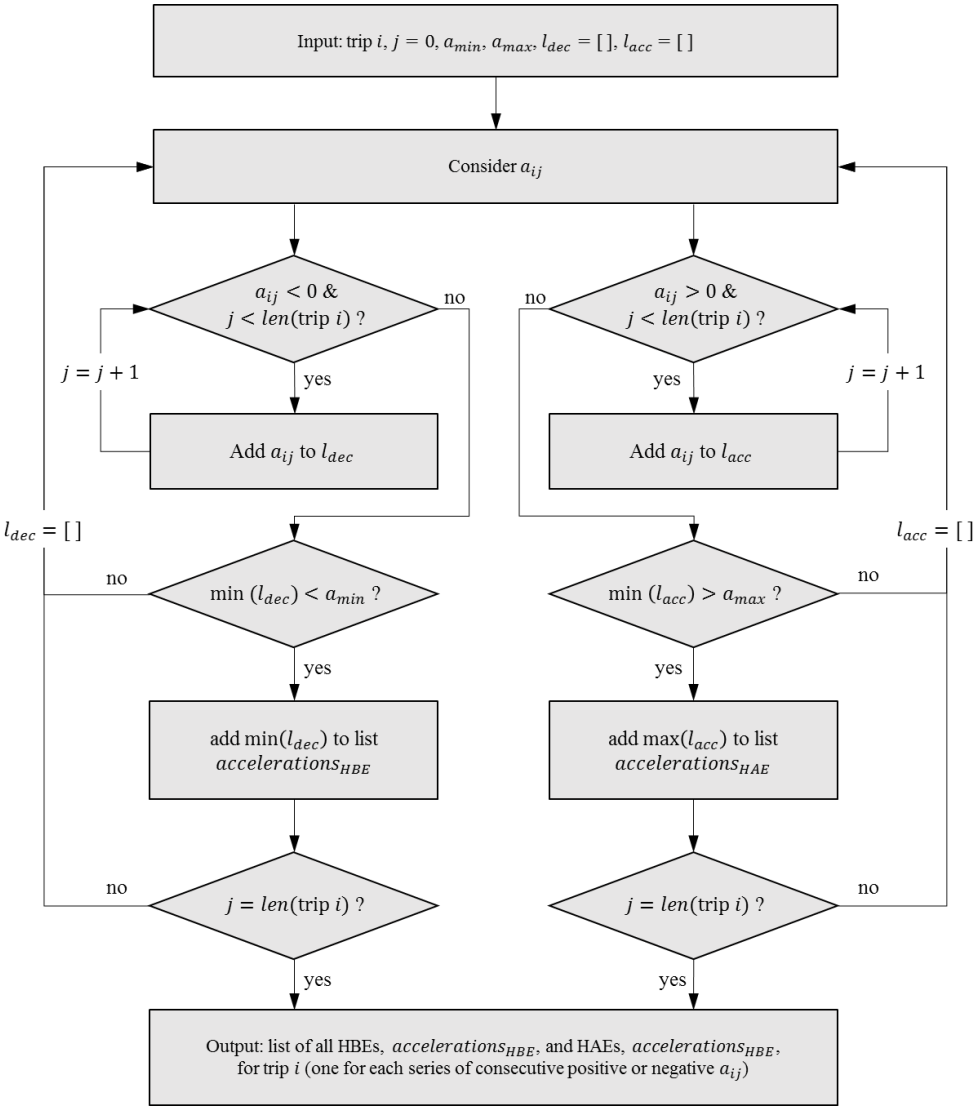

FIGURE 3 Algorithm for extracting vehicle manoeuvres from GPS trip data

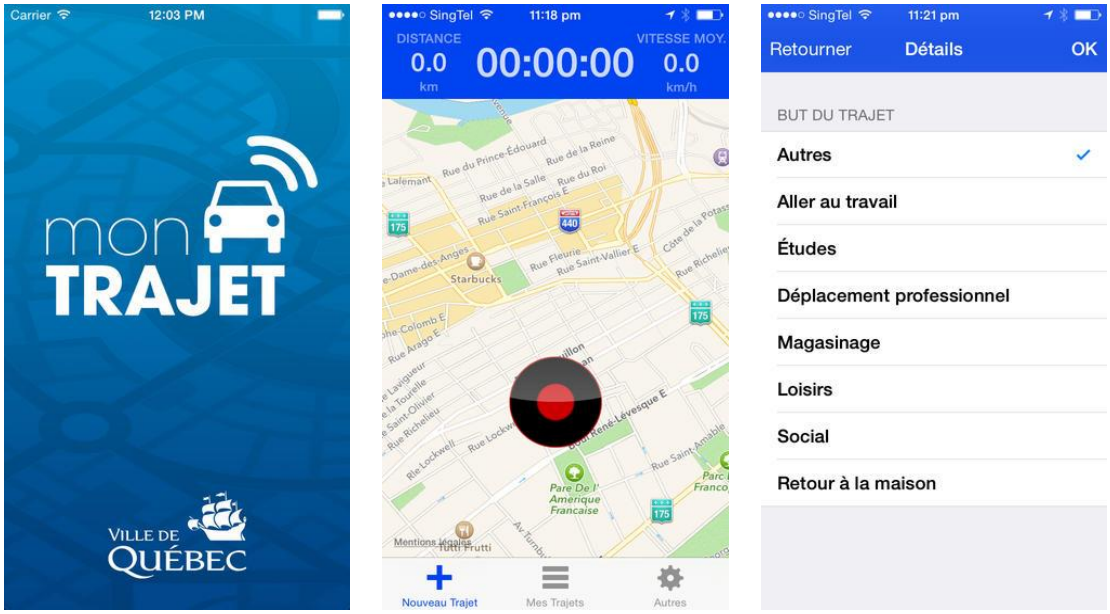

FIGURE 4 Smartphone application interfaces

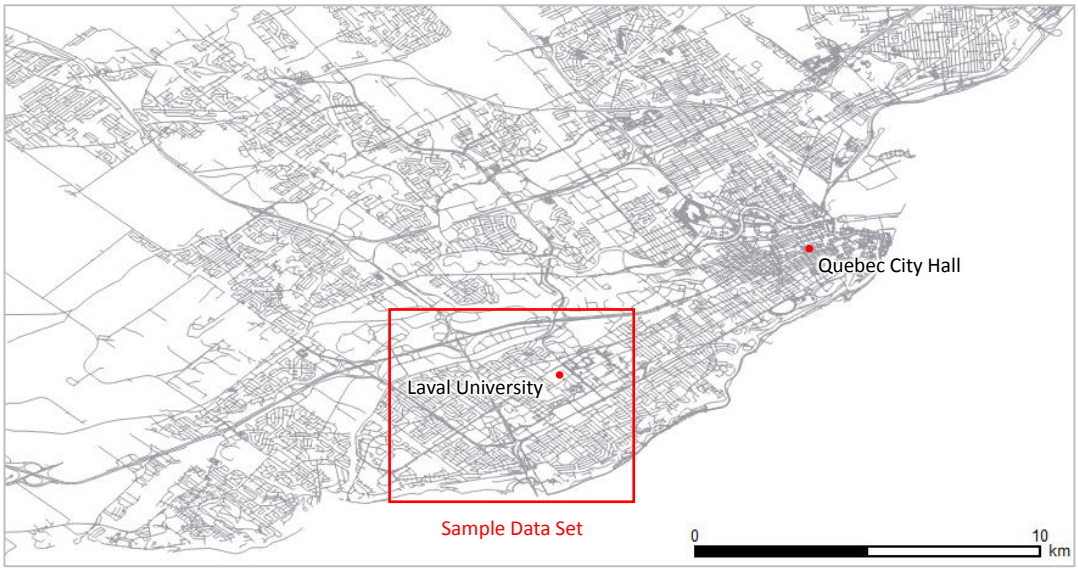

FIGURE 5 Map of study location

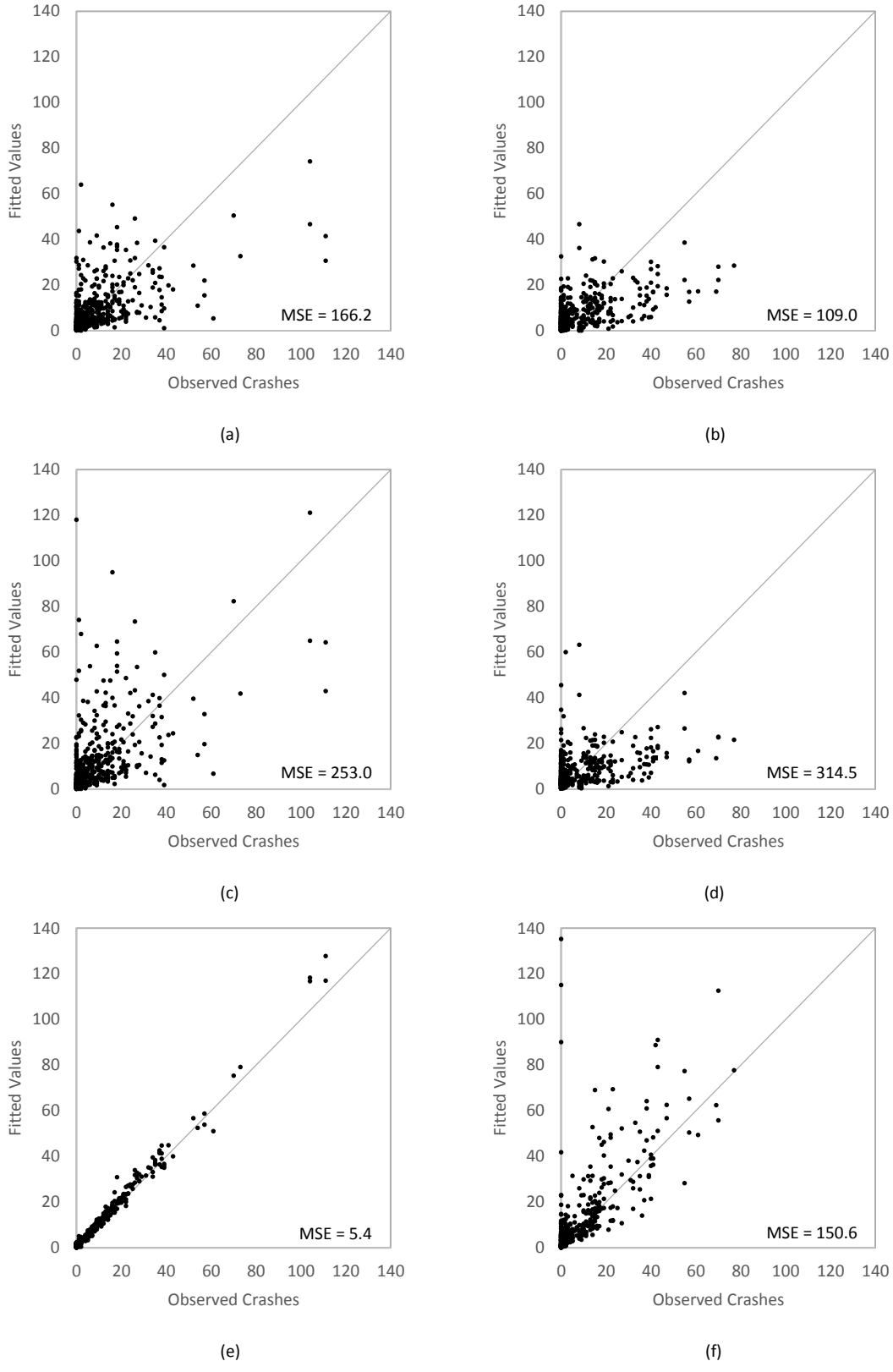

FIGURE 6 Fitted values versus observed crashes for Poisson links (a) and intersections (b), NB links (c) and intersections (d) and NB BYM links (e) and intersections (f)

TABLE 1 Variables and Descriptive Statistics for the Calibration Data Set

	Units	Mean	Minimum	Maximum	Std. Dev.
Links					
Crashes	-	10.16	0.00	111.00	15.07
Length	m	150.55	4.49	1046.34	131.64
Trips	-	205.07	4.00	2023.00	266.51
HBEs/Trip	-	0.12	0.00	1.50	0.18
Congestion Index	-	0.18	0.00	0.78	0.15
CVS	-	0.28	0.01	0.80	0.14
Average Speed	m/s	13.09	3.90	29.84	5.46
Intersections					
Crashes	-	7.08	0.00	77.00	12.56
Trips	-	393.99	4.00	3718.00	424.27
HBEs/Trip	-	0.20	0.00	1.78	0.24
Congestion Index	-	0.16	0.00	0.78	0.15
CVS	-	0.39	0.08	1.04	0.15
Average Speed	m/s	11.07	2.51	27.70	4.48

TABLE 2 Link Model Results for Poisson and Negative Binomial Models

	Poisson				NB			
Explanatory variables	**mean**	**std dev**	**95% CI**		**mean**	**std dev**	**95% CI**	
Intercept	-2.103	0.13	-2.352	-1.854	-2.045	0.44	-2.894	-1.186
Trips	0.001	0.00	0.001	0.001	0.002	0.00	0.001	0.003
HBEs/Trip	0.373	0.09	0.193	0.550	0.156	0.45	-0.681	1.073
Congestion Index	0.897	0.12	0.658	1.135	1.279	0.48	0.358	2.226
CVS	0.492	0.16	0.168	0.814	1.213	0.64	-0.029	2.466
Average Speed	-0.154	0.01	-0.171	-0.138	-0.178	0.03	-0.235	-0.121
Motorway	-0.111	0.09	-0.294	0.069	-0.107	0.32	-0.730	0.540
Arterial/Collector	1.269	0.05	1.176	1.363	1.284	0.16	0.973	1.592
Number of cases		429				429		
DIC		5711.1				2658.4		
MSE		166.2				253.0		

TABLE 3 Intersection Model Results for Poisson and Negative Binomial Models

	Poisson				NB			
Explanatory variables	**mean**	**std dev**	**95% CI**		**mean**	**std dev**	**95% CI**	
Intercept	3.593	0.16	3.285	3.901	1.816	0.60	0.637	2.997
Trips	0.001	0.00	0.001	0.002	0.002	0.00	0.001	0.002
HBEs/Trip	0.403	0.07	0.257	0.546	0.836	0.48	-0.065	1.833
Congestion Index	-1.511	0.17	-1.852	-1.171	-0.852	0.76	-2.312	0.666
CVS	-0.412	0.22	-0.842	0.017	0.745	0.98	-1.182	2.683
Average Speed	-0.234	0.01	-0.255	-0.214	-0.125	0.03	-0.193	-0.058
Motorway	-0.392	0.10	-0.580	-0.206	-0.474	0.34	-1.129	0.208
Arterial/Collector	0.949	0.04	0.876	1.024	0.870	0.19	0.502	1.242
Number of cases		534				534		
DIC		6730.5				2677.3		
MSE		109.0				314.5		

TABLE 4 Negative Binomial BYM Model Results for Links and Intersections

	Links				Intersections			
Explanatory variables	**mean**	**std dev**	**95% CI**		**mean**	**std dev**	**95% CI**	
Intercept	-1.474	0.49	-2.443	-0.506	-0.082	0.47	-1.002	0.838
Trips	0.000	0.00	-0.001	0.002	0.001	0.00	0.000	0.001
HBEs/Trip	0.844	0.50	-0.122	1.823	1.283	0.52	0.267	2.311
Congestion Index	0.725	0.72	-0.691	2.126	-0.671	1.10	-2.843	1.485
CVS	-0.184	0.65	-1.461	1.085	0.455	0.85	-1.205	2.112
Average Speed	-0.115	0.03	-0.172	-0.058	-0.053	0.03	-0.116	0.009
Motorway	-1.066	0.47	-1.989	-0.155	-0.741	0.49	-1.703	0.211
Arterial/Collector	0.452	0.26	-0.057	0.957	0.854	0.25	0.376	1.339
Number of cases		429				534		
DIC		2374.7				2570.0		
MSE		5.4				150.6		