Accident Analysis and Prevention 120 (2018) 174–187

Contents lists available at ScienceDirect

Accident Analysis and Prevention

journal homepage: www.elsevier.com/locate/aap

Surrogate safety and network screening: Modelling crash frequency using GPS travel data and latent Gaussian Spatial Models

Joshua Stipancic[a,*], Luis Miranda-Moreno[b], Nicolas Saunier[c], Aurelie Labbe[d]

[a] Department of Civil Engineering and Applied Mechanics, McGill University, Room 391, Macdonald Engineering Building, 817 Sherbrooke Street West, Montréal, Québec, H3A 0C3, Canada
[b] Department of Civil Engineering and Applied Mechanics, McGill University, Room 268, Macdonald Engineering Building, 817 Sherbrooke Street West, Montréal, Québec, H3A 0C3, Canada
[c] Department of Civil, Geological and Mining Engineering, Polytechnique Montréal, C.P. 6079, Succ. Centre-Ville, Montréal, Québec, H3C 3A7, Canada
[d] Department of Decision Sciences, HEC Montréal, 3000 Chemin de la Côte-Sainte-Catherine, Montréal, Québec, H3T 2A7, Canada

ARTICLE INFO

Keywords:
Crash frequency
GPS
Spatial
Full Bayes
Latent Gaussian Model
INLA

ABSTRACT

Improving road safety requires accurate network screening methods to identify and prioritize sites in order to maximize the effectiveness of implemented countermeasures. In screening, hotspots are commonly identified using statistical models and ranking criteria derived from observed crash data. However, collision databases are subject to errors, omissions, and underreporting. More importantly, crash-based methods are reactive and require years of crash data. With the arrival of new technologies including Global Positioning System (GPS) trajectory data, proactive surrogate safety methods have gained popularity as an alternative approach for screening. GPS-enabled smartphones can collect reliable and spatio-temporally rich driving data from regular drivers using an inexpensive, simple, and user-friendly tool. However, few studies to date have analyzed large volumes of smartphone GPS data and considered surrogate-safety modelling techniques for network screening. The purpose of this paper is to propose a surrogate safety screening approach based on smartphone GPS data and a Full Bayesian modelling framework. After processing crash data and GPS data collected in Quebec City, Canada, several surrogate safety measures (SSMs), including vehicle manoeuvres (hard braking) and measures of traffic flow (congestion, average speed, and speed variation), were extracted. Then, spatial crash frequency models incorporating the extracted SSMs were proposed and validated. A Latent Gaussian Spatial Model was estimated using the Integrated Nested Laplace Approximation (INLA) technique. While the INLA Negative Binomial models outperformed alternative models, incorporating spatial correlations provided the greatest improvement in model fit. Relationships between SSMs and crash frequency established in previous studies were generally supported by the modelling results. For example, hard braking, congestion, and speed variation were all positively linked to crash counts at the intersection level. Network screening based on SSMs presents a substantial contribution to the field of road safety and works towards the elimination of crash data in evaluation and monitoring.

1. Introduction

Network screening is the low-cost examination of an entire road network to identify a smaller subgroup of sites for detailed investigation or site diagnosis (Hauer et al., 2002). This smaller subgroup (hotspots, blackspots, hazardous road locations, or high risk sites) is expected to include locations where design or operation "create an increased risk of unforeseeable accidents" (Algerholm and Lahrmann, 2012) with potential for crash reduction (Aguero-Valverde and Jovanis, 2009). Considering parties involved in road safety management have limited budgets, sites should be identified and prioritized to maximize the efficiency of implemented countermeasures, the specifics of which are determined in the diagnosis phase. Methods for network screening commonly use statistical (regression) models or safety performance functions and Bayesian statistics to estimate the expected number of crashes at each location of interest in the road network (Park and Sahaji, 2013). Through these techniques, the risk factors contributing to crash occurrence can also be uncovered (Chang and Wang, 2006). In addition to the expected number of crashes, various risk measures can be derived including the posterior probability of excess and posterior of

* Corresponding author.
E-mail addresses: joshua.stipancic@mail.mcgill.ca (J. Stipancic), luis.miranda-moreno@mcgill.ca (L. Miranda-Moreno), nicolas.saunier@polymtl.ca (N. Saunier), aurelie.labbe@hec.ca (A. Labbe).

https://doi.org/10.1016/j.aap.2018.07.013
Received 17 January 2018; Received in revised form 3 July 2018; Accepted 5 July 2018
0001-4575/

ranks among others (Miranda-Moreno et al., 2007; Heydecker and Wu, 2001).

In network screening, safety must be objectively quantified. Most existing techniques use posterior ranking criteria based on Bayesian methods (Huang et al., 2009). Using statistical models, the relationships between attributes of traffic, geometry, environment, and driver (Abdel-Aty and Pande, 2005) and crash frequency and severity is first established (Lu, 2007). Then, Bayesian posterior analysis is used to quantify risk for each location. Though popular, methods derived from crash data are subject to errors in collision databases and are sensitive to underreporting (Kockelman and Kweon, 2002). Furthermore, as traffic collisions are relatively rare events (the low mean problem), long collection periods are required to accumulate sufficient crash data for analysis (Lee et al., 2006; Lord and Miranda-Moreno, 2008). For this reason, crash-based network screening cannot be carried out continuously, but is often performed periodically (for example, once every few years) so that crashes can accrue and databases can be updated. This highlights what is perhaps the most critical flaw of existing models. Crash-based methods are reactive, requiring crashes to occur before hazardous sites are identified and improvements are made (Algerholm and Lahrmann, 2012). For these reasons, alternative screening methods would be valuable for identifying high risk sites more quickly, more accurately, and with limited reliance on crash data (Cafiso and Di Graziano, 2011).

An alternative screening method requires an alternative data source from which crash risk can be constantly and systematically estimated throughout the network. Instrumented, or probe, vehicles that act "as moving sensors, continuously feeding information about traffic conditions" (El Faouzi et al., 2011) are an important source of such data. Vehicles instrumented with Global Position Systems (GPS) or other sensor types can assist in reducing dependence on crash databases by supporting the development of screening methods based on surrogate safety measures (SSMs) rather than collision statistics, a concept previously explored by Strauss et al. (2015) using instrumented bicycles. SSMs are any non-crash measures that are predictably related to crashes (Tarko et al., 2009). Although such methods rely on crash data for calibration, the application of the developed models to continuously monitor safety depends only on the input probe data that is continuously available. 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 driving data can now be collected unobtrusively from regular drivers (Jun et al., 2007) in crashes, near crashes, and under normal conditions (Bagdadi, 2013; Wu and Jovanis, 2013). Smartphones are inexpensive and user-friendly, minimally impact behaviour, eliminate the need for external sensors (Eren et al., 2012; Johnson and Trivedi, 2011), take advantage of widespread technology, and exploit existing communication infrastructure (Herrera et al., 2010).

Despite advances in 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 across a large urban road network. Other than previous studies by the authors (Stipancic et al., 2018, 2017a, 2017b), even fewer have considered the link between GPS-derived SSMs and large volumes of historical crash data at the network level. Finally, few studies to date have considered advanced modelling techniques, including recent developments in Bayesian inference and spatial models, for screening in road networks using GPS-derived SSMs. The purpose of this paper is to propose a surrogate safety screening approach based on smartphone GPS data and a Full Bayesian modelling framework. Various SSMs, including vehicle manoeuvres and traffic flow measures, are first extracted. A Full Bayesian Spatial Latent Gaussian Model (LGM) is then developed with SSMs as covariates using the Integrated Nested Laplace Approximation (INLA) technique. Finally the model is validated on an independent data set.

2. Literature review

Bayesian techniques have been widely used within the field of transportation for modelling the likelihood of crashes and their frequency throughout road networks. Empirical Bayes (EB) models were popular in the 1980s and 1990s. In EB models, the probability of a crash is determined, in part, by using observed historical crash data (Jiang et al., 2014). Hauer (1992) described the EB process, noting that the safety of a site is described by both its characteristics and historical crash record, and presented applications of the model for estimating crashes in the US and Ontario, Canada. Mountain et al. (1996) applied the EB technique to a series of at grade crossings in the UK, showing an improvement over naïve regression models. Later, Miaou and Lord (2003) compared EB and Full Bayes (FB) techniques for the specific application of crash modelling, noting that EB estimates deviated from the FB estimates, and that those deviations could become significant for some data sets. FB techniques for complex problems (such as non-conjugate models including LGMs) typically determine the posterior distributions by first assuming a prior distribution and then iteratively computing and updating the posterior marginal using a Monte Carlo Markov Chain (MCMC) simulation. Many examples of complex FB models can be found in the existing literature, including models which incorporate random effects (Jiang et al., 2014; El-Basyouny and Sayed, 2009) and/or spatial correlations (Song et al., 2006; Miaou and Song, 2005). Of most interest to this study are the relatively rare examples of INLA used to approximate FB models of crash risk. For example, Hu et al. (2013) studied the patterns of highway crashes over time using a dynamic generalized linear model approximated using INLA, and found that safety was improving over time. Similarly, Serhiyenko et al. (2014) used INLA to approximate Bayesian models describing temporal trends in pedestrian crashes. Serhiyenko et al. (2016) also developed a multivariate Poisson Lognormal model estimated using INLA to model various types of crash counts.

Several studies have attempted to extract SSMs from probe vehicle data. Event-based techniques identify individual driver manoeuvres including steering, braking, or accelerating (Dingus et al., 2006). Fazeen et al. (2012) 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 et al. (2007) analyzed the relationship between temporal-spatial driving behaviour activity 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 Lahrmann (2012) correlated jerk and crash occurrence both across drivers and across sites. Using GPS, accelerometer, radar, and self-reported collision data, Bagdadi (2013) proposed a jerk-based surrogate measure that correctly identified 86% of near misses. Traffic flow techniques rely on aggregate volume, speed, and density to measure risk (Yan et al., 2008). Though speed, flow, and variation in speed and flow have been suggested as potential SSMs in several studies (Abdel-Aty and Pande, 2005), traffic flow SSMs have rarely been studied using GPS data. Speed profiles from GPS devices were considered by Moreno and Garcia (2013) and Boonsiripant (2009).

Despite previous work on crash frequency modelling and extracting SSMs from GPS data, several shortcomings remain in the existing literature. 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 travel data alone. Few studies have used large amounts of data and validated the proposed SSMs using crash data. Instead, results are often compared to self-reported safety data or to near misses. More effort is needed to compare any proposed SSM with a reasonable amount of historical crash data to demonstrate that such a relationship exists. In terms of crash modelling, although more complex models continue to improve estimates of road traffic crashes, very few