Accident Analysis and Prevention 51 (2013) 274–291

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.

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