An index for safety management of road networks

Prasad Buddhavarapu and Jorge Prozzi

Department of Civil, Architectural and Environmental Engineering
The University of Texas at Austin

ABSTRACT

Recently, highway safety management is gaining popularity among transportation agencies and becoming an important component of transportation asset management programs. Identification and prioritization of the safety needs of road networks is critical for optimal utilization of limited safety funds. This research commends on model-based safety management approach and proposes a safety index. The index is calculated using a hierarchical negative- binomial specification accounting for spatial correlation. Model parameters are estimated using computationally efficient data augmentation-based method rather than typically utilized Metropolis-Hastings algorithm based Bayesian estimation. The influence of road geometric features, pavement surface condition attributes including surface dis- tresses, skid resistance and ride quality on the crash frequency is reported.

1. INTRODUCTION

Recently, roadway safety management is gaining ample attention of transportation agencies and becoming an important component of transportation asset management programs. The primary goal of any roadway safety management system is to reduce the frequency and severity of road crashes. An annual report from National Highway Traffic Safety Administration (NHTSA) (NHTSA, 2011) reported an overall decline of 22.7 percent in the occupant fatality rate (per 100,000 population) from 1975 to 1992, which further decreased by 30.3 percent from 1992 to 2010; a similar trends were reported for occupant injury rate. Although a substantial improvement is evident in overall safety levels during last 35 years, about 35,000 fatalities and about 1.7 million injuries are still being reported annually in highway vehicle crashes in the recent years (2005-10) (NHTSA, 2011). In 2000, NHTSA estimated traffic crashes in the United States accounted for over $ 230 billion in economic losses (Blincoe et al., 2002). The cost of traffic crashes is reportedly more than two and one-half times the cost of congestion in urban areas (Herbel et al., 2010). The alarming crash statistics and the associated economic and social costs call for impending safety countermeasures across the United States. Federal Highway Association (FHWA) is actively performing research to improve the safety performance of roadways to reduce the number of fatalities and injuries by identifying and analyzing critical road safety issues.

Roadway safety management has evolved over the past several years into data-driven or evidence-based, rather than adherence to the standards, experiences, beliefs and intuitions (Herbel et al., 2010). In 1966, United States Congress enacted Highway Safety Act, a major Federal initiative towards improving roadway safety, that required individual states to establish and monitor a highway safety program in conformity with uniform standards constituted by Secretary of Transportation. FHWA and NHTSA shared the responsibility of implementing 18 essential standards that are established under the act. The 1966 Highway Safety Act was modified further in 1973 to replace the established standards with five priority safety improvement program areas. In 1978, Surface Transportation Assistance Act coalesced the five different areas into the Railway-Highway Grade Crossing and Hazard Elimination Programs. Hazard elimination program was primarily targeted at reducing the frequency of fatalities and serious injuries caused by road crashes on all public roads. The program provided funding for implementing projects to allay or eradicate the hazardous public road segments. Subsequently, individual states required to develop a Safety Management Sys- tem (SMS) under Intermodal Surface Transportation Efficiency Act (ISTEA) of 1991. SMS promoted the culture of maintaining crash database upon which safety decisions and performance measures may be established. In2005, Safe, Accountable, Flexible, Efficient Transportation Equity Act, established Highway Safety Improvement Program (HSIP), a federally funded, state-administered program with sole purpose of reducing traffic fatalities. As part of HSIP, states are recommended to implement Strategic Highway Safety Plans (SHSP) that are data-driven and promote the use of crash data analyses. SHSPs shall include strategies to identify and prioritize the safety needs of states' road network, and to establish performance-based goals for optimal utilization of limited safety funds and thereby to max- imize roadway safety. The enactment of Moving Ahead for Progress

in the 21st century act (MAP-21), a federal law approved in 2012, required each state-level transportation agencies to actively develop and modify SHSP. MAP-21 act dramatically increased the size of the successful Highway Safety Improvement Program (HSIP) with an average annual funding of $ 2.4 billion, thereby supporting the aggressive safety agenda of FHWA. According to the HSIP manual (Herbel et al., 2010), the goals of a scientific safety management framework are to understand and quantify the changes in the expected crash consequences; the future safety decisions are indeed reliant on such quantifiable evidence or experience. HSIP's data-driven strategic approach to improve highway safety emphasizes the need for comprehensive database management systems and state-of-the-art data analysis methodologies to identify crash prone zones and for network screening.

The integration and maintenance of reliable data sources including crash information, socio and demographic at- tributes, road features, pavement condition and road safety maintenance history (previously used safety treatments, if any) assist state transportation agencies in building superior SHSPs. Currently, Highway Safety manual (HSM) provides assistance to integrate safety into transportation asset management systems, mainly intended for practition- ers at the state, county, metropolitan planning organization (MPO), or local level (see (Gao, 2012) for discussion on integrating safety elements into transportation asset management systems). Generally, majority of state transportation agencies do maintain massive and detailed road databases. Accessibility of a comprehensive database is not beneficial unless statistical tools that facilitate learning about roadway safety from such massive datasets are available. Na¨ıve safety measures such as observed crash counts and descriptive crash statistics are indeed deceptive in depicting the road safety due to rareness and highly stochastic nature of the crash event. Hauer (1997) described roadway safety in crash counts or consequences, by kind and severity, expected to occur on a road segment during a specific period. HSM highlights the strengths of implementing predictive analysis over na¨ıve descriptive analysis methods to learn from roadway crash data. Predictive methods facilitate estimation of the expected frequency and severity of crashes on a given road segment with specific features over a particular time period as well as the precision and reliability of the estimates. The predicted/expected crash counts are useful to identify crash-vulnerable zones followed by a proba- bilistic ranking, thereby to prioritize safety maintenance projects across the network. Crash prediction accommodates common statistical issues such as "Regression to Mean " (RTM3) bias and natural fluctuations. Furthermore, crash prediction methods allows for incorporating any potential factors4 that influence the crash frequency and severity into the analyses; thus, the relationship between these factors and the crash statistics is learned. The knowledge on the sensitivity of the road features and socio and demographic attributes on crash statistics is arguably a valuable resource for framing future safety maintenance decisions and policy development.

1.1 Earlier Works

Bayesian predictive methods have been widely used for the analysis of road safety data and recently gaining further popularity due to the availability of efficient algorithmic techniques for the parameter estimation (Gilks et al., 1998). Bayesian methods allows to conveniently combine prior knowledge5 on crash outcomes, if any, and the observed crash data during a specific time period on a given road segment; this is often desired in road safety management. Among the roadway safety studies on the application of the Bayesian methods in the recent decades, the works of Ezra Hauer (Hauer, 1986, 1996a,b, 1997; Hauer et al., 2002) are noteworthy. HSM (2008), an important practice-oriented guide in safety management, mentions two varieties of Bayesian predictive methodologies: Empirical Bayes (EB) and Full Bayes (FB); both combine prior and current information to derive an estimate (posterior) of the expected safety level at a given road segment under investigation. It is to be noted that the manual currently recommends EB method for crash prediction. Several earlier researchers (Persaud et al., 2001, 1999; Elvik, 2008) extensively discussed the significance of Empirical Bayes (EB) method in crash prediction, particularly in before-after safety evaluations and site ranking applications. FB method is often considered methodologically superior as it offers several statistical advantages over the EB method. It allows for incorporating variability of the prior knowledge into the crash prediction, while EB method assumes a deterministic (point estimate) prior knowledge. A vast body of literature on road safety emphasized the benefits of FB methods and demonstrated their application in safety management applications(Persaud et al., 2010; Shively et al., 2010; Miranda-Moreno and Fu, 2007; Carriquiry and Pawlovich, 2004).

The reliability of a statistical inference of any crash predictive method relies largely on the underlying specification of the dependent variables i.e. crash frequency and severity. A realistic specification that best fits the observed data and allows learning maximum possible information from the relevant databases is desired. Currently, HSM (2008) models annual crash frequency as negative binomially distributed population using Annual Average Daily Traffic (AADT) and segment length as model attributes;the population mean/expected value is termed as Safety Performance Function(SPF) within the manual. Mannering and Bhat (2013) described that HSM is outdated and several method- ological generations behind the state-of-art statistics in the safety research. Mannering and Bhat (2013) presented a historical overview of the evolution of statistical specifications in the roadway safety research and identified several methodological barriers. They highlighted that the adoption of new methodologies is essential in the field of roadway safety research to address several statistical issues including unobserved heterogeneity, spatial and temporal corre- lation that may potentially impact the precision of resulting crash predictions, thereby affect budget allocation and relevant policies(Mannering and Bhat, 2013).

Although not currently practiced, a vast body of the literature over the last two decades addressed majority of the aforementioned statistical issues. Lord and Mannering (2010) documented the consistent progress of the existing liter- ature addressing various statistical methodological challenges in crash-frequency data analysis over the years. A recent work by Castro et al. (2012) developed a modeling framework to accommodate spatial and temporal correlations in crash frequency

prediction at intersections using classical likelihood maximization methods. (Narayanamoorthy et al., 2013) proposed a multi-variate count model while accommodating spatial correlation using Generalized Ordered Re- sponse (GOR) framework, previously proposed by Castro et al. (2012). Likelihood construction of complex statistical models that accommodate above statistical issues is often not straightforward and occasionally infeasible. Fully Bayesian or Hierarchical Bayesian methods are an alternative that greatly enhances the flexibility of model specification 6. For instance, it is relatively easier to accommodate correlation across crash frequencies of different seventies, spatial and temporal (or longitudinal) correlations (see Lord and Mannering, 2010; Lan, 2010 for discussion) in a hierarchical modeling framework. Numerous earlier research studies implemented Hierarchical Bayesian frameworks for modeling crash frequencies. For instance, Miaou and Song (2005) utilzied a Poisson-Gamma hierarchical speci- fication that accounts for spatial correlation to identify the vulnerable road segments or intersections. Inferences and site rankings were obtained through computer programs coded in WinBUGS language7. Another site ranking applica- tion utilized a multi-variate poisson log-normal model that accommodates potential spatial and temporal correlation of the crash counts Wang et al. (2011); the model estimates were obtained using WinBUGS coded programs.Ma et al. (2008) presented a hierarchical modeling framework accounting for correlation across the categories of crash counts; the model estimation was implemented using Gibbs sampling and Metropolis-Hastings algorithms, within an Marcov Chain Monte Carlo (MCMC) simulation framework. Another study by Aguero-Valverde and Jovanis (2008) presented a hierarchical Poisson-lognormal specification including spatial random effects to model the observed crash counts; the parameter estimates were obtained using computer program coded in OpenBUGS language8. Ahmed et al. (2011) and Yu et al. (2013) also utilized WinBUGS to estimate the parameters of poisson-gamma models with spatial and. random effects for predicting crash counts on mountainous freeway segments.

1.2. Current paper

In summary, the literature suggests a tremendous research emphasis on constructing the underlying specification of crash outcomes to address complex statistical concerns using Hierarchical modeling frameworks. A well-known drawback of the MCMC simulation based estimation methods is larger computational times, often accompanied by convergence issues and ad hoc algorithmic tuning. Achieving stationarity in sampling chains may take considerably longer necessitating larger burn-in periods9, particularly while using Metropolis-Hastings algorithm within an MCMC framework; this is primarily due to slow mixing while implementing Metropolis-Hasting algorithm directly on the product of prior and likelihood (see Van Dyk and Meng, 2001, for discussion). WinBUGS, the most commonly used programming environment for Bayesian parameter estimation in the road safety literature, uses Metropolis-Hastings algorithm to handle intractable posterior distributions, thereby facilitating Gibbs sampling in a typical MCMC sim- ulation framework. Burn-in periods spanning over first few thousands (up to 10,000) of MCMC iterations are typi- cally reported in the aforementioned road safety literature. The availability of

analytical/closed conditional posterior distributional forms (only up to proportionality constant is sufficient) avoids computationally expensive Metropolis- Hastings algorithms for obtaining parameter posterior draws. Data Augmentation techniques involving intermediate latent variables is a commonly used technique to construct analytically tractable posteriors in the Statistics literature (see (Van Dyk and Meng, 2001) for discussion on the art of data augmentation).

The primary goal of this paper is to demonstrate a procedure to construct safety index for road networks using computationally efficient Bayesian estimation algorithms. The proposed safety index is the expected crash frequency obtained through the model. The validity of the safety index is highly dependent on the underlying specification; implementation of sophisticated statistical techniques ensure a reliable and realistic safety index that accounts for underlying heterogeneity. A model for predicting annual crash frequency is developed assuming a negative binomial population while incorporating spatial correlation using neighborhood based approach. The spatial correlation is accommodated through Intrinsic Conditional Auto Regressive (ICAR)priors, originally proposed by Besag and Kooperberg (1995);the benefits of ICAR priors over most commonly used Conditional Auto Regressive (CAR) priors are provided. A recent study by Aguero-Valverde and Jovanis (2008) also modeled spatial random effects using a similar improper prior structure. Because of the hierarchical nature of proposed specification, it is relatively easier to extend the proposed data augmentation methodology to accommodate correlation across categories (or multi-variate count modeling) and spatio-temporal correlations. Broadly speaking, the paper identifies the need for research on improving the efficiency of the simulation based methodologies, particularly the parameter estimation of the hierarchical Bayesian specifications that are commonly used in crash modeling. Any such efforts (similar to this paper) arguably enhance the accessibility of statistically flexible and superior hierarchical (or fully) Bayesian approaches to state transportation authorities, thereby ensuring the feasibility of their implementation in safety management applications.

The rest of the paper is structured as follows. Section 2 outlines the development of hierarchical specification for crash counts. Subsequently, a Bayesian estimation procedure based on data augmentation technique for obtaining posterior distributions of model parameters is presented in section 3. Section 4 provides simulation results that demonstrate the computational efficiency and comment on accuracy of retrieving true model parameters of the pro- posed method. Section 5 describes an empirical example of crash count modeling using crash and road condition data from Harris county in Texas, USA. Section 6 highlights important findings and concludes the paper with a note on potential future extensions.

2. Model development

A hierarchical specification to model count data (crash frequency) while incorporating over-dispersion and spatial correlation is progressively developed in the beginning of the section. Subsequently, Bayesian inference of the model parameters is carried out using data-augmentation techniques via Gibbs sampling in a typical

MCMC simulation framework. The later part of the section discusses model evaluation and selection procedures.

Poisson regression is the most widely used, elementary specification for modeling non-negative discrete count data (Cameron and Trivedi, 1998). Single parameter is sufficient to fully characterize a Poisson regression model—the rate10 or intensity parameter (Casella, 2002). Expected value and variance of a Poisson distributed crash counts are equal to the intensity parameter, which is often modeled as a function of relevant regression features. An exponentiated linear combination of the regression features is typically used in order to ensure positivity of the underlying rate parameter. The weights (or regression coefficients) on the individual regression features within the linear combination are estimated using the observed data.

$$y_i \sim Poi(\lambda_i), \qquad \lambda_i = e^{x_i^T \beta}$$

where,

i is road segment index; i $\in \{1, 2, ...n\}$

y_i is crash frequency on i^{th} road segment

x_i is $k \times 1$ feature vector, $x = [x_1, x_2, ...x_n]^T$

β is $k \times 1$ weight (or coefficient) vector

$$E(y_i|x_i, \beta) = Var(y_i|x_i, \beta) = \lambda_i$$

The crash rate of each road segment is predicted by the respective road features in a deterministic fashion under Poisson regression framework. In other words, unobserved component of the crash rate is not included within Poisson regression. One way to incorporate the unobserved portion of the crash rate is through the multiplication of a strictly positive random effect to the deterministic portion of the crash rate. Lord and Mannering (2010) described over- dispersion as the notable characteristic of the crash-frequency data, although under-dispersion is rarely observed. Inclusion of the random effect into the crash rate removes the underlying equidispersion restriction (equality of mean and variance) imposed by Poisson regression model. Over-dispersion is induced due to the inclusion of random effect within the underlying rate parameter of the Poisson regression model. The random effects (ϵ_i) may be assumed to be identically and independently distributed, and generated from a probabilistic distribution that yields strictly positive random numbers. For instance, assuming a Gamma population for ϵ_i produces a Poisson-Gamma mixture model (shown below), which has been extensively utilized for modeling crash counts.

$$\lambda_i = e^{x_i^T \beta} \epsilon_i$$

$$\epsilon_i \sim Ga(r, 1)$$

$$E[y_i|x_i] = re^{x_i^T \beta} \qquad Var[y_i|x_i] = E[y_i|x_i] + \frac{1}{r} E^2[y_i|x_i]$$

The Poisson-Gamma mixture model can be reformulated as a negative binomial model by marginalizing the rate parameter (λi). The analytical expression for negative binomial likelihood is.

$$P(y_i|\psi_i, r) = \frac{\Gamma(y_i + r)}{\Gamma r y_i!}\left(\frac{e^{\psi_i}}{1 + e^{\psi_i}}\right)^{y_i}\left(\frac{1}{1 + e^{\psi_i}}\right)^{r} \qquad \text{where, } \psi_i = x_i^T\beta$$

The negative binomial likelihood parametrized in log-odds allows to implement a recently developed data-augmentation technique using Polya-Gamma random variable, which is further discussed in the later part of the document.

Spatial Correlation

The underlying assumption of the negative-binomial model is that the random crash rates of the road segments are independent of each other across the whole road network. However, relatively closer road segments arguably possess common unobserved features, thereby inducing a correlation between the crash rates of road segments within a neighborhood —spatial correlation. Spatial correlation among independent and negative binomially distributed crash frequencies may be incorporated using spatially correlated random effects; these are hierarchically generated using any neighborhood-based prior distributions. At this point, two types of random effects are introduced into the crash rate or rate parameter of the base Poisson regression model: 1) ϵ_i– aspatial random effect and 2) ϕ_i– spatial random effect. ϵ_i is directly multiplied with the deterministic component of the base crash rate as it is generated using strictly positive Gamma distribution; whereas φi is generated form normal distribution with real support (or $\phi_i \in (-\infty, \infty)$), hence exponentiated to ensure positivity of the rate parameter.

$$E(y|x_i) = e^{\psi_i + \phi_i}\epsilon_i$$

Conditionally Auto Regressive (CAR) priors, originally proposed by Besag (1974) are increasingly being used in the context of hierarchical spatial models (Banerjee et al., 2004) to generate spatially correlated random effects. Employing CAR priors facilitate relatively easier and computationally efficient implementation of the Gibbs sampling. CAR priors have been utilized extensively in the road safety literature for incorporating spatial correlation into crash frequency modeling (Miaou and Song, 2005; Wang and Kockelman, 2013). A Gaussian or autonormal CAR prior (shown below) specifies prior probability distribution of the spatial random effect corresponding to i^{th} road segment, given that of the remaining road segments.

$$p(\phi_i|\phi_{-i}) \sim N\left(\sum_j b_{ij}\phi_j, \tau_i^2\right), \qquad i \in \{1, 2, \ldots n\}$$

The n-dimensional joint prior distribution of the spatial random effects is obtained by combining their individual full conditional prior distributions (shown below) using Brooks Lemma (8).

$$p(\phi) \propto exp\left(-\frac{1}{2}\phi^T D^{-1}(I - B)\phi\right)$$

where,

$\phi = [\phi_1, \phi_2, ...\phi_n]^T$, D is a diagonal matrix with $D_{ii} = \tau_i^2$ and $B = \{b_{ij}\}$

A proximity matrix (W) that govern the extent of spatial dependence across the road network is defined. Either a neighborhood-based (binary) or distance-based proximity matrix (need not be a row-stochastic matrix) with zero diagonal elements is typically utilized. In order to ensure the symmetry of the covariance matrix, $b_{ij} = \frac{w_{ij}}{w_{i+}}$ and $b_{ij} = \frac{\tau_c^2}{w_{i+}}$ are commonly used. A neighborhood-based binary proximity matrix is adopted in this study.That is, $w_{ij} = 1$ if i & j are either first-order or second-order11 neighboring road segments, otherwise $w_{ij} = 0$. Hyper prior can be used to learn τc rather than providing a fixed prior value, which is often unavailable.

The propriety of the joint density of the spatial random effects is another concern as the matrix $D^{-1}(I - B)$ is clearly singular. Many earlier studies in road safety have introduced ρ parameter in the full conditional mean of the spatial random effects to circumvent the issue of improper joint distribution. Also, the ρ parameter has been extensively utilized as a proxy for the strength of underlying spatial correlation among the crash counts across road network. On the other hand, it is reported that ρ parameter does not calibrate very well with other descriptive measures of spatial association such as Moran's I or Geary's C (see Banerjee et al. (2004) page 78 for discussion). Banerjee et al. (2004) mentioned that ρ can mislead the analyst regarding the inferences on the strength of spatial correlation, particularly in the context of CAR priors. Interestingly, Banerjee et al. (2004) did not provide any guidelines on the inclusion of the ρ parameter, but remained neutral. We opted not to include the ρ parameter; specifically, spatial random effects are modeled using an improper CAR (often termed as Intrinsic CAR (ICAR)) prior. The joint distribution of the spatial random effects under the ICAR prior is shown below.

$$p(\phi) \propto exp\left(-\frac{1}{2\tau_c^2}\sum_{i \neq j} W_{ij}(\phi_i - \phi_j)^2\right)$$

In fact, the impropriety of the prior joint distribution of spatial random effects arising due to the absence of ρ parameter is no more a problem in regard with the posterior; the posterior precision is the sum of the likelihood and the prior precision, thereby ensuring the propriety of the posterior probability distribution of ϕ. However, ICAR being a pairwise difference prior identifies random effects only upto an additive constant; a sum-to-zero constraint is one way to evade the issue. The constraint is numerically imposed by recentering the draws of ϕ_i around its own mean in each Gibbs sampling iteration.

In summary, the following hierarchical specification is utilized for crash frequency modeling within this study. The posterior summaries of the model parameters are obtained via Gibbs sampling through data-augmentation technique.

$$y_i \sim NB(r, p_i), \qquad i \in \{1, 2, ...n\}$$

$$p_i = \frac{1}{1 + e^{\psi_i}} \qquad \psi_i = x_i^T \beta + \phi_i$$

$$\phi_i | \phi_{-i} \sim N\left(\sum_j \frac{w_{ij}}{w_{i+}} \phi_j, \frac{\tau_c^2}{w_{i+}} \right)$$

$$r \sim Ga(r_0, h) \qquad h \sim Ga(a_0, b_0)$$

$$\beta \sim N(b_0, B_0)$$

$$P_c = 1/\tau_c \qquad P_c \sim Ga(c_0, d_0)$$

2.2 Model estimation

Posterior summaries of the model parameters are constructed via typical MCMC simulation framework using Gibbs sampling. The joint distribution of all the model parameters is generated iteratively drawing from full condi- tional posterior distributions of the individual parameters (see Gamerman and Lopes (2006) for details). Full con- ditional posterior distributions are often complex and sampling is infeasible. Although, Metropolis-Hastings (M-H) algorithm (using random walk or importance sampling) is a great alternative sampling technique (used by WinBUGS), it is computationally expensive and often accompanied by convergence issues requiring ad hoc algorithmic tuning. Implementing M-H algorithm directly on the product of prior and likelihood results in slow mixing, thereby delaying attainment of stationarity in MCMC chains(see Van Dyk and Meng, 2001, for discussion). To avoid M-H algorithm, we used data augmentation technique in this paper and constructed analytical posterior probability distributions. Data augmentation involves introducing additional model parameters (latent variables) into the hierarchical specification that facilitate the construction of analytical full conditional posteriors for the desired model parameters. The procedure for constructing posterior analytical form of each of the model parameters is provided below.

Inferring Negative Binomial Regression Coefficients (β)

A recently developed data augmentation strategy for fully Bayesian inference in models with negative binomial likelihoods using Polya-Gamma random variables is adopted in this study (see Polson et al. 2013). Polya-Gamma random variable are introduced into the hierarchical specification, which in turn assist in building analytical conditional posterior of negative binomial regression coefficients.

A random variable ω has a PG(b,c) distribution (Polya-Gamma distribution with parameters b and c) if

$$\omega \overset{D}{=} \frac{1}{2\pi^2} \sum_{k=1}^{\infty} \frac{g_k}{(k - 1/2)^2 + c^2/(4\pi^2)}$$

where g1, g2, ...gk , ... are independent, Gamma(b,1) random variables.

Polson et al. (2013) proved that binomial likelihoods parametrized by log-odds can be written as mixtures of Gaussians with respect to Polya-Gamma distribution. We directly adopted this finding to construct a Gibbs sampling framework for inferring

negative binomial regression coefficients (β). The full conditional distribution of ωi (given the data and β) also turns out to be a Polya-Gamma distribution (see Polson et al. (2013) for discussion). Thus, the posterior draws of the regression coefficients are obtained via a Gibbs sampling framework or repeatedly sampling between the following two posterior forms.

$$P(\beta|\omega, r, y, x, \phi) \propto N(m_\beta, V_\beta)$$

$$m_\beta = V_\beta(x^T\Omega(z-\phi) + B_0^{-1}b_0) \qquad V_\beta = (x^T\Omega x + B_0^{-1})^{-1}$$

where,

z is $n \times 1$ vector; $z_i = \frac{y_i - r}{2\omega_i}$ for $i \in \{1, 2, ...n\}$

Ω is a diagonal matrix; $\Omega_{ii} = \omega_i$ for $i \in \{1, 2, ...n\}$

$$P(\omega_i|\beta, r, y, x, \phi) \propto PG(y_i + r, \psi_i + \phi_i), \qquad i \in \{1, 2, ...n\}$$

Inferring dispersion parameter (r)

Zhou et al. (2012) recently developed a Bayesian inference procedure for dispersion parameter of negative binomial likelihoods using compound Poisson representation. Negative binomial random variables can also be generated using Poisson and logarithmic distributed random variables as follows (Quenouille, 1949).

$$y_i = \sum_{l=1}^{L_i} u_l, \qquad u_l \overset{iid}{\sim} Log(p_i), \qquad L_i \sim Poisson(-r\ln(1-p_i)), \qquad i \in 1, 2, ...n$$

Data augmentation is achieved through the random variable Li with the Poisson prior. Full conditional posterior distribution of r, given Li is constructed using the conjugacy of the Poisson likelihood of the Li (contains r) and the Gamma prior of r.

$$p(r|L, -) \propto Gamma\left(r_0 + \sum_{i=1}^{N} L_i, h - \sum_{i=1}^{N} ln(1-p_i)\right)$$

The conditional posterior of the Li, given r is essential to complete Gibbs sampling sequence for inferring the dispersion parameter (r). Zhou et al. (2012) derived a closed form expression for the conditional posterior of the Li. The likelihood of observing yi given a particular realization of Li is constructed using Probability Generating Function (PGF) of the Logarithmic distribution, which is combined with the aforesaid Poisson prior with respect to the Li. We refer the reader to Zhou et al. (2012) for a detailed derivation of the conditional posterior of Li. Finally, the hyper parameter h is learned by constructing the full conditional posterior using conjugacy between the Gamma likelihood of r and the Gamma prior

Inferring spatial random effects

The posterior of the spatial random effects is constructed by exploiting the conditional independence of aforemen- tioned ICAR prior. The binary neighborhood-based proximity matrix induces a Gaussian Markov Random Field (GMRF) (see Rue and Held, 2005); in other words, spatial random effects, given the random effects within the respective neighborhoods are Gaussian distributed and independent of each other. This avoids the need for multivari- ate draws during MCMC simulation

and thereby gaining considerable computational efficiency during the Bayesian estimation. Full conditional posterior of the individual spatial random effects (conditioning on the posterior draws of the neighborhood spatial random effects) is constructed rather than constructing for the vector of spatial random effects. As discussed previously, we utilize the likelihood of the zi instead of yi in order to exploit the advantages of conditional Gaussian likelihood, given Polya-Gamma random variables (ωi), thereby to construct a closed form posterior (avoiding M-H algorithm). subsequently, the ICAR prior is combined with the likelihood of the zi, which yields a Gaussian posterior for individual spatial random effects conditional on the spatial random effects of the neighboring road segments. The posterior of spatial random effect of i^{th} road segment is provided below.

$$p(\phi_i|\phi_{-i},-) \sim N(m_\phi, V_\phi)$$

$$m_\phi = \left((z_i - x_i^T\beta)\omega_i + \left(\sum_j \frac{w_{ij}}{w_{i+}}\phi_j\right)\frac{w_{i+}}{\tau_c^2}\right)V_\phi$$

$$V_\phi = \left(\omega_i + \frac{w_{i+}}{\tau_c^2}\right)^{-1}$$

The posterior draws of all the spatial road segments is iteratively obtained rather a multi-variate draw. Also, the obtained posterior draws are recentered around their own mean during each MCMC iteration for the identification reasons described earlier.

A full conditional posterior facilitate learning the ICAR precision parameter, Pc from the data and avoids the need for ad-hoc tuning. The Gamma hyper prior of the Pc is combined with the likelihood of the previously generated spatial random effects within a single MCMC iteration, which results in a Gamma posterior.

$$p(P_c|\phi,-) \sim Ga\left(c_0 + \frac{n}{2}, d_0 + \sum_{i=1}^{n}\frac{w_{i+}}{2}\left(\phi_i - \sum_j b_{ij}\phi_j\right)^2\right)$$

It is to be noted that τ_c is associated with conditional distribution of the spatial random effects and does not represent the strength of the spatial correlation. Multiplying the τ_c by a constant does not change the strength of the spatial correlation.

Measuring spatial correlation

The absence of ρ parameter appears to complicate measurement of the strength of spatial correlation among the crash frequencies of the road segments across the road network. We use the empirical proportion of variability (α) in the random effects due to clustering as the strength of the spatial clustering (see Banerjee et al., 2004, pg. 160). The remaining portion (1 – α) of the variability is due to unstructured heterogeneity. The spatial random effects enter the crash rate through an exponential function, while the random effects due to the unstructured heterogeneity are directly multiplied with the deterministic component. The random effects due to clustering are exponentiated before calculating α to ensure a sensible comparison with unstructured random effects.

$$\alpha = \frac{\sigma_\phi}{\sigma_\phi + \sigma_\epsilon}$$

σ_ϕ is the empirical standard deviation of the exponential of the spatial random effect posterior draws in each MCMC iteration. σ_ϵ is the standard deviation of the unstructured random effects generated by the Gamma mixing in the negative binomial model. Thus, empirical posterior distribution of α is constructed and utilized for the Bayesian inference.

3. Empirical example

State governments typically maintain detailed information describing historical crash incidents and surface condition of the road network across the State as part of regular road management programs. Safety information typically includes characteristics of individual units associated with a specific crash such as vehicle, passenger and driver; whereas, pavement surface condition information contains level of road cracking, rutting, smoothness of the ride, skid resistance and geometric features. Database applications that can effectively merge the crash and pavement surface condition information facilitates the incorporation of safety into annual road management programs and assists in project prioritization. An integrated database was developed by merging the crash data with road condition and feature information and utilized for building the proposed hierarchical specification to model the crash counts on highway segments. The empirical example is intended to demonstrate a framework for developing an index that depicts the current safety level of individual road segments using localized/county-level databases; this example utilizes databases that belongs to Harris county, Texas.

The number of severe crashes16 occurred on half-a-mile long road segments of ten different road facilities around Houston area in the year 2009; such observed crash count is the dependent variable (y). A map of the high and low severity crash counts are shown in Figures 1 respectively. The high severity crash map is sparser than low severity crash map as expected. Spatial clustering of crash incidence locations is clearly visible in the figures, which supports the importance of spatial correlation in the proposed hierarchical model. Network-level measurements of surface distress condition (such as cracking, rutting), International Roughness Index (IRI), skid resistance along with road features such as shoulder width, surface width, number of lanes, imposed speed limit, Annual Average Daily traffic (AADT) and Truck traffic percentage are utilized as the predictors in the proposed model. The crash level information such as occupant and vehicle characteristics cannot be incorporated in this analysis for trivial reasons.

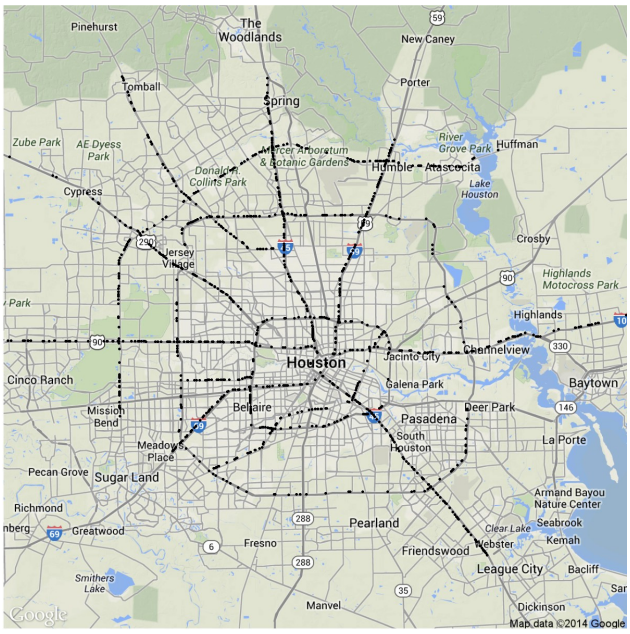

Figure 1: Crash map of selected road facilities in the Houston area

The descriptive statistics of the dataset are provided in Table 1. Traffic or exposure level is one of the important and most intuitive predictors in crash frequency modeling. The dataset covers a wide range of traffic levels across the road segments under the study. The truck traffic percentage varied between 2.2% to 24.8% with a mean value of 6.32%. Increase in truck traffic is intuitively related to the crash counts, hence investigated in this study. Several road features are investigated for a potential relationship with the crash counts. The descriptive statistics (see Table 1) indicate a wide variety of road segments in the study. The speed limit of the road segments varied between 35 and 70 miles/hour with a mean value of 56 miles/hour. In Texas, skid resistance (resistance offered by the pavement surface to tire slippage) is not measured at network-level; skid resistance information is collected only on about 20% of the network. Skid measurement indicator indicates the availability of the skid resistance measurement in the pavement management databases. Inclusion of the the skid measurement indicator facilitates to understand whether the skid resistance measurements are associated with high crash prone regions. As mentioned in the Table 1, 50% of the road segments possessed skid resistance information. Distress score represents the extent of surface distresses (such as cracking, rutting, edge drop offs, raveling, etc.) within a road segment. A utility value ranging from 0 (worst) to 1 (best) is given to each surface distress, depending on its impact on road serviceability. The Distress score is defined as the product of the utility values corresponding to each of the prevailing pavement distresses across a given road segment multiplied by 100. It ranges between 1 (worst pavement condition with major distresses) and 100 (best pavement condition with minor distresses). Table 1 reports a mean value of 93 with a standard deviation of 14. Such a high mean demonstrates the existence of an active and effective pavement management program in Texas. IRI is a measure of roughness obtained

from longitudinal road profiles. A higher IRI value indicates a lower ride quality on a given pavement. Table 1 shows that the IRI corresponding to the road segments included in this study ranged between 27 and 299 inch/mile with a mean value of 126 inch/mile. Condition score describes the overall condition of the pavement in terms of surface distress and ride quality, an indicator of an average persons perception of pavement quality. Typical Condition score values range between 1 and 100. Table 1 reports a mean value of 85 along with a standard deviation of 21.

Table 1: Descriptive Statistics

Variable Description	Minimum	Maximu	Mean	St.De
Traffic level (AADT)	3150	164075	51467	3535
Truck AADT	2.2	24.	6.32	3.7
Number of lanes	2	8	3.6	1.3
Length of road segment	0.10	1.0	0.47	0.1
Left shoulder width	0	27	8	4
right shoulder width	0	12	9	3
Total surface way width	20	119	5	19
Speed Limit	35	70	5	6
Skid measurement	0	1	0.50	0.5
Distress Score	8	100	9	14
IRI (inch/mile)	27	29	126	39 Condition
10	10	85	21	

Routes: FM1093, FM1960, IH0010, IH0045, IH0610, SH0006, SH0249, SL0008, UA0090, US0059, US0290

Table 2: Estimation results

Variable Description	Mean	St.Dev	5th	95th
Constant	-0.412	0.174	-0.727	-
Traffic level (AADT)	0.242	0.108	0.070	0.425
IRI (inch/mile)	0.136	0.060	0.040	0.236
Distress Score	-0.100	0.048	-0.178	-
Skid measurement indicator	0.549	0.221	0.165	0.895
Skid measurement indicator*Skid	-0.110	0.164	-0.392	0.162
Speed Limit	-0.097	0.066	-0.202	0.010
Dispersion parameter (r)	1.996	0.302	1.554	2.559
τ_C	0.236	0.065	0.142	0.348
Spatial correlation proportion	0.500	0.049	0.423	0.583
DIC	2510.0			

The hierarchical model is estimated using the proposed data augmentation based Bayesian estimation. It is important to note that the spatial dependence among road segments belonging to different road facilities is ignored in this analysis. The spatial dependence is only imposed among the road segments belonging to the same highway route. Intuitively, the spatial dependence is possible among the road segments belonging to different highway routes, particularly at their intersections. However, the grade separated freeway intersections are quite different from the

urban intersections, thereby allowing to ignore the correlation among the respective crash counts. Although we are ignoring the spatial dependence, modeling the pooling of the crash data on different facilities is statistically beneficial regarding the parameter estimation; a larger dataset always produces relatively better crash predictions.

Table 2 reports the model parameter estimates: mean, standard deviation, 5th and 95th quantiles are provided. The model goodness of fit or DIC is also reported. An intuitive interpretation of the model regression coefficients is provided below. The expected crash counts on a given road segment is exponentially related to the regression features as shown in the Equation 4. Consequently, the interpretation of the regression coefficients is straightforward. The possitive coefficient on the Traffic level is expected, which indicates that high traffic areas are associated with the higher crash frequencies. The positive sign on the IRI (a measure of ride quality) indicates that smoother roads are associated with lower crash frequencies; this is an important finding which suggests safety benefits of pavement ride quality management programs. A negative sign on the distress score indicates that roads with considerable distresses (such as cracking, rutting, pot holes etc.) are associated with higher crash frequencies. The positive sign on the skid measurement indicator reflects that the areas of skid measurements indeed witnessed larger crash frequencies, which is intuitive. Among the road segments with skid resistance measurements, road segments with larger skid resistance are associated with lower crash frequencies; this provides an empirical evidence for the influence of the skid resistance on the crash frequency. The spatial correlation proportion is estimated as 0.5, which indicates the presence of considerable spatially correlated heterogeneity in the dataset.

4. Conclusions

The primary objective of this paper is to develop a safety index while incorporating spatial correlation. The paper commends the Bayesian approaches in crash predictive modeling, particularly in road safety management. Bayesian approaches provide incorporation of the prior knowledge into the model building process, which is often desired in road safety. The paper identifies the need for better computational tools within the Bayesian estimation in order to avoid popularly used Metropolis-Hastings algorithm based approaches. The availability of faster computational tools enhances the accessibility of the sophisticated statistical methods, which are indeed necessary in crash prediction. For instance, this paper demonstrates implementing a spatial negative binomial model using data-augmentation based Bayesian estimation technique.

The paper demonstrated the proposed methodology using a dataset from Harris county, Texas in USA. Data was gathered by integrating pavement management databases with crash data from a total of ten different routes around Houston area. The influence of road condition and various road features was studies in this empirical analysis. Model estimation results suggested that roads with smoother ride, higher skid resistance and little surface distresses (such as a cracking, rutting, pot holes, etc.) are associated with the lower crash frequencies. The traffic was also a statistically significant predictor in the empirical model. The road features including speed limit, shoulder width, lane width, number of lanes were

not statistically significant predictors in this dataset. Similarity of the road facilities under the study is a potential reason for such weak relationship.

The crash count prediction of the spatial negative binomial model on each of the road segment is the proposed safety index of the road segment. In addition to the crash prediction, the reliability of the crash prediction is also important and shall thus be reported. We also recommend incorporating the probability of ensuring a particular safety threshold level or a threshold crash count along with the actual crash prediction into the project prioritization. A map of safety index can be constructed and reported to the transportation agency for a potential use in annual road safety management. Current model can be extended to incorporate temporal correlation of the crash counts and correlation across crash severity categories. Subsequently, the proposed safety index can be formulated as a weighted combination of predicted crash counts of multiple severity levels on a given road segment at any given time; the weights may be selected based on relative economic losses associated with the corresponding severity level.

REFERENCES

1. Aguero-Valverde, J., Jovanis, P. P., Dec. 2008. Analysis of road crash frequency with spatial models. Transportation Research Record: Journal of the Transportation Research Board 2061 (-1), 55–63.
2. Ahmed, M., Huang, H., Abdel-Aty, M., Guevara, B., Jul. 2011. Exploring a bayesian hierarchical approach for developing safety performance functions for a mountainous freeway. Accident Analysis & Prevention 43 (4), 1581–1589.
3. Banerjee, S., Carlin, B. P., Gelfand, A. E., 2004. Hierarchical modeling and analysis for spatial data. No. 101 in Monographs on statistics and applied probability. Chapman & Hall/CRC, Boca Raton, Fla.
4. Besag, J., 1974. Spatial interaction and the statistical analysis of lattice systems (with discussion). Journal of Royal Statistical Society 36 (Ser.B), 192–236.
5. Besag, J., Kooperberg, C., 1995. On Conditional and Intrinsic Autoregressions.
6. Blincoe, L., Seay, A., Zaloshnja, E., Miller, T., Romano, E., Luchter, S., Spicer, R., May 2002. The economic impact of motor vehicle crashes, 2000.
7. Cameron, A. C., Trivedi, P. K., 1998. Regression analysis of count data. No. 30 in Econometric society monographs. Cambridge University Press, Cambridge, UK ; New York, NY, USA.
8. Carriquiry, A., Pawlovich, M. D., 2004. From empirical bayes to full bayes: methods for analyzing traffic safety data. White Paper, Iowa State University.
9. Casella, G., 2002. Statistical inference, 2nd Edition. Thomson Learning, Australia ; Pacific Grove, CA.
10. Castro, M., Paleti, R., Bhat, C. R., Jan. 2012. A latent variable representation of count data models to accommodate spatial and temporal depen- dence: Application to predicting crash frequency at intersections. Transportation Research Part B: Methodological 46 (1), 253–272.
11. Elvik, R., Nov. 2008. The predictive validity of empirical bayes estimates of road safety. Accident Analysis & Prevention 40 (6), 1964–1969.

12. Gamerman, D., Lopes, H. F., 2006. Markov chain Monte Carlo: stochastic simulation for Bayesian inference, 2nd Edition. No. 68 in Texts in statistical science series. Taylor & Francis, Boca Raton.
13. Gao, J., Sep. 2012. Asset management and safety: A performance perspective. Final Report MTC Project 2009-01, Center for Transportation Research and Education.
14. Gilks, W. R., Richardson, S., Spiegelhalter, D. J., 1998. Markov chain Monte Carlo in practice. Chapman & Hall, Boca Raton, Fla. Hauer, E., 1986. On the estimation of the expected number of accidents. Accident Analysis & Prevention 18 (1), 112.
15. Hauer, E., 1996a. Detection of safety deterioration in a series of accident counts. Transportation Research Record: Journal of the Transportation Research Board 1542 (1), 3843.
16. Hauer, E., 1996b. Identification of sites with promise. Transportation Research Record: Journal of the Transportation Research Board 1542 (1), 5460.
17. Hauer, E., 1997. Observational before–after studies in road safety: estimating the effect of highway and traffic engineering measures on road safety, 1st Edition. Pergamon, Oxford, OX, U.K. ; Tarrytown, N.Y., U.S.A.
18. Hauer, E., Harwood, D. W., Council, F. M., Griffith, M. S., 2002. Estimating safety by the empirical bayes method: a tutorial. Transportation Research Record: Journal of the Transportation Research Board 1784 (1), 126131.
19. Herbel, S., Laing, L., McGovern, C., Jan. 2010. Highway safety improvement program (HSIP) manual. Report No. FHWA-SA-09-029.HSM, Dec. 2008. Highway safety manual. Tech. rep., Transportation Research Board.
20. Lan, B., 2010. Exploration of theoretical and application issues in using Fully Bayesian methods for road safety analysis.
21. Lord, D., Mannering, F., Jun. 2010. The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives.Transportation Research Part A: Policy and Practice 44 (5), 291–305.
22. Ma, J., Kockelman, K. M., Damien, P., May 2008. A multivariate poisson-lognormal regression model for prediction of crash counts by severity, using bayesian methods. Accident Analysis & Prevention 40 (3), 964–975.
23. Mannering, F. L., Bhat, C. R., Oct. 2013. Analytic methods in accident research: Methodological frontier and future directions. Analytic Methods in Accident Research.
24. Miaou, S.-P., Song, J. J., Jul. 2005. Bayesian ranking of sites for engineering safety improvements: Decision parameter, treatability concept, statistical criterion, and spatial dependence. Accident Analysis & Prevention 37 (4), 699–720.
25. Miranda-Moreno, L. F., Fu, L., 2007. Traffic safety study: Empirical bayes or full bayes? In: Transportation Research Board 86th Annual Meeting.
26. Narayanamoorthy, S., Paleti, R., Bhat, C. R., Sep. 2013. On accommodating spatial dependence in bicycle and pedestrian injury counts by severity level. Transportation Research Part B: Methodological 55, 245–264.
27. NHTSA, 2011. Traffic safety facts 2010. Tech. Rep. DOT HS 811 659, National Highway Traffic Safety Administration, Washington, DC. URL http://www-nrd.nhtsa.dot.gov/Pubs/811659.pdf
28. Persaud, B., Lan, B., Lyon, C., Bhim, R., Jan. 2010. Comparison of empirical bayes and full bayes approaches for beforeafter road safety evalua- tions. Accident Analysis & Prevention 42 (1), 38–43.

29. Persaud, B., Lyon, C., Nguyen, T., Jan. 1999. Empirical bayes procedure for ranking sites for safety investigation by potential for safety improve- ment. Transportation Research Record 1665 (1), 7–12.
30. Persaud, B., Retting, R., Garder, P., Lord, D., Jan. 2001. Safety effect of roundabout conversions in the united states: Empirical bayes observational before-after study. Transportation Research Record 1751 (1), 1–8.
31. Polson, N. G., Scott, J. G., Windle, J., Aug. 2013. Bayesian inference for logistic models using plya-gamma latent variables. Journal of the American Statistical Association, 130808174755007.
32. Quenouille, M., Jun. 1949. A relation between the logarithmic, poisson, and negative binomial series. Biometrics 5 (2), 162–164.
33. Rue, H., Held, L., 2005. Gaussian Markov random fields: theory and applications. No. 104 in Monographs on statistics and applied probability.Chapman & Hall/CRC, Boca Raton.
34. Shively, T. S., Kockelman, K., Damien, P., Jun. 2010. A bayesian semi-parametric model to estimate relationships between crash counts and roadway characteristics. Transportation Research Part B: Methodological 44 (5), 699–715.
35. Spiegelhalter, D. J., Best, N. G., Carlin, B. P., Van Der Linde, A., 2002. Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 64 (4), 583639.
36. Van Dyk, D. A., Meng, X.-L., 2001. The art of data augmentation. Journal of Computational and Graphical Statistics 10 (1).
37. Wang, C., Quddus, M. A., Ison, S. G., Nov. 2011. Predicting accident frequency at their severity levels and its application in site ranking using a two-stage mixed multivariate model. Accident Analysis & Prevention 43 (6), 1979–1990.
38. Wang, Y., Kockelman, K. M., 2013. A poisson-lognormal conditional-autoregressive model for multivariate spatial analysis of pedestrian crash counts across neighborhoods. Accident Analysis & Prevention 60, 7184.
39. Yu, R., Abdel-Aty, M., Ahmed, M., Jan. 2013. Bayesian random effect models incorporating real-time weather and traffic data to investigate mountainous freeway hazardous factors. Accident Analysis & Prevention 50, 371–376.
40. Zhou, M., Li, L., Dunson, D., Carin, L., 2012. Lognormal and gamma mixed negative binomial regression. arXiv preprint arXiv:1206.6456.

Author Biography

Prasad Buddhavarapu

Prasad Buddhavarapu is currently a doctoral student at The University of Texas at Austin in the Department of Civil Engineering. He received his M.S. in Civil Engineering from the University of Texas at Austin in 2011. He is also currently working on another Master's degree in Statistics along with his doctoral degree. Prasad's primary research interests are in the areas of statistical modeling of transportation data and pavement management. He is currently working on developing a network level safety index which includes modeling of historical crash count data while accounting for spatial and temporal correlation using Hierarchical Bayesian modeling techniques. He has previously worked on a few research projects including pavement material engineering, quality management in pavement construction, pavement performance data collection and diamond grinding.