Infrastructure Asset Management

A Bayesian network model to predict accidents on Swiss highways
Deublein, Schubert, Adey and García de Soto

Infrastructure Asset Management
http://dx.doi.org/10.1680/jinam.15.00008
Paper 1500008
Received 28/05/2015; accepted 16/09/2015
Keywords: infrastructure planning/management/risk & probability analysis

ice Institution of Civil Engineers
publishing

A Bayesian network model to predict accidents on Swiss highways

Markus Deublein PhD
Project Leader, Ernst Basler + Partner AG, Zollikon, Switzerland
Matthias Schubert PhD
CEO, Matrisk GmbH, Affoltern am Albis, Switzerland
Bryan T. Adey PhD
Professor and Head of the Infrastructure Management Group, Institute of Construction and Infrastructure Management, ETH Zürich, Switzerland
Borja García de Soto PhD, PE
Research Associate, Institute of Construction and Infrastructure Management, ETH Zürich, Switzerland

Although in 2014, Switzerland had an average of less than two fatalities per billion vehicle-kilometres, making its roads among the safest in Europe, still more than 17 000 traffic accidents occurred on Swiss communal roads, cantonal roads and national highways. On the highway network of approximately 1800 km alone, there were almost 1700 accidents involving personal injuries. In order to further reduce this number of accidents, it is important that accident risks are assessed as accurately as possible. A state-of-the-art methodology is used to develop a Bayesian probabilistic network model to estimate the number of accidents involving personal injury on the Swiss highway network. The developed model predicts the number of accidents on a given highway segment and can be used to identify segments with a high expected number of accidents. During validation, the number of accidents was correctly predicted on 86% of the segments with a tolerance of 25%. The model can also be used to conduct parametric studies, which help to ensure that the risk reduction interventions are effective and efficient. Road traffic and road infrastructure engineers and managers can use the model during the decision-making processes in the planning, construction and maintenance of road networks.

Notation

AADT	annual average daily traffic
$\hat{\mathbf{B}}$	matrix of regression coefficients
l	length of segment (km)
n	sample size
p	number of risk indicators considered
t	tolerated uncertainty for verification of compliance (0·25)
v	exposure, represented by the number of vehicles travelling on a road segment with a specific length
v_i	exposure on segment i
$\tilde{v}_i$	observed exposure on segment i
X	design matrix of $j = 1,..., u$ different indicator variables
X	$= \begin{pmatrix} x_{11} & \dots & x_{1u} \\ \vdots & \ddots & \vdots \\ x_{n1} & \dots & x_{nu} \end{pmatrix}$
Y_{ik}	vector with numbers of accidents of type k on segment i
$\tilde{y}_{ik}$	observed number of accidents of type k on segment i
y	number of accidents (per year)
$\hat{y}$	predicted number of accidents
$\dot{y}$	observed number of accidents
$\alpha_{ik}^{'}$	prior parameter for the expected number of accidents per accident type k per segment i
$\alpha_{ik}^{''}$	posterior parameter for the expected number of accidents per accident type k per segment i
$\beta_i^{'}$	prior scale parameter representing the weighted exposure for each segment i
$\beta_i^{''}$	posterior scale parameter representing the weighted exposure for each segment i
$\boldsymbol{\varepsilon}_{ik}$	vector of error terms per accident type k and per segment i
Δt	observation period (year)
λ	mean accident rate
λ_{ik}	mean accident rate for type k accidents on segment i
$\lambda_{ik}^{'}$	prior accident rate
$\lambda_{ik}^{''}$	posterior accident rate
$\dot{\lambda}_k$	background rates for each accident type considered
$\boldsymbol{\Lambda}$	matrix of $k = 1,..., z$ different target variables, in this case the posterior accident rates $\lambda_{ik}^{''}$
$\boldsymbol{\Lambda}$	$= \begin{pmatrix} \lambda_{11}^{''} & \dots & \lambda_{1z}^{''} \\ \vdots & \ddots & \vdots \\ \lambda_{n1}^{''} & \dots & \lambda_{nz}^{''} \end{pmatrix}$
$\boldsymbol{\Xi}$	matrix of the error terms
$\boldsymbol{\Xi}$	$= \begin{pmatrix} \varepsilon_{11} & \dots & \varepsilon_{1z} \\ \vdots & \ddots & \vdots \\ \varepsilon_{n1} & \dots & \varepsilon_{nz} \end{pmatrix}$
ω	binary variable with values of 1 (match) and 0 (does not match)

Introduction

Although the number of accidents in Switzerland is lower than in most European countries, road safety is a major societal concern. In 2013, 26 025 people died on the roads of the European Union (EU28) due to road accidents (ETSC, 2014). In 2013, 24 out of the 32 countries monitored by the Performance Index Report Program showed a decrease in the number of road deaths compared to the previous year. Slovakia (−24%) and Switzerland (−21%) registered the highest drops in 2013 against 2012 (ETSC, 2014).

Despite these comparatively positive results, according to the evaluation of the road traffic accident records of the Federal Roads Office (FEDRO) in 2014, on the entire Swiss road networks, there were over 17 800 traffic accidents, involving personal injuries (ASTRA, 2015). During 2014, the number of fatal accidents on Swiss roads has decreased from 269 to 243 persons when compared to 2013, a reduction of 10%. A more modest reduction in accidents with seriously injured persons also occurred in 2014, when the number decreased from 4129 in 2013 to 4043 in 2014, a reduction of 2% (ASTRA, 2015). In order to make further reductions in the number of accidents, it is necessary to assess traffic risks as accurately as possible, so that they can be appropriately considered during decision-making processes in the planning, construction and maintenance of road sections and road networks.

Bayesian inference and updating algorithms have gradually become more relevant in the field of accident risk assessment, and modern accident risk analysis is often based on the Bayesian interpretation of probability (e.g., Persaud *et al*., 2010). Empirical Bayesian (EB) methods were investigated first and are still frequently applied (e.g., AASHTO, 2010; Persaud *et al*., 1999). The EB method is considered as the most common state-of-the-art method for the development of accident prediction models, and considerable research has been conducted using it, including that by Carlin and Louis (1997), Cheng and Washington (2005), Elvik (2008), Hauer (1986, 1992), Hauer *et al*. (1991, 2002), Persaud and Dzbik (1993), Persaud *et al*. (1999) and Tunaru (2002). The EB method has already been compared to methods based on the full Bayes (FB) approaches. Persaud *et al*. (2010), for instance, came to the conclusion that the differences between the two methods were too small to be of any statistical significance or practical relevance. However, the magnitude of uncertainties connected to the predictions is different between the approaches and indicate that estimates based on methods which use the FB approach are more precise. Bayesian probabilistic networks (BNs) can be used as a helpful tool to apply Bayesian inference and updating algorithms in an intuitive, understandable and illustrative manner. An overview of the current developments in accident research can be found in Mannering and Bhat (2014) and Lao *et al*. (2014). In this paper, the methodology described by Schubert *et al*. (2007) and advanced by Deublein *et al*. (2013) is used to develop a BN model to predict the number of injury accidents that are likely to occur on the Swiss highway network (class 1 and 2).

Bayesian inference and updating algorithms are used to establish a full BN 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 Ang and Tang (2007), Benjamin and Cornell (1970), Congdon (2006) and Pearl (1988). For a detailed description of BNs, reference is given to Cowell *et al*. (1999), Jensen and Nielsen (2007) and Kjaerulff and Madsen (2008).

BNs 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 BN in terms of measured observations of the risk-indicating variables. The inference calculation of the BN uses the structure and the conditional probability tables (CPTs) for propagating the observed information of the evidence 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 Der Kiureghian and Ditlevsen (2009) and Faber and Maes (2005) as it allows for capturing both aleatory and epistemic uncertainties in accident modelling. Information provided by the Federal Road Office (FEDRO) was used to develop the network, to establish the CPTs of the BN, to learn the BN and to test the model.

Methodology

The methodology can be subdivided into five main steps (Deublein *et al*., 2013). These steps are shown in Figure 1 and briefly explained in the subsequent sections.

Since the methodology is exclusively based on data, a large and reliable data set with response variables (e.g., injury accidents, number of fatalities) and risk variables (e.g., road design parameters, traffic volume) is required. During the model development, the data are used for two complementary, but not overlapping, modelling steps. First, the information from the data is used to establish a multivariate Poisson-lognormal regression model which forms the basis for the prior BN. Predictions of the prior BN 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 (annual average daily traffic (AADT)) in the data set is 80 000 vehicles/d but the model covers a range up to 100 000 vehicles/d). Second, the information of the prior BN is updated by means of parameter-learning algorithms using the observations of response variables and risk-indicating variables as contained in the available data set. 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

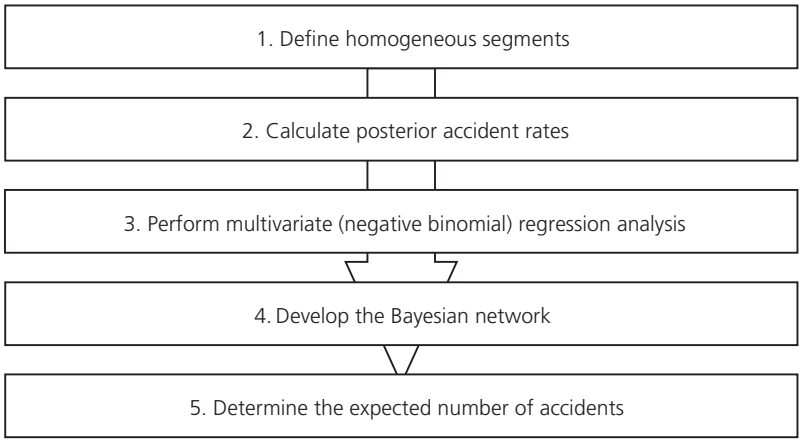

Figure 1. Main steps of methodology

the prior model probabilities for which observations of the response and risk variables are available are replaced. 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 from the applied data is implemented into the model development process in a complementary manner solely.

Define homogeneous segments

In the first step, the highway network is subdivided into homogeneous segments (HSs), taking into account the values of the variables considered to be significant in the prediction of the number of accidents, e.g. AADT, percentage of heavy good vehicles (HGV), curvature (BEND), slope/gradient (SLP). A homogeneous segment is defined as a road segment in which the values of the considered variables remain constant. For example, if the only key variable was curvature, then, every time the value of the curvature changed, an HS would end and another would start, independent of their length, until the entire road section or network is evaluated. From here on, the word 'homogeneous' is dropped as all segments are homogenous and only the word 'segments' is used instead.

Calculate posterior accident rates

In the second step, the accident rates for the different types of accidents are calculated for each segment. A two-level hierarchical approach is used.

Level one

On the first hierarchical level, the prior distributions of the accident rates are assessed. This is done by (*a*) estimating the background accident rates based on the accident counts, traffic volume and length of the entire network, and then by (*b*) updating these rates using the observed number of accidents on each segment. The background accident rate, which is assessed based on the accident observations divided by the average exposure of the entire road network, can be understood to be the 'best guess' for the accident rate of a specific road segment as long as nothing is known about the risk indicators of that segment.

It is assumed that the number of accidents on a segment is best represented by means of a Poisson distributed random variable (Equation 1), as done by many other researchers, such as Park and Lord (2007) and Song *et al*. (2006)

$$1. \quad Y_{ik}|\lambda_{ik} \sim \text{Poisson}(\nu_i \cdot \lambda_{ik})$$

Accident data observations on road sections are often characterised by very large deviations, which are referred to as over-dispersion (Berk and MacDonald, 2008; Cox, 1983; Dean and Lawless, 1989; Gschloessl and Czado, 2006; Hauer, 2001; Karlis and Meligkotsidou, 2005). This means that the sample variance frequently has a value greater than the sample average. For this reason, the previously mentioned assumption of a Poisson distribution is not suitable for modelling the expected accident rate because it has only one parameter and does not allow modelling the variance independently of the mean. As an alternative approach to the Poisson distribution, it can be assumed that the number of accidents can be best represented by a negative binomial distribution. This is a mixture of (1) a Poisson distribution, which reflects the probability of a certain number of accident events on a given segment over a defined period, and (2) the natural conjugated Gamma distribution with the probability distribution of the Poisson parameter λ (Gelman *et al*., 2004).

For this reason, the mean accident rate per segment i and accident type k is represented as Equation 2

$$2. \quad \lambda_{ik} \sim \text{Gamma}(\alpha_{ik}, \beta_i)$$

where α is the shape parameter and β is the inverse scale parameter of the Gamma distribution.

The combination of the Poisson and Gamma distribution results in a negative binomial distribution with parameters α, β and λ (Equation 3). Notice that Equation 3 does not contain the exact probability distribution functions, but it shows how the Poisson

and Gamma distribution functions are connected as elements of the negative binomial distribution. Definitions of the Poisson and Gamma probability distribution functions can be found in Faber (2012). The derivation of Equation 3 can be found in Gelman *et al*. (2004).

$$3.\quad Y|\alpha,\beta \sim NB(y|\alpha,\beta) \sim \frac{\text{Poisson}(y|\lambda)\cdot\text{Gamma}(\lambda|\alpha,\beta)}{\text{Gamma}(\lambda|\alpha+y,1+\beta)}$$

where the corresponding expected value E [−] and variance VAR [−] are computed in accordance with Equation 4.

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

Level two

On the second hierarchical level, the probability distributions used to represent the prior and posterior Gamma parameters (see level one) are determined. The probability distribution of the prior Gamma parameters corresponds to Equation 5.

$$5.\quad \lambda_{ik}^{'} \sim \text{Gamma}(\alpha_{ik}^{'},\beta_{i}^{'})$$

The expected value of the Gamma-distributed prior accident rate is determined using Equation 6

$$6.\quad E[\lambda_{ik}^{'}]=\alpha_{ik}^{'}\cdot\frac{1}{\beta_{i}^{'}}$$

where $\lambda_{ik}^{'}$ and $\alpha_{ik}^{'}$ are given by Equations 5 and 7, respectively. The Poisson parameter λ itself is modelled using the hyper-parameters of a Gamma distribution as described in Gelman *et al*. (2004).

$$7.\quad \alpha_{ik}^{'}=\dot{\lambda}_k\cdot\beta_{i}^{'}$$

$\beta_{i}^{'}$ can be seen as a way to counteract the influence of very long segments by reducing their weight in relation to their length (Equation 8).

$$8.\quad \beta_{i}^{'}=v_i\cdot\frac{1}{l^2}$$

The background rates are to be either taken from the literature, expert opinion or calculated empirically from historical data. In the latter case, the observation period should be sufficiently large to obtain a good approximation of the background rate, for example, 5 years. It must also be considered that the data are non-stationary, that is, they do not change significantly over time, so that the accident rates are changing over time, for example, due to changes in demography, legislative measures or technical/automotive developments. If the analysis of historical data is to be done for a less than sufficiently large observation period, it can, however, be assumed to be representative if it is based on mean values of the indicator variables, and the accident rates for each accident type and known trends are appropriately taken into consideration in the analysis.

The estimated values of the scale parameter are the same for all segments of one type for all accident rates. The accident rate varies from segment to segment due to only the number of vehicles travelling on the segment and the length of the segment, that is, the weighted exposure.

The posterior accident rates are calculated by updating the prior accident rates per segment based on the background rates and using the observed number of accidents on each segment (Equation 9). The parameters of the Gamma distribution for the posterior accident rates are calculated as shown in Equations 10 and 11 (Gelman *et al*., 2004).

$$9.\quad \lambda_{ik}^{''}|\tilde{y}_{ik},\tilde{v}_i \sim \text{Gamma}(\alpha_{ik}^{''},\beta_{i}^{''})$$

$$10.\quad \alpha_{ik}^{''}=\alpha_{ik}^{'}+\tilde{y}_{ik}$$

$$11.\quad \beta_{i}^{''}=\beta_{i}^{'}+\tilde{v}_i$$

These posterior, or updated, accident rates are then used in the multivariate regression analysis in the next step, as target variables. This updating process essentially alleviates, or drastically reduces, the problematic issues in the evaluation of rare events, such as excess dispersion and regression to the mean, since the effect of individual outlier values on the regression analysis (as estimated by the correlations), is reduced, and none of the accident rates for individual segments remains at zero, albeit they remain very small. The latter conceptually means that the use of updating ensures that not observing any accidents on a segment does not mean that no accidents can ever occur there.

Perform multivariate (negative binomial) regression analysis

In this step, the differences between the expected accident rates, $\hat{\mathbf{\Lambda}}$, and the observed accident rates, $\mathbf{\Lambda}$, that is, the residuals, $\hat{\mathbf{R}}$ (Equation 12), and the covariance of the error terms (Equation 14) are determined using multivariate regression analysis. Multivariate regression analysis is a special multifactorial form of regression analysis, in which not only different risk indicators enter into the regression equation at the same time, but also several target variables are estimated at the same time. Thus, potential dependencies on both sides of the regression equation can be taken into account, which is not possible with other types of regression analysis,

that is, linear or multivariable regression analysis. The general form of the multivariate regression equation, in matrix form, is given in Equation 15 (Gelman *et al*., 2004).

$$12. \quad \hat{\mathbf{R}} = \mathbf{\Lambda} - \mathbf{X}\hat{\mathbf{B}} = \mathbf{\Lambda} - \hat{\mathbf{\Lambda}}$$

The matrix with the regression coefficients $\hat{\mathbf{B}}$ corresponds to the case of normally distributed target variables (or log-transformed normal distributed target variables) estimated using the method of least squares. These regression coefficients are calculated using the latter method as follows (Gelman *et al*., 2004) (Equation 13).

$$13. \quad \hat{\mathbf{B}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{\Lambda}$$

The covariance matrix of the error terms can be estimated on the basis of the residuals (Equation 14).

$$14. \quad \hat{\mathbf{\Sigma}} = \frac{1}{n-p-1}\hat{\mathbf{R}}^T\hat{\mathbf{R}}$$

$$15. \quad \ln(E[\mathbf{\Lambda}|\mathbf{X}]) = \hat{\mathbf{B}}\mathbf{X} + \mathbf{\Xi} \quad \triangleq \quad E[\mathbf{\Lambda}|\mathbf{X}] = \exp(\hat{\mathbf{B}}\mathbf{X} + \mathbf{\Xi})$$

The multivariate regression model can be easily modified to take into account indicator variables and different functional relationships.

Development of the Bayesian network

In this step, a BN model is developed and learned so that it can be used to determine the predictive probability distributions for the expected accident rates of each investigated accident (injury) type, for example, accidents with a maximum of light, serious and fatal personal injuries. In the BN model, all indicators and target variables are represented as nodes and the relationships between the variables are represented as directed edges The possible values of each variable are grouped in intervals (hereinafter referred to as states), and the occurrence probabilities of the states are represented with discretised distributions. For given probability distributions of the indicator variables, the resulting probability distributions of the accident rates can be calculated using a CPT. More information about BNs can be found in Jensen and Nielsen (2007), Pearl (1988) and Pearl (1997), among others.

The values used in the prior BN model are determined exclusively using the multivariate regression analysis, and it is considered to be the basic model. This model is then learned, or updated, using the expectation-maximisation algorithm (EM algorithm) (Box and Tiao, 1992; Heckman, 1995) with the observed number of accidents per segment within a specified time period. It is used to update the prior conditional probabilities and to adjust the causal relationships to reflect the new information. The learned BN is referred to as the posterior BN.

The observed accidents are used to establish contingency tables for the EM-learning algorithm. An example of a contingency table for light injury accidents (LINJ) is given in Table 1 (refer to Table 2 for more information on the different variables used). By preparing the data in that way, it is relatively easy to see potential relationships between risk-indicating variables, that is, the covariance of different indicator and target variables, and to update the relationships in the conditional probabilities of the target variables.

Changes in the probability distributions compared to the prior model are done by learning only in the areas of CPTs in which there are observations. For all other areas, the model results are still based on the models of the regression analysis.

The posterior BN model is the model to be used to calculate the conditional probabilities of the number of accidents per road segment.

Year	Segment	Type	AADT	HGV	RAD	GRAD+	—	LINJ
2010	1	Tunnel	30 000	4	1000	0		0·0560
	2	Open road	30 000	4	4000	6		0·0573
	3	Open road	30 000	4	2000	3		0·0583
	—		—	—	—	—		—
2011	1	Tunnel	30 000	4	1000	0		0·0560
	2	Open road	30 000	4	4000	6		0·0573
	3	Open road	30 000	4	2000	3		0·0583
	—		—	—	—	—		—
2012	1	Tunnel	30 000	4	1000	0		0·0560
	2	Open road	30 000	4	4000	6		0·0573
	3	Open road	30 000	4	2000	3		0·0583
	—		—	—	—	—		—

Table 1. Example of a contingency table

Variable	Description	Units	Classification
AADT	Annual average daily traffic based on the 2010 traffic model (ARE, 2010)	Vehicles/day	10 000, 20 000,…, 100 000
HGV	Percentage of heavy traffic (heavy good vehicles) with respect to the AADT based on the 2010 traffic model (ARE, 2010)	%	5, 10,…, 25
RAD	Middle curve radius based on author's calculations	m	1000, 4000,…, 8000
GRAD+	Amount of the average slope (up) based on the authors' own calculations (GRAD1 in regression equation)	%	0, 1, 3, 6
GRAD–	Amount of the average slope (down) based on the authors' own calculations (GRAD2 in regression equation)	%	0, 1, 3, 6
SPEED	Signalled maximum speed	km/h	60, 80, 100, 120
LANES	Number of lanes in each direction	—	1, 2, 3, 4
EVEN (I2)	Road surface texture: evenness in the longitudinal direction, according to VSS standard SN 640 925b (VSS, 2003). In this case, a mark between 0 and 2 is good, between 2 and 4 is sufficient, and 5 is poor surface condition.	—	2, 4, 5
ROUGH (I4)	Road surface texture: grip, according to VSS standard SN 640 925b (VSS, 2003). In this case, a mark between 0 and 2 is good, between 2 and 4 is sufficient, and 5 is poor surface condition.	—	2, 4, 5
TYPE	Distinction between 'open road' including bridges, and 'tunnel' including galleries.	—	1, 2

Table 2. Risk-indicator variables

Determine the expected number of accidents

In this step, the expected number of accidents is estimated for each segment. This is done by first entering the values of the indicator variables as input into the posterior BN model (i.e., the evidence is set on the observed states of the input indicator variables) for each segment. The inference algorithms of the BN model are then used to propagate the probabilities of the input nodes through the network, which results in updated predictive probability distributions of the target variables that are conditional on the entered evidence. The mean values of the predictive probability distributions are then multiplied by the exposure of the segment, to obtain the expected number of accidents per year, instead of the accident rates.

Data

The prediction model to estimate the accidents involving personal injury was developed solely using available data from FEDRO for the Swiss highway network from 3 years (2010, 2011 and 2012). The road network was divided in two groups of road types, open roads (including bridges) and tunnels (including galleries). The data were extracted from four modules in the FEDRO information management system, MISTRA. The data were processed using ArcGIS software, developed by the Environmental Systems Research Institute (ESRI, 2011). This involved creating a network of equidistant, georeferenced points of reference, and associating all data to these points of reference, that is, all infrastructure and transportation-related information to be used to divide the network into segments. All accident data were also assigned to these points of reference by means of the nearest neighbour method. The resolution of this network corresponded to the constant spacing of the equidistant points of reference. The original 82 road sections were divided using 147 323 reference points, that is, before the network was divided into segments (see 'Case study' section).

Variables

The target variables are the accident rates of (1) accidents with no more than light injuries (LINJ), (2) accidents with no more than heavy personal injuries, (SINJ) and (3) accidents with fatalities (FAT). Accidents involving only material damage were excluded because the data could not be regarded as complete since collecting these data was not mandatory for the police departments in Switzerland. The accident rates are expressed as the ratio between the number of observed accidents in a given period and the exposure (Equation 16), where exposure is defined as the product of segment length (km), the observation period (years) and the annual average daily traffic (AADT) (Equation 17).

$$\lambda = \frac{y}{v} \tag{16}$$

$$v = l \cdot \Delta t \cdot \mathrm{AADT} \tag{17}$$

The indicator variables to be used are either infrastructure or traffic characteristics, that is, characteristics which can be controlled and/or changed by the infrastructure managers of FEDRO. The effects of road users (e.g., age, influence of

narcotics and drugs, traffic violations, etc.), influences of the vehicle (e.g. occupant protection, assistance systems, etc.) and the weather (rain, ice, fog, etc.) are not considered in the current investigations; however, such surrogate variables are important for the overall explanation and estimation of the accident occurrences on road networks. Consideration of such variables is planned for future research projects. The indicator variables are summarised in Table 2.

Case study

Define homogeneous segments

The 147 323 equidistant reference points (see 'Data' section) were grouped, on the basis of the indicator values, into 13 298 segments (see section 0), each containing a constant value for the indicator variables. Based on the combination of the values of the indicators, there are 691 200 possible types of segments. The shortest segment has a length of 20 m (corresponding to the minimum possible resolution) and the longest segment has a length of 5300. The average segment length is 221 m and 75% of all segments are longer than 60 m.

The relative frequencies of the categorised values of the indicator variables (Table 2) and target variables for all 13 298 segments are shown in Figures 2 and 3, respectively.

Note that the intervals of the bins in Figure 3 have different sizes. For light and severe injuries and for rates smaller than 0·01, the interval size is 0·001. For rates between 0·01 and 0·02, the interval size is 0·01. The reason is that the most relevant region for accident prediction is between 0 and 0·01, and this region should thus be modelled using a higher level of detail. The same applies for the histogram showing the fatal accident rates. Here, an interval size of 0·0001 was used from 0·0 to 0·001 and an interval size of 0·001 between 0·01 and 0·02.

Using a higher resolution in the histogram for the region with higher accident rates would lead to a much higher computational effort but would not yield improvement in the accuracy of the result. Of course, this problem could in general be formulated as an objective function and the bin size could be optimised for the problem at hand. However, the less formal approach used here was shown to be sufficient in this project.

Calculate posterior accident rates

In order to develop the risk model and describe the causal relationships between the indicator and target variables, the number of observed accidents was converted into an annual average per million vehicle kilometres (mvk). It was assumed that the values of the AADT and HGV remained constant over the observation period.

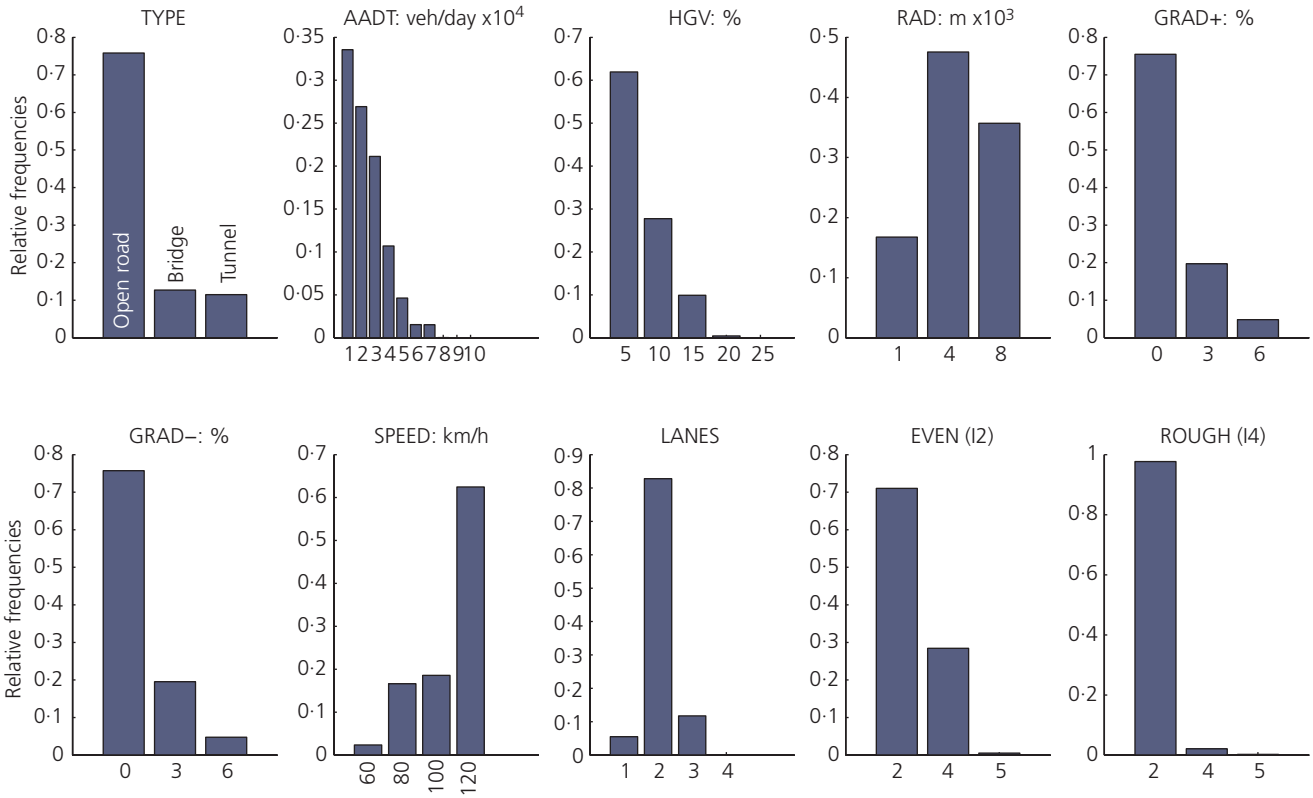

Figure 2. Histograms of the relative frequencies of the indicator variables of the 13 298 segments

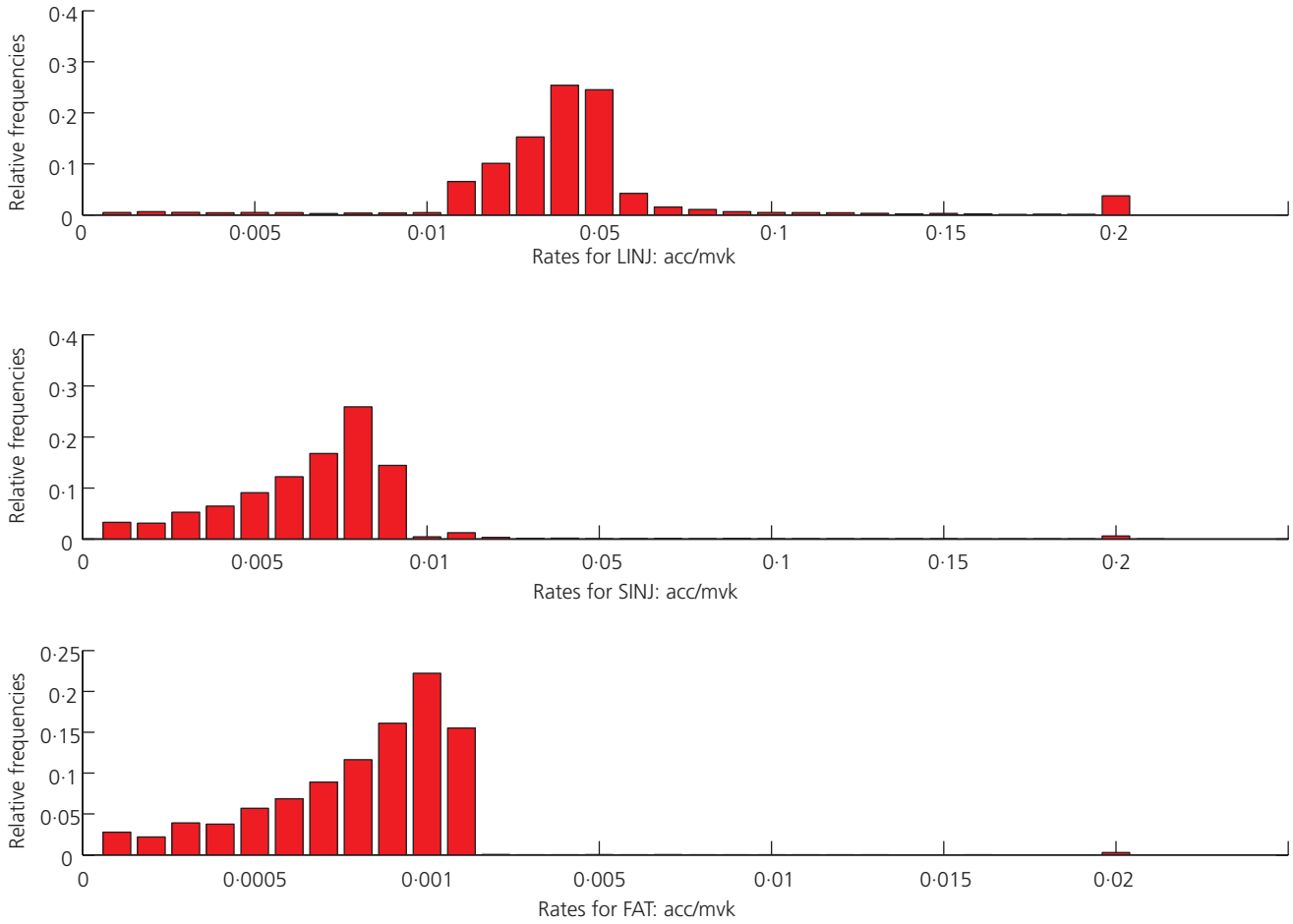

Figure 3. Histograms of the relative frequencies of the target variables of the 13 298 segments

The background rates used in this investigation have been computed empirically based on the historical data from the entire network. The background accident rates are shown in Table 3. These rates are used when there is no knowledge about the number of accidents in a given segment; hence they are used as the *best guess* for predicting the accident rate.

Determine residuals and covariance of error terms

To account for the different road types distinguished in the model (open roads (including bridges) and tunnels (including galleries)), an individual multivariate regression analysis was performed for each road type. The general regression equation is given in Equation 18. All indicator variables were used without transformations, for example, squaring each indicator variable, to strengthen their influence on the regression results, as these were found not to provide a significant benefit in the accuracy of the prediction model, and only added to the complexity of the interpretation of the results.

$$\begin{aligned}\ln(\mathrm{LIN}|\mathrm{TYP}) = {} & b_0 + b_1 \cdot \mathrm{AADT} + b_2 \cdot \mathrm{HGV} \\ & + b_3 \cdot \mathrm{RAD} + b_4 \cdot \mathrm{GRAD1} + b_5 \cdot \mathrm{GRAD2}... \\ & + b_6 \cdot \mathrm{SPEED} + b_7 \cdot \mathrm{LAN} + b_8 \cdot \mathrm{EVEN} \\ & + b_9 \cdot \mathrm{ROUGH} + \varepsilon \end{aligned} \quad (18)$$

$\dot{\lambda}_{U_{LV}}$	$\dot{\lambda}_{U_{SV}}$	$\dot{\lambda}_{U_G}$
[LINJ/mfk]	[SINJ/mfk]	[FAT/mfk]
0·058443	0·008717	0·001088

Table 3. Background accident rates

The expected values of the regression coefficients for all three target variables are shown in Tables 4 and 5. The statistical significance of the regression results was checked using a Student's t test at a significance level of $\alpha = 0{\cdot}05$ and with *n-u-1* degrees of freedom. In Tables 4 and 5, a high significance is denoted with '++', a low significance with '+' and no significance with '–'.

Regression coefficient	λ LINJ			λ SINJ			λ FAT		
	E[x]	STDEV[*x*]	Sig	*E*[*x*]	STDEV[*x*]	Sig	*E*[*x*]	STDEV[*x*]	Sig
β_0 (intercept)	−2·90	$3{\cdot}79 \times 10^{-2}$	++	−4·89	$5{\cdot}46 \times 10^{-2}$	++	−7·00	$5{\cdot}55 \times 10^{-2}$	++
β_1 (AADT)	$1{\cdot}19 \times 10^{-6}$	$2{\cdot}64 \times 10^{-7}$	++	$7{\cdot}91 \times 10^{-7}$	$3{\cdot}80 \times 10^{-7}$	++	$-1{\cdot}92 \times 10^{-6}$	$3{\cdot}86 \times 10^{-7}$	++
β_2 (HGV)	$-9{\cdot}74 \times 10^{-3}$	$1{\cdot}26 \times 10^{-3}$	++	$-1{\cdot}87 \times 10^{-2}$	$1{\cdot}81 \times 10^{-3}$	++	$-9{\cdot}50 \times 10^{-3}$	$1{\cdot}84 \times 10^{-3}$	++
β_3 (RAD)	$5{\cdot}48 \times 10^{-6}$	$1{\cdot}60 \times 10^{-6}$	++	$-3{\cdot}22 \times 10^{-5}$	$2{\cdot}30 \times 10^{-6}$	++	$-6{\cdot}82 \times 10^{-5}$	$2{\cdot}33 \times 10^{-6}$	++
β_4 (GRAD+)	$2{\cdot}43 \times 10^{-2}$	$4{\cdot}32 \times 10^{-3}$	+	$4{\cdot}58 \times 10^{-2}$	$6{\cdot}21 \times 10^{-3}$	++	$7{\cdot}49 \times 10^{-2}$	$6{\cdot}31 \times 10^{-3}$	++
β_5 (GRAD−)	$3{\cdot}43 \times 10^{-2}$	$4{\cdot}34 \times 10^{-3}$	−	$4{\cdot}15 \times 10^{-2}$	$6{\cdot}24 \times 10^{-3}$	+	$7{\cdot}74 \times 10^{-2}$	$6{\cdot}34 \times 10^{-3}$	+
β_6 (SPEED)	$-2{\cdot}13 \times 10^{-3}$	$2{\cdot}52 \times 10^{-4}$	++	$-2{\cdot}32 \times 10^{-3}$	$3{\cdot}63 \times 10^{-4}$	++	$-2{\cdot}95 \times 10^{-3}$	$3{\cdot}69 \times 10^{-4}$	++
β_7 (LANES)	$2{\cdot}51 \times 10^{-2}$	$1{\cdot}05 \times 10^{-2}$	+	$3{\cdot}36 \times 10^{-2}$	$1{\cdot}51 \times 10^{-2}$	++	$2{\cdot}15 \times 10^{-2}$	$1{\cdot}53 \times 10^{-2}$	++
β_8 (EVEN)	$6{\cdot}88 \times 10^{-2}$	$5{\cdot}29 \times 10^{-3}$	++	$1{\cdot}62 \times 10^{-1}$	$7{\cdot}61 \times 10^{-3}$	++	$2{\cdot}19 \times 10^{-1}$	$7{\cdot}73 \times 10^{-3}$	++
β_9 (ROUGH)	$7{\cdot}17 \times 10^{-2}$	$9{\cdot}21 \times 10^{-3}$	+	$-2{\cdot}03 \times 10^{-2}$	$1{\cdot}33 \times 10^{-2}$	−	$-1{\cdot}25 \times 10^{-2}$	$1{\cdot}35 \times 10^{-2}$	−

Table 4. Regression coefficients for open roads (including bridges)

Develop and learn the BN

Figure 4 shows the two components of the BN. A structural component (showing the nodes and edges and how these are connected in correspondence to the causalities of the problem) and the parameter component, which quantifies the strengths of the connections. Notice that Figure 4 only refers to the structural component and is graphically representing the considered nodes and edges. The structure of the developed Bayesian network is convergent because all the edges run directly from the individual nodes of the indicator variables to the nodes of the target variables. Each node represents one of the risk-indicator or target variables. The conditional probability distributions of the accident rates are calculated for the prior BN from the results of the multivariate regression analysis, entered into the CPTs, and transferred to the respective nodes of the network.

The comparison of the expected number of accidents and the observed numbers of light injury accidents are shown in Figure 5. It can be seen that the expected results obtained by multivariate non-linear regression analysis (Figure 5(a)) are improved when the learning algorithms are used (Figure 5(b)). This can be seen as the correlation coefficient between the expected and observed values improves from $r = 0{\cdot}7935$ (coefficient of determination $R^2 = 0{\cdot}6296$) (prior BN model) to $r = 0{\cdot}8334$ (coefficient of determination $R^2 = 0{\cdot}6952$) (posterior BN model).

In addition to the increase in the correlation coefficient, it can be seen that the regression lines are close to the optimum line ($y = 0 + 1 * x$), which implies that there is little bias in the predictions. All points do not lie on the optimum line because the number of observed accidents cannot be predicted with 100% accuracy using the selected indicator variables. It is likely that this variability can be reduced by using larger data sets (longer observation periods) and additional indicator variables. It should be noted that variables describing driver characteristics are presently not considered in the model.

Regression coefficients	λ LINJ			λ SINJ			λ FAT		
	E[*x*]	STDEV[*x*]	Sig	*E*[*x*]	STDEV[*x*]	Sig	*E*[*x*]	STDEV[*x*]	Sig
β_0 (intercept)	−2·78	$1{\cdot}14 \times 10^{-1}$	++	−4·50	$1{\cdot}47 \times 10^{-1}$	++	−6·70	$1{\cdot}27 \times 10^{-1}$	++
β_1 (AADT)	$6{\cdot}94 \times 10^{-6}$	$9{\cdot}43 \times 10^{-7}$	++	$2{\cdot}72 \times 10^{-6}$	$1{\cdot}22 \times 10^{-6}$	++	$3{\cdot}25 \times 10^{-6}$	$1{\cdot}05 \times 10^{-6}$	++
β_2 (HGV)	$-1{\cdot}52 \times 10^{-2}$	$3{\cdot}91 \times 10^{-3}$	+	$-2{\cdot}76 \times 10^{-2}$	$5{\cdot}04 \times 10^{-3}$	+	$-2{\cdot}31 \times 10^{-2}$	$4{\cdot}35 \times 10^{-3}$	++
β_3 (RAD)	$-5{\cdot}26 \times 10^{-5}$	$5{\cdot}25 \times 10^{-6}$	+	$-7{\cdot}93 \times 10^{-5}$	$6{\cdot}78 \times 10^{-6}$	+	$-8{\cdot}30 \times 10^{-5}$	$5{\cdot}84 \times 10^{-6}$	+
β_4 (GRAD+)	$-1{\cdot}36 \times 10^{-2}$	$1{\cdot}53 \times 10^{-2}$	−	$-2{\cdot}71 \times 10^{-2}$	$1{\cdot}97 \times 10^{-2}$	−	$1{\cdot}84 \times 10^{-2}$	$1{\cdot}70 \times 10^{-2}$	−
β_5 (GRAD−)	$-6{\cdot}50 \times 10^{-3}$	$1{\cdot}48 \times 10^{-2}$	−	$4{\cdot}65 \times 10^{-4}$	$1{\cdot}91 \times 10^{-2}$	−	$4{\cdot}20 \times 10^{-2}$	$1{\cdot}65 \times 10^{-2}$	−
β_6 (SPEED)	$-4{\cdot}49 \times 10^{-3}$	$7{\cdot}94 \times 10^{-4}$	−	$-5{\cdot}64 \times 10^{-3}$	$1{\cdot}03 \times 10^{-3}$	−	$-6{\cdot}74 \times 10^{-3}$	$8{\cdot}84 \times 10^{-4}$	+
β_7 (LANES)	$1{\cdot}16 \times 10^{-1}$	$3{\cdot}12 \times 10^{-2}$	−	$1{\cdot}03 \times 10^{-1}$	$4{\cdot}02 \times 10^{-2}$	−	$8{\cdot}42 \times 10^{-2}$	$3{\cdot}47 \times 10^{-2}$	−
β_8 (EVEN)	$5{\cdot}50 \times 10^{-2}$	$1{\cdot}81 \times 10^{-2}$	−	$1{\cdot}00 \times 10^{-1}$	$2{\cdot}34 \times 10^{-2}$	−	$1{\cdot}54 \times 10^{-1}$	$2{\cdot}01 \times 10^{-2}$	+
β_9 (ROUGH)	$2{\cdot}80 \times 10^{-2}$	$1{\cdot}42 \times 10^{-2}$	+	$9{\cdot}13 \times 10^{-2}$	$1{\cdot}83 \times 10^{-2}$	+	$7{\cdot}10 \times 10^{-2}$	$1{\cdot}58 \times 10^{-2}$	+

Table 5. Regression coefficients for tunnels (including galleries)

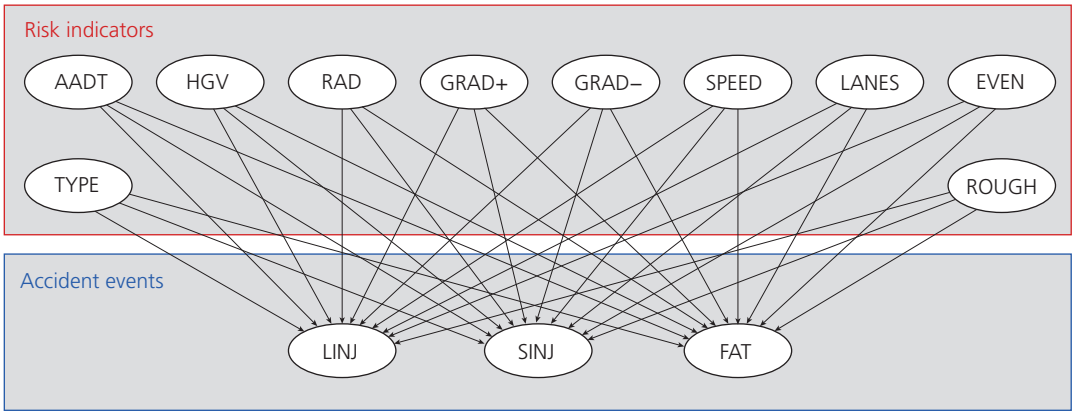

Figure 4. BN model structure

The comparison between the expected and the observed numbers of severe and fatal injury accidents are shown in Figure 6. The severe and fatal injury accidents were combined as there were too few observations for each to be treated alone. It can be seen that the expected results obtained by multivariate non-linear regression analysis (Figure 6(a)) are improved when the learning algorithms are used (Figure 6(b)), as shown by the improvement in the correlation coefficient between the expected and observed values (from $r = 0{\cdot}4568$ (coefficient of determination $R^2 = 0{\cdot}2087$) (prior BN model) to $r = 0{\cdot}6227$ (coefficient of determination $R^2 = 0{\cdot}3878$) (posterior BN model)). In these scatter plots, however, there is a division of the point clouds for accidents involving severe and fatal injuries, which was not the case for the light injury accidents. Both the relatively low correlation coefficient and the abnormalities in the point clouds are due to the small sample size of the observations of accidents with severe and fatal personal injury in the Swiss national roads.

The relationships are dominated by two groups: (1) the set of observations for which values of the accident events and the accident rate are actually observed and (2) the set of segments on which no accidents with severe or fatal personal injuries were observed. In the latter case, although the zero observations of the individual segments with the network-wide background rate are methodically updated, their effect on the prediction quality is low.

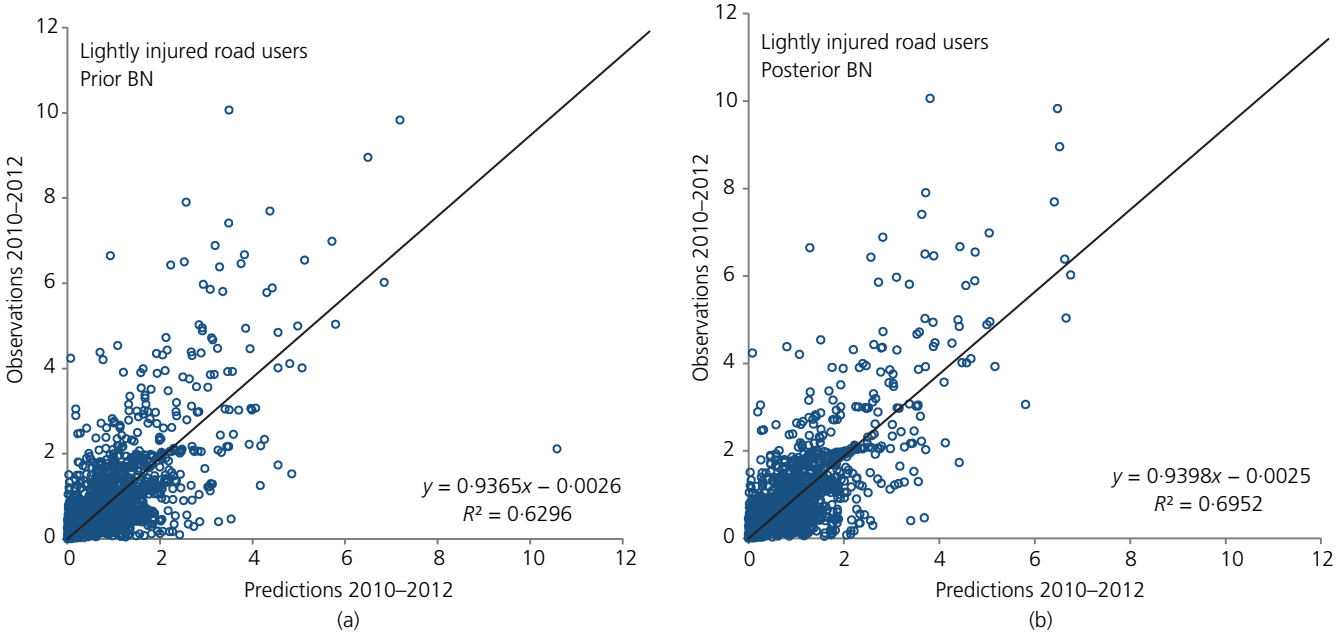

Figure 5. Comparison of predicted against observed number of accidents with light personal injury with prior BN model (a) and with posterior BN model (b)

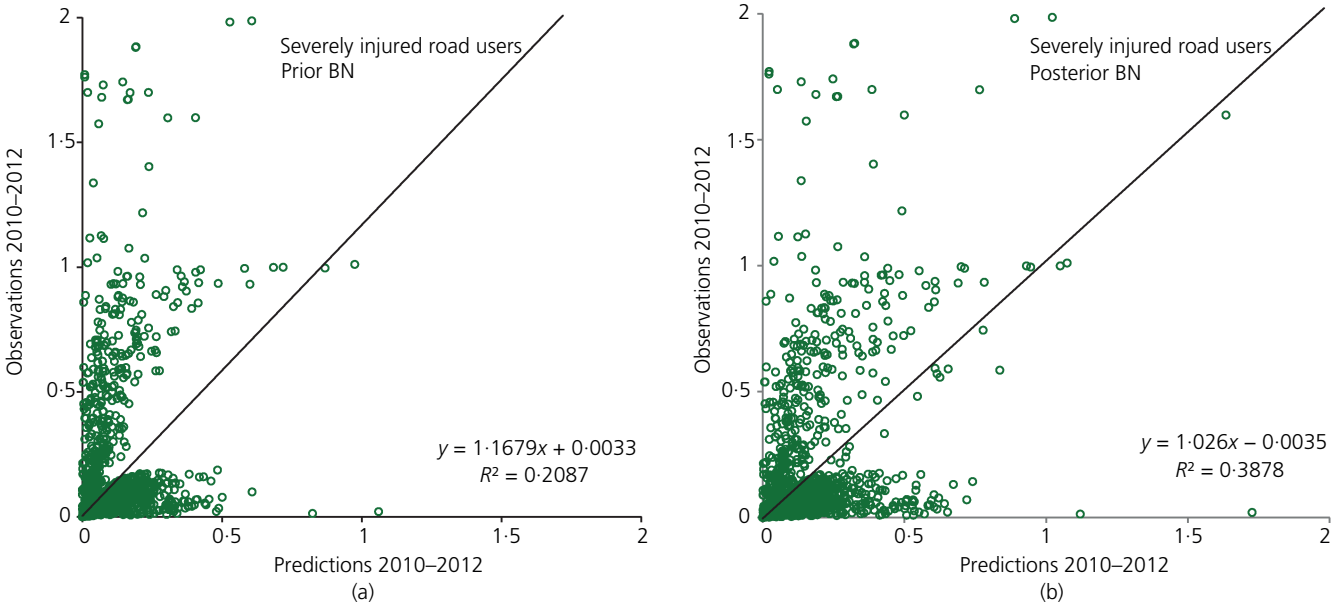

Figure 6. Comparison of predicted against observed number of severe and fatal injury accidents with prior BN model (a) and with posterior BN model (b)

Determine the expected number of accidents

Using the learned BN model, the expected number of accidents for each type of accident in 2009 was predicted. A backward prediction, rather than a forward prediction, was selected as available data were more complete in the years 2010, 2011 and 2012 than in 2009, and it was considered to be wiser to use the more complete years to learn the BN model and use 2009 data for testing the model.

Results

The results are presented using geo-referenced network graphics. The coordinates in Figures 7 and 8 correspond to the Swiss coordinate system CH1903 LV03. Scatter diagrams, like the ones shown in Figures 5 and 6, are not suitable because the observations from 2009 are integer values, while the predictions belong to the set of natural numbers that have decimal places. This was not the case in the earlier figures because averages over 3 years were displayed.

For the entire network, there were only 9380 segments (instead of 13 298) for which observations had been made. On these road segments, in 2009, there were a total of 979 light injury accidents, 140 severe injury accidents and 16 fatal injury accidents. Figure 7 shows the observed (Figure 7(a)) and predicted (Figure 7(b)) number of light injury accidents for a part of the network (i.e., HW1 between cities A and B).

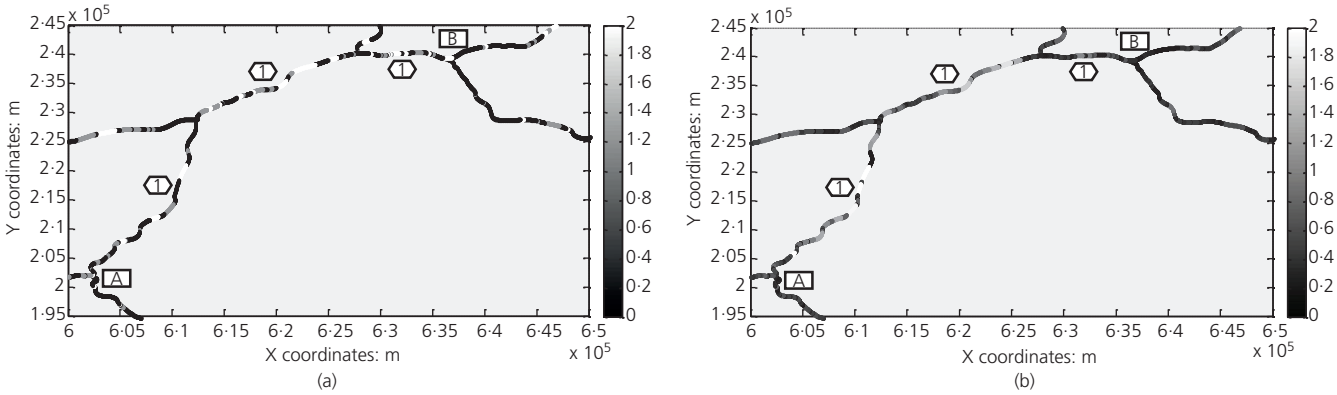

Figure 7. Enlarged view for section of the HW1 between City A and City B (shading scale indicates the number of accidents) with observed (a) and predicted (b) number of light injury accidents in 2009

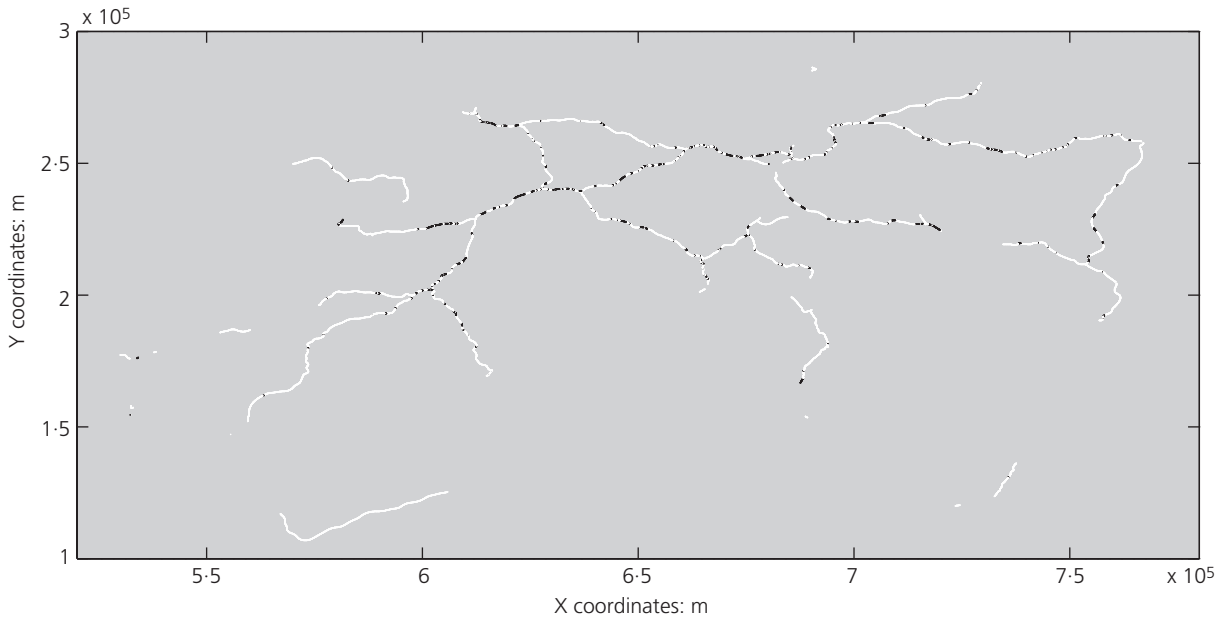

Figure 8. Comparison between predicted and observed number of light injury accidents in 2009 for all roads in the network with observations; white = 1 matching values (with tolerance of 25%), black = 0 no matching values

A comparison of observations and predictions in Figure 7 shows that, despite the limited amount of data for many segments, the shading (indicating the number of accidents) in Figure 7(b) matches quite well with Figure 7(a), that is, the predicted number of accidents is quite close to the observed number of accidents. It can, however, also be seen that there are differences.

From Figure 7, it can be seen that the model performs reasonably well. This can, however, be better seen by the fact that the number of accidents are correctly predicted on 86·53% of the road segments with a tolerance (t) of 25%, using Equation 19. This is illustrated in Figure 8, which shows the agreement between the predicted and the observed number of accidents for all road segments in the network with observations.

$$\begin{aligned} &\text{For } \dot{y} = 0 \\ &\omega = \begin{cases} 1, & \text{if } \hat{y} \in [\dot{y} - t, \dot{y} + t] \\ 0, & \text{if } \hat{y} \notin [\dot{y} - t, \dot{y} + t] \end{cases} \\ &\text{for } \dot{y} \geq 1 \\ &\omega = \begin{cases} 1, & \text{if } \hat{y} \in [\dot{y} - t \cdot \dot{y}, \dot{y} + t \cdot \dot{y}] \\ 0, & \text{if } \hat{y} \notin [\dot{y} - t \cdot \dot{y}, \dot{y} + t \cdot \dot{y}] \end{cases} \end{aligned} \tag{19}$$

It can, however, also be seen that not all accidents were correctly predicted by the model. This is because accident occurrence can only be partially explained using the selected indicator variables, as there are many more factors that affect whether or not an accident occurred, for example, the presence of road works, ice or fog on the highway, confusing sections with frequent congestion events.

Conclusion

In this paper, it was shown how an existing methodology can be used to develop models to predict the number of light, severe and fatal injury accidents that will occur on the road segments that comprise the Swiss highway network, as well as the conditional probability distributions of the accident rates and thus the uncertainties associated with the estimates. It was also shown that the developed model, which is based on easily observable indicator variables, can be used to identify road segments that are likely to have relatively high numbers of accidents; information that is useful in the planning of risk-reducing intervention.

Additionally, it was found that

1. when combined, simultaneous consideration of all indicator variables show that they can be used to predict the number of accidents reasonably well
2. using a multivariate regression analysis, the accident rates can be extrapolated to larger sample spaces
3. the updating of the BN model increases the accuracy of accident predictions by about 5–10% and
4. it is possible to obtain a high degree of agreement between the estimated and observed numbers of light injury accidents even with a relatively small data set. That said, using a purely data-based prediction, it is not currently possible to accurately predict fatal injury accidents, as the number of observations is too small. A combination of severe and fatal injury accidents is, therefore, recommended.

Acknowledgement

The work presented in this paper was conducted with financial support from the Swiss Federal Roads Office (FEDRO). FEDRO also provided the data for the case study. Special thanks are given to Dr Anja Simma and Mr Gerhard Schuwerk and their team at FEDRO.

REFERENCES

AASHTO (American Association of State Highway and Transportation Officials) (2010) *Highway Safety Manual*. AASHTO, Washington, DC, USA.

Ang AHS and Tang WH (2007) *Probability Concepts in Engineering: Emphasis on Applications in Civil & Environmental Engineering (Vol. 1)*. Wiley, New York, NY, USA.

ARE (Bundesamt für Raumentwicklung) (2010) *Nationales Personenverkehrsmodell des UVEK*. Bundesamt für Raumentwicklung, Switzerland (in German).

ASTRA (Bundesamt für Strassen) (2015) *Verkehrsunfallstatistik 2014 – Tabellen*, 26.03.2015 (in German). See http://www.news.admin.ch/NSBSubscriber/message/attachments/38807.pdf (accessed 02/04/2015).

Benjamin JR and Cornell C (1970) *Probability, Statistics, and Decision for Civil Engineers*. McGraw-Hill, New York, NY, USA.

Berk R and MacDonald JM (2008) Overdispersion and Poisson regression. *Journal of Quantitative Criminology* **24(3)**: 269–284.

Box GE and Tiao GC (1992) *Bayesian Inference in Statistical Analysis*. John Wiley & Sons, New York, NY, USA.

Carlin BP and Louis TA (1997) Bayes and empirical Bayes methods for data analysis. *Statistics and Computing* **7(2)**: 153–154.

Cheng W and Washington SP (2005) Experimental evaluation of hotspot identification methods. *Accident Analysis & Prevention* **37(5)**: 870–881.

Congdon P (2006) *Bayesian Statistical Modelling*. John Wiley & Sons, Chichester, UK.

Cowell RG, Dawid AP and Lauritzen SL (1999) *Probabilistic Networks and Expert Systems*. Springer, New York, NY, USA.

Cox DR (1983) Some remarks on overdispersion. *Biometrika* **70(1)**: 269–274.

Dean C and Lawless JF (1989) Tests for detecting overdispersion in Poisson regression models. *Journal of the American Statistical Association* **84(406)**: 467–472.

Der Kiureghian AD and Ditlevsen O (2009) Aleatory or epistemic? Does it matter? *Structural Safety* **31(2)**: 105–112.

Deublein M, Schubert M, Adey BT, Köhler J and Faber MH (2013) Prediction of road accidents: a Bayesian hierarchical approach. *Accident Analysis & Prevention* **51**: 274–291.

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

ESRI (Environmental Systems Research Institute) (2011) *ArcGIS Desktop: Release 10*. ESRI, Redlands, CA, USA.

ETSC (European Transport Safety Council) (2014) *Ranking EU Progress on Road Safety. 8th Road Safety Performance Index (PIN) Report. June 2014*. See http://etsc.eu/wp-content/uploads/ETSC-8th-PIN-Report_Final.pdf (accessed 20/09/2015).

Faber MH (2012) *Statistics and Probability Theory: In Pursuit of Engineering Decision Support (Vol. 18)*. Springer Science & Business Media, New York, NY, USA.

Faber MH and Maes MA (2005) Epistemic uncertainties and system choice in decision making. *Proceedings ICOSSAR2005, 9th International Conference on Structural Safety and Reliability, June 19–23 2005, Rome, Italy*, pp. 3519–3526.

Gelman A, Carlin JB, Stern HS and Rubin DB (2004) *Bayesian Data Analysis*. Chapman & Hall/CRC, Boca Raton, FL, USA.

Gschloessl S and Czado C (2006) *Modelling Count Data with Overdispersion and Spatial Effects*. Center of Mathematical Sciences, TU Munich, Munich, Germany.

Hauer E (1986) On the estimation of the expected number of accidents. *Accident Analysis & Prevention* **18(1)**: 1–12.

Hauer E (1992) Empirical Bayes approach to the estimation of "unsafety": the multivariate regression method. *Accident Analysis & Prevention* **24(5)**: 457–477.

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

Hauer E, Persaud BN, Smiley A and Duncan D (1991) Estimating the accident potential of an Ontario driver. *Accident Analysis & Prevention* **23(2)**: 133–152.

Hauer E, Harwood D, Council F and Griffith M (2002) Estimating safety by the empirical Bayes method: a tutorial. *Transportation Research Record* **1784**: 126–131.

Heckman D (1995) *A Tutorial on Learning with Bayesian Networks*. Microsoft Research, Advanced Technology Division, Redmond, WA, USA.

Jensen FV and Nielsen TD (2007) *Bayesian Networks and Decision Graphs*. Springer, New York, NY, USA.

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

Kjaerulff UB and Madsen AL (2008) *Bayesian Networks and Influence Diagrams*. Springer Science+ Business Media, New York, NY, USA.

Lao Y, Zhang G, Wang Y and Milton J (2014) Generalized nonlinear models for rear-end crash risk analysis. *Accident Analysis & Prevention* **62**: 9–16.

Mannering FL and Bhat CR (2014) Analytic methods in accident research: methodological frontier and future directions. *Analytic Methods in Accident Research* **1**: 1–22.

Park ES and Lord D (2007) Multivariate Poisson-lognormal models for jointly modeling crash frequency by severity. *Transportation Research Record* **2019**: 1–6.

Pearl J (1988) *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Kauffman, San Mateo, CA, USA.

Pearl J (1997) Bayesian networks. In *MIT Encyclopedia of the Cognitive Sciences* (Wilson RA and Keil FC (eds)). MIT Press, Cambridge, MA, USA.

Persaud B and Dzbik L (1993) Accident prediction models for freeways. *Transportation Research Record* **1401**: 55–60.

Persaud B, Lyon C and 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**: 7–12.

Persaud B, Lan B, Lyon C and Bhim R (2010) Comparison of empirical Bayes and full Bayes approaches for before–after road safety evaluations. *Accident Analysis & Prevention* **42(1)**: 38–43.

Schubert M, Köhler J and Faber MH (2007) Analysis of tunnel accidents by using Bayesian Networks. *International Probabilistic Symposium 2006, Ghent, Belgium, 28–29 November 2007*. See http://archiv.ibk.ethz.ch/emeritus/fa/people/schuberm/Schubert_Tunnel_Accidents_web.pdf (accessed 09/10/2015).

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

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

VSS (Swiss Association of Road and Traffic Experts) (2003) *SN 640 925b: Pavement Maintenance Management – Condition Assessment and Index Valuing*. Swiss Association of Road and Traffic Experts (VSS), Technical Committee 7: Maintenance Management, Zurich, Switzerland (in German).