Accident Analysis and Prevention 43 (2011) 1979–1990

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

Predicting accident frequency at their severity levels and its application in site ranking using a two-stage mixed multivariate model

Chao Wang*, Mohammed A. Quddus, Stephen G. Ison

Transport Studies Group, Department of Civil and Building Engineering, Loughborough University, Loughborough, Leicestershire, LE11 3TU, United Kingdom

ARTICLE INFO

Article history:
Received 3 December 2010
Received in revised form 12 May 2011
Accepted 15 May 2011

Keywords:
Road accidents
Site ranking
Bayesian spatial model
Mixed logit model
Two-stage mixed multivariate model

ABSTRACT

Accident prediction models (APMs) have been extensively used in site ranking with the objective of identifying accident hotspots. Previously this has been achieved by using a univariate count data or a multivariate count data model (e.g. multivariate Poisson-lognormal) for modelling the number of accidents at different severity levels simultaneously. This paper proposes an alternative method to estimate accident frequency at different severity levels, namely the two-stage mixed multivariate model which combines both accident frequency and severity models. The accident, traffic and road characteristics data from the M25 motorway and surrounding major roads in England have been collected to demonstrate the use of the two-stage model. A Bayesian spatial model and a mixed logit model have been employed at each stage for accident frequency and severity analysis respectively, and the results combined to produce estimation of the number of accidents at different severity levels. Based on the results from the two-stage model, the accident hotspots on the M25 and surround have been identified. The ranking result using the two-stage model has also been compared with other ranking methods, such as the naïve ranking method, multivariate Poisson-lognormal and fixed proportion method. Compared to the traditional frequency based analysis, the two-stage model has the advantage in that it utilises more detailed individual accident level data and is able to predict low frequency accidents (such as fatal accidents). Therefore, the two-stage mixed multivariate model is a promising tool in predicting accident frequency according to their severity levels and site ranking.

1. Introduction

Accident prediction models (APMs) are widely used to estimate the frequency of accidents for a given spatial unit over a certain period of time. One of the important practical applications of APMs is site ranking which aims to identify hazardous sites or locations with underlying safety problems. Site ranking is essential in designing engineering programmes to improve safety of a road network. After identification of accident hotspots, necessary engineering improvements could be applied to the selected sites with limited highway funds. This improves road safety and ensures cost-effectiveness in resource allocation. APMs are required in site ranking given the regression-to-the-mean problem as accidents are rare and random events (Elvik, 2007; Persaud and Lyon, 2007). Site ranking is also referred to as network screening (Persaud et al., 2010); and the sites with potential for safety treatments are also known as sites with promise, accident black-spots or hotspots in the literature (Hauer et al., 2004; Maher and Mountain, 1988; Elvik, 2007; Cheng and Washington, 2005; Huang et al., 2009). The terms "site ranking" and "accident hotspots" are used in this paper for consistency.

Accident data are often provided with classification according to the accident types (e.g. head-on; rear-end) or severities (e.g. fatal, serious and slight). It is particularly important to take into account accident severities in site ranking, because the cost of accidents could be hugely different at different severity levels. This means that, for instance, a road segment with higher frequency of fatal accidents may be considered more hazardous than a road segment with fewer fatal accidents but more serious or slight injury accidents, therefore it is necessary to estimate accident frequency for each severity category. A straightforward and traditional approach to this problem is to apply an accident frequency model on different types of accidents separately (i.e. a univariate modelling approach). For example, Noland and Quddus (2005) disaggregated road casualties into three categories by their severity levels – i.e. fatalities, serious injuries and slight injuries, and they applied negative binomial (NB) models on each category of road casualties separately, resulting in three independent univariate models.

Recently researchers have explored the multivariate modelling approach which can model the number of different types of accidents simultaneously (instead of separately). Several multivariate

* Corresponding author. Tel.: +44 0 1509 564682; fax: +44 0 1509 223981.
E-mail addresses: c.wang@lboro.ac.uk, excelwang@gmail.com (C. Wang), m.a.quddus@lboro.ac.uk (M.A. Quddus), s.g.ison@lboro.ac.uk (S.G. Ison).

0001-4575/$ – see front matter.
doi:10.1016/j.aap.2011.05.016

models have been employed such as multivariate spatial models (Song, 2004; Song et al., 2006), multivariate Poisson (MVP) models (Ma and Kockelman, 2006), and multivariate Poisson-lognormal (MVPLN) models (Park and Lord, 2007; Ma et al., 2008; Aguero-Valverde and Jovanis, 2009; El-Basyouny and Sayed, 2009). Compared to the univariate modelling approach, the multivariate models (i.e. MVP or MVPLN) are argued to be superior since multivariate models can take account of correlation between different types of accidents, or in other words to "borrow strength" from similar sources (Song et al., 2006). However, as pointed out by Ma et al. (2008), the superiority of the multivariate models compared to univariate models is not "theoretical" but rather "empirical". By comparing several Poisson based models using both the multivariate and univariate approach, Lan and Persaud (2010) found that univariate models fit the data better and outperform the multivariate models, and thus univariate models were recommended. Another limitation of the classical multivariate regression is that the same set of explanatory variables is required for each type of response (Frees, 2004). This is a concern as factors affecting one type of accidents may have no effect on the other. Accident data also suffers from an under-reporting problem, especially for less serious accidents such as slight injury accidents. This means that the data qualities of different types of accident vary, and thus different types of accident may more suitably be modelled separately.

This paper proposes an alternative method to estimate accident frequency at different severity levels. Accidents are, essentially *mutually exclusive* and *collectively exhaustive* events. In other words, an accident is in, and can only be in, one category of different severities (i.e. either fatal or serious or slight). Such data involving two types of discrete outcomes (i.e. count and discrete choice) can be modelled using a *mixed multivariate model* (Cameron and Trivedi, 1998). There are several approaches for estimating a mixed multivariate model, for instance a mixed multinomial (logit) Poisson model, or alternatively simply estimating the Poisson based models for each category of events independently. These two approaches are equivalent (Cameron and Trivedi, 1998). Another approach of estimating a mixed multivariate model is using a two-stage model, in which count data models (e.g., a NB regression) and discrete choice models (e.g., a multinomial logit regression) are estimated in two stages (Cameron and Trivedi, 1998). While this modelling approach appears to be less used by safety researchers, it has been employed by Hausman et al. (1995) in modelling the number of trips to alternative recreational sites, in which the model was referred to as a "combined discrete choice and count data model".

This paper develops and presents the two-stage mixed multivariate model in accident prediction and its application to site ranking. It should be noted that several road safety researchers have proposed a similar approach. For instance, Milton et al. (2008) used a mixed logit model to assess severity distribution of accidents on road segment and pointed out the possibility of combining the severity model with the frequency model. Geedipally et al. (2010) employed a multinomial logit (MNL) model to estimate the proportions of different types of accidents and a NB model to estimate the total number of accidents on a road segment. As such, the counts of various types of accidents could be determined. Both the studies by Milton et al. (2008) and Geedipally et al. (2010) were however based on the road segment level. In other words, the proportions of types of accidents on a road segment were directly estimated in their studies. This paper differs in the sense that the proportions of accidents on a road segment were estimated using a model at an *individual accident level*. This approach has certain advantages over the road segment level estimation which is discussed below.

The paper is organised as follows: firstly the methodology employed in this paper is described. This includes both accident frequency and severity models that are used in the two-stage model and site ranking. It is then followed by the description of the data and the results of the two-stage model and site ranking. Discussion is then provided and finally conclusions are drawn.

2. Methodology

As discussed above, accident data involving two types of discrete outcomes (i.e. count of accident and discrete choice of accident severity) are analysed using a two-stage mixed multivariate model. In the two-stage model, a count (accident frequency) model is used to estimate the total number of events (accidents); and then a discrete choice (accident severity) model is used to "allocate" these events (accidents) into different categories (severities) (Cameron and Trivedi, 1998). The accident frequency and severity models used in each of the stages are discussed below.

2.1. Accident frequency model

Several models that are suitable for count data have been considered. Negative binomial (NB) models are among those popular types of models employed to estimate accident frequency (see Lord, 2000; Ivan et al., 2000; Graham and Glaister, 2003; Noland and Quddus, 2005). Empirical Bayes (EB) method which utilises NB models has been successfully used in identifying accident hotspots (Elvik, 2007). The EB method however is allegedly using the data twice and inadequate to account for all uncertainties associated with road accidents and their contributing factors (Huang et al., 2009). Recently more advanced models have been developed such as full Bayesian spatial models (Miaou et al., 2003; Aguero-Valverde and Jovanis, 2006; Quddus, 2008; Wang et al., 2009). The full Bayesian method has also been used in site ranking (e.g. Miaou and Song, 2005) and it has been shown to outperform the EB method (Huang et al., 2009). This paper adopts full Bayesian spatial models that controls for spatial correlation. The model can be expressed as follows:

$$Y_{it} \sim \text{Poisson}(\mu_{it}) \tag{1}$$

$$\log(\mu_{it}) = \alpha + \boldsymbol{\beta}\mathbf{X}_{it} + v_i + u_i + \delta_t + e_{it} \tag{2}$$

where Y_{it} is the annual number of observed accidents that occurred on a road segment i at year t; μ_{it} is the expected accident count on a road segment i at year t; α is the intercept; $\mathbf{X}_{it}$ is the vector of explanatory variables for a road segment i at year t; $\boldsymbol{\beta}$ is the vector of coefficients to be estimated; v_i is a random term which captures the heterogeneity effects for road segment i; u_i is a random term which captures the spatially correlated effects for neighbouring road segment i and assumed a conditional autoregressive (CAR) prior[1]; δ_t is the term representing time effects (i.e. year-to-year effects); e_{it} is a random term for extra space-time interaction effects.

Models can be estimated using the Markov chain Monte Carlo (MCMC) method under the full hierarchical Bayesian framework using the software – WinBUGS (Spiegelhalter et al., 2003).[2] The

[1] This random term controls for the potential spatial correlation which may be due to unobserved similar traffic, infrastructure or environment conditions among neighbouring road segments. As detailed by Quddus (2008) and Wang et al. (2009), $u_i|u_j, i \neq j \sim N\left(\frac{\sum_j u_j w_{ij}}{w_{i+}}, \frac{\tau_u^2}{w_{i+}}\right)$, where w_{ij} denotes the weight between road segment i and j; $w_{ij} = 1$ if road segment i and j are adjacent (i.e. shared vertex) and $w_{ij} = 0$ otherwise; $w_{i+} = \sum_j w_{ij}$; and τ_u^2 is a scale parameter assumed as a gamma prior.

[2] Readers who are interested in the details of the model specification (e.g. prior distributions) are directed to Wang et al. (in press). Generally non-informative priors were used.

deviance information criterion (DIC), which can be thought of as a generalisation of the Akaike information criterion (AIC), can be used to compare goodness-of-fit and complexity of different models estimated under a Bayesian framework (Spiegelhalter et al., 2002). As with AIC, in terms of model fit and complexity, the lower the DIC the better the model.

2.2. Accident severity model

Accident severity is often measured categorically, for instance, the severity level of an accident can be classified as fatal, serious injury, slight injury or no injury (property damage only). As such, statistical models that are suitable for categorical data, such as logistic and probit models, have been used to analyse accident severities.

Since the accident severity is ordered in nature (ranging from non-injury to fatality), it seems natural to choose discrete ordered response models (such as ordered logit and probit models) for analysing accident severity data. Examples of previous studies utilising ordered response models include O'Donnell and Connor (1996), Eluru et al. (2008) and Quddus et al. (2010). However, as discussed in Kim et al. (2007), Savolainen and Mannering (2007) and Yamamoto et al. (2008), ordered response models have two limitations which are related to the constraint on the variable influence (e.g. a variable would either increase or decrease accident severity) and under-reporting, especially for low severity levels in accident data. This led to the use of alternative and more flexible unordered nominal response models such as multinomial logit (MNL) models. Compared to ordered response models, unordered nominal response models offer more flexibility in terms of the functional form and consistent coefficient estimates with under-reporting data (Kim et al., 2007; Savolainen and Mannering, 2007).

This paper adopts the unordered nominal response models for analysing accident severity. Two types of such models were considered: a standard MNL model and a mixed logit model. The MNL model has been widely used in previous research (e.g. Shankar and Mannering, 1996; Kim et al., 2007). The MNL model can be written as (Long and Freese, 2006):

$$Pr(y_n = j) = \frac{\exp(\boldsymbol{\beta}_{j|b}\mathbf{X}_n)}{\sum_{m=1}^{M}\exp(\boldsymbol{\beta}_{m|b}\mathbf{X}_n)}, j = 1, 2, 3\cdots M \quad (3)$$

where $\mathbf{X}_n$ is a vector of explanatory variables related to accident n; b is the base outcome that other severity outcomes (j) are compared with; $\boldsymbol{\beta}_{j|b}$ is a vector of injury-specific coefficients and $\boldsymbol{\beta}_{b|b} = 0$; m indicates a certain category of accident severity. In this paper, the observed accident severity y is coded as follows: 1 = slight injury accident; 2 = serious injury accident; and 3 = fatal accident.

One potential problem of a MNL model is that it assumes that the unobserved components (effects) associated with each accident severity category are independent, which is referred to as the *independence of irrelevant alternatives* (IIA) property (Train, 2003). If the IIA assumption is violated, i.e. different accident severity categories share unobserved effects, the model estimation results would be incorrect. Previous research has shown that accident severity types may be correlated (i.e. sharing unobserved effects) (Milton et al., 2008). To circumvent this limitation, a more generalised modelling approach has been proposed by adding a more general mixing distribution of error component to the model. This model, which is referred to as the mixed logit model, is flexible and powerful. It can accommodate complex patterns of correlation among accident severity outcomes and unobserved heterogeneity (Train, 2003; Milton et al., 2008). The mixed logit can be expressed as follows:

$$Pr(y_n = j) = \int \frac{\exp(\boldsymbol{\beta}_{j|b}\mathbf{X}_n)}{\sum_{m=1}^{M}\exp(\boldsymbol{\beta}_{m|b}\mathbf{X}_n)} f(\boldsymbol{\beta})d\boldsymbol{\beta}, j = 1, 2, 3\cdots M \quad (4)$$

where $f(\boldsymbol{\beta})$ is a density function.

The mixed logit probability is then a weighted average with weights given by $f(\boldsymbol{\beta})$. Some parameters of the vector $\boldsymbol{\beta}$ may be fixed or randomly distributed. The standard MNL model is a special case of the mixed logit model when $\boldsymbol{\beta}$ are fixed parameters. For random parameters, the coefficients $\boldsymbol{\beta}$ are allowed to vary over different accidents and assumed randomly distributed. In this paper the random coefficients are specified to be normally distributed, e.g. $\beta_1 \sim N(b,W)$ where b is the mean and W is the variance.

The MNL model can be estimated using the standard maximum likelihood method. The estimation of mixed logit models however is difficult as the probability function is involved with integration and hence is not in a closed form. One solution is to use the maximum simulated likelihood (MSL) method in which Halton draws[3] can be used to achieve convergence more efficiently (Bhat, 2003; Train, 2003). MSL is also shown to be more efficient to achieve the same degree of accuracy than other estimation methods such as the classical Gauss-Hermite quadrature or adaptive quadrature (Haan and Uhlendorff, 2006). In this paper the mixed logit model is estimated using a user written Stata program (–mixlogit–) developed by Hole (2007). The Akaike information criterion (AIC) are used to compare goodness-of-fit and complexity of MNL and mixed logit model.

2.3. Predicting accident frequency at different severity levels and site ranking

The two-stage model combines results from both the accident frequency model and accident severity model described above. At the first stage, the total number of accidents on a road segment for a given year is estimated using an accident frequency model (the full Bayesian spatial model in this paper). Then at the second stage, the expected proportions of accidents at different severity levels on a road segment for a given year is estimated using an accident severity model (the MNL and mixed logit models in this paper), which then 'allocates' the number of accidents to different severity levels. Finally, the number of accidents at different severity levels can be obtained. The proportions of each accident category can be obtained by aggregating the predicted probabilities for each severity category across all individual accidents on a road segment for a given year. Suppose there are a number of N accidents on a road segment for a given year, and P_{nj} is the predicted probability of accident n at severity level j, then the proportion of accidents for severity j on this road segment for the given year is:

$$\hat{S}(j) = \frac{1}{N}\sum_{n=1}^{N} P_{nj} \quad (5)$$

where $\hat{S}(j)$ is the predicted proportion of accident for severity j.

Note that as mentioned by Geedipally et al. (2010), proportions of different types of accidents can also be assumed fixed and directly calculated from the observed data, rather than estimated from an accident severity model. This method is referred to as the "fixed proportion method" (Geedipally et al., 2010) and will be compared

[3] Halton draws are generated from number theory to create a sequence of quasi-random numbers, which is generally more efficient to compute integrals compared to a purely random sequence (see Train, 2003).

with the two-stage mixed multivariate model described in this study.

The results from both the accident frequency and severity models can then be combined to estimate the number of accidents at each severity level. The accuracy of the two-stage model through goodness-of-fit can be determined by a number of statistics such as mean absolute deviation (MAD), and mean squared error (MSE). For example, Oh et al. (2003) and Xie et al. (2007) employed the MAD statistics and indicated that a lower MAD characterises a better model in term of predicting accuracy. After obtaining the expected number of accidents at each severity level, road segments can then be ranked by an appropriate decision parameter (Θ) for further engineering examination and treatment. The choice of decision parameter (Θ) depends on the context under which the rank is to be used, especially the range of safety treatments to be implemented (Miaou and Song, 2005). Therefore, inputs from decision makers can be useful, for their interests can be taken into consideration for ranking. Since accident data used in this paper are classified into different categories according to their severity levels, monetary costs of accidents are used as an illustration. The decision parameter Θ_i in this paper is defined as the total accident cost per vehicle-kilometre for road segment i:

$$\Theta_i = \frac{\sum_t \sum_j \text{cost}_j \hat{\mu}_{ijt}}{365 \times \textit{length}_i \times \sum_t AADT_{it}} \quad (6)$$

where cost_j is the monetary cost of an accident at severity level j; $\hat{\mu}_{ijt}$ is the posterior estimate of count of accidents at severity level j on road segment i at time (year) t, estimated from the two-stage model; $\textit{length}_i$ is the length of road segment i; $AADT_{it}$ is the annual average daily traffic on road segment i at time (year) t.

The decision parameter (Θ_i) above provides a direct measurement of expected accident cost rate for the time period of interest. A road segment with higher expected accident cost per vehicle-kilometre is considered more hazardous, and thus is ranked higher as an accident hotspot for further safety treatment.

3. Data description

To demonstrate the applicability of the two-stage model, relevant data have been collected from the M25 motorway and its surrounding major roads (other motorways and A roads). The M25 motorway is an orbital motorway that encircles London, England.

Traffic and road infrastructure data were obtained from the UK Highways Agency (HA). The HA collects hourly traffic characteristics and road infrastructure data for major motorways and A roads at the road segment level (a road segment is a stretch of road that starts or ends at a junction and has one direction[4]) in the UK. The data obtained include the hourly traffic characteristics data for road segments on the M25 and surround during the years 2003–2007, including traffic flow, average travel time, and total vehicle delay.[5] Road infrastructure data such as segment length, number of lanes, radius of curvature and gradient[6] for each road segment have also been obtained.

Accident data for the years 2003–2007 were derived from the STATS19 UK national road accident database. The database contains information on the direction of the vehicles just before an accident, and this information has been used to match the accidents onto the correct road segments using the method described in Wang et al. (2009). Only accidents recorded as occurring on the M25 motorway and surround are retained. Accidents coded as junction accidents (around 30% of total accidents within the study area) in the STATS19 database were excluded from the analysis. This is because major road junctions are complicated in terms of road design (such as fly-overs and slip roads) compared to road segments and it is also difficult to obtain a single measure of traffic flow at fly-over and/or slip roads merging to and diverging from the main roads. One road segment with a minimum radius of 4.94 m is viewed as an outlier and has been excluded from the dataset. Three road segments with speed limits of less than 64.4 km/h (i.e. 40 mph) have also been excluded from the dataset since these road segments are also viewed as outliers in the context of the major road network. As such, the analysis is based on 262 road segments.

For the accident frequency analysis, counts of accidents and traffic characteristics data were aggregated at a road segment level (e.g. total traffic volume per segment per year) and eventually a panel dataset containing 262 cross-sectional observations for all road segments during a five year period was created (2003–2007).[7] Summary statistics of the accident, traffic characteristics and road infrastructure data on the M25 motorway and surround for the accident frequency models are presented in Table 1.

The total number of observations is 1310. Motorway indicator is a dummy variable with 1 representing motorway or A roads with motorway standard such as A1(M); and 0 representing other major A roads.

For the accident severity analysis, the analysis was conducted at an individual accident level rather than at a segment level. In addition to accident location and severity information, other relevant data have been derived from the STATS19 database. This includes date, time, lighting, weather conditions, number of vehicles and number of causalities for each accident. The accident data have also been integrated into traffic and road geometry data based on the information of the accident (location, time and date); and the corresponding segment-based characteristics for an accident have been obtained. As a result, traffic and road geometry data such as traffic

[4] The primary reason for employing variable segments (i.e. between two consecutive junctions) is that traffic data (e.g. traffic flow) are only available for such segmentation. The advantage of such segmentation is that the traffic is homogeneous. Similar segmentation method was used in the literature (Tanaru, 2002; Aguero-Valverde and Jovanis, 2009). Other segmentation methods may be arguably better and can be used in safety analysis such as dynamic segmentation, if the required data are available (Ogle et al., 2011). This seems not a serious issue in this study however, as suggested by El-Basyouny and Sayed (2009), the use of a spatial model (i.e. controlling for spatial correlation) could ease the issues relevant to segment selection.

[5] Delay is defined as the difference between the actual travel time and the travel time at a reference speed (often free flow speed). See DfT (2009) and Wang et al. (in press).

[6] The minimum radius and the maximum gradient for a road segment were used in the model. While this or similar measurement was used in previous studies (such as Shankar et al., 1995), this measurement has a limitation in that it cannot take into account overall curvature of a road segment. Other curvature measurements used in the literature include: number of sharp horizontal curves, sharp curve indicator (1 if curve radius is less than 868 m, 0 otherwise), bend density, detour ratio, straightness index, cumulative angle, mean angle (see Milton and Mannering, 1998; Miaou et al., 2003; Haynes et al., 2007). Another alternative measurement has also been suggested by an anonymous reviewer: $DC_{segment} = \sum_{p=1}^{k} 2\frac{l_p}{L} \sin^{-1}\left(\frac{50}{R_p}\right)$, where DC is the degree of curvature of a road segment, l_p is the length of a curved section p on the road segment, L is the total length of the road segment, R_p is the radius of the curved section p on the road segment and k is the number of total curved sections on the road segment. Generally speaking, using one single measurement alone may not be sufficient as each measurement has its limitations. As suggested by Haynes et al. (2007), "a single measure of road curvature does not capture all the properties that might be of interest". Since the purpose of this paper is not to re-investigate the effect of various measurements of curvature on road safety, minimum radius and maximum gradient were used.

[7] Due to missing values (e.g. traffic flow) for some road segments at a certain year, some road segments were removed from the original data, resulting in 262 road segments.

Table 1
Summary statistics of the variables for accident frequency analysis.

Variable	Mean	Standard deviation	Min	Max	Sum[a]
Annual number of accidents	9.062	10.022	0	97	11871
Traffic characteristics					
Annual average daily traffic (AADT)	46167.1	20616.280	5.918	98394.83	6.05E+07
Annual total vehicle delay (sec per km)	196036.7	241008.100	622.865	1900374	2.57E+08
Road segment characteristics (same direction)					
Segment length (km)	5.065	3.675	0.32	22.08	6635.7
Minimum radius (m)	681.084	364.541	20.38	2000	–
Maximum gradient (%)	3.169	1.326	0.6	8	–
Number of lanes	2.909	0.709	1	6	–
Speed limit (km/h)	110.015	6.704	77	112	–
Dummy variables					
Motorway indicator	1 = motorway (count = 915); 0 = otherwise (count = 395)				

[a] This includes 5 years' (2003–2007) data.

Table 2
Summary statistics of the variables for accident severity analysis.

Variable	Obs	Mean	Standard deviation	Min	Max
Level of accident severity[a]	12254	1.140	0.394	1	3
Traffic characteristics					
Traffic flow (veh/h)	11722	3222.276	1672.409	0	8116
Traffic delay (min per 10 km)	11936	8.374	22.252	0	751.935
Road segment infrastructure					
Minimum radius (m)	12254	729.709	292.151	20.38	2000
Maximum gradient (%)	12254	3.221	1.113	0.6	8
Speed limit (km/h)	12254	110.457	6.532	80	112
Number of casualties per accident	12254	1.606	1.187	1	42
Dummy variables					
Road segment infrastructure					
Number of lanes $\leq$ 3 indicator		1 = 3 lanes or less (count = 9768); 0 = otherwise (count = 2486)			
Number of lanes = 4 indicator (reference case)		1 = 4 lanes (count = 2090); 0 = otherwise (count = 10,164)			
Number of lanes $\geq$ 5 indicator		1 = 5 lanes or more (count = 310); 0 = otherwise (count = 11,944)			
Motorway indicator		1 = motorway (count = 10,261); 0 = A road (count = 1993)			
Environment indicators					
Lighting condition (darkness)		1 = darkness (count = 3950); 0 = daylight (count = 8304)			
Weather (fine, reference case)		1 = fine (count = 10,048); 0 = otherwise (count = 2206)			
Weather (raining)		1 = raining (count = 1742); 0 = otherwise (count = 10,512)			
Weather (snowing)		1 = snowing (count = 58); 0 = otherwise (count = 12,196)			
Other weather conditions (e.g. fog/mist)		1 = others (count = 406); 0 = otherwise (count = 11,848)			
Other factors					
Weekday indicator		1 = weekday (count = 9213); 0 = otherwise (count = 3041)			
Single vehicle accident indicator		1 = single vehicle (count = 2374); 0 = otherwise (count = 9880)			

[a] 1 = slight injury accident (count = 10,748), 2 = serious injury accident (count = 1293), 3 = fatal accident (count = 213).

flow, traffic delay and road curvature for each accident has been determined. In order to avoid the impact of an accident itself on traffic conditions, hourly traffic data corresponding to a time period that is 30 min prior to the occurrence of an accident are used. For example, if an accident happened at 15:20 then hourly traffic data for 14:00–15:00 were used.

Finally, a dataset containing various traffic, road and environment information for each accident record on the M25 and surround during 2003–2007 was established. The summary statistics of the variables for the accident severity analysis are presented in Table 2.

Table 3
Average values of prevention of road accidents (£ per accident).

	Fatal	Serious	Slight
2003	1,492,910	174,520	17,540
2004	1,573,220	184,270	18,500
2005	1,645,110	188,960	19,250
2006	1,690,370	196,020	20,120
2007	1,876,830	215,170	22,230

Source: UK Department for Transport.

As can be seen from Table 2 there were a total number of 12,254 accidents on the M25 and surround, over the period 2003–2007 with approximately 2450 accident records each year. The mean value of the accident severity variable is 1.14, meaning that the majority of accidents are slight injury accidents. To be more precise, 87.71% (10,748) of total accidents were slight injury accidents; 10.55% (1293) were serious injury accidents; and only 1.74% (213) were fatal accidents.

The monetary costs of accidents at each severity level for a given year are obtained from the UK Department for Transport (DfT, 2008),[8] which are presented in Table 3.

It is interesting to note from Table 3 that the cost of accidents increased gradually from 2003 to 2007, for all severity levels. This may reflect inflation over the years in question.

[8] According to the DfT (2008) the cost of an accident, or in other words the value of preventing an accident includes: the human costs (e.g., willingness to pay to avoid pain, grief and suffering); the direct economic costs of lost output; the medical costs associated with road accident injuries; costs of damage to vehicles and property; police costs; and administrative costs of accident insurance.

Table 4
Accident frequency model.

Variables	Mean	Standard deviation (S.D.)	95% credible sets
log(AADT)	0.124**	0.036	(0.064, 0.209)
log(segment length in m)	0.958**	0.065	(0.831, 1.084)
log(delay in sec per km)	0.043*	0.026	(−0.005, 0.097)
log(minimum radius)	0.126	0.067	(−0.032, 0.240)
Maximum gradient (%)	0.065	0.043	(−0.022, 0.142)
Number of lanes	0.436**	0.073	(0.291, 0.565)
Speed limit (km/h)	0.009**	0.004	(0.002, 0.018)
Motorway	0.221	0.141	(−0.068, 0.499)
Year 2003	0	–	
Year 2004	0.075**	0.035	(0.007, 0.144)
Year 2005	0.044	0.036	(−0.026, 0.113)
Year 2006	−0.020	0.036	(−0.091, 0.052)
Year 2007	−0.079**	0.037	(−0.152, −0.005)
Intercept	−11.450**	0.718	(−12.820, −10.030)
S.D. (u)	0.229**	0.060	(0.110, 0.351)
S.D. (e)	0.178**	0.016	(0.145, 0.210)
S.D. (v)	0.492**	0.045	(0.406, 0.583)
DIC	6275.02		
N	1310		

* Statistically significant from zero (90% credible sets show the same sign).
** Statistically significant from zero (95% credible sets show the same sign).

4. Results

4.1. Accident frequency analysis

A spatio-temporal Bayesian hierarchical count model that controls for spatially correlated effects has been developed to model the total number of accidents on road segments. First-order contiguity based neighbouring structures and fixed-time effects are used (see Wang et al., in press). Two MCMC chains were used to ensure convergence. The initial 180,000 iterations were discarded as burn-ins to achieve convergence and a further 30,000 iterations for each chain were performed and kept to calculate the posterior estimates of interested parameters. The Monte Carlo (MC) errors (i.e. the Monte Carlo standard error of the mean) were also monitored, and they were less than 0.005 for most parameters. Using the guide from the WinBUGS user manual (Spiegelhalter et al., 2003), MC errors less than 0.05 indicate that convergence may have been achieved. The Gelman–Rubin statistics are also generally below 1.2 which indicates convergence (Brooks and Gelman, 1998). The model estimation results are presented in Table 4.

It can be seen from Table 4 that the effects of various variables are generally found to be consistent with previous studies (Milton and Mannering, 1998; Kononov et al., 2008). The model estimation results indicate that there is significant spatially correlated effects (u). As expected, both *AADT* and *road segment length* are statistically significant and positively associated with accidents. The coefficient of log(segment length in metre) is approximately 1 suggesting that the elasticity of *road segment length* with respect to accidents is about 1. This means a 1% increase in *road segment length* would increase accident frequency by 1%. *Traffic delay per km* is positively (at the 90% confidence level) associated with the number of accidents, which may be due to the higher speed variance among vehicles within and between lanes and erratic driving behaviour in the presence of congestion (Wang et al., in press). This result is consistent with the study undertaken by Kononov et al. (2008) who also found that fatal and injury accidents increase with the increase in traffic congestion. *Number of lanes* is positive and statistically significant, suggesting more accidents would occur on roads with more lanes, which may be due to increased chance of lane-changing related conflicts on roads with more lanes. *Speed limit* is positively associated with the number of accidents, which suggests that segments with higher speed limits would result in more accidents. *Motorway*, *minimum radius* of horizontal curvature and *maximum gradient* however are statistically insignificant which means that they have little impact on the frequency of road accidents.

4.2. Accident severity analysis

A standard multinomial logit (MNL) model and a mixed logit model have been developed to model accident severity. For the mixed logit model, generally coefficients are considered to be random parameters if they produce statistically significant standard deviations for their assumed normal distributions (Milton et al., 2008). In this study, the results have been obtained from 150 Halton draws.[9] Slight injury accidents were used as the base outcome. A two-level mixed logit model (accident and road segment levels) has also been tested and produced similar results to the normal mixed logit model in terms of the signs of the coefficients and AIC values (the difference of AIC values is less than 4.5), which means the two-level model does not significantly improve the goodness-of-fit. Therefore the results from the two-level model are not presented in this paper for brevity. Model estimation results for the MNL and mixed logit model are presented in Table 5.

As can be seen from Table 5, the estimation results from the MNL and mixed logit models are similar in terms of the set of statistically significant variables and the signs of their coefficients. As suggested by Haque and Chin (2010), a likelihood ratio (LR) test can be performed to compare the mixed logit model with MNL. The test result indicates that the inclusion of the random parameters in the mixed logit model significantly improves the model fit (LR test statistic = 22.57). This is also confirmed by the lower AIC values obtained by the mixed logit model. Considering that the mixed logit model provides a lower AIC value (i.e. better model performance) and the fact that the mixed logit model can control for the unobserved correlated effects and heterogeneity, it is believed that the mixed logit model is more accurate and fits the data better than the

[9] Train (2003) suggested that the parameter estimation would be more consistent in the MSL if a high number of Halton draws could be used. Our initial test has shown that the number of draws above 100 would produce reasonably stable estimations and the results are generally consistent between 100 and 150 draws in terms of the set of significant estimators. Haan and Uhlendorff (2006) also showed that 100–150 Halton draws may be sufficient for stable results. Also as discussed below, the mixed logit model produced a significantly better statistical fit than the standard MNL model. Since the main purpose of this paper is accident prediction, the specification of the mixed logit model used seems appropriate.

Table 5
Estimation results for MNL and mixed logit models.

Variables	MNL		Mixed logit[a]	
	Coefficient	*z* value	Coefficient	*z* value
Serious injury accident				
log(Traffic flow in veh/h)	−0.246**	−4.48	−0.244**	−4.18
Traffic delay (min per 10 km)	0.0002	0.15	−0.020* (0.036**)	−1.65 (2.42)
log(minimum radius)	0.101	1.64	0.120*	1.81
Maximum gradient (%)	0.066**	2.22	0.071**	2.24
Number of lanes ≤ 3 indicator	0.189**	2.08	0.214**	2.20
Number of lanes ≥ 5 indicator	0.181	0.76	0.134	0.50
Motorway indicator	−0.195**	−2.22	−0.211**	−2.25
Speed limit (km/h)	0.001	0.27	0.002	0.37
Lighting condition (darkness)	−0.160**	−1.99	−0.174**	−2.03
Weather (raining)	−0.329**	−3.42	−0.328**	−3.21
Weather (snowing)	−0.250	−0.52	−0.256	−0.51
Other weather conditions (e.g. fog/mist)	−0.319	−1.62	−0.332	−1.59
Peak time indicator	−0.255**	−2.14	−0.262**	−2.08
Weekday indicator	−0.022	−0.31	−0.0004	−0.01
Single vehicle accident indicator	0.474**	6.19	0.484**	5.96
Number of casualties per accident	0.315**	13.52	0.270** (0.261**)	4.65 (2.31)
Year 2004	−0.261**	−2.81	−0.271**	−2.74
Year 2005	−0.300**	−3.2	−0.320**	−3.2
Year 2006	−0.367**	−3.78	−0.392**	−3.76
Year 2007	−0.198**	−2.00	−0.210**	−2.00
Intercept	−1.338*	−1.78	−1.548*	−1.9
Fatal accident				
log(Traffic flow in veh/h)	−0.560**	−5.96	−0.576**	−5.88
Traffic delay (min per 10 km)	−0.003	−0.68	−0.003	−0.75
log(minimum radius)	0.117	0.79	0.129	0.85
Maximum gradient (%)	−0.102	−1.49	−0.106	−1.51
Number of lanes ≤ 3 indicator	0.042	0.19	0.043	0.19
Number of lanes ≥ 5 indicator	−0.475	−0.63	−0.555	−0.71
Motorway indicator	−0.252	−1.32	−0.277	−1.40
Speed limit (km/h)	0.025	1.52	0.025	1.52
Lighting condition (darkness)	0.232	1.24	0.26	1.34
Weather (raining)	−0.490**	−2.10	−0.510**	−2.12
Weather (snowing)	−12.795	−0.03	−18.551	−0.00
Other weather conditions (e.g. fog/mist)	−1.258*	−1.84	−1.275*	−1.78
Peak time indicator	−0.279	−1.15	−0.247	−0.98
Weekday indicator	0.201	1.22	0.22	1.28
Single vehicle accident indicator	0.725**	4.36	0.772**	4.44
Number of casualties per accident	0.424**	11.51	0.352** (0.256**)	4.79 (3.1)
Year 2004	−0.428*	−1.79	−0.440*	−1.78
Year 2005	−0.074	−0.34	−0.048	−0.21
Year 2006	−0.137	−0.60	−0.125	−0.53
Year 2007	0.012	0.05	0.007	0.03
Intercept	−3.467*	−1.66	−3.522*	−1.65
Statistics				
Log likelihood	−4553.108		−4541.825	
AIC	9190.216		9173.65	
N	11501		11501	

Slight injury accident is the base outcome.
[a] Standard deviations and their associated *z* values of random parameters in parentheses.
* $p < 0.1$.
** $p < 0.05$.

MNL model. Therefore, the results from the mixed logit model are preferred.

The coefficient of log(*Traffic flow*) has been modelled as a fixed parameter, and it has been found to be negative and statistically significant for both serious injury accidents and fatal accidents. This indicates that an increase in traffic flow would decrease the probability of serious injury and fatal accidents. This finding is in line with the previous study by Quddus et al. (2010) who employed ordered response models to analyse accident severity. With regard to the results of road infrastructure factors, *minimum radius* is positive and significant (at the 90% confidence level) for the case of serious injury accidents in the mixed logit model, suggesting that horizontally straighter roads tend to increase accident severity. This may be due to the lower speed and increased driver vigilance in the presence of a horizontal curve (Haynes et al., 2007). Increased vertical *gradient* however is found to increase the likelihood of serious injury accidents compared to slight injury accidents. It has been found that motorways tend to decrease the accident severity compared to A roads, which may be due to the higher engineering standard and better road designs on motorways. This finding is consistent with the previous study by Chang and Mannering (1999) who found that interstate highways are more likely to result in property damage

Table 6
Mean absolute deviation (MAD) values.

	Two-stage model	MVPLN model	Fixed proportion method
Fatal	0.249	0.247	0.261
Serious	0.755	0.718	0.794
Slight	1.688	1.633	1.716

Table 7
Ranking of segments.

Road number	Segment description	Ranking using two-stage model		Naïve ranking		Ranking using MVPLN model		Ranking using fixed proportion method	
		Rank	Cost rate[a]	Rank	Cost rate[a]	Rank	Cost rate[a]	Rank	Cost rate[a]
M1	M1 J10 to M1 J9	1	3.62	6	2.48	1	2.76	1	4.38
A3	A3100 to A3100	2	2.61	5	2.49	6	1.70	15	1.31
M1	M1 J8 to M1 J9	3	2.48	24	1.48	4	1.82	2	2.84
A1	A5135 to M25 J23	4	2.35	8	2.27	2	2.07	4	1.77
M25	M25 J19 to A41	5	2.19	135	0.53	8	1.52	6	1.73
M25	M25 J21 to M25 J21A	6	1.97	25	1.46	5	1.77	7	1.65
M1	M1 J9 to M1 J8	7	1.64	22	1.51	10	1.41	3	1.78
M25	A41 to M25 J19	8	1.58	125	0.58	87	0.79	29	1.10
A3	A320 to A322	9	1.55	95	0.73	9	1.48	11	1.38
A1 M	A1(M) J8 to A1(M) J7	10	1.54	4	2.66	7	1.59	10	1.38
A3	A247 to A3100	11	1.50	9	2.25	12	1.30	61	0.89
A20	A20 to M25 J3	12	1.48	176	0.37	22	1.10	52	0.92
M23	M23 J8 to M23 J7	13	1.47	68	0.89	25	1.07	28	1.11
A13	M25 J30 to A1306	14	1.45	65	0.90	18	1.16	34	1.07
A30	M25 J13 to A3044	15	1.44	54	1.02	15	1.28	44	0.99
M23	M23 J7 to M23 J8	16	1.42	179	0.37	30	1.05	19	1.22
M10	M10 J1 to M1 J7	17	1.40	92	0.75	44	0.99	35	1.07
A3	A244 to A309	18	1.40	37	1.25	29	1.06	51	0.93
M25	M25 J26 to M25 J25	19	1.35	41	1.15	40	1.00	27	1.13
A2	A2018 to A2	20	1.32	60	0.94	14	1.29	20	1.20

[a] Cost rate is in £ per 100 vehicle-kilometres travelled in 2003–2007.

only accidents instead of possible injury or injury/fatal accidents. *Raining* weather has been found to decrease the probability of serious injury and fatal accidents and increase the probability of slight injury accidents, which may be due to lower driving speed in rainy weather. This finding is consistent with the study by Savolainen and Mannering (2007) who found that accidents on wet pavements are more likely to be "no injury" accidents. *Single vehicle accident* has been found to be statistically significant and positively associated with both serious injury and fatal accidents, suggesting that a *single vehicle accident* is more likely to be serious or fatal. The effects of other variables have also generally been found to be consistent with previous research (e.g., Shankar et al., 1996; Chang and Mannering, 1999; Quddus et al., 2010).

An interesting finding from the mixed logit model is the effect of traffic congestion (i.e. traffic delay). The coefficient of the *traffic delay* has been taken as a random parameter (assuming a normal distribution) for serious injury accidents in the mixed logit model. The estimated mean values of the coefficients associated with serious injury accidents are statistically significant (at the 90% confidence level). This means that overall traffic congestion tends to decrease the severity of an accident given that the accident has occurred. The standard deviation of the coefficient for the case of serious injury accidents is statistically significant at the 95% confidence level, which means that the effect of congestion varies across different accidents. From the estimated parameters (mean −0.02 and standard deviation 0.036), it can be seen that for 71% of the accidents, an increased level of congestion decreases the probability of a serious injury accident occurring (compared to the probability of a slight injury accident occurring); and for 29% of the accidents, an increased level of congestion increases the likelihood of a serious injury accident occurring. The results suggest the complexity of the effect of traffic congestion on accident severity.

4.3. Two-stage model

The two-stage model combines both accident frequency and severity models and their estimation results have been presented above. In the two-stage process, two types of data are computed: (1) the total expected number of accidents and (2) the expected proportions of accidents for different severity levels (i.e. fatal, serious and slight).

Based on the total number of accidents and the proportions for each severity level, it is straightforward to calculate the predicted number of accidents at different severity levels on a road segment. Based on the segment-level observed and predicted number of accidents, the MAD values are calculated for different categories of severity and presented in Table 6. For comparison, the traditional multivariate Poisson-lognormal (MVPLN) model (with fixed-time effects) and fixed proportion method have also been tested and the corresponding MAD values are also reported in Table 6.

As can be seen from Table 6, all three methods produced comparable results in terms of MAD values. Both two-stage and MVPLN models outperform the fixed proportion method, while the MVPLN model seems to be slightly better than the two-stage model though the difference is marginal.

4.4. Site ranking

After obtaining the expected number of accidents per segment at each severity level using the two-stage model, monetary costs can then be applied to the accidents to calculate the total costs of accidents on road segments for the purpose of site ranking. Sites (road segments) can then be ranked by the total accident cost rate for the period 2003–2007. The higher accident cost rate of a road segment, the more hazardous it is considered to be. The top 20 most hazardous road segments ranked by the accident cost rate are listed in Table 7. For comparison, naïve ranking using pure observed accident count data and ranking using the multivariate Poisson-lognormal (MVPLN) model and the fixed proportion method have also been produced and presented.

As can be seen from Table 7, the two-stage model produces significantly different rankings from the naïve ranking method. 15 out of the top 20 road segments in the model based ranking are not in the top 20 in the naïve ranking. The differences between the ranking using the two-stage model and naïve ranking are significant. Accident cost rates for the majority of the top 20 road segments ranked by the two-stage models are higher than the naïve estimates. This implies that the naïve ranking method underestimated the accident costs for road segments.

On the other hand, ranking results among the two-stage model, MVPLN model and fixed proportion method are more comparable. Comparison of different model based rankings for the top 20 road

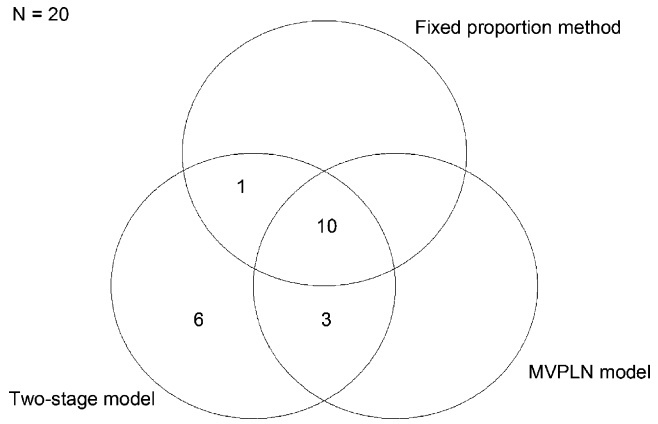

Fig. 1. Comparison of different models based rankings for the top 20 road segments.

segments is presented in Fig. 1. As can be seen, only 7 out of the top 20 road segments in the two-stage model ranking are not in the top 20 in the MVPLN model ranking; and 9 out of the top 20 road segments in the two-stage model ranking are not in the top 20 in the fixed proportion method ranking. A total number of 10 road segments are ranked in the top 20 in all three model based rankings. This means that the model based rankings are generally consistent to each other, compared to the naïve ranking.

The differences between the ranking using the two-stage model and other ranking methods are presented in Fig. 2. It is clear that there are significant differences between the two-stage model and naïve ranking method (Fig. 2(a)). This result is consistent with previous studies (e.g. see Miaou and Song, 2005; Huang et al., 2009 for the comparison between model based ranking and naïve ranking). The differences between the two ranking methods are mainly due to the high stochastic and sporadic nature of accidents, and the fact that considerably higher costs are given to fatal accidents than the other two types of accidents (Miaou and Song, 2005). As discussed, due to the regression-to-the-mean problem, the ranking results using the naïve method may be biased and inaccurate, and as such the model based ranking method is preferred. As can be seen in Fig. 2(b) and (c), model based ranking using the two-stage model, MVPLN model and fixed proportion method are more consistent compared to the naïve ranking. This confirms that a model based ranking should be used instead of naïve ranking to obtain consistent ranking results. It should be noted that, although all model based methods have similar goodness-of-fit performance in terms of MAD values as presented above, there are still notable differences in the ranking results as suggested in Fig. 2(b) and (c). Thus more information (e.g. inputs from policy makers) may be useful in selecting the sites for further safety examination and remedial treatment.

Based on the ranking results using the two-stage model, the locations of the top 20 most hazardous road segments listed in Table 7 are highlighted in Fig. 3, in which the rank, road number and direction information is shown. It can be seen that the top ranked segments are found scattered throughout the road network.

After identifying the hazardous road segments, further safety examination and treatment can be applied on these road segments. The higher ranked segments can be given higher priorities for safety treatment with a limited budget. A cost-benefit analysis of potential safety treatment can also be performed by policy makers based on the predicted accident costs on the road segments (Miaou and Song, 2005).

5. Discussion and conclusions

This paper proposed a two-stage mixed multivariate model which combines both accident frequency and severity models

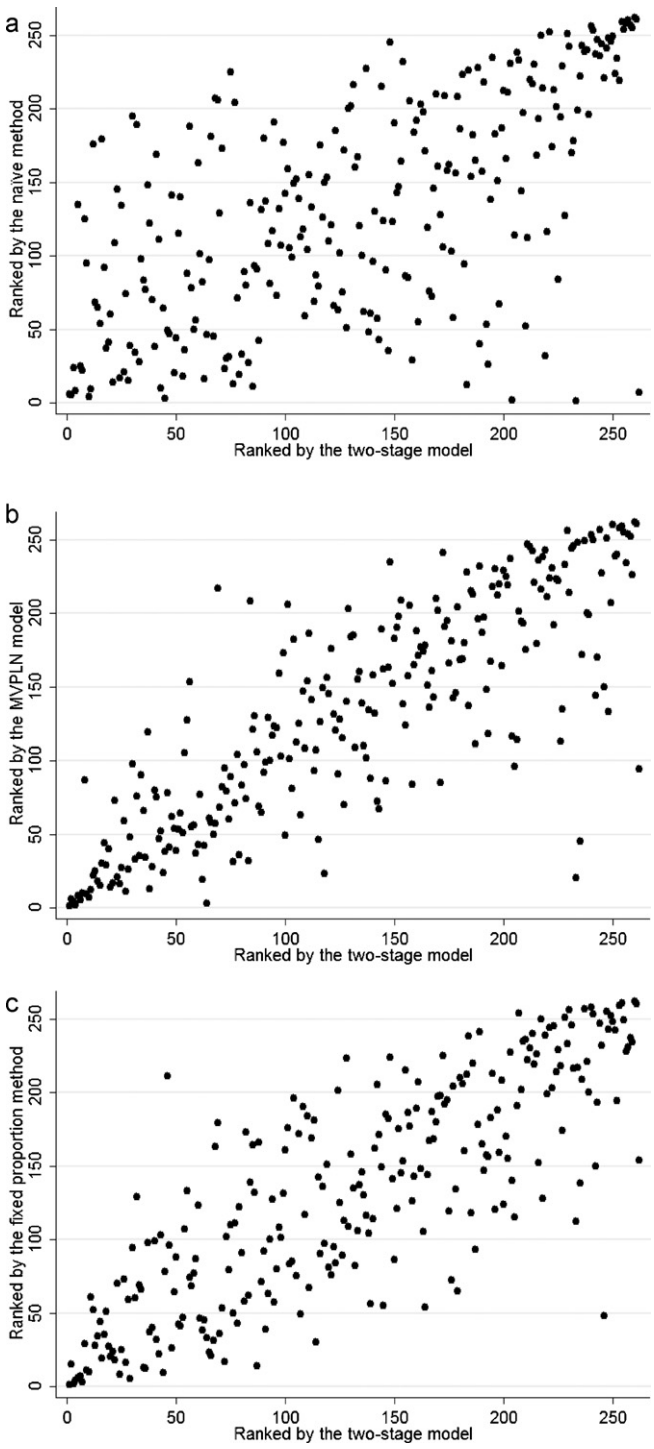

Fig. 2. (a) Comparison of ranking results: two-stage model vs. naïve method. (b) Comparison of ranking results: two-stage model vs. multivariate Poisson-lognormal (MVPLN) model. (c) Comparison of ranking results: two-stage model vs. fixed proportion method. Comparison of ranking results.

to predict the number of accidents in different categories (e.g., severities). The practical application of the two-stage model is illustrated in site ranking, which aims to identify hazardous road segments (i.e. accident hotspots) on a road network (i.e. M25 and surround). Based on the accident prediction results from the two-stage model, road segments on the M25 and surround were ranked by their monetary cost rate (£ per 100 vehicle km) of accidents. The ranking using the two-stage model was also compared with the naïve rankings using observed accident data and

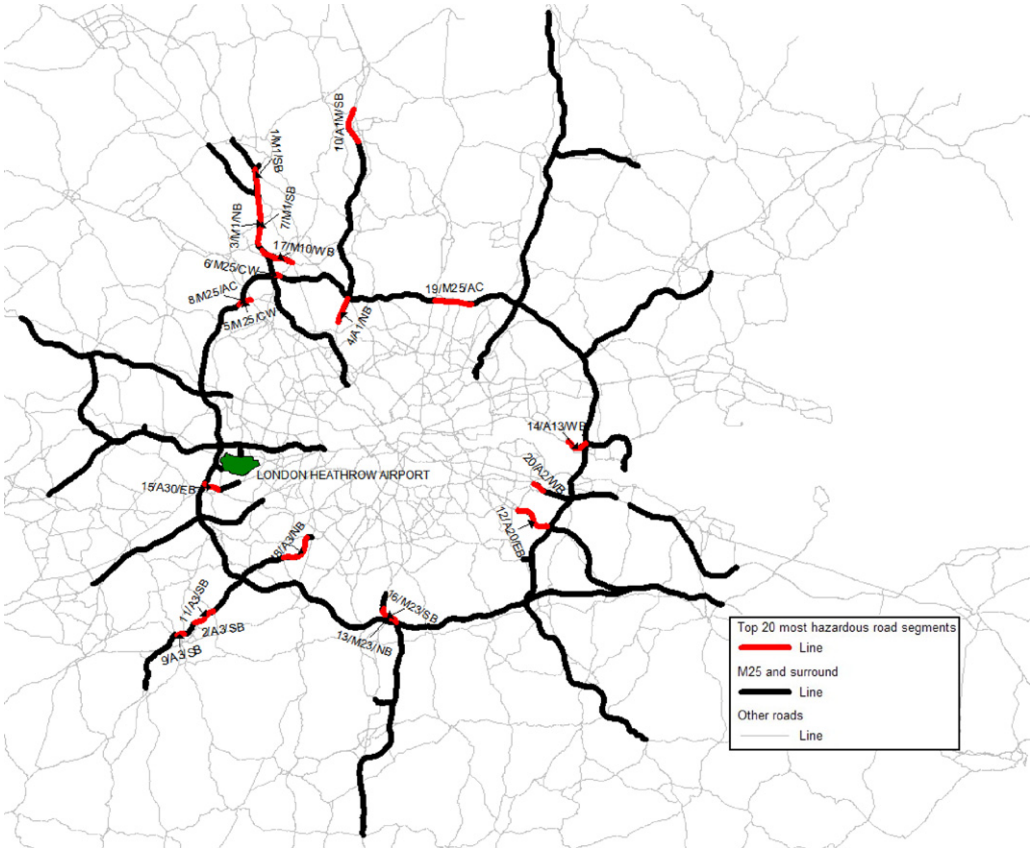

Fig. 3. Top ranked 20 most hazardous road segments using the two-stage model.

two other model based rankings (i.e. MVPLN model and fixed proportion method). It was found that there were significant differences in terms of ranking results between the naïve ranking and model based rankings. Naïve ranking method tends to underestimate the cost of accidents on road segments. The two-stage model is generally comparable to the MVPLN model and fixed proportion method. Top ranked hazardous road segments were also identified and located based on the results from the two-stage model.

Compared to the traditional road safety analysis using only accident frequency models, the two-stage model has several distinct advantages:

First, detailed data associated with individual accidents are normally available and can be incorporated into accident severity models to accurately estimate the proportions of accidents at different severity levels, in addition to the aggregated segment level data. In the case of the data used in this paper, as shown in the data description section, only traffic and road characteristics data are available at the aggregated road segment level for accident frequency models. On the other hand, in addition to the aggregated traffic and road characteristics data (e.g. road geometry), more detailed data are available at the individual accident level for accident severity models such as lighting and weather conditions, time when the accident occurred, the number of vehicles involved in an accident and hourly traffic flow just before an accident. It is expected that the additional data in accident severity analysis would allow a better understanding of the severity outcome of an accident, and subsequently the distribution (proportion) of accidents at different severity levels on a given road segment. This is also the benefit of this method (individual accident level severity analysis) compared to road segment level severity analysis by Milton et al. (2008) and Geedipally et al. (2010). In addition, this method avoids the potential aggregation bias (Davis, 2004) in an accident severity analysis.

Individual accident level data can be conveniently obtained from the STATS19 database in the UK, which enables researchers to develop an insight into the severity distribution of accidents. This paper has shown how results from an accident severity model at a disaggregate individual accident level can be aggregated to predict the proportions of types of accidents on a road segment in the two-stage modelling process.

Second, there are cases that some categories of accident severities, due to many zero or low accident counts at an aggregated road segment (or an area) level, cannot be analysed using accident frequency models (e.g. MVPLN) directly. This is particularly an issue for high severity level accidents (such as fatal accidents). This issue can be addressed using the accident severity models as there may be enough observations for each category of severities at a disaggregate individual accident level. The two-stage model may still be possible to predict the expected number of accidents at different severity levels even when there are many zero or low accident counts at an aggregated road segment (or area) level. In the case of this paper, there are only 213 fatal accidents on the 262 road segments during 2003–2007, resulting in many zero (more than 85% cases) and low count of fatal accidents (per road segment per year). Therefore, it may not always be statistically feasible to use accident frequency models to directly predict the number of fatal accidents. Traditionally a researcher avoids this problem by combining two or several categories of accidents, for instance combining fatal accidents with injury accidents (e.g., El-Basyouny and Sayed, 2009). This issue however can be addressed using the two-stage model, as there are enough cases of fatal accidents to develop an accident severity model which can predict the expected proportion of fatal accidents on a road segment.

Finally, the two-stage model is flexible in terms of the model specification and estimation. A researcher is not constrained to one type of model, but can choose the appropriate modelling method at each stage. For example, when sample size is relatively small, which is often the case for an accident frequency analysis, a Bayesian approach may be used to obtain robust estimates; when sample size is very large, which is often the case for severity analysis (11,501 observations in the severity analysis in this paper), a frequentist inference can be chosen as its estimation results are equivalent to the Bayesian approach (Train, 2003).[10] In fact, it is not essential for a researcher to employ a regression model at all in any of the two stages. For instance, one can use a neural network model in the accident frequency analysis (Xie et al., 2007; Lord and Mannering, 2010) and a data mining technique such as the classification and regression tree approach in the severity analysis (Chang and Wang, 2006). This may also benefit the practitioners in that two teams are able to work on the frequency and severity analyses separately and the results can then be combined.

Therefore in the scenarios that disaggregate individual accident level data are available and it is required to predict a certain type of low frequency accident, the two-stage mixed multivariate model can be recommended. As such the two-stage model is a promising alternative to accident frequency models in predicting counts of accidents in different categories and site ranking. Future research may focus on validating this method with other data samples or models.

Acknowledgements

The authors would like to thank the UK Highways Agency for providing traffic and road characteristics data for the M25 motorway and surrounding major roads. The content of the paper however does not necessarily express the views of the Highways Agency and the authors take full responsibility for the content of the paper and any errors or omissions. The authors are also indebted to the Editor and three anonymous referees for detailed and informative comments on earlier drafts of the paper.

References

Aguero-Valverde, J., Jovanis, P.P., 2006. Spatial analysis of fatal and injury crashes in Pennsylvania. Accident Analysis & Prevention 38 (3), 618–625.
Aguero-Valverde, J., Jovanis, P.P., 2009. Bayesian multivariate Poisson log-normal models for crash severity modeling and site ranking. In: Proceedings of the Paper Presented at the 88th Annual Meeting of the Transportation Research Board.
Bhat, C.R., 2003. Simulation estimation of mixed discrete choice models using randomized and scrambled Halton sequences. Transportation Research Part B: Methodological 37 (9), 837–855.
Brooks, S.P., Gelman, A., 1998. General methods for monitoring convergence of iterative simulations. Journal of Computational and Graphical Statistics 7 (4), p434–455.
Cameron, A.C., Trivedi, P.K., 1998. Regression Analysis of Count Data. Cambridge University Press, Cambridge.
Chang, L., Mannering, F., 1999. Analysis of injury severity and vehicle occupancy in truck- and non-truck-involved accidents. Accident Analysis & Prevention 31 (5), 579–592.
Chang, L., Wang, H., 2006. Analysis of traffic injury severity: an application of non-parametric classification tree techniques. Accident Analysis & Prevention 38 (5), 1019–1027.
Cheng, W., Washington, S.P., 2005. Experimental evaluation of hotspot identification methods. Accident Analysis & Prevention 37 (5), 870–881.
Davis, G.A., 2004. Possible aggregation biases in road safety research and a mechanism approach to accident modeling. Accident Analysis & Prevention 36 (6), 1119–1127.
Department for Transport, 2008. Road Casualties Great Britain: 2007. Transport Statistics.
Department for Transport, 2009. Road Statistics 2008: Traffic, Speeds and Congestion. National Statistics.
El-Basyouny, K., Sayed, T., 2009. Collision prediction models using multivariate Poisson-lognormal regression. Accident Analysis & Prevention 41 (4), 820–828.
Eluru, N., Bhat, C.R., Hensher, D.A., 2008. A mixed generalized ordered response model for examining pedestrian and bicyclist injury severity level in traffic crashes. Accident Analysis & Prevention 40 (3), 1033–1054.
Elvik, R., 2007. State-of-the-art approaches to road accident black spot management and safety analysis of road networks.
Frees, E.W., 2004. Longitudinal and Panel Data: Analysis and Applications in the Social Sciences. Cambridge University Press.
Geedipally, S.R., Patil, S., Lord, D., 2010. Examining methods for estimating crash counts according to their collision type. In: Proceedings of the Paper Presented at the 89th Annual Meeting of the Transportation Research Board.
Graham, D.J., Glaister, S., 2003. Spatial variation in road pedestrian casualties: the role of urban scale, density and land-use mix. Urban Studies 40 (8), 1591–1607.
Haan, P., Uhlendorff, A., 2006. Estimation of multinomial logit models with unobserved heterogeneity using maximum simulated likelihood. Stata Journal 6 (2), 229–245 (17).
Haque, M.M., Chin, H.C., 2010. A mixed logit analysis on the right-angle crash vulnerability of motorcycles at signalized intersections. In: Proceedings of the Paper Presented at the 89th Annual Meeting of the Transportation Research Board.
Hauer, E., Allery, B., Kononov, J., Griffith, M., 2004. How best to rank sites with promise. Transportation Research Record: Journal of the Transportation Research Board 1897, 48–54.
Hausman, J.A., Leonard, G.K., McFadden, D., 1995. A utility-consistent, combined discrete choice and count data model assessing recreational use losses due to natural resource damage. Journal of Public Economics 56 (1), 1–30.
Haynes, R, Jones, A., Kennedy, V., Harvey, I., Jewell, T., 2007. District variations in road curvature in England and Wales and their association with road-traffic crashes. Environment and Planning A 39 (5), 1222–1237.
Hole, A.R., 2007. Fitting mixed logit models by using maximum simulated likelihood. Stata Journal 7 (3), 388–401.
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. Transportation Research Record: Journal of the Transportation Research Board 2103, 32–41.
Ivan, J.N., Wang, C., Bernardo, N.R., 2000. Explaining two-lane highway crash rates using land use and hourly exposure. Accident Analysis & Prevention 32 (6), 787–795.
Kim, J., Kim, S., Ulfarsson, G.F., Porrello, L.A., 2007. Bicyclist injury severities in bicycle–motor vehicle accidents. Accident Analysis & Prevention 39 (2), 238–251.
Kononov, J., Bailey, B., Allery, B., 2008. Relationships between safety and both congestion and number of lanes on urban freeways. Transportation Research Record: Journal of the Transportation Research Board 2083, 26–39.
Lan, B., Persaud, B., 2010. Evaluation of multivariate Poisson log normal fully Bayesian methods. In: Proceedings of the Paper Presented at the 89th Annual Meeting of the Transportation Research Board.
Long, J.S., Freese, J., 2006. Regression Models for Categorical Dependent Variables Using Stata, Second ed. Stata Press, College Station, Texas.
Lord, D., 2000. The prediction of accidents on digital networks: characteristics and issues related to the application of accident prediction models. Ph.D. Dissertation. Department of Civil Engineering, University of Toronto, Toronto.
Lord, D., Mannering, F., 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.
Ma, J., Kockelman, K., 2006. Bayesian multivariate Poisson regression for models of injury count, by severity. Transportation Research Record: Journal of the Transportation Research Board 1950, 24–34.
Ma, J., Kockelman, K.M., Damien, P., 2008. A multivariate Poisson-lognormal regression model for prediction of crash counts by severity, using Bayesian methods. Accident Analysis & Prevention 40 (3), 964–975.
Maher, M.J., Mountain, L.J., 1988. The identification of accident blackspots: a comparison of current methods. Accident Analysis & Prevention 20 (2), 143–151.
Miaou, S., Song, J.J., Mallick, B., 2003. Roadway traffic crash mapping: a space-time modeling approach. Journal of Transportation and Statistics 6 (1), 33–57.
Miaou, S., Song, J.J., 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.
Milton, J.C., Shankar, V.N., Mannering, F.L., 2008. Highway accident severities and the mixed logit model: an exploratory empirical analysis. Accident Analysis & Prevention 40 (1), 260–266.
Milton, J., Mannering, F., 1998. The relationship among highway geometrics, traffic-related elements and motor-vehicle accident frequencies. Transportation 25 (4), 395–413.
Noland, R.B., Quddus, M.A., 2005. Congestion and safety: a spatial analysis of London. Transportation Research Part A: Policy and Practice 39 (7–9), 737–754.
O'Donnell, C.J., Connor, D.H., 1996. Predicting the severity of motor vehicle accident injuries using models of ordered multiple choice. Accident Analysis & Prevention 28 (6), 739–753.

[10] The Bayesian theorem indicates that the posterior distribution is proportional to the prior distribution times the likelihood of the observed data. For large sample size, the prior becomes irrelevant and the maximum of the likelihood function becomes the same as the maximum and also the mean of the posterior. It can be shown that the resulting estimator under Bayesian inference is asymptotically equivalent to the classical maximum likelihood estimator when large sample size is used: the mean of the posterior distribution of a parameter can be seen as the classical point estimate; and the standard deviation of the posterior distribution can be seen as the standard error of the estimate (for more details see Train, 2003).

Ogle, J.H., Alluri, P., Sarasua, W., 2011. MMUCC and MIRE: the role of segmentation in safety analysis. In: Proceedings of the Paper Presented at the 90th Annual Meeting of the Transportation Research Board.

Oh, J., Lyon, C., Washington, S., Persaud, B., Bared, J., 2003. Validation of FHWA crash models for rural intersections: lessons learned. Transportation Research Record: Journal of the Transportation Research Board 1840, 41–49.

Park, E., Lord, D., 2007. Multivariate Poisson-lognormal models for jointly modeling crash frequency by severity. Transportation Research Record: Journal of the Transportation Research Board 2019, 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. Accident Analysis & Prevention 42 (1), 38–43.

Persaud, B., Lyon, C., 2007. Empirical Bayes before–after safety studies: lessons learned from two decades of experience and future directions. Accident Analysis & Prevention 39 (3), 546–555.

Quddus, M.A., 2008. Modelling area-wide count outcomes with spatial correlation and heterogeneity: an analysis of London crash data. Accident Analysis & Prevention 40 (4), 1486–1497.

Quddus, M.A., Wang, C., Ison, S.G., 2010. Road traffic congestion and crash severity: econometric analysis using ordered response models. ASCE Journal of Transportation Engineering 136 (5), 424–435.

Savolainen, P., Mannering, F., 2007. Probabilistic models of motorcyclists' injury severities in single- and multi-vehicle crashes. Accident Analysis & Prevention 39 (5), 955–963.

Shankar, V., Mannering, F., Barfield, W., 1995. Effect of roadway geometrics and environmental factors on rural freeway accident frequencies. Accident Analysis & Prevention 27 (3), 371–389.

Shankar, V., Mannering, F., 1996. An exploratory multinomial logit analysis of single-vehicle motorcycle accident severity. Journal of Safety Research 27 (3), 183–194.

Shankar, V., Mannering, F., Barfield, W., 1996. Statistical analysis of accident severity on rural freeways. Accident Analysis & Prevention 28 (3), 391–401.

Song, J.J., 2004. Bayesian multivariate spatial models and their applications. Ph.D Dissertation. Texas A&M University.

Song, J.J., Ghosh, M., Miaou, S., Mallick, B., 2006. Bayesian multivariate spatial models for roadway traffic crash mapping. Journal of Multivariate Analysis 97 (1), 246–273.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., Linde, A.v.d., 2002. Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society. Series B (Statistical Methodology) 64 (4), 583–639.

Spiegelhalter, D., Thomas, A., Best, N. and Lunn, D., 2003. WinBUGS user manual version 1.4.

Tanaru, R., 2002. Hierarchical Bayesian models for multiple count data. Austrian Journal of Statistics 31 (3.), 221–229.

Train, K.E., 2003. Discrete Choice Methods with Simulation. Cambridge University Press.

Wang, C., Quddus, M.A., Ison, S.G., 2009. Impact of traffic congestion on road accidents: a spatial analysis of the M25 motorway in England. Accident Analysis & Prevention 41 (4), 798–808.

Wang, C., Quddus, M.A., Ison, S.G., in press. A spatio-temporal analysis of the impact of congestion on traffic safety on major roads in the UK. *Transportmetrica*.

Xie, Y., Lord, D., Zhang, Y., 2007. Predicting motor vehicle collisions using Bayesian neural network models: an empirical analysis. Accident Analysis & Prevention 39 (5), 922–933.

Yamamoto, T., Hashiji, J., Shankar, V.N., 2008. Underreporting in traffic accident data, bias in parameters and the structure of injury severity models. Accident Analysis & Prevention 40 (4), 1320–1329.