Contents lists available at SciVerse ScienceDirect

Accident Analysis and Prevention

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

Prediction of road accidents: A Bayesian hierarchical approach

Markus Deublein[a,*], Matthias Schubert[b], Bryan T. Adey[a], Jochen Köhler[c], Michael H. Faber[d]

[a] Institute of Construction and Infrastructure Management, Swiss Federal Institute of Technology ETH, Zurich, Switzerland
[b] Matrisk GmbH, Managing Technical Risks, Zurich, Switzerland
[c] Department of Structural Engineering, Norwegian University of Science and Technology NTNU, Trondheim, Norway
[d] Department of Civil Engineering, Technical University of Denmark, DTU, Kgs. Lyngby, Denmark

ARTICLE INFO

Article history:
Received 22 June 2012
Received in revised form 16 October 2012
Accepted 26 November 2012

Keywords:
Road safety assessment
Accident prediction
Injury accidents
Bayesian Probabilistic Networks
Accident risk modelling
Multivariate regression analysis
Hierarchical Bayes

ABSTRACT

In this paper a novel methodology for the prediction of the occurrence of road accidents is presented. The methodology utilizes a combination of three statistical methods: (1) gamma-updating of the occurrence rates of injury accidents and injured road users, (2) hierarchical multivariate Poisson-lognormal regression analysis taking into account correlations amongst multiple dependent model response variables and effects of discrete accident count data e.g. over-dispersion, and (3) Bayesian inference algorithms, which are applied by means of data mining techniques supported by Bayesian Probabilistic Networks in order to represent non-linearity between risk indicating and model response variables, as well as different types of uncertainties which might be present in the development of the specific models.

Prior Bayesian Probabilistic Networks are first established by means of multivariate regression analysis of the observed frequencies of the model response variables, e.g. the occurrence of an accident, and observed values of the risk indicating variables, e.g. degree of road curvature. Subsequently, parameter learning is done using updating algorithms, to determine the posterior predictive probability distributions of the model response variables, conditional on the values of the risk indicating variables.

The methodology is illustrated through a case study using data of the Austrian rural motorway network. In the case study, on randomly selected road segments the methodology is used to produce a model to predict the expected number of accidents in which an injury has occurred and the expected number of light, severe and fatally injured road users. Additionally, the methodology is used for geo-referenced identification of road sections with increased occurrence probabilities of injury accident events on a road link between two Austrian cities. It is shown that the proposed methodology can be used to develop models to estimate the occurrence of road accidents for any road network provided that the required data are available.

© 2012 Elsevier Ltd. All rights reserved.

1. Introduction

Despite significant improvements in vehicle technology and road engineering over the last 40 years, on a world-wide scale road accidents are still one of the main accidental causes of death and injury (WHO, 2004). The assessment of the occurrence of road accidents and the management of infrastructure to deal with this risk are therefore research areas of considerable interest. Numerous studies have been performed to identify the most important risk indicating variables that contribute to the occurrence of road accidents. Comprehensive overviews of the different research approaches can be found e.g. in Hauer (2009), Elvik (2011), Lord and Mannering (2010) and Savolainen et al. (2011). The most common approach applied in early works is to model the interaction between road geometry, traffic characteristics and accident frequencies by means of conventional (multiple) linear regression models. In such studies univariate counting models for only one single model response variable are used, implying, for example that the number of accidents corresponding to different degrees of injury severity are modelled separately without taking into account the dependencies that exist between them (Park and Lord, 2007; Ma et al., 2008). Such dependencies are considered in more recent studies where the different response variables are modelled jointly using multivariate modelling techniques (Bijleveld, 2005; Song et al., 2006; Elvik, 2011). Multivariate data analysis based on multivariate normal distributions has often been used to analyse continuous data. However, when only discrete multivariate data on accident numbers are available, the assumption of multivariate normal distributions may be misleading since accident data is often characterized by small observed mean values and a large number of zero counts leading to the

* Corresponding author at: Swiss Federal Institute of Technology, ETH Zurich, Institute of Construction and Infrastructure Management, IBI, HIL F 27.1, Wolfgang-Pauli-Strasse 15, CH-8093 Zurich, Switzerland. Tel.: +41 44 633 71 31; fax: +41 44 633 10 88.
E-mail address: deublein@ibi.baug.ethz.ch (M. Deublein).

0001-4575/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.aap.2012.11.019

well discussed phenomenon of over-dispersion (Cox, 1983; Dean and Lawless, 1989; Hauer, 2001; Karlis and Meligkotsidou, 2005; Gschloessl and Czado, 2006; Berk and Macdonald, 2008). Some of the existing research dealing with the joint modelling of discrete accident count data for different degrees of injury severity use multivariate Poisson regression analysis as done by Tsionas (2001), Tunaru (2002), Bijleveld (2005), Miaou and Song (2005), Song et al. (2006) and Ma and Kockelman (2006). The multivariate Poisson models, however, do not appropriately account for over-dispersion and covariance between the response variables. In Park and Lord (2007), Ma et al. (2008) and El-Basyouny and Sayed (2009b) multivariate Poisson-lognormal regression approaches are introduced which are capable to cope with both, the full covariance structure of the response variables and the aspect of over-dispersion.

With increasing computing capacities, Bayesian inference and updating algorithms have gradually become more relevant in the field of accident risk assessment. Empirical Bayesian methods were investigated first and are still frequently applied (Persaud et al., 1999; Carlin and Louis, 2000; Hauer et al., 2002; Cheng and Washington, 2005; Elvik, 2008). The step from empirical Bayes to full Bayes approaches is taken e.g. by Schlüter et al. (1997), Heydecker and Wu (2001), Macnab (2003), Ying (2004), Carriquiry and Pawlovich (2005), Miaou and Song (2005), Qin et al. (2005), Song et al. (2006), Maes et al. (2007), Persaud et al. (2010), Park et al. (2010) and Huang and Abdel-Aty (2010). The full Bayesian approach facilitates the consistent consideration of aleatory and epistemic uncertainties, non-linear dependencies amongst the indicator variables and the updating of the developed risk models based on new available data (Faber and Maes, 2005; Der Kiureghian and Ditlevsen, 2009). Bayesian Probabilistic Networks (BPN) can be used as a helpful tool to apply Bayesian inference and updating algorithms in an intuitively, understandable and illustrative manner. However, the application of BPNs for the analysis of accidents and accident related injury severity levels is still rather scarce. BPNs are applied for accident reconstruction modelling by Davis and Pei (2003) with the purpose to update prior physical models with observations made at accident sites. Marsh and Bearfield (2004) used BPNs for accident modelling on the UK railway network and Ozbay and Noyan (2006) applied them to investigate incident clearance duration time on road links. Simoncic (2004) developed a two car accident injury severity model based on BPNs using information of road user attributes, environmental conditions and road characteristics. In Schubert et al. (2007, 2011) the development of a generic methodology for the risk assessment of road tunnels is described, and BPNs are used to construct hierarchical indicator based risk models. For modelling accident injury severities on Spanish roads De Oña et al. (2011) and Mujalli and De Oña (2011) applied 18 risk indicating variables related to driver, vehicle, road properties and environmental characteristics in the development of a BPN. BPNs are also used in Karwa et al. (2011) to investigate the potential use of causal inference methods in transportation safety. Hossain and Muromachi (2012) are using BPNs for real-time accident risk prediction on urban.

The methodology presented in this paper is based on a combination of both, (1) a hierarchical multivariate Poisson-lognormal regression analysis, which facilitates taking into account the covariance structure of the model response variables as well as over-dispersion effects, and (2) BPNs that take into account aleatory and epistemic uncertainties as well as possibly non-linear dependencies between the risk indicating variables and the response variables. In the subsequent sections, the methodology for the development of models to be used to predict the occurrence frequencies of injury accidents and injury severities of road users is explained, and the methodology is demonstrated through a case study using the Austrian road network.

2. Methodology

In accordance with the definitions of risk in Kaplan and Garrick (1981), accident risk can be understood as the product of the occurrence probability and the corresponding consequences. In the subsequent paragraphs, however, the definition of accident risk is constricted just to the occurrence frequencies of accidents. The assessment of the consequences in terms of monetary equivalents is left to future investigations.

The proposed methodology is composed of six major steps: (1) identification and determination of the response variables and risk indicating variables (Section 2.1), (2) subdivision of the road network into homogenous segments (Section 2.2), (3) Gamma-updating of the response variables (Section 2.3), (4) the development of a multivariate Poisson-lognormal regression model for the description of the relationships between risk indicating variables and the response variables (Section 2.4), (5) the construction and parameter learning of the BPN (Section 2.5) and (6) the prediction of the expected number of response variable events, i.e. the expected number of injury accidents (Section 2.6).

2.1. *Use of data*

The methodology is exclusively based on data. A sufficiently large and reliable data set with information about observations of response variables (e.g. injury accidents, number of fatalities) and risk indicating variables (e.g. road design parameters, traffic volume) is required. During the model development the data is applied for two complementary but not overlaying modelling steps:

First, the information of the data is used to establish a multivariate Poisson-lognormal regression model which forms the basis for the prior BPN. Predictions of the prior BPN are exclusively based on the results of the regression analysis. The regression parameters and covariance structures between response variables and risk indicating variables are assessed probabilistically allowing the interpolation and extrapolation of the information of the data into model domains for which no data are available (e.g. maximum traffic volume (AADT) in the dataset is 80,000 vehicles/day but the model covers a range up to 100,000 vehicles/day).

Second, the information of the prior BPN is updated by means of parameter learning algorithms using the observations of response variables and risk indicating variables as contained in the available dataset. The updating of the prior model can be considered as a replacement of the prior model probabilities with the values of the updated posterior model probabilities. However, only the prior model probabilities are replaced for which observations of the response variables and risk indicating variables are available. The replacement is incorporated into the updating process by assigning a very low weight to the prior model information. This ensures that the use of the information of the applied data is implemented into the model development process in a complementary manner solely.

2.2. *Determination of model variables*

Step 1 concerns the determination of the model variables. The methodology used to determine appropriate accident risk models is based on defined sets of explanatory risk indicating variables and dependent response variables. The risk indicating variables are observable road and traffic variables (e.g. number of lanes, degree of slope, number of vehicles, etc.), that are considered to influence the conditional occurrence probability of the response variables (e.g. number of injury accidents and different levels of injury severity of the road users being involved in injury accidents). It is advantageous to identify risk indicating variables that are relevant for the prediction of accident events also of any road sections, which

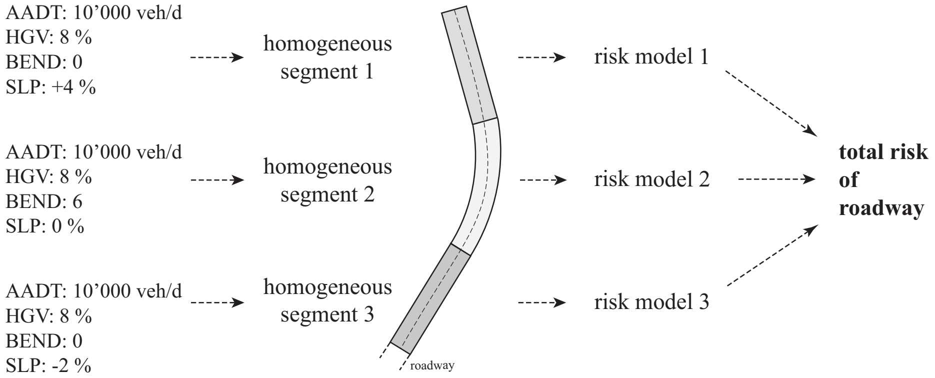

Fig. 1. Example for segregation of a road section into three homogenous segments based on values of the observed risk indicating variables.

might not be part of the initial model development. They, however, have to be sufficiently specific to enable reasonable conclusions on the response variables. A balanced trade-off needs to be found and the selection of model variables in itself comprises already a strong Bayesian element.

2.3. Construction of homogeneous segments

Step 2 concerns the sub-division of the road network into so-called homogeneous segments based on available data of the risk indicating variables. A homogenous segment is a segment of road over which it can be assumed that the values of all risk indicating variables to be included in the model are constant, and thus also the risk is uniform. Fig. 1 illustrates the process of sub-dividing a road section into three homogeneous segments based on the values of four risk indicating variables. The change in the value of any risk indicating variable results in the start of a new homogeneous segment (e.g. AADT (annual average daily traffic) and HGV (fraction of heavy good vehicles) are constant but changes are observed in BEND (curvature) and SLP (slope)).

One generic accident risk model is developed which is becoming specific when applied for every individual homogeneous segment. The risk assessed for each homogenous segment is summed to estimate the total accident risk for the entire considered road link or network. The development of the generic risk model is described in the subsequent paragraphs and its application to homogeneous segments is evaluated in the case studies (Section 3).

2.4. Gamma-updating of model response variables

Step 3 concerns the Gamma-updating of the response variables. A two-level hierarchical approach is used for modelling the response variables of the model. On the first level the parameters for the probability distributions of the expected number of accident events are estimated. These parameters themselves are assumed to be random variables and are described at the second level of the hierarchy by means by probability distributions with so-called hyper-parameters.

2.4.1. First level of hierarchy

At the first level of the hierarchy the probability of a specific number of observations y_{ik} of the $k = 1, \ldots, z$ different response variables on the ith homogeneous segment is assumed to be Poisson distributed as suggested e.g. by Song et al. (2006) and Park and Lord (2007).

$$Y_{ik} | \mu_{ik} \sim \text{Poisson}\,(\mu_{ik}) \tag{1}$$

with mean occurrence frequency

$$\mu_{ik} = \nu_i \cdot \lambda_{ik} \tag{2}$$

where ν_i is the exposure (in million vehicle kilometres of travel (mvk) per year) and λ_{ik} is the occurrence rate of the response variables. Accident count data are often characterized by over-dispersion (sample variance larger than sample mean) and hence, the assumption for the single-parameter Poisson model that sample variance and sample mean are the same is not fulfilled. As a consequence, the distribution of the count data is assumed to be negative binomial (NB) distributed, being a mixture of (1) a Poisson distribution describing the probability of having a defined number of accidents or injuries in one particular homogeneous segment over a defined period of time (e.g. per year), and (2) the natural conjugate Gamma distribution describing the probability distribution of the Poisson parameter λ itself defined by the parameters α and β (Gelman et al., 2004). The NB distribution is capable of having different values for the mean and variance and the expected numbers of events are given by:

$$\begin{aligned} NB(y \,|\alpha, \beta) &= \frac{\text{Poisson}\,(y \,|\lambda) \cdot \text{Gamma}\,(\lambda \,|\alpha, \beta)}{\text{Gamma}\,(\lambda \,|\alpha + y, 1 + \beta)} \\ &= \binom{y + \alpha - 1}{y} \cdot \left(\frac{\beta}{\beta + 1}\right)^{\alpha} \cdot \left(\frac{1}{\beta + 1}\right)^{y} \end{aligned} \tag{3}$$

with expected value and variance

$$E[y] = \frac{\alpha}{\beta} \text{ and } \mathrm{VAR}[y] = \frac{\alpha}{\beta^2}(\beta + 1) = \frac{\alpha}{\beta} \cdot \frac{\beta + 1}{\beta} \tag{4}$$

with shape (dispersion) parameter $\alpha > 0$ and inverse scale parameter $\beta > 0$. $E[.]$ is the expectation operator and VAR[.] the variance operator. According to Eq. (4) the variance is always greater than the expected value. As β approaches infinity with $E[y]$ remaining constant the variance of the Gamma distribution approaches zero and hence the NB distribution approaches the Poisson distribution.

To estimate the occurrence rates so-called prior distributions of λ'_{ik} are first calculated based on averaged information over all homogeneous segments of the network. The prior distribution parameters of λ'_{ik} are then updated on the second level of hierarchy using observations of the response variables for the individual homogeneous segments.

2.4.2. Second level of hierarchy

At the second level of hierarchy the probability distribution of the prior and posterior Gamma parameters are described as:

$$\lambda'_{ik} \sim \text{Gamma}\,(\alpha'_{ik}, \beta'_i) \tag{5}$$

with probability density

$$p(\lambda'_{ik}) = \frac{\beta_i'^{\alpha'_{ik}}}{\Gamma(\alpha'_{ik})} \cdot \lambda'_{ik}{}^{\alpha'_{ik}-1} \cdot e^{-\beta'_i \cdot \lambda'_{ik}} \tag{6}$$

and expected value of λ'_{ik} as

$$E\left[\lambda'_{ik}\right] = \alpha'_{ik} \cdot \frac{1}{\beta'_i} \tag{7}$$

The prior shape parameter α'_{ik} represents the expected number of accidents and is calculated for the ith homogeneous segment and the kth response variable as $\alpha'_{ik} = \dot{\lambda}_k \cdot \beta'_i$. β'_i is the prior inverse scale parameter of the Gamma distribution representing the weighted exposure as given by $\beta'_i = \nu_i \cdot \omega_i = \nu_i \cdot \psi / l_i$. The values of $\beta_i{}'$ are the same for the z different response variables but vary between homogeneous segments according to their lengths and exposures. The exposure ν_i is multiplied by the prior weight ω_i being the fraction of the weighting factor ψ and the individual homogenous segment length l_i. ψ is introduced in a Bayesian sense to give weight to the prior parameter β'_i in order to take into account simultaneously the time period based on which the prior information has been gathered, experts experience and appraisal of the quality of the prior information. The exposure of each homogeneous segment is inversely normalized by its length, i.e. the longer a homogeneous segment, the larger the reduction of the weight given to the values of the prior variables. So-called background rates $\dot{\lambda}_k$ are used for the assessment of the prior Gamma parameters α'_{ik} and $\beta_i{}'$. These background rates can either be determined based on experts' knowledge or based on the analysis of available historical data. In case historical data is used from a large time span (e.g. >10 years) the data has to be assumed non-stationary, since demographical trends and technical developments may have influenced the relationships between risk indicating variables and response variables over time. In smaller time spans (e.g. ≤10 years) analysis of historical data is considered to be representative when it is based on average values of the risk indicating variables and response variables. The determination of the background rates based on analysis of available historical data can, for example, be done by means of a multi-objective optimization algorithm, to find both: the optimal background rates $\dot{\lambda}_k$ for the $k = 1, \ldots, z$ different response variables and the optimal value for the weighting factor ψ, which is the same for all response variables and all homogeneous segments. The objective is to minimize the difference between the observed ($\tilde{y}_{ik}$) and the NB distributed ($\hat{y}_{ik}$) number of events of the response variables, the latter being assessed using Eqs. (2) and (7) as $\hat{y}_{ik} = \nu_i \cdot (\alpha'_{ik} / \beta'_i)$. The optimization problem may be formulated as

$$arg \min_{\psi^*, \dot{\lambda}^*_k} \{f(\psi, \dot{\lambda}_k)\} \tag{8}$$

with

$$f(\psi, \dot{\lambda}_k) = \sum_{i=1}^{n} \tilde{y}_{ik} - \sum_{i=1}^{n} \hat{y}_{ik} \tag{9}$$

subject to $0 < \psi \leq 1$ and $\dot{\lambda}_k > 0$.

As soon as observations $\tilde{y}_{ik}$ become available for the counts of the response variables in the homogeneous segments the prior occurrence rates λ'_{ik} are updated and the Gamma distribution of the posterior rates λ''_{ik} is assessed as (Gelman et al., 2004):

$$\lambda''_{ik} | \tilde{y}_{ik}, \tilde{\nu}_i \sim \text{Gamma}\,(\alpha''_{ik}, \beta''_i). \tag{10}$$

with

$$\alpha''_{ik} = \alpha'_{ik} + \tilde{y}_{ik} \text{ and } \beta''_i = \beta'_i + \tilde{\nu}_i \tag{11}$$

where $\tilde{y}_{k,i}$ represents the observed sum of response variable counts and $\tilde{\nu}_i$ represents the exposure in the ith homogeneous segment over the observed time period. The updated (posterior) rates of the response variables being the result of the Gamma-updating procedure can directly be used as dependent variables for the multivariate regression analysis. By using the posterior rates of the response variables the common problems mentioned with using the Poisson distribution, i.e. over-dispersion and regression-to-the-mean, as discussed earlier are eliminated. This procedure also dilutes the effects of individual outliers of exceedingly high occurrence counts in the dataset through the embedded weighting process, and avoids the preponderance of zero values since the posterior rates of the response variables are always larger than zero. This conforms to the author's assumption that observing zero events on a road segment over a defined period of time certainty does not mean that no accidents may ever occur on the same segment.

2.5. Development of regression models

The fourth step of the proposed methodology is the development of a multivariate Poisson-lognormal regression model for the description of the relationships between the risk indicating variables and the response variables. Regression models in general are comprised of two main components: a structural and a random component. The first specifies the interrelationship between the expected response variables and risk indicating variables. The latter specifies the error terms of the regression analysis by describing the probability distributions of the response variables around their expected value. The error terms represent the heterogeneity and randomness of the modelled response variables and are assumed not to be correlated with the risk indicating variables. In the regression analysis of the proposed methodology the posterior rates are used as response variables with error terms assumed to be lognormal distributed (Ma et al., 2008). Based on that assumption the values of the posterior rates are converted into logarithmic values and subsequently considered as z-dimensional normal distributed random variables (for z different response variables). This allows the application of the multivariate log-linear regression analysis with, now, normal distributed response variables, regression coefficients and error terms. The error terms ε_k have zero means and an estimated standard deviation for the $k = 1, \ldots, z$ different response variables. The normal assumption can be tested (e.g. by means of probability plots of the residuals and linear quantile–quantile plots) and the constant variance of the error term over the entire sample range can be proved.

The model has a hierarchical structure with non-time-dependent random effects and is applied to every homogeneous segment (Lan et al., 2009). The posterior rates λ''_{ik} for every ith homogenous segment are applied as multi-dimensional response variables taking into consideration their co-variances. The structural component of the proposed multivariate log-linear regression model is

$$\ln(E[\mathbf{\Lambda} \,|\mathbf{X}]) = \mathbf{BX} + \mathbf{\Xi} \mathrel{\hat{=}} E[\mathbf{\Lambda} \,|\mathbf{X}] = \exp(\mathbf{BX} + \mathbf{\Xi}) \tag{12}$$

where $\mathbf{X}$ is the design-matrix of $j = 1, \ldots, u$ different risk indicating variables, $\mathbf{\Lambda}$ the response matrix of the $k = 1, \ldots, z$ different response variables, $\mathbf{B}$ the matrix of regression coefficients and $\mathbf{\Xi}$ the matrix of the error terms:

$$\mathbf{\Lambda}=\begin{pmatrix}\lambda''_{11} & \cdots & \lambda''_{1z}\\ \vdots & \ddots & \vdots\\ \lambda''_{n1} & \cdots & \lambda''_{nz}\end{pmatrix},\ \mathbf{X}=\begin{pmatrix}x_{11} & \cdots & x_{1u}\\ \vdots & \ddots & \vdots\\ x_{n1} & \cdots & x_{nu}\end{pmatrix},\ \text{and } \mathbf{\Xi}=\begin{pmatrix}\varepsilon_{11} & \cdots & \varepsilon_{1z}\\ \vdots & \ddots & \vdots\\ \varepsilon_{n1} & \cdots & \varepsilon_{nz}\end{pmatrix} \quad (13)$$

$\boldsymbol{\varepsilon}_{ik}$ represents the vector of the normal distributed random variation for each homogeneous segment and for the kth different response variable with $\mathbf{M}_{\varepsilon_{ik}}$ as the mean and $\mathbf{\Sigma}_{\varepsilon_{ik}}$ as the covariance matrix. For normally distributed response variables (logarithmically transformed), the regression coefficients $\hat{\mathbf{B}}$ estimated by means of the Maximum Likelihood Method correspond to the least squares estimates assessed as

$$\hat{\mathbf{B}}=(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{\Lambda}. \quad (14)$$

with a matrix of residuals $\hat{\mathbf{R}}$ being computed as the difference between the observations and the model predictions of the event occurrence rates, $\mathbf{\Lambda}$ and $\hat{\mathbf{\Lambda}}$, respectively: $\hat{\mathbf{R}}=\mathbf{\Lambda}-\mathbf{XB}=\mathbf{\Lambda}-\hat{\mathbf{\Lambda}}$. The predicted covariance matrix of the error term is assessed based on the residuals as

$$\hat{\mathbf{\Sigma}}=\frac{1}{n-p-1}\hat{\mathbf{R}}^T\hat{\mathbf{R}} \quad (15)$$

where n is the sample size and p the number of considered variables.

Related forms of the proposed regression model can be found in Tunaru (2002), Bijleveld (2005), Park and Lord (2007), Song et al. (2006), Tsionas (2001), Miaou and Lord (2003), Karlis (2003), Karlis and Meligkotsidou (2005), Qin et al. (2005), Ma et al. (2008), El-Basyouny and Sayed (2009a) and El-Basyouny and Sayed (2011). The proposed multivariate Poisson-lognormal regression model can be modified straightforwardly in order to take into account additional risk indicating variables and different functions of dependencies, e.g. time trends,[1] although currently, no temporal and demographic effects are considered. The aggregation of accident counts over a specified period of time may also help to avoid confounding effects as e.g. changes of traffic volume or regression-to-the-mean which might have significant impacts on the results of the models (Elvik, 2002; Cheng and Washington, 2005). The multiplicative structure of the proposed regression model is supported by the investigations of Hauer (2004) where it is recognized that the effect of explanatory variables that influence the probability of accident occurrences over a longer proportion of the road link is more effectively represented by multiplicative terms.

2.6. Construction and parameter learning of BPNs

Step 5 concerns the construction and parameter learning of the BPN. Bayesian inference and updating algorithms are used to establish a full BPN which represents the joint probability density function of all random variables of which the model consists in a compact manner. For general concepts of Bayesian inference calculations the reader is referred to Benjamin and Cornell (1970), Pearl (1988), Congdon (2006) and Ang and Tang (2007). For a detailed description of BPNs reference is given to Kjaerulff and Madsen (2008), Cowell (1999) and Jensen and Nielsen (2007). BPNs are designed to represent the knowledge of a problem, explicitly encoding the dependency between the variables in the model by causal relationships. So-called evidence can be introduced into the parent (input) nodes of the BPN in terms of measured observations of the risk indicating variables. The inference calculation of the BPN uses the structure and the conditional probability tables for propagating the observed information of the evidences through the network and to assess the conditional predictive probability distribution of the response variables. Non-linear relationships between risk indicating variables and response variables can be implemented and the consideration of uncertainties related to the influence of the risk indicating variables on the response variables is facilitated, which is necessary in the estimation of accident risks according to Faber and Maes (2005) and Der Kiureghian and Ditlevsen (2009) since it allows for more realistic standard errors of the resulting model than would otherwise be determined (Li et al., 2008).

The BPN is used to model the conditional probability distributions of the rates of the response variables λ_{ik} given observations of the different risk indicating variables. The joint probability distribution represented by the BPN can be formulated as

$$p\left(\mathbf{\Lambda}\right)=\prod_{k=1}^{z}p(\mathbf{\Lambda}_k\left|\mathbf{X}\right.) \quad (16)$$

$p(\mathbf{\Lambda}_k\left|\mathbf{X}\right.)$ represents the conditional probability of $\mathbf{\Lambda}_k$ given $\mathbf{X}$.

A BPN is defined by two components: the structural component of BPNs can be considered as a directed acyclic graph containing chance nodes representing different random variables either as continuous random variables or as random variables with discrete states or intervals. The nodes are connected through directed edges (arrows) representing the causal dependencies between the random variables (Pelikan, 2005). The parameter component of a BPN is represented by using multidimensional conditional probability tables. In a conditional probability table all conditional probabilities for a variable are assessed given the probabilities of all variables on which the considered one depends. Given a sufficiently large dataset the BPN can learn by both, structural learning and parameter learning, the former meaning the definition of the conditional dependencies and independencies through the determination of an optimized structure of the network, and the latter meaning the updating of the conditional probability tables based on additionally available data.

Structural learning can be done e.g. by using the hill climbing algorithm which is capable to automatically search for an optimal structure (purely based on statistical measures) of the network. The hill climbing algorithm is applied e.g. by De Oña et al. (2011) and described in Tsamardinos et al. (2006) and Madden (2009). Structural learning can also be performed by means of model-building heuristics e.g. scoring metrics which are typically represented by the AIC (Akaike Information Criterion), BIC (Bayesian Information Criterion) or negative log-likelihood values (Pelikan, 2005). For the methodology introduced in this paper, the structure of the BPN is empirically determined and the causal relationships are evaluated based on the outcomes of the regression analysis and complemented by expert's judgements.

Parameter learning of the BPN is done by constructing so called contingency tables which provide input information for the BPN. The contingency tables contain observations of the risk indicating variables and the response variables for each homogeneous segment in the investigated time period. The Expectation-Maximization algorithm (*EM* algorithm) is used as described e.g. in Cox (1983), Fahrmeir and Osuna (2003) and Karlis (2003) to adapt the prior *BPN* to the new dataset. The internal causal interrelationships and dependencies in the BPN are iteratively updated based on additional data. Hence, purely empirical regression model based probabilities and linear relationships are replaced by observation based posterior probabilities and non-linear relationships. Experience factors are applied to weight the content of the prior BPN during the EM algorithm. The values of the experience factors have to be determined for every investigated

[1] Lord and Persaud (2000) and Anastasopoulos and Mannering (2009) compared models including and excluding temporal effects and concluded that both types of models work well.

problem individually according to the expert experience on how much weight should be given to the available prior information and the informative value of the available data. The result of the parameter learning is an updated conditional probability table for the – now termed – posterior BPN. As during the parameter learning process only the cells of the prior BPN are updated for which new data are available, the remaining cells whose values were initially determined for the prior BPN remain unchanged. The parameter learning included in the proposed methodology takes into account the conclusions of previous research about non-linearity in the relationship between exposure and accident rates (Hauer, 1995).

When new data becomes available in future, the model can be updated by means of the same updating procedure as described in the methodology part. However, the values of the prior weight shall be adjusted in order to appropriately take into account the size and informative value of the new dataset.

2.7. Prediction of the expected number of events

The sixth step is the prediction of the expected number of the response variable counts on specific homogeneous segments. In this step evidences for the road in question, i.e. the road for which predictions of the expected number of injury accidents shall be assessed, are entered into the input nodes of the developed posterior BPN. This allows the estimation of the posterior predictive probability density function of the response variables, conditional on the values of the risk indicating variables. The mean value of the posterior predictive probability density function is then multiplied with the exposure of the homogenous segment as given in Eq. (2) and subsequently used as the Poisson distribution parameter to estimate the expected number of response events on the specific road section over a defined period of time (Eq. (1)).

3. Case study

3.1. General

In order to demonstrate the methodology and its usefulness a case study was conducted. The case study taken was the entire Austrian rural motorway network where data on numbers of injury accident events, injured roadway users and various risk indicating variables were kindly provided by the Austrian Road Safety Board (KFV Austria). Nearly 40,000 geographical coordinates were accessible to represent the total length of 1821 km. Since data was provided for both driving directions separately, the total length of the investigated road network is 3642 km. For all risk indicating variables, the geographical position information was converted into so-called motorway kilometres for every road link with a basic resolution of 50 m. The investigated road network is illustrated in Fig. 2. The terming road and motorway are used equivalently in order to represent rural motorways with one or more driving lanes being separated for the different driving directions.

The methodology introduced above implies modern data mining techniques and issues like e.g. double use of data are accordingly taken into account (Section 2.1). For the model development and model testing, the entire dataset was split into two identical structured but independent data subsets containing different randomly selected road sections. The first subset (development dataset) was exclusively used for the model development and not used again for the model testing procedures. The second data subset (test dataset) was exclusively used for model testing purposes to show the capability of the developed model to predict the number of response variable counts on road sections, which have been randomly excluded from the initial dataset and have not been used for the model development. Additionally, a geo-referenced application of the accident risk model was performed for a specific road link. The road link chosen is the A1 motorway between the Austrian cities Vienna and Salzburg, which is especially labelled in Fig. 2. The data of this road link was also excluded from the initial dataset and was then used for testing the model predictions.

3.2. Determination of model variables

3.2.1. Model response variables

Four different response variables were modelled simultaneously: injury accidents, light injuries, severe injuries and fatalities. Acronyms, descriptions and discretization intervals are provided in Table 1. The discretization intervals are determined based on the rates of the response variables λ_k to be used for the development and parameter learning of the BPN. The interval sizes are defined in three increasing steps for the occurrence rates of the injury accidents and in two steps for the rates of the remaining response variables in order to have a higher resolution for the intervals of lower values of the response variable rates. Definitions of injury accidents and injury levels are in accordance with the Austrian penal code (Bundesrepublik Österreich, 2012).

Definitions of different levels of injury severity are also given in Al-Ghamdi (2002), Shankar et al. (1996), Abdel-Aty (2003), Chang and Wang (2006), Simoncic (2004), De Oña et al. (2011) and Milton et al. (2008). Most of these studies define the injury severity of a road accident according to the injury level of the worst injured vehicle occupant. In difference to these references the current case study is not referring to different degrees of injury accidents but directly to the number of road users being injured by different degrees of severity. All injury accidents together with the corresponding numbers of injured road users have been recorded by the Austrian police authority and were allocable to the road network via *GPS* coordinates. In case the injury could not be assigned to one of the injury levels, injuries “with unknowable magnitude” were merged to the group of severe injuries. Mass accidents (>10 vehicles being involved at one accident site) were excluded from the dataset. Additionally, injury accidents which were caused by drivers who were under the influence of alcohol were not taken into account, as well as data which had no clear specification of the location or could not be allocated to one of the driving directions. Table 2 contains the counts of injury accidents and differently injured road users in the 7 years between 2004 and 2010 for the entire dataset and the two sub-datasets. Additionally, the length of the network is provided as the sum of both driving directions of the road network.

3.2.2. Risk indicating variables

The input variables used in this case study were referred to as a set of observable road-specific risk indicating variables. The amount of traffic and the share of heavy good vehicles were chosen as additional risk indicating variables. The risk indicating variable's acronyms, units and intervals for discretization are given in Table 3.

The sample mean (m), sample standard deviation (s), minimum (min) and maximum (max) values of the eight investigated risk indicating variables for the development and test datasets are given in Table 4.

3.3. Construction of homogeneous segments

The homogeneous segments were determined using the values of the risk indicating variables shown in Table 3. The entire road network was sectioned into $n_{HS} = 6932$ homogeneous segments, which were randomly apportioned into the development dataset ($n_{dev} = 5546$, 75% of data) and the test datasets ($n_{test} =$

Table 1
Definition of response variables.

Acronym	Description	Discretization intervals for BPN
IAC	The response variable IAC represents all injury accidents. An injury accident is an accident event where at least one vehicle is involved and at least one occupant becomes at least slightly injured. In this case study exclusively injury accidents are considered together with the corresponding number of injured road users	Interval size of 0.001 for $0 \le \lambda_{IAC} < 0.01$ 0.01 for $0.01 \le \lambda_{IAC} < 0.2$ 0.1 for $0.2 \le \lambda_{IAC} < 2$
LINJ	The response variable LINJ represents the number of light injured road users being involved in the IACs. Light injuries are bodily harms with less than 24 days of damage to health or incapacity to work	Interval size of 0.001 for $0 \le \lambda_{LINJ} < 0.01$ 0.01 for $0.01 \le \lambda_{LINJ} < 0.2$
SINJ	The response variable SINJ represents the number of severe injured road users being involved in the IACs. Severe injuries are considered as aggravated assault. A road user is severe injured if the damage to health or incapacity to work remains longer than 24 days or the injury is severe in a sense that particular organs or bodily parts are affected with uncertain healing process	Interval size of 0.001 for $0 \le \lambda_{SINJ} < 0.01$ 0.01 for $0.01 \le \lambda_{SINJ} < 0.2$
FAT	The response variable FAT represents the number of fatally injured road users being involved in the IACs. A road user is fatally injured when he has died directly at the accident scene or within 30 days after the accident event in hospital as a consequence of the accident induced injuries	Interval size of 0.0001 for $0 \le \lambda_{FAT} < 0.001$ 0.001 for $0.001 \le \lambda_{FAT} < 0.02$

Table 2
Length of investigated road networks and counts of response variables between 2004 and 2010.

	Length	IAC	LINJ	SINJ	FAT
	[km]	[–]	[–]	[–]	[–]
Entire dataset	3642	12,892	14,482	5861	529
Development dataset	2952	10,282	11,460	4677	409
Test dataset	690	2610	3022	1184	120

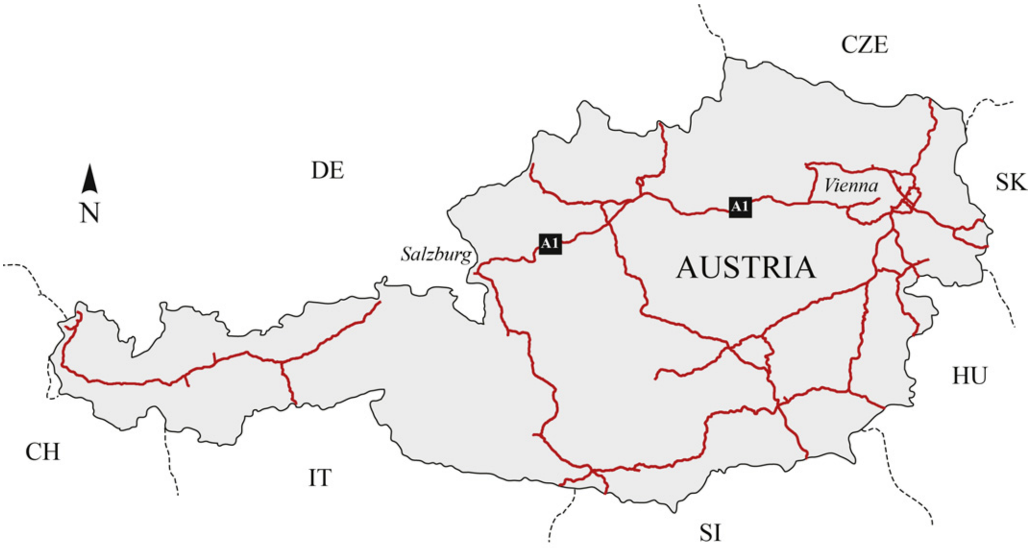

Fig. 2. Austrian motorway network.

1386, 25% of data). Random sampling techniques were applied to allow the consideration of the datasets as representative subsets of the investigated Austrian rural motorway network. Data in both sub-samples were treated identically in terms of pre-assessments and raw data transformations.

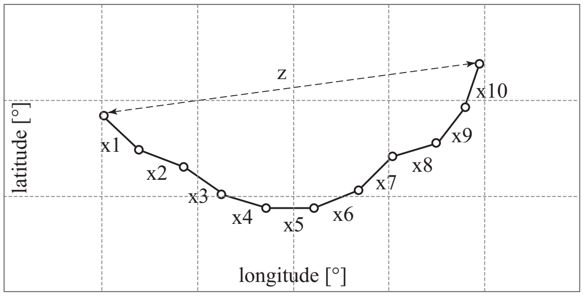

Fig. 3. Calculation of bend factor (curvature).

3.4. Gamma-updating of model response variables

The occurrence frequencies of the response variables were assessed as described in Section 2.3. The parameters of the Gamma distribution were quantified and updated for each homogeneous segment based on the background rates given in Table 5.

Both, the weighting factor ψ and the background rates $\dot{\lambda}_k$ for injury accidents and the different levels of injury were assessed by using a non-linear generalized reduced gradient optimization algorithm to solve the objective function given in Eq. (8). The posterior rates λ''_{ik} were assessed based on the Gamma-updating procedure taking into account the background rates $\dot{\lambda}_k$ and the observed counts of the response variables $\tilde{y}_{ik}$ (likelihoods) for every

Table 3
risk indicating variables for accident risk modelling.

Acronym [units] [intervals]	Description
CHAR [exit corridors, intersections, tunnels, open roads] [1, 2, 3, 4]	The variable CHAR represents different types of road sections, namely (1) exit corridors, (2) intersections, (3) tunnels and (4) normal/open roads. Exit corridors at which vehicles are entering or departing the roads and intersections are defined over a range of one kilometre including a 500 m section before and after the centroid of the exit or intersection
AADT [vehicles/day] [0, 10, . . ., 100] × 10^3	The variable AADT represents the annual average daily traffic pro driving direction. The probability of a road user becoming involved in an injury accident is assumed to be directly connected to their exposure in terms of travel distance, travel time and traffic volume. The travel distances correspond to the length of every homogeneous segment. Based on AADT and length of the homogeneous segments the exposure can be assessed for one year as communicated in terms of million vehicle kilometre travelled (mvk) $v = \text{AADT} \cdot 365 \cdot \text{length},$ (17)
HGV [%] [0,5, . . ., 30]	The variable HGV represents the fraction of heavy good vehicles travelling on the road section with respect to the AADT. with AAHGV being the annual average number of heavy good vehicles $\text{HGV} = \frac{\text{AAHGV}}{\text{AADT}} \cdot 100\%$ (18)
BEND [–] [0, 2, . . .,10]	The variable BEND represents the magnitude of the road curvature in terms of a horizontal bend factor. In general, the curvature of a road segment can be measured by means of the radius, however, for a straight road segment the radius would approximate infinity, which is a value that cannot be used for discrete model assumptions. Thus, the curvature of the road sections was categorized into an integer variable with values between zero (straight road) and ten (very high curvature). The bend factor is calculated as a moving window of ten subsequent road sections being provided by a geographical information system (GIS) in a 50 m grid as illustrated in Fig. 3 using the following equation. This method of calculating the horizontal bend factor is comparable to the detour ratio of Haynes et al. (2007). Discussions about the relationship between road curvature and accident rates can be found in Haynes et al. (2007) and Milton and Mannering (1998) $bend = \left(\frac{\sum_{i=1}^{10} x_i}{z} - 1\right) \cdot 100$ (19) z represents the direct distance between the starting point and the end point. x_i represents the individual road sections taken from the GIS layers having a constant length of 50 m
SLP [%] [−6, −4, −2, 0, . . ., 6]	The variable SLP represents the percentage of the upwards or downwards gradient (slope) of the road separately for the different driving directions. Vehicles slow down, in general, with increasing grade of upwards slope, especially trucks, and this reduced speed often results in an increase in the amount of passing vehicles. Vehicles speed up, in general, with increasing grade of downwards slope and this often results in decreases in the amount of time a driver has to react to an unexpected event and in a reduction in control once a corrective action has been started
LAN [–] [1,2, 3, 4]	The variable LAN represents the number of driving lanes of the road section separately for every driving direction. The minimum value of LAN is 1, the maximum 4 lanes
SPD [km/h] [80, 90, . . ., 130]	The variable SPD represents the signalized speed limit. The Austrian speed limit generally is set to be 130 km/h and road design codes are also based on this speed. On a small fraction of road sections, however, the signalized speed limit is set to different values (e.g. 80 or 100 km/h) in order to adapt the driving speed of the road users to special design characteristics, driver distractions or other accident risk promoting situations
EML [yes/no] = [1/0] 0 (no)/1 (yes)	The binary variable EML represents the existence of road emergency lanes having the values 0 or 1 for "no" and "yes", respectively

Table 4
Numerical summaries of the risk indicating variables based on development and test datasets.

	CHAR	AADT	HGV	BEND	SLP	LAN	SPD	EML
	[–]	[vehicle/day]	[%]	[–]	[%]	[–]	[km/h]	[–]
n = 5546	**Development dataset**							
m	3.08	17,796	12.5	4.2	−0.01	2.2	120	0.56
s	1.26	11,357	4.9	5.74	1.67	0.5	16	0.50
min	1	2800	3	0	−6	1	80	0
max	4	99,950	29	76.3	6	4	130	1
n = 1386	**Test dataset**							
m	3.14	17,521	12.6	4.49	0.01	2.1	120	0.55
s	1.22	10,996	4.9	6.24	1.64	0.5	16	0.50
min	1	2800	3	0	−6	1	80	0
max	4	99,950	29	71	6	4	130	1

Table 5
weighting factor and background rates assessed for model development dataset.

ψ [–]	$\dot{\lambda}_{IAC}$ [IAC/mvk]	$\dot{\lambda}_{LINJ}$ [LINJ/mvk]	$\dot{\lambda}_{SINJ}$ [SINJ/mvk]	$\dot{\lambda}_{FAT}$ [FAT/mvk]
0.3	0.08764	0.09910	0.03705	0.00315

ith homogeneous segment, as well as for every of the $k = 1, \ldots, z$ response variables. Given a time period t, the observed exposure $\tilde{v}_i$ and the length l_i, the posterior parameters were computed according to Eq. (11). The influence of different $\tilde{y}_{ik}$ values on the expected number of injury accidents (based on posterior rates) is shown in Fig. 4.

The Gamma shaped prior probability density function for the expected number of injury accidents is illustrated in Fig. 4 by the dashed line. The probability density functions of the updated posterior distributions are drawn with solid lines. For zero observations only a very small change in the posterior probability density function is observable when it is compared to the prior probability density function. With $\tilde{y}_{ik} = 1$ the posterior probability density function results in a strongly right tailed function with mode around 0.3. For observations of two and three counts of the response variables the tail to the right of the posterior probability density function becomes less concise, however, at the same time the variance is increasing.

3.5. Development of regression model

Regression models, and therefore the values of the regression coefficients and error terms for each, were assessed for every of the four types of road sections (CHAR) simultaneously. Dependent variables of the regression analysis were the Gamma updated posterior rates of the response variables (Table 1). The seven risk indicating variables (excluding CHAR) from Table 3 were used as independent variables. The regression equation used has the multiplicative form as

$$\mathbf{Y}_{ik}\,\Big|\,\mathrm{CHAR} = v_i \cdot \exp\left(\begin{matrix}\boldsymbol{\beta}_{0,ik} + \boldsymbol{\beta}_{1,ik}\cdot\ln(\mathbf{AADT}_i) + \boldsymbol{\beta}_{2,ik}\cdot\ln(\mathbf{HGV}_i) + \boldsymbol{\beta}_{3,ik}\cdot\mathbf{BEND}_i + \ldots \\ \boldsymbol{\beta}_{4,ik}\cdot\mathbf{SLP}_i^2 + \boldsymbol{\beta}_{5,ik}\cdot\mathbf{LAN}_i + \boldsymbol{\beta}_{6,ik}\cdot\mathbf{VEL}_i^2 + \boldsymbol{\beta}_{7,ik}\cdot\mathbf{EML}_i + \varepsilon_k\end{matrix}\right) \tag{20}$$

where $\mathbf{Y}_{ik}$ is a $n \times z$ matrix of the $i = 1, \ldots, n$ homogeneous segments $k = 1, \ldots, z$ response variables. The explanatory variables are the $n \times 1$ vectors of the risk indicating variables and the βs are $n \times z$ matrices. In order to improve the model results, for AADT and HGV logarithms were used instead of the observed raw values of the variables. The values of the variable SLP were used in quadratic form assigning the same accident promoting effects to upwards and downwards slope. Additionally, the signalized speed limit (SPD) was employed in the regression model in a squared format according to previous investigations e.g. Hauer (2009), Malyshkina and Mannering (2008), Aljanahi et al. (1999), Aarts and Van Schagen (2006), Haglund and Åberg (2000) and Nilsson (2004).

Regression analysis was performed on the development dataset with the data of the homogeneous segments weighted according to their individual exposure values. The total size of the weighted dataset was $n_{HS,w} = 73,389$. The statistical significance of the results was tested using a Student's t-test for the individual regression coefficients, where the Null-hypothesis of the t-test with *n-q-1* (q = number of estimated regression coefficients) degrees of freedom was rejected at the significance level of $\alpha = 0.05$.

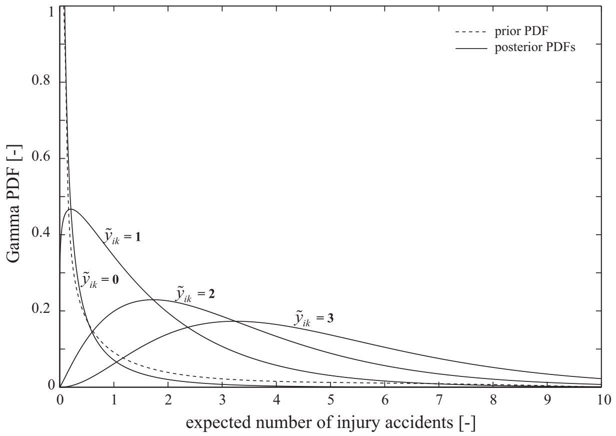

Fig. 4. Updating of Gamma probability density function (pdf) with different values of $\tilde{y}_{ik}$ with constant exposure.

3.5.1. Results and discussion

The values of the multivariate normal maximum likelihood estimates of the expectation operators ($E[.]$) of the regression coefficients together with their values for statistical significance testing (t-statistics) are given in Table 6. Values of the regression coefficients considered to be statistically not significant are marked with asterisks (*). It has to be noted that in Table 6 the estimated regression coefficients are shown for the risk indicating variables and response variables being transformed according to equation 20 of the multivariate regression analysis (Section 3.5).

Even though some regression coefficients in Table 6 appear to be statistically not significant for some combinations of risk indicating variables and response variables, all variables are kept as input parameters for the model since they are significant for other combinations; for example, the variable SLP is less significant on

exit corridors than on other types of road sections (Table 6, *column 5*). It can be observed, that the values of the t-statistics of all risk indicating variables become lower with increasing level of injury, a phenomenon which is assumed to be related to the statistical uncertainties. An increasing level of injury severities leads to a decreasing amount of observations being used for the estimation of the regression coefficients. In the following paragraphs the results of the regression analysis as given in Table 6 are described separately for the particular risk indicating variables. Comparisons are made with results of previous studies, due to the large amount of literature, however, only few example references were selected.

AADT: In Table 6, *column 2* the variable AADT always has a positive influence on the values of the response variables with high values of the corresponding t-statistics. Only for fatalities on open roads (*row 16*) the AADT shows very low negative influence with a t-statistic indicating no statistical significance. This might be due to high statistical uncertainty and large variance (Table 7, "*open roads*") caused by small numbers of fatal injury observations. On open roads only very limited conclusions can be drawn about the dependencies between the risk indicating variable AADT and the occurrence frequencies of fatalities. The lowest impacts are found for open roads and exit corridors with β_1-values mostly smaller than 1. The strongest impact of the AADT on the response variables is observed in tunnels and on intersections with β_1-values mostly larger than 1. In tunnels, exit corridors or on intersections, when the environment requires a higher level of concentration (e.g. due to light changes) than on open roads, an increase of traffic volume seems to additionally increase the accident probabilities. For all road types the β_1-values are higher for injury accidents and light injuries than for severe injuries and fatalities. With increasing exposure of the road users due to higher AADT values the probabilities that one or more vehicles are colliding with each other are increasing. However, such collisions are less likely to result in severe or fatal injuries since the driving speed is considerably reduced. The t-statistics of the variable AADT for all road types are the highest when compared to the ones of the other risk indicating variables. Amongst all risk indicating variables, the AADT is considered to

Table 6
Results of regression analysis. Multivariate normal maximum likelihood estimates of regression coefficients; values of the *t*-statistics in brackets.

Columns		1 Intercept $E[\beta_0]$	2 ln(*AADT*) $E[\beta_1]$	3 ln(*HGV*) $E[\beta_2]$	4 *BEND* $E[\beta_3]$	5 *SLP*^{2} $E[\beta_4]$	6 *LAN* $E[\beta_5]$	7 *SPD*^{2} $E[\beta_6]$	8 *EML* $E[\beta_7]$	Rows
Exits	$\ln(\lambda''_{IAC})$	−11.319 (−43.37)	0.801 (30.81)	0.120 (4.80)	0.014 (7.00)	−4.23E−4* (−0.14)	−0.066 (−3.00)	2.57E−5 (9.38)	−0.190 (−9.05)	1
	$\ln(\lambda''_{LINJ})$	−13.462 (−46.26)	1.027 (35.41)	0.263 (9.39)	0.013 (6.50)	5.83E−3* (1.94)	−0.152 (−6.08)	1.94E−6 (0.64)	−0.137 (−5.96)	2
	$\ln(\lambda''_{SINJ})$	−11.937 (−35.53)	0.573 (17.36)	−0.051* (−1.59)	0.013 (6.50)	7.86E−3* (1.97)	0.240 (8.57)	9.39E−5 (26.60)	−0.007* (−0.26)	3
	$\ln(\lambda''_{FAT})$	−11.708 (−41.96)	0.291 (10.39)	−0.045* (−1.67)	0.016 (8.00)	−2.19E−3* (−0.73)	0.133 (5.54)	4.21E−5 (14.42)	0.052 (2.36)	4
Intersections	$\ln(\lambda''_{IAC})$	−13.877 (−46.26)	1.023 (33.00)	−0.088 (−2.59)	0.023 (23.00)	−0.014 (−3.50)	0.012* (0.46)	5.23E−5 (15.12)	−0.097 (−3.59)	5
	$\ln(\lambda''_{LINJ})$	−22.619 (−61.63)	1.845 (49.86)	0.250 (5.95)	0.031 (15.50)	−0.045 (−9.00)	−0.310 (−9.69)	8.00E−5 (18.91)	−0.013* (−0.39)	6
	$\ln(\lambda''_{SINJ})$	−10.404 (−28.82)	0.452 (12.22)	−0.174 (−4.24)	0.022 (11.00)	−0.004* (−0.80)	0.342 (11.03)	6.76E−5 (16.25)	−0.026* (−0.81)	7
	$\ln(\lambda''_{FAT})$	−20.412 (−61.30)	1.056 (31.06)	0.295 (7.76)	0.005 (2.50)	−0.054 (−13.50)	−0.258 (−8.90)	1.23E−4 (32.03)	0.168 (5.60)	8
Tunnels	$\ln(\lambda''_{IAC})$	−18.811 (−28.63)	1.482 (26.00)	0.947 (11.84)	0.013 (3.25)	0.041 (5.86)	−0.316 (−6.20)	−3.57E−5 (−2.98)	0.058* (1.29)	9
	$\ln(\lambda''_{LINJ})$	−21.314 (−27.33)	1.698 (25.34)	1.048 (11.03)	0.017 (3.40)	0.023 (2.88)	−0.309 (−5.15)	−1.53E−5* (−1.08)	−0.071* (−1.31)	10
	$\ln(\lambda''_{SINJ})$	−13.696 (−16.21)	1.015 (13.90)	0.590 (5.73)	−0.001* (−0.20)	0.060 (6.67)	−0.697 (−10.72)	−8.81E−5 (−5.72)	0.141 (2.43)	11
	$\ln(\lambda''_{FAT})$	−19.655 (−22.36)	1.408 (18.53)	−0.184* (−1.72)	0.001* (0.20)	0.075 (7.50)	−1.938 (−28.50)	2.46E−4 (15.28)	0.525 (8.75)	12
Open roads	$\ln(\lambda''_{IAC})$	−7.609 (−59.91)	0.445 (34.23)	−0.088 (−8.00)	0.011 (11.00)	−0.002* (−2.00)	0.062 (5.64)	1.59E−5 (8.83)	−0.086 (−8.60)	13
	$\ln(\lambda''_{LINJ})$	−7.073 (−48.12)	0.402 (26.80)	−0.166 (−12.77)	0.013 (13.00)	−0.002* (−2.00)	0.118 (9.08)	4.37E−6 (2.10)	−0.055 (−4.58)	14
	$\ln(\lambda''_{SINJ})$	−12.214 (−64.28)	0.676 (35.58)	0.087 (5.44)	−0.010 (−5.00)	−0.005 (−5.00)	0.113 (6.65)	5.33E−5 (19.74)	−0.039 (−2.60)	15
	$\ln(\lambda''_{FAT})$	−8.762 (−38.10)	−0.030* (−1.30)	−0.071 (−3.55)	0.008 (4.00)	−0.007 (−3.50)	0.446 (21.24)	1.04E−5 (3.18)	0.020* (1.11)	16

have the biggest influence on the investigated response variables. The results of the current investigations are in line with the results provided by e.g. Milton and Mannering (1998), Anastasopoulos and Mannering (2009) and Abdel-Aty and Radwan (2000) who are concluding that for a vast majority of road segments the accident frequency is increasing as the AADT increases.

HGV: For the variable HGV, no general conclusions can be drawn about positive or negative influences on the occurrences of the response variables. The signs of the estimated HGV regression coefficients are changing frequently depending on the combination of the different risk indicating variables and response variables. For some of these combinations no statistical significance can be assigned (e.g. for severe injuries and fatalities on exit corridors (Table 6, *column 3*, *rows 3–4*) and fatalities in tunnels (*row 12*)). Considering solely the regression results for HGV on open roads (being the most frequent and representative road type in the investigated road network), the HGV has mainly a negative influence on the values of the response variables. This result is in line with the results of Miaou (1994), Milton and Mannering (1998) as well as Anastasopoulos and Mannering (2009), who conclude that an increased percentage of trucks leads to decreasing frequencies of vehicle overtaking and lane changing behaviour and hence, to a reduced number of accidents. However, as soon as trucks are involved in accidents, the level of injuries is likely to be severe due to the heavy weights and higher impacts. This effect might be the explanation for the positive regression coefficient for SINJ in Table 6, row 15. In tunnels, when the non-significant regression coefficient for fatalities is neglected, a positive impact of the share of HGV on the occurrence of response variable events can be observed. It is assumed that specific factors (e.g. light changes, reduced lane width) might additionally influence the driving confidence of the road users and the risk increasing effect of the HGV might be amplified. Positive impacts also result for light injuries on intersections and on exit corridors, for fatalities on intersections, severe injuries on open roads and injury accidents on exits. This positive influence of HGV on the occurrence frequency of accidents is supported by the results of Joshua and Garber (1990). They show that the involvement of trucks into accidents is mainly triggered by the variables AADT, HGV and SLP.

BEND: In Table 6, *column 4*, the estimated values of the regression coefficients of the variable BEND mostly indicate positive impacts on the response variables (only exception for SINJ in tunnel and on open roads). With increasing radius of the horizontal curve, the values of the response variables are decreasing. These results are supported by the outcomes of the work of Miaou (1994), Shankar et al. (1995), Milton and Mannering (1998), Abdel-Aty and Radwan (2000) as well as Noland and Oh (2004) who state that there is a positive correlation between the horizontal curvature or sharpness of the horizontal curves and the frequency of accidents. In Haynes et al. (2007) and Anastasopoulos and Mannering (2009) an inverse effect of the curvature is found, which they assume to be the result of increased driver alertness in relatively sharper curves.

SLP: The estimated regression coefficients for the variable SLP are given in Table 6, *column 5*. For road intersections and open roads the vertical gradient SLP has a negative influence on all response variables. The same effect is observed in Anastasopoulos and Mannering (2009) and Noland and Oh (2004) investigating the number of vertical curves per mile. However, they also found an inverse effect for the ratio of the vertical curve length over the road segment length. In tunnels, such a positive effect is also observed in the current investigations being in line with the findings of Miaou (1994) and Milton and Mannering (1998). For all response variables on exit corridors (except for severe injuries) the influence of SLP is statistically not significant (Table 6, *rows 1–4*). In the investigations of Abdel-Aty and Radwan (2000) no effect of vertical alignment was observed neither, which might in their opinion be a data-related

Table 7
Covariance matrices for the error terms, showing the covariances (*C*), the variances (VAR) and the correlation coefficients (*r*) for the logarithmic response variables.

$\Sigma_{1,k}$	Exit corridors				$\Sigma_{2,k}$	Intersections			
	$\ln(\lambda''_{IAC})$	$\ln(\lambda''_{LINJ})$	$\ln(\lambda''_{SINJ})$	$\ln(\lambda''_{FAT})$		$\ln(\lambda''_{IAC})$	$\ln(\lambda''_{LINJ})$	$\ln(\lambda''_{SINJ})$	$\ln(\lambda''_{FAT})$
$\ln(\lambda''_{IAC})$	**VAR = 1.215**	*r* = 0.855	*r* = 0.581	*r* = 0.218	$\ln(\lambda''_{IAC})$	**VAR = 0.960**	*r* = 0.877	*r* = 0.511	*r* = 0.249
$\ln(\lambda''_{LINJ})$	*C* = 1.157	**VAR = 1.508**	*r* = 0.291	*r* = 0.194	$\ln(\lambda''_{LINJ})$	*C* = 1.030	**VAR = 1.436**	*r* = 0.241	*r* = 0.271
$\ln(\lambda''_{SINJ})$	*C* = 0.909	*C* = 0.507	**VAR = 2.012**	*r* = 0.236	$\ln(\lambda''_{SINJ})$	*C* = 0.590	*C* = 0.340	**VAR = 1.390**	*r* = 0.195
$\ln(\lambda''_{FAT})$	*C* = 0.283	*C* = 0.280	*C* = 0.394	**VAR = 1.381**	$\ln(\lambda''_{FAT})$	*C* = 0.265	*C* = 0.353	*C* = 0.250	**VAR = 1.183**
$\Sigma_{3,k}$	Tunnels				$\Sigma_{4,k}$	Open roads			
	$\ln(\lambda''_{IAC})$	$\ln(\lambda''_{LINJ})$	$\ln(\lambda''_{SINJ})$	$\ln(\lambda''_{FAT})$		$\ln(\lambda''_{IAC})$	$\ln(\lambda''_{LINJ})$	$\ln(\lambda''_{SINJ})$	$\ln(\lambda''_{FAT})$
$\ln(\lambda''_{IAC})$	**VAR 1.290**	*r* = 0.800	*r* = 0.593	*r* = 0.192	$\ln(\lambda''_{IAC})$	**VAR 1.007**	*r* = 0.822	*r* = 0.569	*r* = 0.178
$\ln(\lambda''_{LINJ})$	*C* = 1.225	**VAR 1.817**	*r* = 0.301	*r* = 0.044	$\ln(\lambda''_{LINJ})$	*C* = 0.960	**VAR 1.355**	*r* = 0.283	*r* = 0.106
$\ln(\lambda''_{SINJ})$	*C* = 0.983	*C* = 0.592	**VAR 2.133**	*r* = 0.100	$\ln(\lambda''_{SINJ})$	*C* = 0.861	*C* = 0.496	**VAR 2.270**	*r* = 0.151
$\ln(\lambda''_{FAT})$	*C* = 0.332	*C* = 0.091	*C* = 0.223	**VAR 2.311**	$\ln(\lambda''_{FAT})$	*C* = 0.326	*C* = 0.225	*C* = 0.415	**VAR 3.338**

effect connected to the topographical flat region where the observations have been recorded. For the investigations presented in this paper, it is assumed that vehicles on exit corridors are decelerating or accelerating in order to leave or enter the motorway and a general higher alertness of the road users is given diluting the effect of road gradients.

LAN: On open roads the variable LAN has positive impact on all response variables (Table 6, *column 6*). A higher number of lanes implies more traffic and hence, more lane changing behaviour and overtaking manoeuvers of the road users. The accident risk is increasing. The values of the investigated regression coefficients are in line with the outcomes provided by Milton and Mannering (1998), Noland (2003) as well as Noland and Oh (2004). The results are different for tunnels in which higher numbers of lanes help to reduce the values of the response variables.

SPD: The variable SPD is positively correlated with all response variables (Table 6, *column 7*). This means that higher signalized speed limits lead to higher injury accident and injury severity probabilities, which is in agreement with the work of Nilsson (2004). The power functions of Nilsson, however, suggest impacts a magnitude larger than the ones observed in the current investigation (Fig. 5(g)). Only in tunnels, speed limits appear to be counteracting the occurrence frequencies of injury accidents and injuries (*rows 9–12*) which is in agreement with the results of Milton and Mannering (1998). This phenomenon is assumed to be caused by the situation that the speed limit is on many road segments already adapted to the combination of safety enhancing elements and disturbing ambient factors (e.g. light changes) on such road segments. The investigations, presented in this paper, show that the variance of the SPD is in general very low over the entire considered motorway network. Deviations of the signalized speed limit from the design speed (e.g. reduction of speed limit from 130 to 100 km/h) are often connected to local segments with higher numbers of observed injury accidents e.g. due to distraction. These conditions might contribute to the non-intuitive negative effects between increased SPD and decreased number of injury accidents.

EML: The occurrence frequencies of the response variables are only loosely influenced by the presence of emergency lanes and hence, the corresponding estimated regression coefficients for the variable EML are statistically not significant for many of the road type – response variable combinations (Table 6, *column 8*, *rows 6*, *7*, *9*, *10* and *16*). This might be supported by the binary definition of EML with values of 0 and 1. For open roads, however, statistical significant impacts of emergency lanes can be observed where the presence of an emergency lane is decreasing the expected numbers of injury accidents, light injuries and severe injuries. For fatalities, however, an increase is shown which might be the result of road users stopping their vehicle on the emergency lanes and thereby creating the situation where other vehicles may crash into it; something that happens with a lower probability in the absence of emergency lanes. Not much literature can be found about the variable EML and hence, results for the outside shoulder width has been used to compare the results with the outcomes of other studies. In Milton and Mannering (1998), Abdel-Aty and Radwan (2000) as well as Noland and Oh (2004) the effect of shoulder width is discussed and it is stated that larger outside shoulder widths are decreasing the accident frequencies.

The covariance matrices for the error terms with $\Sigma_{CHAR,k}$ for the different road characteristics and response variables are given in Table 7. The variances are represented in the diagonal cells (VAR), the covariances (*C*) in the lower left diagonal part and the correlation coefficients (*r*) in the upper right diagonal part of the matrices.

3.5.2. Sensitivity analysis

In order to illustrate the variation of accident rates with variation of the value of the risk indicating variables, a sensitivity analysis was performed where the values of the risk indicating variables were changed one at a time according to their discretization intervals as provided in Table 3. When the value of one risk indicating variable was changed all other variables were kept constant to their most probable value (Table 8), i.e. the interval with the highest relative occurrence frequency of one risk indicating variable determined using the entire dataset.

The results of the sensitivity analysis are shown in Fig. 5. Accident modification factors (AMF) were calculated. The AMF of the *k*th response variable was assessed as

$$\mathrm{AMF}_k = \frac{\tilde{\lambda}_k}{\dot{\lambda}_k} \tag{21}$$

The results of the sensitivity analysis show for the different road characteristics in Fig. 5(a) that there are almost no differences in the AMF for injury accidents, light and severe injuries and the AMF values are all below 1. When the type of road characteristic are varied however, the numbers of fatalities go down considerably for exit corridors and intersections which might be related to reduced driving speed and the fact that vehicle drivers are more concentrated in these sections. Increased values of AADT, HGV and BEND (Fig. 5(b)–(d)) appear to increase the AMF for injury accidents and the different levels of injured road users. For example, increases in AADT result in increases in the injury accident rate but with decreases in injury severity when accidents occur. This can be explained by higher traffic densities and hence, lower speeds and crash impacts. Previous studies state that the influence of AADT on accident rates should not be modelled linearly since investigations showed non-linear relationships (Hauer, 1995) some with

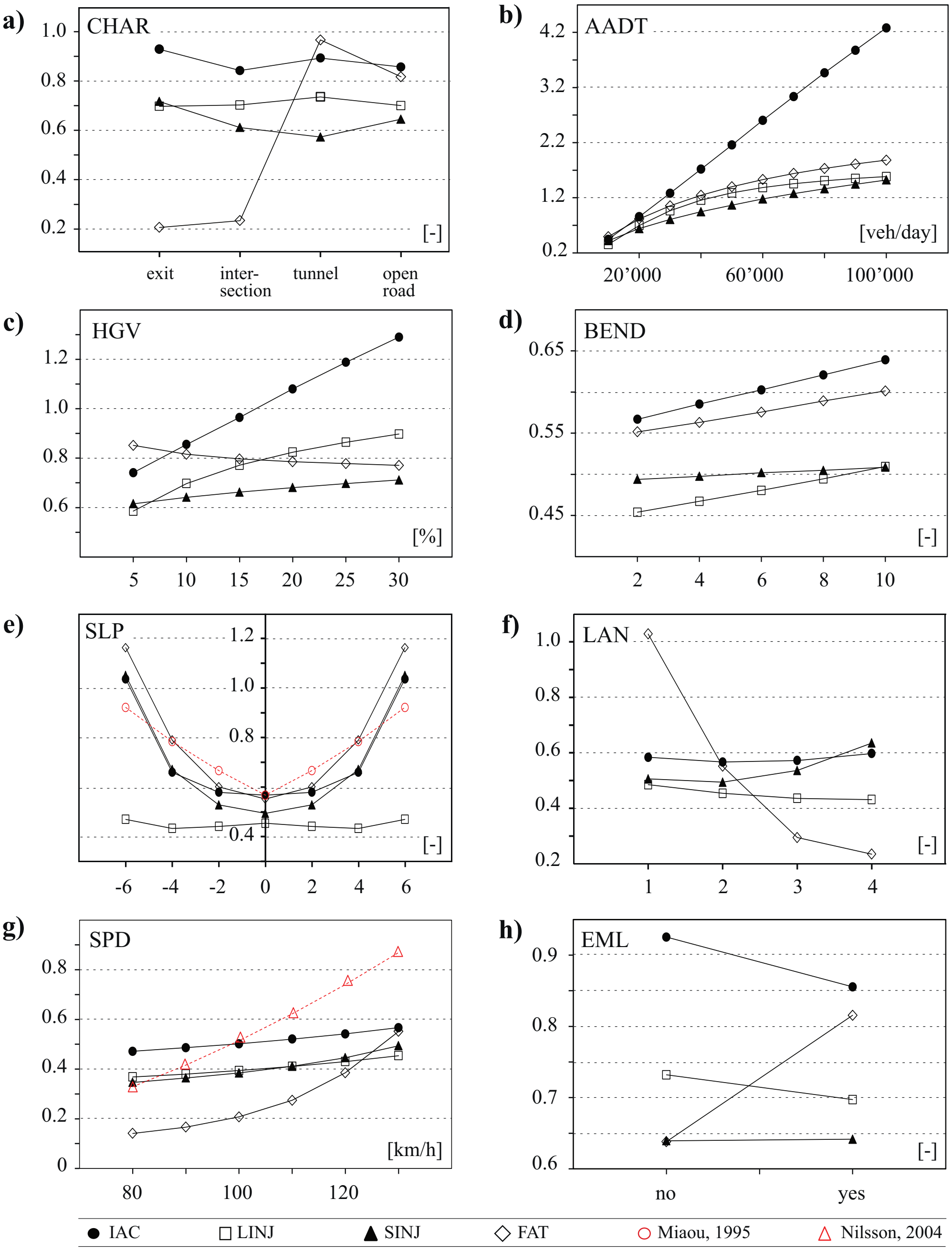

Fig. 5. Sensitivity analysis by means of the AMF when the different risk indicating variables were varied.

the pattern of increasing AMF to a certain amount of traffic which then turns to decreasing AMF values (Schubert et al., 2011). The quadratic values of SLP prove to fit well to previously established models like the one of Miaou (1995) which is adopted to the reference AMF value given SLP = 0. Fig. 5(e) shows the impacts of the slope of the current investigations in comparison to Miaou's model for road gradients. Only for the light injuries the SLP seems to have no impact. For SPD (Fig. 5(g)) the power model of Nilsson (2004) assessed with reference speed of 100 km/h appears to have a much steeper gradient of the function for injury accidents than the model assessed based on the dataset of the current investigation. The function between fatalities and SPD is approaching an exponential form. The impact of the number of lanes in every driving direction (Fig. 5(f)) is rather low except of the case when fatalities are considered since the AMF of the fatalities is reduced remarkably for roads with increasing number of lanes. This trend seems

Table 8
Most probable values of the risk indicating variables according to the relative frequencies in the defined intervals.

CHAR [–]	AADT [vehicles/day]	HGV [%]	BEND [–]	SLP [%]	LAN [–]	SPD [km/h]	EML [no/yes]
Open road	20,000	10	2	0	2	130	Yes

at first glance to be contradicting the results presented in Table 6, *column 6* but it has to be kept in mind that due to the exclusive consideration of the most probable values the sensitivity analysis covers only a very small share of the entire analysis results. In accordance with the results of the regression analysis in Table 6, *column 8*, it can be observed in Fig. 5(h) that the existence of EML has only faint negative influence on the AMF of the injury accidents but noticeable positive effect on fatalities. There is an increase in the number of fatalities when emergency lanes do exist. This effect could be caused by breakdown vehicles which are stopped at the emergency lane. Occupants may step out of the vehicles carelessly and may get run over; or rear-end collision accidents may happen more frequently with higher impacts.

3.6. *Construction and parameter learning of BPNs*

Both the estimates of the regression coefficients, as well as the distribution of the error term, were used to assess the predictive distribution of the response variables for establishing the prior BPN.

3.6.1. *Prior BPN*

The outcomes of the regression analysis were used to develop a prior BPN. The inference engine of Genie 2.0 (Decision-Systems-Laboratory-Pittsburgh, 2006) was applied to construct the network and to calculate the marginal probability distribution functions. The values of the random response variables were discretized in order to be used in a straightforward manner for the development and parameter learning procedures of the BPN. Structural learning was not applied in this investigations since the causal relationships were evaluated and determined based on the outcomes of the regression analysis and expert judgement. The structure of the developed BPN is given in Fig. 6.

The BPN contains eight parent nodes for the different risk indicating variables and four child nodes for the response variables. The response variables are the rates of injury accidents and the different injury severity levels being connected to all parent nodes by directed edges. Only discrete state BPNs were considered. Based on the estimated distributions of the regression coefficients and the covariance matrices of the error terms, Monte Carlo simulations of all regression coefficients and error terms were performed in order to establish the predictive probability density functions of the response variables and to fill the conditional probability tables of the prior BPN. Simulations were also used to extrapolate the outcome of the regression analysis to the entire modelling space and to provide accident rates also in those domains of the conditional probability tables where no observations were currently available. A particular homogeneous segment of the road was then described by putting evidence in the different parent nodes by selecting the appropriate states (e.g. AADT = 40,000, HGV = 12%, etc.). Each combination of the node states has its own predictive probability density function of the response variables. The distributions of the response variables were discretized into 48 states of the predictive injury accident rates and 30 states for each of the different injury severities. The sizes of the four conditional probability tables (one for each response variable) correspond to the products of the number of states of the response variables and the risk indicating variables. As an example, the size of the injury accidents conditional probability table is 19,353,600 cells, assessed as:

$$q = \lambda_{IAC,states} \cdot \mathrm{CHAR}_{states} \cdot \mathrm{AADT}_{states} \cdot \mathrm{HGV}_{states} \cdot \mathrm{BEND}_{states} \cdot \mathrm{SLP}_{states} \cdot \mathrm{LAN}_{states} \cdot \mathrm{SPD}_{states} \cdot \mathrm{EML}_{states}$$
$$q = 48 \cdot 4 \cdot 10 \cdot 6 \cdot 5 \cdot 7 \cdot 4 \cdot 6 \cdot 2 = 19,353,600 \tag{22}$$

In the values of the conditional probability tables are allocated to every of the cells describing the probability that particular injury accident rates or injury rates are occurring, conditional on the evidence values of the risk indicating variables in the parent nodes.

3.6.2. *Posterior BPN*

Using the EM-algorithm the prior BPN was updated to the posterior BPN based on observations of the risk indicating variables and response variables being recorded in so-called contingency tables which were established using the information of the development dataset. The contingency table features twelve columns representing the twelve nodes of the prior BPN, eight for the different risk indicating variables plus four for the observations of the response variables. The parameter learning was performed assuming a value for the experience factor for the EM-algorithm of 0.1 which gives almost no weight to the prior information and hence, the posterior distribution represents essentially the observed data. During the parameter learning process only these domains of the prior BPN for which information were available in the development dataset were updated using Bayesian inference and the EM-algorithm. For the domains of the prior BPN for which no information was available in the development dataset, the parameter learning could not be performed and the prior probabilities assessed by means of the multivariate regression analysis were kept (Section 2.5). The updating of the prior model can be considered as a replacement of the prior model probabilities with the values of the updated posterior model probabilities. The outcome of the learning process is the posterior BPN providing the predictive probability density function of the accident rate, conditional on the observations of the risk indicating variables.

The mean value of predictive probability density functions of the response variables was used as the expected value of the Poisson parameter $\hat{\lambda}_{ik}$ being multiplied with the observed exposure $\tilde{v}_i$ (Eq. (2)) to calculate the expected number of injury accidents and injured road users according to Eq. (1). The observed and predicted numbers of injury accidents and light injuries are plotted in Fig. 7 separately for the prior and posterior BPN.

The model predictions of the numbers of injury accidents and lightly injured road users for the development dataset are shown in the scatter plots of Fig. 7. The prior BPN predictions are exclusively based on the multivariate regression analysis. By means of the parameter learning procedure the posterior BPN was established and the posterior BPN predictions were assessed. It can be seen that both, the correlation coefficients and the regression equation between the observations and the predictions are improved by the parameter learning procedure. The correlation coefficients (*r*-values) for injury accident predictions are increased from $r = 0.71$ to $r = 0.80$ and for the predicted number of light injured road users from $r = 0.61$ to $r = 0.70$. The correlation coefficients for the number of severe injured road users and for fatalities were increased from $r = 0.59$ to $r = 0.67$ and $r = 0.26$ to $r = 0.48$, respectively. The regression equation is improved to the point that there is almost perfect accordance ($y = 0 + 1x$) between the model predictions and observations. The results of the prior BPN indicate considerable bias for the regression line that is remarkably reduced in the scatter plots of the posterior BPN after the parameter learning process.

3.7. *Prediction of expected number of events*

3.7.1. *Model application to randomly selected road segments*

The developed posterior BPN was applied to assess the predictive distribution of the response variables for every homogeneous segment of the test dataset. The test dataset contained randomly selected homogeneous segments which have not been used for the model development. The model predictions of the numbers of

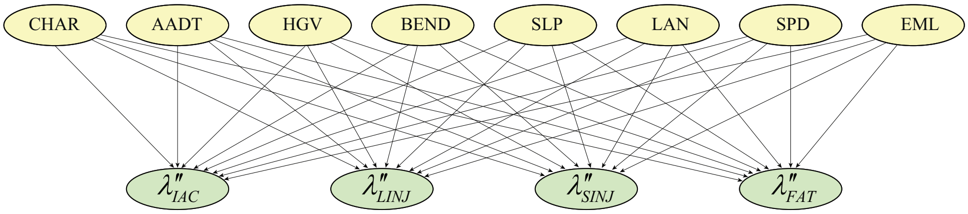

Fig. 6. Structure of the developed BPN.

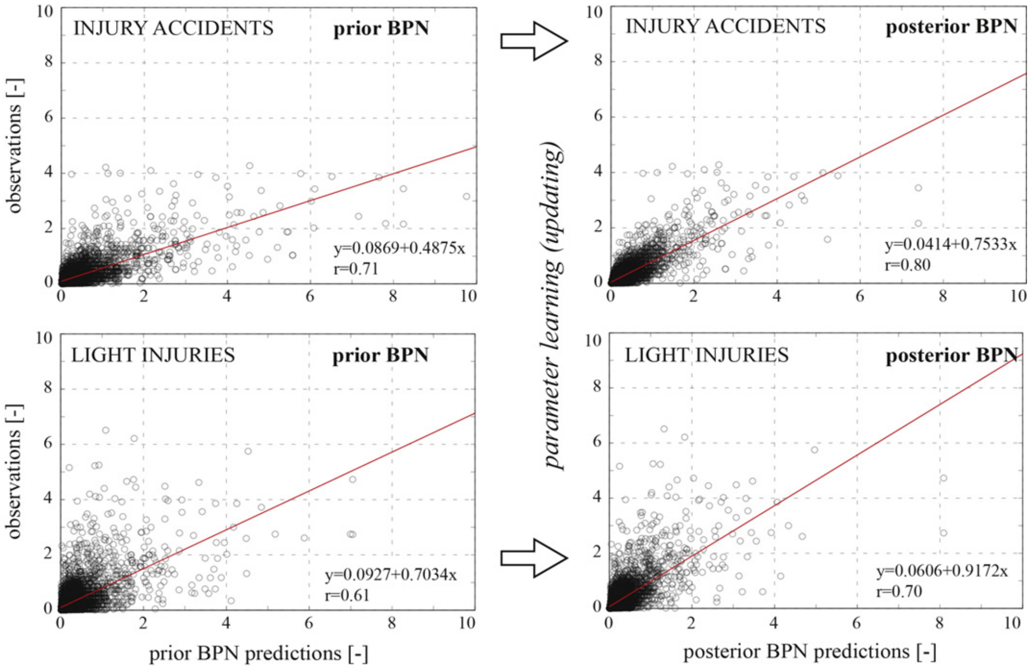

Fig. 7. Comparison of predicted and observed number of injury accidents and light injuries for the prior and posterior BPN modelling results.

injury accidents and lightly injured road users were compared to the real observations of the response variables on the homogeneous segments of the test dataset. The results are illustrated in the scatter plots of Fig. 8.

The r-values of the test dataset are lower than those assessed based on the development dataset, which is reasonable since the models have been established based on the observations of the development dataset. For injury accidents and light injuries comparatively high correlations are achieved with values of $r = 0.73$ and $r = 0.67$, respectively. For the severe injured road users and for the number of fatalities the correlation coefficients between model predictions and real observations are comparatively low with $r = 0.50$ and $r = 0.31$, respectively. It is assumed that the low r-values for severe injuries and fatalities are mainly due to the small number of observations of these injury categories. Additionally, accidents with a high level of injury severity might often be caused by very special circumstances and confounding variables like distraction, health problems or weather related phenomena as e.g. dense fog, clear ice, strong rain etc. Unfortunately information about such confounding variables was not provided in the available dataset and hence, could not be considered and incorporated into the model structure. As soon as such data would become

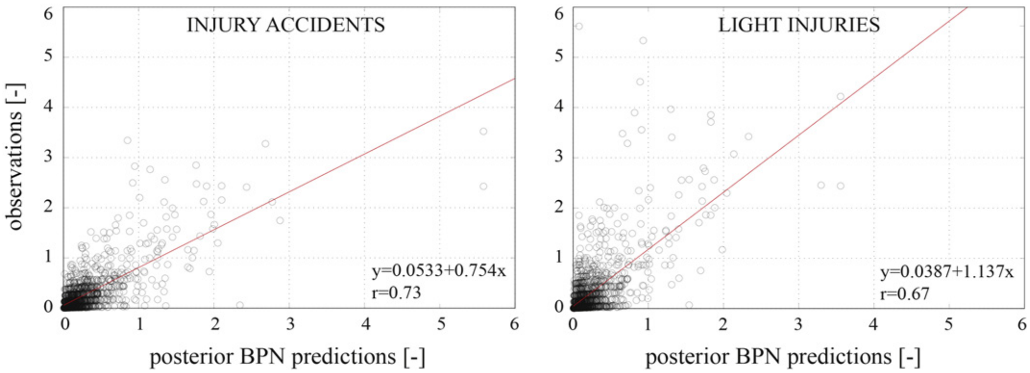

Fig. 8. Comparison of predicted and observed number of injury accidents and light injuries for the prior and posterior BPN modelling results being applied to homogeneous segments of the test dataset.

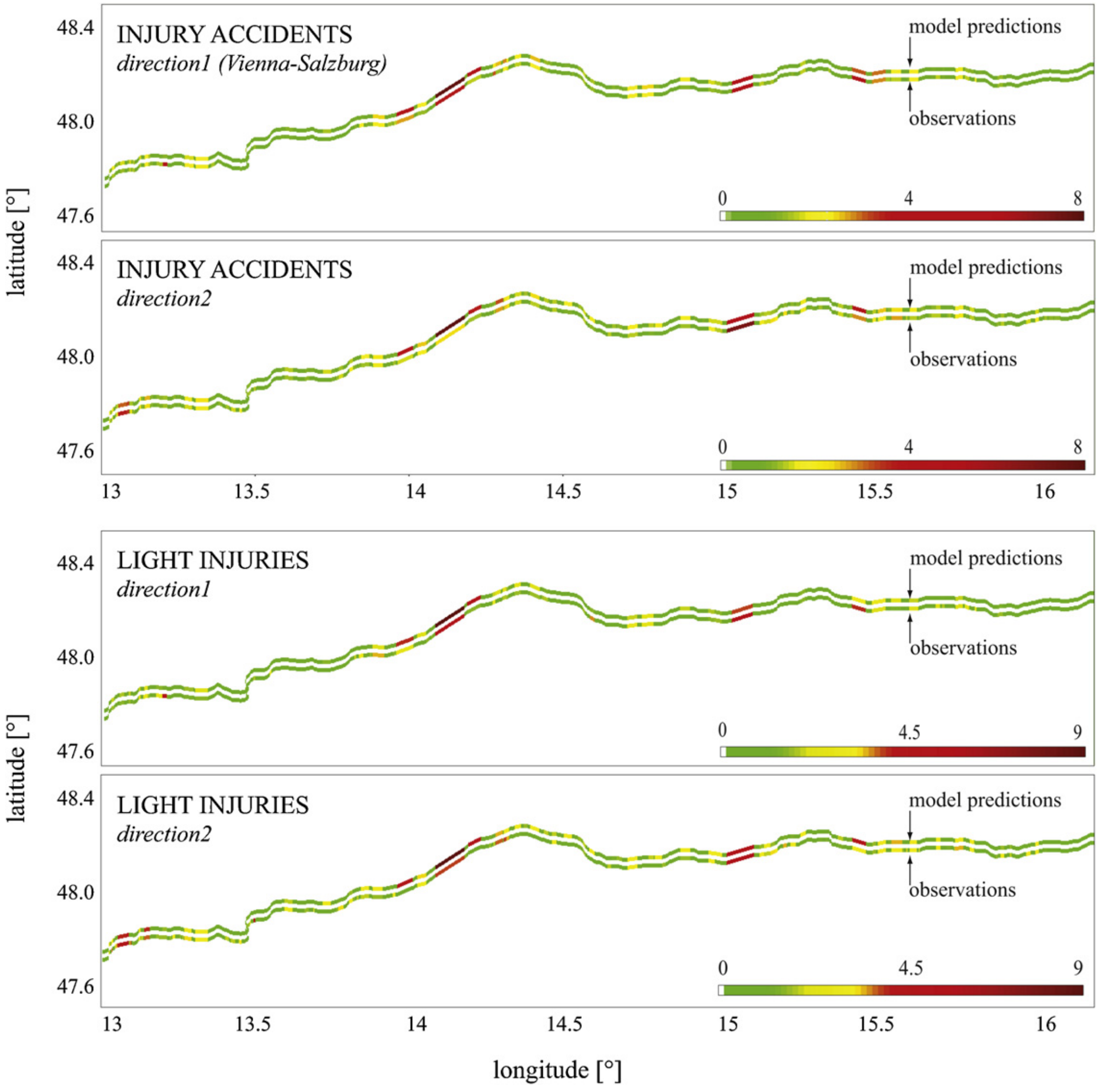

Fig. 9. Graphical comparison between the observed and the predicted numbers of injury accidents and injured road users on the A1 road link over a time period of seven years.

available the implementation of additional variables, e.g. related to weather phenomena into the accident prediction model might have a beneficial effect on the capability of the model to predict also accidents of high injury severity. However, at the same time the model complexity would be increased and so the difficultness for defining the causal relationships in the model structure. Related investigations and discussions have effectually been done in previous studies (Milton and Mannering, 1998; Milton et al., 2008). Another option for reducing the statistical uncertainties would be to combine severe and fatal accidents to one group of response variables in order to increase the number of observed data and decrease statistical uncertainty as done before e.g. in Shankar et al. (1996). This step appears to be reasonable in case of sparse data. However, combination of fatalities and injuries might be potentially misleading as results will show that the risk indicating variables associated with fatalities could be quite different than those associated with injuries. This is also observed by Noland and Quddus (2003).

3.7.2. Model application to specific road link

On a second step, the expected numbers of accident events were predicted for the road link A1 between the Austrian cities Vienna and Salzburg. This link has a length of 292 km and was segregated into 353 homogeneous segments following the same segmentation technique as described above. The length of the A1 road link corresponds to approximately 14% of the entire Austrian road network length. The expected numbers of injury accidents and injured road users were predicted for both directions separately. For every homogeneous segment, information about the risk indicating variables and about the number of observed injury accidents as well as different injury severities were available for the years 2004–2010.

The results are shown in Fig. 9, separate for the injury accidents and the lightly injured road users. Fig. 9 is sectioned into four sub-graphics, each of which contains two coloured lines. The shape and position of the upper line in Fig. 9 correspond to the geographical course of the road link A1 with respect to the longitudinal and latitudinal coordinates as given at the x- and y-axis, respectively. For illustrational purposes the lower line of the predictions is shifted by -0.01 latitudinal degrees. The lines consist of the sequence of homogeneous segments (ordered according to their motorway kilometre) along the A1 road link and are colour coded to indicate the total number of observed or expected number of events (a relatively low number of events in light shades (green) and a relatively high number of events in dark shades (red)). The sum of the real observations over the years 2004–2010 are represented by the upper lines and the predictions are represented by the lower lines.

It can be seen in a qualitative manner that the developed risk model is capable to predict both, the homogeneous segments with a relatively high number of events and those with a relatively low number of events. It can also be seen that the predicted numbers of events for many homogeneous segments are consistent with the observed values. Some deviations are visible on some of the segments which might be caused due to the consideration of only non-specific risk indicating variables (transferable to other road networks) and neglecting individual characteristics of the road users or specific local environmental conditions. It might also be because of the existence of confounding variables for the A1 link, variables, which may be unique on that road link and are therefore, not completely represented in the overall risk model, which has been developed without the data of the A1 road link.

Summing up the results of the case studies, it can be observed that the combination of the Gamma-updating, the multivariate regression analysis and the parameter learning in the Bayesian Probabilistic Network make it possible to predict the expected numbers of injury accidents and injured road users and to deal with both, over-dispersion and a very general covariance structure in the available data.

4. Discussion

For the development of the BPN, the risk indicating variables were assumed to be random variables which could be treated as continuous or discrete distributed variables. In the current investigations the risk indicating variables were discretized into defined intervals into which the observed values have been allocated. In general, it is possible for the methodology of BPNs to treat the risk indicating variables as continuous random variables however discrete values facilitate to remarkably increase computing speed. This advantage is considered to overweigh the small loss of information caused by using discretized data. The methodology allows every time to modify the BPN nodes for continuous distributions of the risk indicating variable's as soon as this is required.

Although it is acknowledged that there may be spatial correlations between the consecutive road-sections that also influence the accident rates, these correlations were not considered in the development of the accident risk model in this case study. It was decided that the additional accuracy that might be obtained by including the explanatory information based on aggregated data and unobserved heterogeneity required to consider such spatial correlations, did not outweigh the desire for a robust, simple and easily understandable model. Validation of this assumption will, however, require future research. Another potentially significant factor in prediction of accident rates that was left out of this model was the presence of construction sites. Construction sites were not considered due to lack of information with respect to their locations and timing.

The mentioned attributes are only related to the developed risk model of the case study with the intention to keep the model on a reasonable level of complexity. They do not affect the proposed methodology for the development of the risk models since they might be overcome by a more sophisticated structure of the accident risk models based on more information in the available datasets.

5. Summary and conclusions

In this paper, a methodology is proposed to determine models that can be used to predict the number of injury accidents and injury severities of road users that occur on roads, where no or little data exist for the specific road segment in question. The usefulness of the proposed methodology is demonstrated in a case study using the Austrian road network.

The risk models developed are formulated in terms of risk indicating variables using Bayesian Probabilistic Networks for homogeneous road segments. The networks are developed by combining both, hierarchical multivariate regression analyses for the assessment of prior inferences and modern data mining techniques to adapt the Bayesian Probabilistic Networks to the available data. The developed models are both, generic and precise in their predictive ability. The generic character allows the developed models to be easily adapted for use on different road networks and to be easily modified to include additional risk indicating variables, if deemed necessary.

In the case study, the model response variables considered are the updated occurrence rates of injury accidents and involved injured road users having no more than light injuries, severe injuries and fatal injuries. The risk indicating variables are selected taking into consideration traffic characteristics and the design parameters of the road, such as traffic volume, traffic composition, speed, curvature and number of lanes. The developed risk model was verified by testing its ability to predict the values of the model response variables for randomly selected road sections the data of which has been excluded from the dataset used for initial model development. Compliance was found between the predicted and observed numbers of response variables with correlation coefficient up to $r = 0.73$. Based on this compliance it was also shown that the model could be used in a geo-referenced manner of application to predict locations of injury accident black spots.

The proposed methodology facilitates the development of accurate models to predict the number of injury accidents and differently injured road users that might occur on a specific road section. The combination of the Gamma-updating, the multivariate Poisson-lognormal regression analysis and the parameter learning in the Bayesian Probabilistic Network make it possible to deal with both, over-dispersion and a very general covariance structure in the available data. When injured road users being involved in the injury accidents are classified by their injury severity these models may also be thought of as accident risk models being relevant for risk informed decision making in the context of traffic and road network management. Relevant potential is also seen for applying the developed methodology for road safety assessment on planned but not yet constructed roads.

Acknowledgements

The work presented in this paper was conducted with financial support by the Austrian Road Safety Board (KFV Austria). The KFV Austria also provided and pre-processed the data for the case study.

References

Aarts, L., Van Schagen, I., 2006. Driving speed and the risk of road crashes: a review. Accident Analysis & Prevention 38 (2), 215–224.

Abdel-Aty, M., 2003. Analysis of driver injury severity levels at multiple locations using ordered probit models. Journal of Safety Research 34 (5), 597–603.

Abdel-Aty, M.A., Radwan, A.E., 2000. Modeling traffic accident occurrence and involvement. Accident Analysis & Prevention 32 (5), 633–642.

Al-Ghamdi, A.S., 2002. Using logistic regression to estimate the influence of accident factors on accident severity. Accident Analysis & Prevention 34 (6), 729–741.

Aljanahi, A.a.M., Rhodes, A.H., Metcalfe, A.V., 1999. Speed, speed limits and road traffic accidents under free flow conditions. Accident Analysis & Prevention 31 (1–2), 161–168.

Anastasopoulos, P.C., Mannering, F.L., 2009. A note on modeling vehicle accident frequencies with random-parameters count models. Accident Analysis & Prevention 41 (1), 153–159.

Ang, A.H.-S., Tang, W.H., 2007. Probability Concepts in Engineering: Emphasis on Applications to Civil & Environmental Engineering, 2nd ed. Wiley, Hoboken.

Benjamin, J.R., Cornell, C., 1970. Probability, Statistics and Decision for Civil Engineers. McGraw-Hill, New York.

Berk, R., Macdonald, J., 2008. Overdispersion and poisson regression. Journal of Quantitative Criminology 24 (3), 269–284.

Bijleveld, F.D., 2005. The covariance between the number of accidents and the number of victims in multivariate analysis of accident related outcomes. Accident Analysis & Prevention 37 (4), 591–600.

Bundesrepublik Österreich, 2012. Strafgesetzbuch: §84. Schwere Körperverletzung.

Carlin, B.P., Louis, T.A., 2000. Bayes and Empirical Bayes Methods for Data Analysis, second ed. Chapman and Hall, Boca Raton.

Carriquiry, A., Pawlovich, M., 2005. From empirical Bayes to full Bayes: methods for analyzing traffic safety data. http://www.iowadot.gov/crashanalysis/pdfs/eb_fb_comparison_whitepaper_october2004.pdf

Chang, L.-Y., Wang, H.-W., 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.

Congdon, P., 2006. Bayesian Statistical Modelling. John Wiley & Sons Ltd., Chichester.

Cowell, R.G., 1999. Probabilistic Networks and Expert Systems. Springer, New York.

Cox, D.R., 1983. Some remarks on overdispersion. Biometrika 70 (1), 269–274.

Davis, G.A., Pei, J., 2003. Bayesian networks and traffic accident reconstruction. In: Proceedings of the 9th International Conference on Artificial Intelligence and Law, ACM, Scotland, United Kingdom, pp. 171–176.

De Oña, J., Mujalli, R.O., Calvo, F.J., 2011. Analysis of traffic accident injury severity on spanish rural highways using Bayesian networks. Accident Analysis & Prevention 43 (1), 402–411.

Dean, C., Lawless, J.F., 1989. Tests for detecting overdispersion in Poisson regression models. Journal of the American Statistical Association 84 (406), 467–472.

Decision-Systems-Laboratory-Pittsburgh, 2006. Genie 2.0, 2nd ed. Decision Systems Laboratory, Pittsburgh, USA http://dsl.sis.pitt.edu

Der Kiureghian, A.D., Ditlevsen, O., 2009. Aleatory or epistemic? Does it matter? Structural Safety 31 (2), 105–112.

El-Basyouny, K., Sayed, T., 2009a. Accident prediction models with random corridor parameters. Accident Analysis & Prevention 41 (5), 1118–1123.

El-Basyouny, K., Sayed, T., 2009b. Collision prediction models using multivariate Poisson-lognormal regression. Accident Analysis & Prevention 41 (4), 820–828.

El-Basyouny, K., Sayed, T., 2011. A full Bayes multivariate intervention model with random parameters among matched pairs for before–after safety evaluation. Accident Analysis & Prevention 43 (1), 87–94.

Elvik, R., 2002. The importance of confounding in observational before-and-after studies of road safety measures. Accident Analysis & Prevention 34 (5), 631–635.

Elvik, R., 2008. The predictive validity of empirical Bayes estimates of road safety. Accident Analysis & Prevention 40 (6), 1964–1969.

Elvik, R., 2011. Assessing causality in multivariate accident models. Accident Analysis & Prevention 43 (1), 253–264.

Faber, M.H., Maes, M.A., 2005. Epistemic uncertainties and system choice in decision making. In: Proceedings of the ICOSSAR2005, 9th International Conference on Structural Safety and Reliability, Rome, Italy, pp. 3519–3526.

Fahrmeir, L., Osuna, L., 2003. Structured count data regression. In: Sonderforschungsbericht. Ludwig-Maximilians-Universität München, Munich.

Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B., 2004. Bayesian Data Analysis. Chapman & Hall/CRC, Boca Raton.

Gschloessl, S., Czado, C., 2006. Modelling Count Data With Overdispersion and Spatial Effects. Center of Mathematical Sciences, TU Munich.

Haglund, M., Åberg, L., 2000. Speed choice in relation to speed limit and influences from other drivers. Transportation Research Part F: Traffic Psychology and Behaviour 3 (1), 39–51.

Hauer, E., 1995. On exposure and accident rate. Traffic Engineering and Control.

Hauer, E., 2001. Overdispersion in modelling accidents on road sections and in empirical Bayes estimation. Accident Analysis & Prevention 33 (6), 799–808.

Hauer, E., 2004. Statistical road safety modeling. Statistical Methods and Safety Data Analysis and Evaluation (1897), 81–87.

Hauer, E., 2009. Speed and safety. Transportation Research Record (2103), 10–17.

Hauer, E., Harwood, D.W., Council, F.M., Griffith, M.S., 2002. Estimating safety by the empirical Bayes method: a tutorial. Transportation Research Record: Journal of the Transportation Research Board 1784 (−1), 126–131.

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. Environmental and Planning A 39 (5).

Heydecker, B.G., Wu, J., 2001. Identification of sites for road accident remedial work by Bayesian statistical methods: an example of uncertain inference. Advances in Engineering Software 32 (10–11), 859–869.

Hossain, M., Muromachi, Y., 2012. A Bayesian network based framework for real-time crash prediction on the basic freeway segments of urban expressways. Accident Analysis & Prevention 45 (0), 373–381.

Huang, H., Abdel-Aty, M., 2010. Multilevel data and Bayesian analysis in traffic safety. Accident Analysis & Prevention 42 (6), 1556–1565.

Jensen, F.V., Nielsen, T.D., 2007. Bayesian Networks and Decision Graphs, 2nd ed. Springer, New York, NY.

Joshua, S.C., Garber, N.J., 1990. Estimating truck accident rate and involvements using linear and Poisson regression models. Transportation Planning and Technology 15 (1), 41–58.

Kaplan, S., Garrick, B.J., 1981. On the quantitative definition of risk. Risk Analysis 1 (1), 11–27.

Karlis, D., 2003. An em algorithm for multivariate Poisson distribution and related models. Journal of Applied Statistics 30 (1), 63–77.

Karlis, D., Meligkotsidou, L., 2005. Multivariate Poisson regression with covariance structure. Statistics and Computing 15 (4), 255–265.

Karwa, V., Slavkovic, A.B., Donnell, E.T., 2011. Causal inference in transportation safety studies: comparison of potential outcomes and causal diagrams. Annals of Applied Statistics 5 (2B), 1428–1455.

Kjaerulff, U.B., Madsen, A.L., 2008. Bayesian Networks and Influence Diagrams: A Guide to Construction and Analysis. Springer, New York, NY.

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. Accident Analysis & Prevention 41 (3), 574–580.

Li, W., Carriquiry, A., Pawlovich, M., Welch, T., 2008. The choice of statistical models in road safety countermeasure effectiveness studies in Iowa. Accident Analysis & Prevention 40 (4), 1531–1542.

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.

Lord, D., Persaud, B., 2000. Accident prediction models with and without trend: application of the generalized estimating equations procedure. Transportation Research Record: Journal of the Transportation Research Board 1717 (1), 102–108.

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 (1), 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.

Macnab, Y.C., 2003. A bayesian hierarchical model for accident and injury surveillance. Accident Analysis & Prevention 35 (1), 91–102.

Madden, M.G., 2009. On the classification performance of TAN and general Bayesian networks. Knowledge-Based Systems 22 (7), 489–495.

Maes, M., Dann, M., Sarkar, S., Midtgaard, A., 2007. Event occurrences within a spatial network using hierarchical Bayes. In: International Forum on Engineering Decision Making, Shoal Bay, Australia, p. 10.

Malyshkina, N., Mannering, F., 2008. Effect of increases in speed limits on severities of injuries in accidents. Transportation Research Record: Journal of the Transportation Research Board 2083 (1), 122–127.

Marsh, W., Bearfield, G., 2004. Using Bayesian networks to model accident causation in the UK railway industry. In: Proceedings of the 7th International Conference on Probabilistic Safety Assessment and Management, PSAM7, Berlin, Germany.

Miaou, S.-P., 1994. The relationship between truck accidents and geometric design of road sections: Poisson versus negative binomial regressions. Accident Analysis & Prevention 26 (4), 471–482.

Miaou, S.-P., 1995. Development of Adjustment Factors for Single Vehicle Run-off-the-road Accident Rates by Horizontal Curvature and Grade. Oak Ridge National Laboratory.

Miaou, S.-P., Lord, D., 2003. Modeling traffic crash-flow relationships for intersections: dispersion parameter, functional form, and Bayes versus empirical Bayes methods. Transportation Research Record: Journal of the Transportation Research Board 1840 (1), 31–40.

Miaou, S.-P., 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., Mannering, F., 1998. The relationship among highway geometrics, traffic-related elements and motor-vehicle accident frequencies. Transportation 25 (4), 395–413.

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.

Mujalli, R.O., De Oña, J., 2011. A method for simplifying the analysis of traffic accidents injury severity on two-lane highways using Bayesian networks. Journal of Safety Research 42 (5), 317–326.

Nilsson, G., 2004. Traffic safety dimensions and the power model to describe the effect of speed on safety. Lund Institute of Technology. Bulletin 221.

Noland, R.B., 2003. Traffic fatalities and injuries: the effect of changes in infrastructure and other trends. Accident Analysis & Prevention 35 (4), 599–611.

Noland, R.B., Oh, L., 2004. The effect of infrastructure and demographic change on traffic-related fatalities and crashes: a case study of Illinois country-level data. Accident Analysis & Prevention (36), 525–532.

Noland, R.B., Quddus, M.A., 2003. A spatially disaggregate analysis of road causalities in England. In: 82nd Annual Meeting of the Transportation Research Board, Washington, USA.

Ozbay, K., Noyan, N., 2006. Estimation of incident clearance times using Bayesian networks approach. Accident Analysis & Prevention 38 (3), 542–555.

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), 1–6.

Park, E.S., Park, J., Lomax, T.J., 2010. A fully Bayesian multivariate approach to before–after safety evaluation. Accident Analysis & Prevention 42 (4), 1118–1127.

Pearl, J., 1988. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Kaufmann, San Mateo, CA.

Pelikan, M., 2005. Hierarchical Bayesian Optimization Algorithm. Springer, Berlin/Heidelberg, pp. 31–48.

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., Nguyen, T., 1999. Empirical Bayes procedure for ranking sites for safety investigation by potential for safety improvement. Transportation Research Record: Journal of the Transportation Research Board 1665 (1), 7–12.

Qin, X., Ivan, J.N., Ravishanker, N., Liu, J., 2005. Hierarchical Bayesian estimation of safety performance functions for two-lane highways using Markov Chain Monte Carlo modeling. Journal of Transportation Engineering 131 (5), 345–351.

Savolainen, P.T., Mannering, F.L., Lord, D., Quddus, M.A., 2011. The statistical analysis of highway crash-injury severities: a review and assessment of methodological alternatives. Accident Analysis & Prevention 43 (5), 1666–1676.

Schlüter, P.J., Deely, J.J., Nicholson, A.J., 1997. Ranking and selecting motor vehicle accident sites by using a hierarchical Bayesian model. Journal of the Royal Statistical Society: Series D (The Statistician) 46 (3), 293–316.

Schubert, M., Hoj, N.P., Kohler, J., Faber, M.H., 2011. Development of a best practice methodology for risk assessment in road tunnels. ASTRA, Federal Road Office, Switzerland, Research Report 1351, p. 157.

Schubert, M.E., Kohler, J., Faber, M.H., 2007. Presentation: analysis of tunnel accidents by using Bayesian networks. In: Proceedings of the International Probabilistic Symposium, Ghent, Belgium.

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

Simoncic, M., 2004. A Bayesian network model of two-car accidents. In: Bauer, L. (Ed.), Proceedings of the 22nd International Conference on Mathematical Methods in Economics 2004. Masarykova Univ., Brno, pp. 282–287.

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

Tsamardinos, I., Brown, L., Aliferis, C., 2006. The max-min hill-climbing bayesian network structure learning algorithm. Machine Learning 65 (1), 31–78.

Tsionas, E.G., 2001. Bayesian multivariate poisson regression. Communications in Statistics – Theory and Methods 30 (2), 243–255.

Tunaru, R., 2002. Hierarchical Bayesian models for multiple count data. Austrian Journal of Statistics 31 (2, 3), 221–229.

WHO, 2004. World Report on Road Traffic Injury Prevention. World Health Organization, Geneva.

Ying, C.M., 2004. Bayesian spatial and ecological models for small-area accident and injury analysis. Accident Analysis & Prevention 36 (6), 1019–1028.