Accident Analysis and Prevention xxx (xxxx) xxx-xxx

Contents lists available at ScienceDirect

Accident Analysis and Prevention

journal homepage: http://ees.elsevier.com

Exploring the modeling and site-ranking performance of Bayesian spatiotemporal crash frequency models with mixture components

Wen Cheng [a,*], Gurdiljot Singh Gill [a], Yongping Zhang [a], Tom Vo [b], Frank Wen [b], Yihua Li [c]

[a] Department of Civil Engineering, California State Polytechnic University, Pomona, 3801 W. Temple Ave., Pomona, CA 91768, United States
[b] Southern California Association of Governments, 818 West 7th Street, 12th Floor, Los Angeles, CA 90017, United States
[c] Department of Logistics Engineering, Logistics and Traffic College, Central South University of Forestry and Technology, Hunan 410004 30, China

ARTICLE INFO

Keywords:
Mixture
Spatiotemporal
Interaction
Predictive accuracy
Cross validation
Site ranking

ABSTRACT

The current study introduces the flexible approach of mixture components to model the spatiotemporal interaction for ranking of hazardous sites and compares the model performance with the conventional methods. In case of predictive accuracy based on in-sample errors (posterior deviance), the Mixture-5 demonstrated superior performance in majority of the cases, indicating the advantage of mixture approach to accurately predict crash counts. LPML (log pseudo marginal likelihood) was also calculated as a cross-validation measure based on out-of-sample errors and this criterion also established the dominance of Mixture-5, further reinforcing the superiority of the mixture approach from different perspectives.The site ranking evaluation results demonstrated the advantages of adopting the mixture approach. In terms of total rank difference (TRD) results, there were several discrepancies between the two approaches, suggesting that two approaches designate unsafe sites differently. Another site ranking criterion, site consistency test (SCT), was employed to explore the difference in identification of unsafe sites based on two datasets: estimated crash count (traditional) and the spatial variability across time. The advantage of mixture models to act as a complimentary approach for site ranking was revealed by the spatial variability SCT results. The method consistency test (MCT) results also indicate the advantages of mixture models over the Base one. These findings suggested that mixture approach may prove helpful in the network screening step of safety management process to identify sites which may turn unsafe in the future and escape the detection from traditional methods.

1. Introduction

The detection of high-risk sites forms the core of overall safety management process. These sites are commonly referred to as hotspots, sites with promise, or black-spots, and the procedure of network screening is adopted for their identification (Hauer et al., 2004; Cheng and Washington, 2005; Huang et al., 2009). Site ranking is essential to realize the goal of safer traffic environment as it allows efficiency in resource allocation for implementing safety countermeasures to mitigate crashes at different spatial levels of road network.

In the past decades, various site ranking methods have been adopted by the transportation safety agencies such as crash frequency method (Deacon et al., 1975), crash rate method (Norden et al., 1956), rate quality control method (Stokes and Mutabazi, 1996), and so on. However, such methods were limited in handling the diverse issues pertaining to crash data such as regression-to-the-mean bias (Hauer, 1997), small sample size and mean of crash frequency (Miaou and Lord, 2003), and the unobserved heterogeneity, which is the tendency to ignore the "uncertainty" of associations of covariates and underlying crash risk (Lord and Mannering, 2010).

Recent studies focused site ranking have employed the statistical advancements in the Markov chain Monte Carlo (MCMC) technique to develop crash prediction models using the Full Bayesian (FB) approach. This method allows the flexibility to incorporate complex space–time and site-specific heterogeneities by means of random-effects and random parameters to address the unobserved heterogeneity inherent to crash data (Huang et al., 2009; Tunaru, 2002; Miaou et al., 2003; Park and Lord, 2007; Lan et al., 2009; Persaud et al., 2010; Wang et al., 2011; Sacchi et al., 2015; Cheng et al., 2017). The ignorance of the unobserved heterogeneity in crash prediction models may lead to serious implications such as biased parameter estimates, erroneous inferences, and inaccurate crash predictions (Mannering et al., 2016).

* Corresponding author.
E-mail addresses: wcheng@cpp.edu (W. Cheng); gurdiljotg@cpp.edu (G.S. Gill); yongpingz@cpp.edu (Y. Zhang); Vo@scag.ca.gov (T. Vo); Wen@scag.ca.gov (F. Wen); yhli@csuft.edu.cn (Y. Li)

https://doi.org/10.1016/j.aap.2019.105357
Received 31 December 2018; Received in revised form 10 October 2019; Accepted 4 November 2019
Available online xxx
0001-4575/© 2019.

UNCORRECTED PROOF

To mitigate the problems resulting from ignorance of unobserved heterogeneity, some studies utilized the correlation between neighboring roadway entities such as intersections (Wang and Abdel-Aty, 2006; Cheng et al., 2017), segments (Aguero-Valverde and Jovanis, 2008), corridors (Abdel-Aty and Wang, 2006), TAZs (Traffic Analysis Zone) (Xu and Huang, 2015), county (Miaou et al., 2003; Gill et al., 2017) and so on. These studies observed that the inclusion of spatial random effects improved the model fit and precision due to the capability to "borrow information" and "pool strength" from the neighboring road entities (Aguero-Valverde and Jovanis, 2006). Similar advantages associated with model performance were observed by many studies which employed different temporal treatments to address the serial correlations within crash data across a time period (Andrey and Yagar, 1993; El-Basyouny et al., 2014; Mannering, 2018; Jiang et al., 2014).

Moreover, some studies observed that crash data tend to be clustered both spatially and temporally since the site-specific effects may remain constant over time or fluctuate periodically (Huang et al., 2009; Aguero-Valverde and Jovanis, 2006; Aguero-Valverde et al., 2016). Hence, both space and time effects are important for consideration for development of crash prediction models). For the purposes of site ranking from different perspectives, many recent studies adopted the spatial and temporal correlation structures to account for unobserved heterogeneity and demonstrated their advantage at superior performance for site ranking (Cheng et al., 2017; Bernardinelli et al., 1995).

Although the aforementioned studies incorporated the spatial and temporal effects, they ignored the interaction of space and time by assuming the impact of time to be constant over space. However, more precise model estimates could be generated by adopting a flexible approach which accommodates the variations of time-trend across space (Richardson et al., 2006). The recent macro level HSID study by Dong et al. (Dong et al., 2016) addressed this issue by employing an autoregressive lag-1 dependence for space-time interaction. This study observed that the proposed model, compared to the models without correlations structures and isolated space and time effects, was significantly superior at identification of high risk sites. The new approach also offered the advantage of monitoring the evolvement of crash risk of sites over time.

The above-discussed spatiotemporal models employed in safety literature have attempted to cover the space-time profile of underlying crash risk of sites. However, they are based on the assumption that all the sites under consideration follow a global space-time pattern. This limiting assumption restricts the identification of sites which show larger fluctuations with respect to expected space-time pattern. The larger fluctuations may serve as an indication of future safety concerns but the prevailing space-time models in network screening may not capture the sites experiencing low crash frequency but large variability. Along with the traditional methodology which screens the high-risk sites based on global space-time pattern, a complementary proactive approach should also be explored which highlights the ones with site-specific variability. For example, there is a general decline in economic activity during a recession which then leads to overall less driving and crashes. However, to reduce household spending, some families might switch to other transportation modes such as active transportation from usual vehicle-driving. Such transportation modal shift may increase the human activities near the public transportation stations in certain TAZs yielding more pedestrian or bicyclist involved collisions, a phenomenon against the global space-time pattern within this time interval.

To this end, the current study borrows a spatiotemporal model with flexible mixture components from an epidemiologic study (Abellan et al., 2008) which documented many benefits of this approach for detection of sites with unstable pattern of diseases. The mixture component model allows the flexibility to capture the predictable (stable) patterns of overall spatial and temporal risk surface and simultaneously capture the sites which show greater variability to depart from the stable pattern of the space-time interaction. This framework explores the full space-time profile of the underlying structure of crash risks by accounting for the unobserved heterogeneity by means of site-specific space-time interactions. This exploration allows the establishment of a spatial crash risk pattern that sustains over a period of time as well as detection of sites which show unusual fluctuation in crash risk over this time. The stable and unstable patterns may be interpreted based on the crash risk of specific sites. Stable pattern represents the recurrent external factors which influence crash risk such as higher traffic volume during peak hours. The unstable pattern may address the unobserved heterogeneity by incorporating the non-recurrent events which heighten the crash risk such as changes in enforcement level or weather conditions. To the best knowledge of the authors, the current study is the first to introduce this approach of mixture components in safety literature to provide a new perspective for site ranking. It is expected that this approach may be coupled with the conventional methods as it would allow distinction of sites which exhibit deteriorating expected safety performance so that precautionary measures may be undertaken to mitigate crashes.

Apart from proposing the new approach for site ranking, this study also makes several other contributions by performing an empirical evaluation of the proposed method. First, the study by Abellan et al. (2008) employed the probabilistic approach with different priors for distinction of sites into two groups: unstable and stable. The current study adopts the similar approach by developing the first group of models which separates the sites into these two levels, and also a second group with five levels which provides a subtler classification of unstable sites. Second, a Base model (Model 1) without mixture components is also developed to represent the existing safety research which accommodates independent spatial and temporal random effects without interaction. This model serves for comparison with the proposed mixture component models with space-time interaction (Model 2 is Mixture-2 and Model 3 is Mixture-5). Third, the safety literature demonstrated the advantages associated with different temporal treatments. This study employed linear, quadratic, and autoregressive with lag-1 trend to represent different temporal treatments for development of spatiotemporal models. Fourth, the model performance based on predictive accuracy is assessed by employing different evaluation criteria which are based on direct in-sample errors as well as direct and indirect out-of-sample errors. Finally, this study adopts two cut-off levels (top 5%, and 10% sites) to replicate the different goals of safety agencies for assessment of models based on site ranking performance and employed three criteria which assess model performance from different perspectives.

2. Methodology

The area of focus was the aggregated vehicular crashes at the 203 TAZs in the City of Irvine, California, during a ten-year period (2003–2012). The crash data were collected from SWITRS (California Statewide Integrated Traffic Records System). The relatively long period selected has the benefits of better understanding the temporal trends and having more data allowing data-splitting for the purpose of model evaluation. However, it is also important to note that, given the 10-year span of data containing a great economic recession, it is likely that there would be temporal stability initially followed by a highly unstable period. For instance, Malyshkina and Mannering (2008) revealed the evidence of temporal stability in injury severities in the mid-2000s even in the presence of speed limit increases. But work that was performed in the great recession years later has found significant temporal instability. More details about temporal instability can be found in the subsection of Mixture with Two Levels.

UNCORRECTED PROOF

The Full Bayesian (FB) framework was employed to estimate the crash frequency of vehicles at the TAZ level. Under the FB approach, the Poisson lognormal model is adopted to account for the overdispersion usually associated with crash data and better handle the low sample mean and small sample size due to the heavier tails associated with lognormal distribution (Lord and Miranda-Moreno, 2008). This model assumes that crash count (y_i) at a given TAZ i, obeys Poisson distribution, while the corresponding observation specific error term ε_i follows a normal distribution:

$$y_i \big| \lambda_i \sim Poisson\,(\lambda_i) \tag{1}$$

Where λ_i is the expected Poisson count for location i, and can be expressed as follows:

$$\ln\left(\lambda_i\right) = \beta_0 + \varepsilon_i \tag{2}$$

Where β_0 is the intercept and ε_i is the independent random effect which captures the extra-Poisson heterogeneity among locations. ε_i can be assumed with the following noninformative Normal priors:

$$\varepsilon_i \sim Normal\,(0,\ \tau^2\ ) \tag{3}$$

Where τ^2 is the variance of the normal distribution for ε_i. The inverse of τ^2 is called precision and it can be modeled using the following gamma prior with prior mean equal to one and its prior variance large (equal to one thousand), representing high uncertainty or prior ignorance:

$$\tau^{2^{-1}} \sim Gamma\,(0.001,\ 0.001) \tag{4}$$

The model presented in Eq. (2) serves as the basic formulation for crash prediction in most safety studies. The current study also adopts this model as the basis for development of spatiotemporal models.

2.1. Three groups of models

This section first introduces the Base model specification, with three different temporal treatments, which serves to represent the space-time interactions by conventional traffic safety studies. Discussions are provided regarding the limitations associated with the traditional approach (Base). The next part of this section introduces the formulation of mixture component models for capturing the space-time interaction. The mixture models for two and five levels were developed in the current study and the rationale of these two approaches is discussed. It should be noted that the three temporal treatments associated with Base models were also utilized for the mixture models.

2.2. Base

This study was focused on exploring the network screening capabilities associated with mixture models and the Base group was utilized for reference. This group represents the approach utilized by safety studies for accommodating the space-time interactions. The basic formulation adopted before incorporating the temporal treatment is obtained by modifying the Eq. (2) to account for the spatial correlation among neighboring TAZs:

$$\ln\left(\lambda_i\right) = \beta_0 + \varepsilon_i + u_i \tag{5}$$

Where u_i is the spatially structured random effect which follows the CAR (conditional autoregressive) formulation to incorporate the spatial correlation among crashes occurring at neighboring TAZs. The CAR distribution allows correlation between the random effects of neighboring geographic units, borrowing strength and creating smooth spatial patterns.

$$u_i \sim N\left(\overline{u_i}, \tau_i^2\right)\ \ \overline{u_i} = \frac{\sum_{i\neq k} u_k w_{ik}}{\sum_{i\neq k} w_{ik}}\ \ \tau_i^2 = \frac{\tau_c^2}{\sum_{i\neq k} w_{ik}} \tag{6}$$

Where w_{ik} is the proximity weight matrix and τ_c is the precision parameter in the CAR prior. Similar to the past studies (Xu and Huang, 2015; Gill et al., 2017; Aguero-Valverde and Jovanis, 2010), the current study adopts a distance-based weight structure to account for the spatial correlation, where the weight between two TAZs is calculated as the inverse of distance between their centroids. The distance matrix for TAZ centroids was provided by SCAG (Southern California Association of Governments).

To introduce the three temporal treatments and the space-time interaction term, Eq. (5) is modified as follows:

$$\ln\left(\lambda_{it}\right) = \beta_0 + f(t) + u_i + v_{it} \tag{7}$$

Where $f(t)$ is the temporal treatment, and v_{it} is the space-time interaction parameter. v_{it} can take different forms such as linear (Aguero-Valverde and Jovanis, 2006) and time-varying spatial correlation (Lawson et al., 2003). This study developed three models based on varying complexity of temporal treatments.

2.2.1. Model 1: linear time trend

In this model, a linear trend is introduced where time is regarded as a potential influential covariate and the model estimates the coefficient for the trend. Eq. (7) assumes the following form:

$$\ln\left(\lambda_{it}\right) = \beta_0 + u_i + v_{it} + \beta_t\ * T \tag{8}$$

Where β_t is the scalar parameter for linear yearly trend T ($T = 1$–5 in the current study). It is expected that the passage of time will impact the crash risk on the TAZs which would be captured by the model.

2.2.2. Model 2: quadratic time trend

In this model, a non-linear impact of time is considered. Eq. (7) takes the following form:

$$\ln\left(\lambda_{it}\right) = \beta_0 + u_i + v_{it} + \beta_t\ * T + \beta_{t2}\ * T^2 \tag{9}$$

Where β_{t2} is the coefficient for the quadratic trend term. It is expected that this model will allow more subtle approach compared to the linear trend as it allows flexibility to fit the crash data by virtue of its additional quadratic trend.

2.2.3. Model 3: autoregressive-1 (AR-1)

This model established a temporal dependency from prior year by specifying the distribution of ε_{it} as a lag-1 dependence in errors, where lag-1 means that the time is varying yearly (Miaou and Song, 2005; Gill et al., 2018; Mannering, 2018). Under this model, Eq. (7) assumes the following form:

$$\ln\left(\lambda_{it}\right) = \beta_0 + \varepsilon_{it} + u_i + v_{it} \tag{10}$$

The weighted sum is fixed and the random terms change at every time step following the same distribution, which means this model is homoscedastic. The distributions are given by:

$$\varepsilon_{it} \sim normal\left(0, \frac{\sigma_{it}^2}{\left(1-\gamma^2\right)}\right) \tag{11}$$

$$\varepsilon_{it} \sim normal\left(\gamma\varepsilon_{i,t-1}, \sigma_{it}^2\right)\ \text{for t} > 1 \tag{12}$$

Where γ is the autocorrelation coefficient with the range of $0 < \gamma < 1$. This model addresses the potential correlation between successive time periods and is expected to deliver a more precise estimation of the model parameters.

UNCORRECTED PROOF

2.3. Mixture with two levels

The above discussed Base models represent the traditional approach of network screening which is based on the estimated crash frequency. For example, the 203 TAZs would be arranged in the decreasing order of crash risk (estimated crash frequency) and a predefined threshold (top 5% or 10%) may be utilized to identify the most unsafe sites for further investigation or safety countermeasure implementation. Beyond the practical implications, the technical aspect of this approach assumes a global space-time trend to be accounted for by the corresponding terms while the unobserved heterogeneity is expected to be captured by the error term which follows normal distribution with a constant variance. However, this approach is limiting in exploring the complete space-time profile of crash risk as the variability associated with each site with respect to time is not accounted for. Rather, all sites are assumed to be subjected to a constant global spatiotemporal trend. This limitation may be addressed by using the mixture components which have the capability to incorporate the site-specific variability associated with passage of time. This would immensely benefit the safety research by complementing the traditional network screening process as the sites showing the signs of instability (with respect to crash risk over time) would be captured.

The presence of temporal instability for crash risk has been acknowledged by many studies (refer to Mannering (2018) for comprehensive review of impact of temporal instability in different fields) and different approaches have been employed in an attempt to capture it. The use of random effects and temporal treatments employed in the current study represents the vast majority of literature dealing with temporal instability. Such approaches, though, are based on the assumption that the temporal impact remain constant over all sites under consideration. Also, the impact of explanatory factors is also assumed to be consistent across all sites. To handle this issue, some studies employed the random parameter approach which operated on the assumption that there exist statistical differences for the impact of an explanatory factors across sites for a given time period. Still this approach was restricted in quantifying the varying impact of factors across different time periods for a given site. Adopting the philosophy of random parameter approach, some studies employed time-varying coefficients which allowed the flexibility to accommodate the variations of explanatory variables across time periods. However, this approach still assumed that global space-time trend which impacts the crash risk of all sites. In other words, the variability in the crash risk of each site was not captured for different time periods. The approach of mixture components allows the flexibility to distinguish between the stable and unstable underlying risk pattern, where the stable (or predictable) patterns are represented by the overall time trend ($f(t)$) and spatial crash risk (u_i), and unstable (or unpredictable) patterns are the atypical departures from the stable pattern (Cheng et al., 2018a).

The methodology of mixture component for space-time interaction was introduced by Abellan et al. (2008) in an attempt to develop a framework which has the flexibility to "uncover the full space-time profile of risks". The specific concerned study employed a 2-component mixture which combines variabilities of different levels, where the levels distinguish the unstable and stable pattern based on variability of space-time interaction. The Mixture-2 model specification is obtained by modifying the space-time interaction term (v_{it}) from Eq. (7) as follows:

$$v_{it} \sim p\text{Normal}\left(0, \tau_1^2\right) + (1-p)\text{Normal}\left(0, \tau_2^2\right) \tag{13}$$

The prior for p is uniform on [0, 1] through a Dirichlet distribution. The flexibility of the approach if afforded by the accommodation of discreet and continuous space-time variations in crash data by means of the varied prior distribution. The discreet variations (i.e. the second component in the above equation) represent the atypical patterns in the space-time profile of the crash site which deviates from the stable global pattern. The atypical behavior will be demonstrated by large fluctuations which indicate the higher risk relative to other sites. Since the first component captures the noise, representing stable global space-time pattern, and the second component captures the departure from that continuous pattern, the half-normal hyper-prior distributions were assigned for the standard deviations τ_k, ($k = 1, 2$), which reflect that τ_1 (corresponding to "stable" risk pattern) has to be small to effect shrinkage so as to capture the noise, whereas the prior for τ_2 (corresponding to "unstable" risk pattern) allows the flexibility to capture large fluctuations with a much wider range. The different ranges of variance are shown in the following equations:

$$\tau_1 \sim Normal\ (0,\ 0.01)\ .\mathrm{I}_{(0,\ +\infty)} \tag{14}$$

$$\tau_2 \sim Normal\ (0,\ 0.01)\ .\mathrm{I}_{(0,\ +\infty)} + Normal\ (0,\ 100)\ .\mathrm{I}_{(0,\ +\infty)} \tag{15}$$

Comparing Eqs. (14) and (15), it is known that τ_1 is part of τ_2. Such formulation not only ensures that τ_2 is greater τ_1, but also prevents the label switching issues Abellan et al. (2008). Additionally, the indicator value (*ind*) for each location can be expressed as follows:

$$ind_{it} = \sum_{k=1}^{2} p_{itk} * k\ (k = 1, 2) \tag{16}$$

For the mixture with 2 levels, it is known that ind_{it} range from 1 to 2, the larger the value, the larger probability of the greater space-time variability for the location i of period t tends to be. Therefore, in addition to rely on λ_{it} (estimated crash counts) for network screening purpose, we can also utilize the value ind_{it} for site-ranking need, which tries to flag out locations with unstable space-time pattern.

2.4. Mixture with five Levels

The above-mentioned Mixture with Two Levels model roughly divide the spatiotemporal variability into two levels, stable and unstable. In essence, the division can be done for any levels based on the researcher needs. For illustration purpose, the subsection focuses on a subtler classification of 5-component mixture. Compared with the Mixture-2 model, the Mixtrure-5 specification is different with more components of spatiotemporal variability:

$$v_{it} \sim \sum_{k=1}^{5} p_k \text{Normal}\left(0, \tau_k^2\right) \tag{17}$$

Similarly, the different ranges of variance are expressed as follows:

$$\tau_1 \sim Normal\ (0,\ 0.01)\ .\mathrm{I}_{(0,\ +\infty)} \tag{18}$$

$$\tau_2 \sim \tau_1 + Normal\ (0,\ 0.1)\ .\mathrm{I}_{(0,\ +\infty)} \tag{19}$$

$$\tau_3 \sim \tau_2 + Normal\ (0,\ 1)\ .\mathrm{I}_{(0,\ +\infty)} \tag{20}$$

$$\tau_4 \sim \tau_3 + Normal\ (0,\ 10)\ .\mathrm{I}_{(0,\ +\infty)} \tag{21}$$

$$\tau_5 \sim \tau_4 + Normal\ (0,\ 100)\ .\mathrm{I}_{(0,\ +\infty)} \tag{22}$$

Likewise, the indicator value (*ind*) for each location can be expressed as follows:

$$ind_{it} = \sum_{k=1}^{5} p_{itk} * k\ (k = 1, 2, 3, 4, 5) \tag{23}$$

UNCORRECTED PROOF

Again, the larger the *ind* value, the more unstable of spatiotemporal pattern for the specific location. All denotations in Eqs. (17)–(23) are the same as those in Eqs. (13)–(16).

2.5. Model evaluation criteria

2.5.1. DIC

This study used DIC (Deviance Information Criterion) to assess the complexity and goodness of fit of the models. DIC is a hierarchical modeling generalization of the AIC (Akaike Information Criterion) which was proposed by Spiegelhalter et al. (2003) to account for model fit and complexity. Specifically, DIC is defined as:

$$DIC = D\left(\bar{\theta}\right) + 2p_D = \bar{D} + p_D \tag{24}$$

Where, $D\left(\bar{\theta}\right)$ is the deviance evaluated at the posterior means of estimated unknowns ($\bar{\theta}$), and posterior mean deviance $\bar{D}$ can be taken as a Bayesian measure of fit or "adequacy". p_D denotes the effective number of parameters in a model, as the difference between $D\left(\bar{\theta}\right)$ and $\bar{D}$, i.e., mean deviance minus the deviance of the means.

2.5.2. LPML

This cross-validation approach is implemented by estimating the conditional predictive ordinate (CPO) using the CV-1 (leave-one-out) technique (Gelfand, 1996) which removed the selection bias by employing a continuous approach of selecting all data points, except one, for model development and the left out data point to verify the prediction accuracy of the calibrated model. Under the MCMC framework, the estimate of CPO for each observation *i* can be calculated as:

$$CPO = \left(\frac{1}{T}\sum_{t=1}^{T}\frac{1}{f(Y_i|\beta^{(t)})}\right)^{-1} \tag{25}$$

Where Y_i is the *i*th observation (i = 1, 2, 3,. . . , n) for all 203 TAZs and β is the vector of estimated model parameters. This harmonic mean of density (CPO) may be extended to calculate the goodness-of-fit of models by computing the product of CPOs over all observations, which is known as the pseudo marginal likelihood. For computational convenience, the log pseudo marginal likelihoods (LPML) is calculated (Cheng et al., 2018b; Heydari et al., 2016):

$$LPML = \sum_{i=1}^{n}\log(CPO_i) \tag{26}$$

The larger LPML value indicates a superior performance associated with the candidate model.

2.6. Site ranking criteria

A natural extension of application of the previously mentioned models is to rank safety performance of sites of interest (Aguero-Valverde and Jovanis, 2009; Geedipally and Lord, 2010; Washington et al., 2014). Due to the importance of this task for enhancement of safety of roadway network, many studies have been dedicated to assessing the performance of hotspot detection methodologies (Montella, 2010; Jiang et al., 2014). Three popular evaluation tests were selected which assess the ranking methodologies from different perspectives.

2.6.1. Site consistency test

This test measures the ability of a model to consistently identify a roadway entity as an unsafe site over subsequent observations. It is based on the premise that an unsafe site would remain in that state if no safety treatment delivered. The larger SCT score indicates superior site ranking performance of a model. The expression of SCT is shown as follows (Cheng and Washington, 2008):

$$SCT_j = \sum_{k=1}^{n} Crash\ Statistic_{k(i),method=j,i+1} \tag{27}$$

where $Crash\ Statistic_{k(i),method=j,i+1}$ is the crash statistic in time period i + 1 for a site that is ranked *k* in time period *i* as identified by the model *j*.

2.6.2. Method consistency test

This test is also based on the assumption that an unsafe site remains unsafe for subsequent observations in time provided that no safety treatment is applied. MCT measures the number of same hotspots identified by a model for two different observations of time. By definition, MCT is expressed as (Cheng and Washington, 2008):

$$MCT_j = \{k_1, k_2, \ldots, k_n\}_i \bigcap \{k_1, k_2, \ldots, k_n\}_{i+1} \tag{28}$$

2.6.3. Total rank difference test

This test measures the difference in rankings of sites for successive periods. The smaller value indicates more consistent performance of a model for ranking the hotspots. The corresponding calculation is shown as (Cheng and Washington, 2008):

$$TRDT_j = \sum_{k=1}^{n}\left|R\left(k_{j,i}\right) - R(k_{j,i+1})\right| \tag{29}$$

Where $R\left(k_{j,i}\right)$ is the rank of site *k* in time period *i* identified by the model *j*.

3. Results

3.1. Model estimates

The models were developed in the statistical software WinBUGS (Lunn et al., 2000) to generate MCMC samples for Bayesian posterior inferences. For the model calibration, two chains of 25,000 iterations were set up. Convergence was ensured by visual inspection of chains and observing the desired threshold condition of MC errors to be lower than 5% of the standard deviation of parameters. After ensuring the convergence, first 10,000 samples were discarded as adaptation and burn-in and rest of the samples were used to draw parameter estimates. For ease of reference, the models with different temporal correlation would be denoted by Model # while the models with different levels would be denoted as model types.

The posterior model estimates of the coefficients are shown in Table 1. It is important to note that no explanatory variables are shown in the table as this study intends to compare the impact of temporal trends and the inclusion of covariates may blur the comparison. The parameter estimates illustrate the robustness of the developed models across the three types (Base and two Mixture models) based on the closeness of coefficients. In terms of comparison across temporal trends, the coefficients are observed to vary, and it is expected as different temporal specifications account for different degrees of variability which impact the coefficient values. The intercept is noted to be statistically significant for all models. However, the inclusion of temporal specifications does not turn up to be significant, which suggests that the crash data may have underlying temporal correlation to a lesser extent than expected. The only exception to the non-significant of temporal terms is the linear trend for Before period for Model 1, which shows a statistically significant negative correlation. The reversal of coefficient signs for linear and quadratic trend between Before and After periods suggests the possibility of temporal instability (Mannering, 2018). In terms of advantages associated with precision of posterior estimates, there is no clear trend of superior precision for any tempo-

UNCORRECTED PROOF

Table 1
Posterior Model Estimates.

Variable	Before			After		
β_0						
	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3
Base	1.032 (0.043)	1.045 (0.072)	0.946 (0.071)	0.922 (0.043)	0.812 (0.075)	0.898 (0.030)
Mixture-2	1.030 (0.042)	1.050 (0.090)	0.944 (0.030)	0.913 (0.048)	0.823 (0.079)	0.894 (0.033)
Mixture-5	1.030 (0.043)	1.040 (0.073)	0.949 (0.053)	0.909 (0.044)	0.815 (0.077)	0.893 (0.067)
β_t						
	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3
Base	−0.028 (0.009)	**−0.039 (0.050)**	NA	**−0.006 (0.010)**	**0.086 (0.052)**	NA
Mixture-2	−0.028 (0.010)	**−0.040 (0.068)**	NA	**−0.005 (0.012)**	**0.070 (0.058)**	NA
Mixture-5	−0.029 (0.010)	**−0.037 (0.052)**	NA	**−0.005 (0.011)**	**0.075 (0.055)**	NA
Btsq						
	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3
Base	NA	**0.002 (0.008)**	NA	NA	**−0.015 (0.008)**	NA
Mixture-2	NA	**0.002 (0.011)**	NA	NA	**−0.012 (0.009)**	NA
Mixture-5	NA	**0.001 (0.008)**	NA	NA	**−0.013 (0.009)**	NA
Gamma						
	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3
Base	NA	NA	**0.525 (0.587)**	NA	NA	**0.321 (0.450)**
Mixture-2	NA	NA	**0.023 (0.573)**	NA	NA	**0.520 (0.494)**
Mixture-5	NA	NA	**0.568 (0.603)**	NA	NA	**0.536 (0.599)**

Notes: 1. The bold text represents the variables not statistically significant at 95% confidence level. 2. The numbers in the parentheses represent the associated standard deviation.

ral correlation or model type, as reflected by the standard deviation of coefficients.

3.2. Predictive accuracy evaluation

This study employed DIC and its components for assessment of goodness-of-fit of alternate models. DIC was selected for the model comparison as it is a penalized criterion which acts as a trade-off between model complexity and model fit which are represented by the effective number of parameters (P_d) and posterior deviance ($\bar{D}$). The Dbar may be regarded as a measure to indicate model performance based on in-sample errors and DIC may be regarded as the indirect representation of model performance based on out-of-sample errors. The results for DIC and its components are shown in the upper three sections of Table 2. In case of Dbar, which represents model fit, the model performance across different temporal treatments depicts mixed results. The AR-1 specification (Model 3) is superior in case of Base model while the linear trend indicates advantage at model fit for three out of four cases of superior performance. The model fit across different types shows that majority of Mixture (5 levels) models are superior, followed by Mixture (2 levels). This indicates that the flexibility to accommodate more levels may lend the advantage to fit the crash data better. In case of Pd, which represents the effective number of parameters employed by the model and represents the complexity, the AR-1 specification (Model 3) is consistently observed to employ the least number of parameters. This specification accounts for the correlation from one prior year and seems to be relatively less complex than the addition of linear or quadratic trend. This may be attributed to the absence of statistical significance for temporal terms in general across most models (refer to Table 1). For comparison across model types, the Base model consistently demonstrates the least complexity and this result aligns with the expectations as the mixture models may incorporate higher number of effective parameters to accommodate multiple levels. In case of DIC, which represents the overall goodness-of-fit and is the indirect measure of model performance based on test errors, The AR-1 models are consistently superior to alternative models for both time periods. This superiority may be attributed to the lowest value of Pd observed for AR-1 models which compensated for the higher posterior deviance in some cases. For model comparison across types, the Base model record the lowest DIC which is influenced by their lower values of Pd. The exception to this trend in the Mixture (5 levels) model employing AR-1 specification which records the lowest DIC for After period. It is evident from the Pd values that the superior DIC for this model is also attributed to the lowest Pd. Overall, the results suggest the significant role of addition of effective number of parameters (Pd) to govern the overall goodness-of-fit of models.

This study also adopts the cross-validation approach to compute the conditional predictive ordinate (CPO) and eventually calculate log pseudo marginal likelihood (LPML) for comparison of model fit. The higher value of LPML reflects relatively superior model fit property. The difference between the LPML values of two concerned models is referred to as the log pseudo Bayes factor (LPBF), where a LPBF greater than 5 reflects the superiority of the model of interest (Ntzoufras, 2009). As shown in the last section of Table 2, the AR-1 models are noted to be superior for four out of six cases for comparison between temporal treatments. This suggests that the superiority of this specification observed in case of training errors may be transferable to test errors (or out-of-sample). In terms of model performance across different types, the Mixture (5 levels) is consistently superior within each temporal treatment. This significant advantage of Mixture (5 levels) was not evident for other criteria but this observation for LPML suggests that Mixture (5 levels) models demonstrate the capability to perform superior to other models for a different crash dataset as LPML essentially represents the model performance for out-of-sample errors.

The authors further conducted the ANOVA (analysis of variance) and Tukey's HSD (Honest Significant Difference) analysis to check whether the performance differences between temporal treatments and model types are statistically significant or not. As shown in Table 3, the model types based on different levels demonstrate the presence of statistically significant differences across all three criteria. This corroborates the findings from Table 2 and suggests that the introduction of different levels within the models creates a significant performance difference. However, in case of difference between models based on temporal treatments, only DIC exhibits the presence of statistically significant difference. This finding suggests that different types of temporal correlations may not lead to a significant performance difference.

Table 3 depicts the presence of statistically significant differences between models but fails to identify the models which are different. To distinguish the superior models, box and whisker plots were developed which are shown in Fig. 1. The upper portion of the figure demonstrates the performance difference for different model types. The superiority of mixture models is evident in case of Dbar and LPML where the Base model demonstrates significant difference with inferior performance. This finding corroborates the results from Table 3. However, contrary to the observations from the table, only the model pair of Mixture (2 levels) and Base depict statistically significant difference for DIC. In general, the plots substantiate the presence of statistically significant differences in the model performance of mixtures and

UNCORRECTED PROOF

Table 2
Model Evaluation Results.

Evaluation Criteria	Before			After		
Dbar						
	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3
Base	3763.37	3775.28	**3756.88**	3804.16	3791.6	**3746.28**
Mixture-2	***3691.57***	3724.54	3750.09	***3682.24***	*3686.36*	3711.37
Mixture-5	**3717.94**	*3722.7*	*3733.66*	3698.72	**3687.91**	*3701.9*
P_D						
	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3
Base	*197.916*	*213.277*	***20.866***	*220.72*	*214.05*	**206.595**
Mixture-2	666.06	800.243	**269.783**	713.951	679.019	**262.835**
Mixture-5	359.844	435.609	**83.273**	496.012	404.131	***155.271***
DIC						
	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3
Base	*3961.28*	*3988.55*	***3777.75***	*4024.88*	*4005.65*	**3952.87**
Mixture-2	4357.63	4524.79	**4019.88**	4396.19	4365.38	**3974.21**
Mixture-5	4077.78	4158.31	**3816.93**	4194.73	4092.05	***3857.17***
LPML						
	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3
Base	−1877	−1879	**−1869**	−1885	−1885	**−1869**
Mixture-2	**−1861**	−1864	−1863	−1847	−1848	**−1845**
Mixture-5	***−1855***	***−1855***	*−1856*	*−1842*	*−1840*	***−1839***

Note: For each criterion, the best model performance across different types (different levels) is highlighted in *italics* and across different temporal correlations (Model #) is highlighted in bold.

Table 3
ANOVA and Tukey's HSD Results.

Criterion	Residuals	D.F.	Sum of Square	Mean Square	F	Pr(>F)
Difference among Base, Mixtrue-2, and Mixture-5						
Dbar	Groups	2.00	16,328.00	8,164.00	16.75	0.00
	Residuals	15.00	7,310.00	487.00		
DIC	Groups	2.00	334,817.00	167,408.00	6.09	0.01
	Residuals	15.00	412,545.00	27,503.00		
LPML	Groups	2.00	2,861.40	1,430.70	21.56	0.00
	Residuals	15.00	995.50	66.40		
Difference among three Model 1, 2 and 3						
Dbar	Types	2.00	158.00	78.90	0.05	0.95
	Residuals	15.00	23,480.00	1,565.30		
DIC	Types	2.00	312,907.00	156,453.00	5.40	0.02
	Residuals	15.00	434,454.00	28,964.00		
LPML	Types	2.00	88.00	44.22	0.18	0.84
	Residuals	15.00	3,769.00	251.23		

Base model for majority of the cases while the difference between the two mixture models is ambiguous. The lower portion of Fig. 1 demonstrates the difference between three temporal models for the evaluation criteria. The results align with the findings from Table 3 as the presence of statistically significant difference is only observed in case of DIC where the AR-1 specification is superior.

3.3. Site ranking evaluation

The previous section evaluated the model performance based on the predictive accuracy and goodness-of-fit. This study also intended to compare model performance based on site ranking and explore whether the superiority at model fit or predictive accuracy also subsequently leads to advantages at site ranking. The criteria of Total Ranking Difference (TRD), Site Consistency Test (SCT), and Method Consistency Test (MCT) are employed for evaluation. It is worth mentioning that the SCT and MCT are based on the distinction between safe and unsafe TAZ's. In practice, the zones are ranked in descending order and with the top zones being selected for further investigation. The specific number chosen is based on the agency policy or available funding. In the present research, both top 5% and 10% zones were arbitrarily selected to represent the hazardous zones with relatively more crashes where the common results would yield more confidence regarding the ranking performance.

As shown in Table 4, the TRD evaluation is conducted between before and after Bayesian estimated counts (λ_{it}) and *ind* values (probability of space-time variability). It is worth restating that ind values are only generated by the mixture models and this study intended to compare the ranking performance based on traditional counts method and the ind values based on mixture approach. A smaller value of TRD is desirable as it indicates that there is lesser difference in the ranks of sites for before and after periods. In other words, assuming that the underlying crash risk of a site remains constant, a lower TRD suggests that the model is capable to assigning correct risk-based ranks for before and after periods. As evident from Table 4, there are some discrepancies among the model performance based on count and ind approaches. This finding meets the expectation as the ind approach was perceived to be employed as a complimentary one to the tradi-

UNCORRECTED PROOF

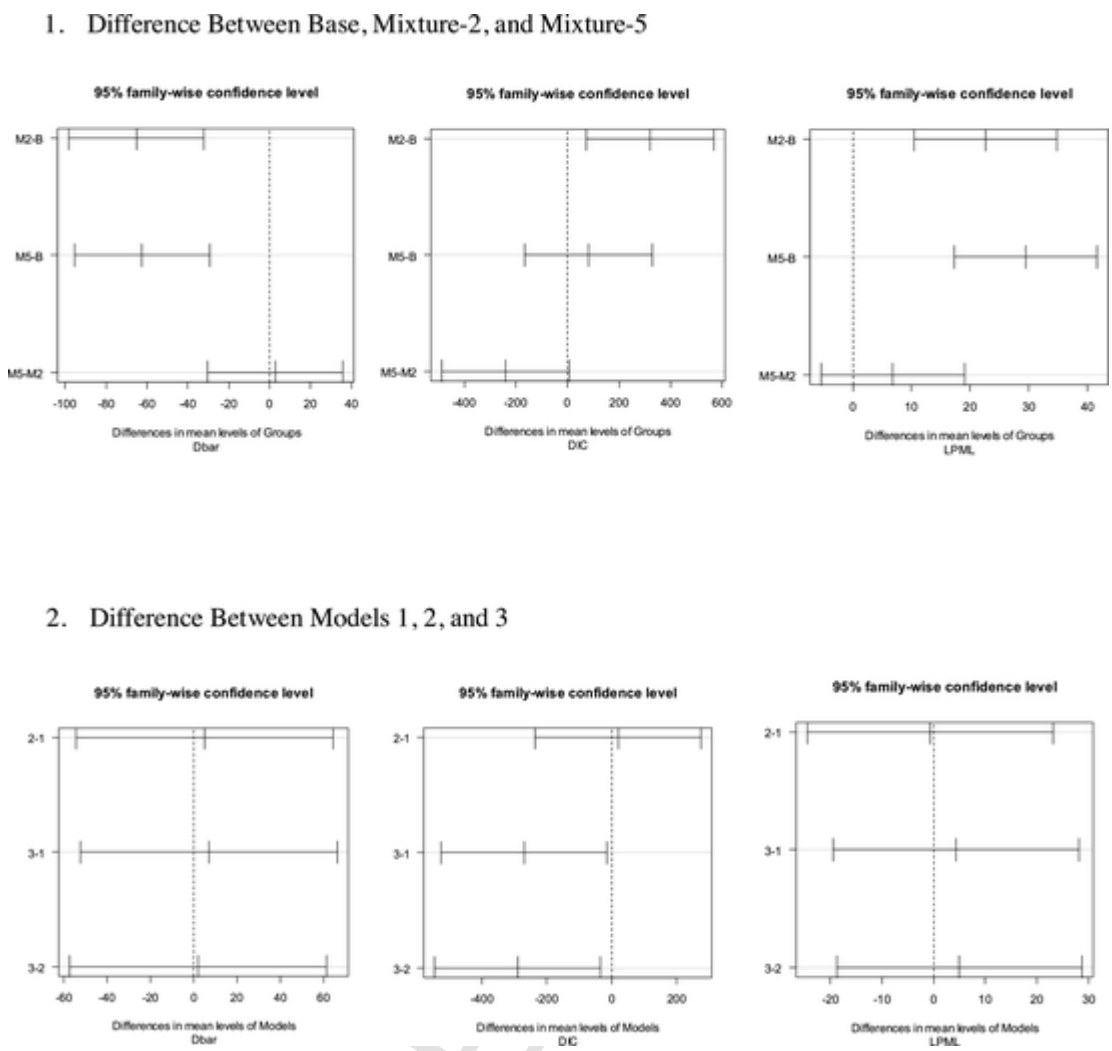

Fig. 1. Box and whisker plots for model performance.

Table 4
Site Ranking Results using TRD and Different Crash Datasets.

	before-λ_{it} and after-λ_{it}			before-ind_{it} and after-ind_{it}		
	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3
Base	***3358***	3360	3438	NA	NA	NA
Mixture-2	3376	***3359***	*3365*	13083	13003	**12998**
Mixture-5	**3361**	3376	3500	*12210*	***11903***	*12724*

Notes: 1. For each criterion, the best model performance across different types (different levels) is highlighted in *italics* and across different temporal correlations (Model #) is highlighted in bold. 2. See Eqs. (8) and (16) for definitions of λ_{it} and ind_{it}.

tional count approach as it may catch the sites which escaped the count approach. In case of count approach, the linear trend Base model is the best overall, closely followed Mixture-2 quadratic model and Mixture-5 linear model. Interestingly, the advantage of AR-1 specification from model fit seems to have limited superiority in terms of site ranking. In case of ind approach, the quadratic trend depicts overall best TRD (11,903), followed by linear and AR-1. It is noteworthy to highlight that all three temporal models for Mixture-5 are superior to Mixture-2, indicating that the flexibility provided by the former approach by introducing five levels may lead to superior site ranking performance as the sites may acquire subtler designation based on probability.

Another evaluation criterion for assessment of site ranking performance is Site Consistent Test. The lower value of SCT essentially indicates that a model consistently identifies similar sites across different time periods. The results based on the traditional approach of posterior crash estimates is shown in the upper portion of Table 5. As evident from the table, all models exhibit exactly the same SCT values (except that Mixture-5 performs slightly in the case of Model 1 and 5%).

Table 5
Site Ranking Results using SCT and Different Datasets.

before-λ_{it} and after-y_{it}						
	5%			10%		
	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3
Base	1189	1189	1189	1855	1855	1855
Mixture-2	1189	1189	1189	1855	1855	1855
Mixture-5	1157	1189	1189	1855	1855	1855
The standard deviation of before-v_{it} and after-v_{it}						
Base	0.002	0.006	0.017	0.004	0.012	0.032
Mixture-2	**0.69**	0.5	0.33	1.12	1.01	0.62
Mixture-5	0.48	**0.55**	**0.43**	**1.14**	**1.18**	**0.92**

Notes: 5% and 10% refers to the top 5% (10) and 10% (20) unsafe sites, respectively. Bold text indicates superior performance.

UNCORRECTED

This suggests that the Base and mixture models with different temporal trends have equivalent competency at site ranking based on the Bayesian estimated crash counts. However, this approach is based on crash counts and offers very limited information regarding the sites which may exhibit higher crash variability in the future. A proactive approach may be desirable in some cases to identify the sites which show smaller risk at present but the variability from the stable space-time pattern may indicate that they are prone to more crashes in the future. Such approach may be complimentary to the traditional one and a collaborative approach may be developed which also identifies the perceived high-risk sites. The mixture models offer such possibility as they isolate the spatial and temporal correlations and the space-time variability may help identify the sites showing higher probability of crash risk based on departure from global space-time pattern. The lower portion of Table 5 illustrates the SCT results based on the standard deviation of v_{it}, which represents the spatiotemporal variability. In this case, the higher value of SCT is desirable as it indicates that the model identified such sites which had lower crash counts in the before period, but the crash risk increased over time. Such sites should also be included in the process of crash mitigation as it is a proactive approach which compliments the traditional one. As depicted by the SCT results, Mixture-5 has higher standard deviation in five out of six cases. This finding indicates that Mixture-5 models offer remarkable superiority in the TAZs crash dataset of this study to identify the sites which may have apparent crash risk, but the underlying risk may be more evident in future.

The authors also employed the Method Consistency Test to quantify the number of sites which each model identifies for both before and after periods. A larger MCT value indicates that the models are able to identify more sites for each time period. This suggests superior site ranking performance as the model depicts consistency to select same sites, assuming the underlying crash risk remains constant. To restate, 5% sites represents top 10 unsafe sites while 10% represents top 20 unsafe sites. As shown in Table 6, the model pair of Mixture-2 and Mixture-5 consistently identify the largest number of common sites across both time periods. Their consistency is more pronounced for the After period as both models share 7 out of 10 and 18 out of 20 sites which is significantly larger than the MCT with Base model. These findings further corroborate the previous findings to suggest that the difference in the fundamental approach between Base and Mixture models leads to varying site ranking performance, but such differences may be exploited by employing Mixture models in collaboration with Base model to compliment the traditional site ranking approach.

The difference in the crash risk based on Bayesian estimated crash count and ind values may be better illustrated using a heatmap. As shown in Fig. 2, the four different heatmaps depict the similarities and differences on assignment of crash risk to sites based on ind values for mixture models, Bayesian estimated crash counts for Base model, and the observed crash count. For illustrative purpose, the heatmaps are developed based on the best models per the evaluation criterion, LPML, which is the Model 3 in the After period. The average values of ind (and other counts) across five years of After period are used for generating the maps. However, such comparison may also be conducted based on individual years, or based on the latest year's data, depending on the objective of study. As depicted in Fig. 2, Mixture-5 and Mixtire-2 models' ranking results are relatively similar. On the other hand, theta and count also seem to be similar in designating the crash risk to sites. This difference clearly shows that the mixture models deviate from traditional site ranking, and rightly so, as they offer another perspective by identifying sites which may tend to have lower crash counts for a given period but their deviation from rest of the sites based on space-time profile should be incorporated for site ranking.

Table 6
Site Ranking Results based on MCT.

MCT						
	Before			After		
	Base	Mixture-2	Mixture-5	Base	Mixture-2	Mixture-5
Base		3	3		3	3
Mixture-2	2	**11**		1		**18**
Mixture-5	1	**5**		1	**7**	

Notes: the number below the main diagonals represents the number of common zones identified by each pair of models for the case of top 5% zones, while the number above the main diagonal represents the one for top 10% situations.

4. Conclusions and recommendations

This study compares the mixture models for space-time interaction with the traditional one in the safety research. The models with the mixture component allow the accommodation of global space-time patterns, which are referred as "stable", while also capturing the atypical departures from the stable pattern which may not be captured by the conventional space-time models in safety research. This study developed two mixture models, namely Mixture-2 and Mixture-5, where the numerical values represents the number of levels. The larger the mixture levels, the subtler categories of the space-time variability. It is important to note that the 5 levels were selected merely for illustration purpose. Technically, any levels could be used based on the research need. A Base model was also developed to accommodate the space and time terms without mixture. Following conclusions were drawn from the model comparison results:

1 The posterior parameter estimates conveyed the robustness of the developed models across the three types (Base and two Mixture models) based on the closeness of coefficients. Conversely, the coefficients were noted to vary in terms of temporal treatments which suggested the presence of discrepancy regarding the temporal impact of independent terms on crash risk. Overall, the model estimates results indicated the presence of temporal instability which may be further explored by incorporating diverse explanatory factors. The inclusion of different temporal treatments to account for the temporal correlation and instability is vital to draw more informed inferences by generating relatively accurate posterior estimates.
2 In case of predictive accuracy based on in-sample errors (Dbar), no clear advantage of a specific temporal treatment was observed. However, in comparison between model types, the Mixture-5 demonstrated superior performance in majority of the cases, followed by Mixture-2. This finding quantified the advantage of mixture approach to accurately predict crashes, as assessed by in-sample errors.
3 DIC values for alternative models were also recorded as DIC represents the overall goodness-of-fit and is the indirect measure of model performance based on test errors. In terms of temporal treatments, AR-1 specification demonstrated superior performance. For model comparison across types, the Base model recorded the lowest DIC. Overall, it was observed that DIC values were largely governed by the model complexity which penalizes the models for addition of effective parameters.
4 To reflect the predictive accuracy based on out-of-sample errors, LPML was also calculated based on leave-one-out cross-validation approach and. In terms of temporal treatments, the AR-1 models were noted to be superior for four out of six cases, indicating the transferability of the advantage of this specification from training errors (in-sample) to test errors (or out-of-sample). In terms of model performance across different types, the Mixture-5 dominated in all

UNCORRECTED PROOF

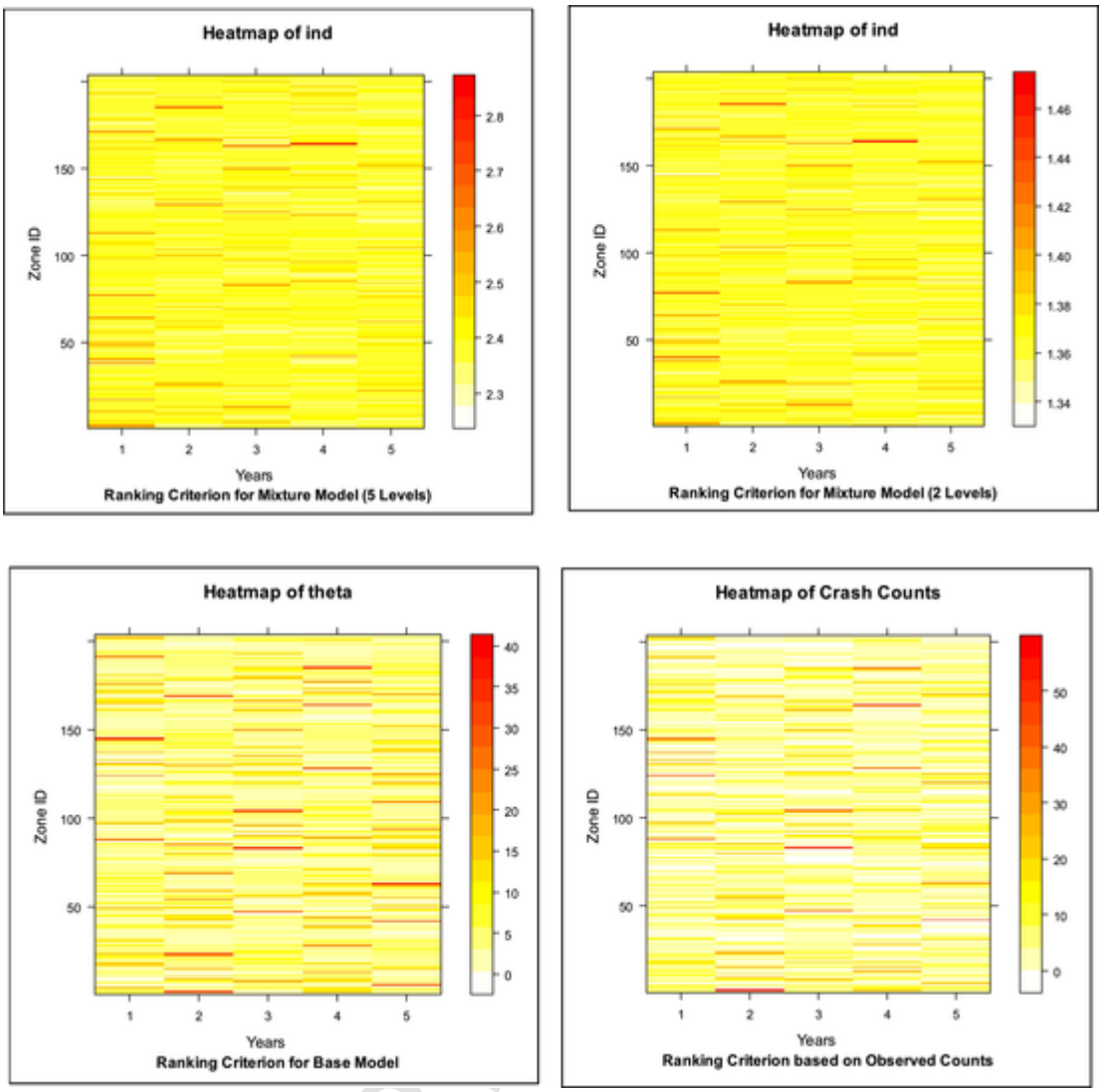

Fig. 2. Heatmaps of various ranking criteria under different models.Notes: ind represents probability of crash risk based on space-time variability and theta represents the crash risk based on Bayesian estimated crash count.

cases, further establishing the superiority of mixture approach. This finding demonstrating Mixture-5′s superiority on cross-validation suggests that this model may be suitable for safety agencies which implement crash prediction models and expect a superior performance under different areas of focus.

5 One of the primary objectives of this study was to explore the possibility of adopting the mixture models as a complimentary approach to traditional site ranking. Different evaluation criteria were employed to assess the site ranking performance. In terms of total rank difference (TRD) results based on estimated crash counts for Base model and probability of space-time variability (ind) for mixture models, there were several discrepancies between the two approaches. This finding established the fundamental difference between traditional and mixture models in designating crash risk. Mixture-5 was superior at site ranking compared to Mixture-2, indicating the advantage of the more flexible approach of larger components.

6 Another site ranking criterion (site consistency test (SCT)) was employed to explore the difference in identification of unsafe sites based on two datasets: estimated crash count and the spatial variability across time. The advantage of mixture models to act as a complimentary approach for site ranking was revealed by the spatial variability SCT results as the mixture models were able to identify the sites which experienced smaller crash counts previously but witnessed higher crash risk in future. This finding offers support for implementing the mixture models as a pro-active approach employed in collaboration with traditional site ranking based on estimated crash counts (whose SCT results offered negligible insights). Such models may prove helpful in the network screening step of safety management process to identify sites which may turn unsafe in future and escape the detection from traditional methods.

7 Overall, the research findings in most cases reveal that the mixture models outperform the traditional ones in both crash prediction and site ranking. As discussed before, such phenomenon indicates the importance of explicitly considering some specific zones where the space-time interactions deviate from the global spatiotemporal pattern. Thus far, most of the existing literature focuses on the models relying on the relationship between covariates and response variables, with very few investigating the variability in the unobserved heterogeneity beyond the general space-time trend. Therefore, it is highly recommended that additional studies using different observed or simulated data be performed to verify the benefits associated with the mixture models as shown in the present study.

Albeit the present study conducts a comprehensive comparison of mixture models from different perspectives, the results observed in this paper require some caveats. First, the planning level data from the TAZs of city of Irvine were used to develop models. Future studies may explore other macro-levels (such as census tracts or counties) or smaller scale roadway entities (such as intersections or segments) to verify and compare the findings of this study. Second, it was clearly stated in the methodology section that this study deliberately avoided the inclusion of explanatory variables as they may have blurred the comparison between concerned models. It would be interesting to see if the findings of this study hold true if the models incorporated diverse covariates as some factors (e.g., exposure) may address the spatial and temporal aspects of crash data, hence influencing the posterior estimates and maybe subsequently model performance. Finally, previous literature has demonstrated there is an obvious relationship between frequency and severity (Kweon and Kockelman, 2003; Malyshkina and Mannering, 2008). Even though the proposed approach represented an alternative efficient way of handling the variability of crash counts within a relatively long time interval, it is worth investigating the adaptation of the approach into injury-severity analysis and evaluating the pertinent performance in the near future.

Acknowledgments

Support for this study was provided by California Department of Transportation (Caltrans, No. 65A0705). The authors would also thank Southern California Association of Governments for providing covariates, shape files and geometric centroid distance data of TAZ used in this study.

References

Abdel-Aty, M., Wang, X., 2006. Crash estimation at signalized intersections along corridors: analyzing spatial effect and identifying significant factors. Transp. Res. Rec.: J. Transp. Res. Board 1953, 98–111.

Abellan, J.J., Richardson, S., Best, N., 2008. Use of space-time models to investigate the stability of patterns of disease. Environ. Health Perspect. 116 (8), 1111.

Aguero-Valverde, J., Jovanis, P.P., 2006. Spatial analysis of fatal and injury crashes in Pennsylvania. Accid. Anal. Prev. 38 (3), 618–625.

Aguero-Valverde, J., Jovanis, P., 2008. Analysis of road crash frequency with spatial models. Transp. Res. Rec.: J. Transp. Res. Board 2061, 55–63.

Aguero-Valverde, J., Jovanis, P.P., 2009. Bayesian multivariate Poisson lognormal models for crash severity modeling and site ranking. Transp. Res. Rec. 2136 (1), 82–91.

Aguero-Valverde, J., Jovanis, P., 2010. Spatial correlation in multilevel crash frequency models: effects of different neighboring structures. Transp. Res. Rec.: J. Transp. Res. Board 2165, 21–32.

Aguero-Valverde, J., Wu, K.F.K., Donnell, E.T., 2016. A multivariate spatial crash frequency model for identifying sites with promise based on crash types. Accid. Anal. Prev. 87, 8–16.

Andrey, J., Yagar, S., 1993. A temporal analysis of rain-related crash risk. Accid. Anal. Prev. 25 (4), 465–472.

Bernardinelli, L., Clayton, D., Pascutto, C., Montomoli, C., Ghislandi, M., Songini, M., 1995. Bayesian analysis of space—time variation in disease risk. Stat. Med. 14 (21–22), 2433–2443.

Cheng, W., Washington, S.P., 2005. Experimental evaluation of hotspot identification methods. Accid. Anal. Prev. 37 (5), 870–881.

Cheng, W., Washington, S., 2008. New criteria for evaluating hotspot identification methods. Transp. Res. Rec. 2083, 76–85.

Cheng, W., Gill, G.S., Dasu, R., Xie, M., Jia, X., Zhou, J., 2017. Comparison of Multivariate Poisson lognormal spatial and temporal crash models to identify hot spots of intersections based on crash types. Accid. Anal. Prev. 99, 330–341.

Cheng, W., Gill, G.S., Choi, S., Zhou, J., Jia, X., Xie, M., 2018. Comparative evaluation of temporal correlation treatment in crash frequency modelling. Transp. A: Transp. Sci. 1–19.

Cheng, W., Gill, G.S., Zhang, Y., Cao, Z., 2018. Bayesian spatiotemporal crash frequency models with mixture components for space-time interactions. Accid. Anal. Prev. 112, 84–93.

Deacon, J.A., Zegeer, C.V., Deen, R.C., 1975. Identification of hazardous rural highway locations. Transp. Res. Rec.: J. Transp. Res. Board 543, 16–33.

Dong, N., Huang, H., Lee, J., Gao, M., Abdel-Aty, M., 2016. Macroscopic hotspots identification: a Bayesian spatio-temporal interaction approach. Accid. Anal. Prev. 92, 256–264.

El-Basyouny, K., Barua, S., Islam, M.T., 2014. Investigation of time and weather effects on crash types using full Bayesian multivariate Poisson lognormal models. Accid. Anal. Prev. 73, 91–99.

Gelfand, A.E., 1996. Model determination using sampling-based methods. In: Markov Chain Monte Carlo in Practice. pp. 145–161.

Gill, G.S., Cheng, W., Xie, M., Vo, T., Jia, X., Zhou, J., 2017. Evaluating influence of neighboring structures on spatial crash frequency modeling and site-ranking performance. Transp. Res. Rec.: J. Transp. Res. Board 2659, 117–126.

Gill, G.S., Cheng, W., Zhou, J., Jia, X., 2018. Comprehensive assessment of temporal treatments in crash prediction models. Transp. Res. Rec. 2672 (38), 93–104.

Geedipally, S.R., Lord, D., 2010. Identifying hot spots by modeling single-vehicle and multivehicle crashes separately. Transp. Res. Rec. 2147 (1), 97–104.

Hauer, E., 1997. Observational before/after Studies in Road Safety. Estimating the Effect of Highway and Traffic Engineering Measures on Road Safety. Transportation Research Board, Washington, D.C..

Hauer, E., Allery, B., Kononov, J., Griffith, M., 2004. How best to rank sites with promise. Transp. Res. Rec.: J. Transp. Res. Board 1897, 48–54.

Heydari, S., Fu, L., Joseph, L., Miranda-Moreno, L.F., 2016. Bayesian nonparametric modeling in transportation safety studies: applications in univariate and multivariate settings. Anal. Methods Accid. Res. 12, 18–34.

Huang, H., Chin, H., Haque, M., 2009. Empirical evaluation of alternative approaches in identifying crash hot spots: naive ranking, empirical Bayes, and full Bayes methods. Transp. Res. Rec.: J. Transp. Res. Board 2103, 32–41.

Jiang, X.M., Abdel-Aty, M., Alamili, S., 2014. Application of Poisson random effect models for highway network screening. Accid. Anal. Prev. 63, 74–82.

Kweon, Y.J., Kockelman, K.M., 2003. Overall injury risk to different drivers: combining exposure, frequency, and severity models. Accid. Anal. Prev. 35 (4), 441–450.

Lan, B., Persaud, B., Lyon, C., Bhim, R., 2009. Validation of a full Bayes methodology for observational before–after road safety studies and application to evaluation of rural signal conversions. Accid. Anal. Prev. 41 (3), 574–580.

Lawson, A.B., Browne, W.J., Rodeiro, C.L., 2003. Disease Mapping With WinBUGS and MLwiN. John Wiley and Sons, p. 11.

Lord, D., Miranda-Moreno, L., 2008. Effects of low sample mean values and small sample size on the estimation of the fixed dispersion parameter of Poisson-gamma models for modeling motor vehicle crashes: a Bayesian perspective. Saf. Sci. 46 (5), 751–770.

Lord, D., Mannering, F., 2010. The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives. Transp. Res. Part A: Policy Pract. 44 (5), 291–305.

Lunn, D.J., Thomas, A., Best, N., Spiegelhalter, D., 2000. WinBUGS – a Bayesian modelling framework: concepts, structure, and extensibility. Stat. Comput. 10 (4), 325–337.

Malyshkina, N.V., Mannering, F., 2008. Effect of increases in speed limits on severities of injuries in accidents. Transp. Res. Rec. 2083 (January 1), 122–127.

Mannering, F., Shankar, V., Bhat, C.R., 2016. Unobserved heterogeneity and statistical analysis of highway accident data. Anal. Methods Accid. Res. 11, 1–16.

Mannering, F., 2018. Temporal instability and the analysis of highway accident data. Anal. Methods Accid. Res. 17, 1–13.

Miaou, S., Lord, D., 2003. Modeling traffic crash-flow relationships for intersections: dispersion parameter, functional form, and bayes versus empirical bayes methods. Transp. Res. Rec.: J. Transp. Res. Board 1840, 31–40.

Miaou, S., Song, J.J., Mallick, B.K., 2003. Roadway traffic crash mapping: a space–time modeling approach. J. Transp. Stat. 6 (1), 33–57.

UNCORRECTED PROOF

Miaou, S.P., Song, J.J., 2005. Bayesian ranking of sites for engineering safety improvements: decision parameter, treatability concept, statistical criterion, and spatial dependence. Accid. Anal. Prev. 37 (4), 699–720.
Montella, A., 2010. A comparative analysis of hotspot identification methods. Accid. Anal. Prev. 42 (2), 571–581.
Norden, M., Orlansky, J., Jacobs, H., 1956. Application of statistical quality-control techniques to analysis of highway-accident data. Highw. Res. Board Bull. 117, 17–31.
Ntzoufras, I., 2009. Bayesian Modeling Using WinBUGS. John Wiley and Sons.
Park, E.S., Lord, D., 2007. Multivariate Poisson-lognormal models for jointly modeling crash frequency by severity. Transp. Res. Rec. 2019 (1), 1–6.
Persaud, B., Lan, B., Lyon, C., Bhim, R., 2010. Comparison of empirical Bayes and full Bayes approaches for before–after road safety evaluations. Accid. Anal. Prev. 42 (1), 38–43.
Richardson, S., Abellan, J.J., Best, N., 2006. Bayesian spatio-temporal analysis of joint patterns of male and female lung cancer risks in Yorkshire (UK). Stat. Methods Med. Res. 15 (4), 385–407.
Sacchi, E., Sayed, T., El-Basyouny, K., 2015. Multivariate full Bayesian hot spot identification and ranking: new technique. Transp. Res. Rec.: J. Transp. Res. Board 2515, 1–9.
D. SpiegelhalterA. ThomasN. BestD. LunnWinBUGS Version 1.4 User ManualMRC Biostatistics UnitCambridgehttp://www.mrc-cam.ac.uk/bugs2003
Stokes, R., Mutabazi, M., 1996. Rate-quality control method of identifying hazardous road locations. Transp. Res. Rec.: J. Transp. Res. Board 1542, 44–48.
Tunaru, R., 2002. Hierarchical Bayesian models for multiple count data. Aust. J. Stat. 31 (3), 221–229.
Xu, P., Huang, H., 2015. Modeling crash spatial heterogeneity: random parameter versus geographically weighting. Accid. Anal. Prev. 75, 16–25.
Wang, X., Abdel-Aty, M., 2006. Temporal and spatial analyses of rear-end crashes at signalized intersections. Accid. Anal. Prev. 38 (6), 1137–1150.
Wang, C., Quddus, M.A., Ison, S.G., 2011. Predicting accident frequency at their severity levels and its application in site ranking using a two-stage mixed multivariate model. Accid. Anal. Prev. 43 (6), 1979–1990.
Washington, S., Haque, M.M., Oh, J., Lee, D., 2014. Applying quantile regression for modeling equivalent property damage only crashes to identify accident blackspots. Accid. Anal. Prev. 66, 136–146.

UNCORRECTED PROOF