ROADWAY ACCIDENT RISK PREDICTION
BASED ON BAYESIAN PROBABILISTIC NETWORKS

A dissertation submitted to
ETH ZURICH
for the degree of
DOCTOR OF SCIENCES

presented by
MARKUS KLAUS DEUBLEIN
Dipl.-Ing. Silv. Univ., Technische Universität München
born 15.03.1979
citizen of
Germany

accepted on the recommendation of

Prof. Dr. Bryan Tyrone Adey, examiner
Prof. Dr. Michael Havbro Faber, co-examiner
Finn Harald Amundsen, co-examiner
Dr. Matthias Ernst Schubert, co-examiner

2013
Acknowledgements

First of all I like to thank my supervisor Prof. Bryan Adey. He gave me the freedom to search my way and guidance to find it. I am very thankful for his support, his valuable advices and for giving me the feeling of being on the right track. For his kind and motivating supervision I would like to express my deep gratitude to him.

Many thanks go to my co-supervisor Dr. Matthias Schubert. Whenever necessary, he generously gave me exceptional support, both professionally and personally. Some parts of this thesis have been completed by direct cooperation and I strongly hope that this cooperation will continue.

My colleagues at IBI and the Group of Infrastructure Management have created a friendly and very pleasant atmosphere, conductive to research work. Thanks to all of you! Among others, I want to thank Dr. Amin Jamali who was my office (and tennis) mate over the last two years. We had a great time and whenever I encountered technical or personal problems he gave me helpful advice and his time to share our thoughts.

I also want to thank my former colleagues in the Group of Risk and Safety for the nice time we had together during my first years at ETH. Especially Prof. Miachel Faber significantly influenced my interest and motivation towards research and Dr. Jochen Köhler helped me a great deal with my first inexperienced steps in doing research.

I warmly thank Prof. Michael Faber and Mr. Finn Harald Amundsen who are so kind to act as co-examiners of this work.

My deepest gratitude goes to my parents and my family for their unconditional support. I couldn’t have done this work without them. This thesis is dedicated to them.

Zurich, 14. Februar 2013

Markus Deublein
Abstract

Transportation networks are an important part of a nation’s entire infrastructure asset. They substantially contribute to the economic and social development of that nation but do also entail considerable accident risks for the users. On a world-wide scale, of all transportation infrastructures that people have to deal with every day, the public roads transportation system is the most dangerous and the resultant injuries are a major public health issue.

Only one third of the countries being investigated in the WHO World Report on Road Traffic Injury Prevention (Peden et al., 2004) have a national road safety strategy that is endorsed by their government with specific targets and intervention strategies to reduce the accident risk and that is funded appropriately for its implementation in practice. Although an affirmative trend has been observable in the last years towards lower death and injury rates incurred due to road accidents, traffic injuries remain an important cause of death, injury and disability (NHTSA, 2010, Peden et al., 2009). Road safety is the interest and responsibility of society as a whole and further serious efforts need to be made to reduce the number of road accident victims. There is an urgent need to recognize the human impacts of traffic injuries and deaths as well as their large economic costs to the society (ETSC, 2007). Appropriate steps have to be taken for increasing the sensitivity and attention given to road traffic injury prevention. Significant research is needed to generate more insights into the nature of road traffic problems and into actions which could lead to successful prevention. Models have to be provided that will enable decision makers to understand better how road infrastructure should be developed in order to improve road safety. Such models are relevant for road infrastructure risk-based decision making and can serve as instruments for road infrastructure safety management (European Parliament, 2008).

The aim of this thesis is to establish a methodology that allows developing accurate accident prediction models. The aim is furthermore to support on optimized allocation of available budgets, e.g. for the implementation of accident risk reducing interventions. If the methodology presented in this thesis is used appropriately, it will contribute to improve accident risk analysis and, based on this, to a reduction of the general societal burden of injuries and fatalities incurred by road accidents. Accident risk is herewith the core element. It is considered as the product of three major dimensions: the exposure, the occurrence probabilities of accident and injury events and the resultant consequences. This differentiation helps to deal with the complex issue of accident risk assessment, this differentiation also helps to structure the content of this thesis:

First, an indicator-based generic methodology for the assessment of accident and injury occurrence probabilities on road networks is developed. The proposed methodology is designed to be used by accident analysts and road engineers. The methodology is generic in the sense that it is formulated in terms of observable risk indicating variables. Thus, it can be adapted to other infrastructure systems, given appropriate data is accessible. The exposure of
road users is taken into account by assessing the occurrence rates (e.g. accidents per million vehicle kilometre) for the different investigated accident events instead of their absolute frequencies (e.g. accidents per year). Bayesian probabilistic inference calculations are used to account for the randomness of accident events and the corresponding uncertainties. A multivariate regression analysis is additionally used to provide empirical prior parameter estimates of the linear relationships between risk indicating and model response variables. These estimates help to establish the structure of a Bayesian Probabilistic Network. When observed accident data becomes available, the causal relationships in the Bayesian Probabilistic Network can be updated by means of parameter learning algorithms. Posterior joint probability distributions are determined for the occurrence rates of the accident and injury events.

Second, two Bayesian methods are compared, which can be used to develop accident prediction models. One of these methods is the Bayesian Probabilistic Network method (developed in this thesis) and the other one is the Empirical Bayes method. The latter is currently the state-of-the-art methodology for accident risk assessment and the evaluation of countermeasures. The differences between the two methods are discussed in terms of their predictive precisions and their applicability as supportive instruments for road infrastructure safety management.

Third, the consequences of accident and injury events in terms of economic costs are addressed. The proposed methodology for the development of accident prediction models is imbedded into a framework for road user impact assessment. Such a probabilistic modelling approach in combination with the use of Bayesian Probabilistic Networks is new in the context of road infrastructure decision making. Impact models for different stakeholder groups affected by road networks require various model assumptions. The different types of uncertainties connected to these model assumptions can be considered using the features of the Bayesian Probabilistic Networks. The predicted impacts are provided in terms of joint posterior distribution functions of the expected costs.

The scope for the model development and its application is set on a probabilistic investigation of road design and traffic related risk promoting factors and their influence on the occurrence probabilities of injury accidents and different levels of injuries. The developed models are applied to road segments of a rural motorway network and implemented into impact models for road users.
Zusammenfassung

Verkehrsnetze sind ein wichtiger Teil der gesamten Infrastruktur einer Nation. Sie tragen wesentlich zur wirtschaftlichen und sozialen Entwicklung des Landes bei, sie verursachen aber auch bedeutsame Risiken für die Nutzer. Von allen Verkehrssystemen, die die tägliche Mobilität der Gesellschaft gewährleisten, ist das öffentliche Strassensystem weltweit eines der gefährlichsten Verkehrssysteme. Verletzungen und Todesfälle durch Verkehrsunfälle auf Strassen sind ein beachtliches Gesundheitsproblem für die Gesellschaft.


Das Ziel dieser Arbeit ist die Ausarbeitung einer Methode, die es ermöglicht, genaue Vorhersagemodelle für das Unfallrisiko von Strassenutzern zu entwickeln. Das Unfallrisiko wird als das Produkt dreier Hauptdimensionen verstanden: die Exposition der Strassenutzer, die Wahrscheinlichkeit, dass Unfall- und Verletzungereignisse eintreten, sowie die daraus resultierenden Konsequenzen für die unterschiedlichen Interessensvertreter. Diese Differenzierung hilft, mit dem komplexen Sachverhalt von Unfallrisiken umzugehen, diese Differenzierung hilft auch, um den thematischen Inhalt der vorliegenden Arbeit zu strukturieren:


Modellansatz im allgemeinen und durch die Verwendung von Bayes’schen Wahrscheinlichkeitsnetzen in der vorliegenden Arbeit berücksichtigt werden. Auf dieser Basis werden die vorhergesagten Auswirkungen anhand von gemeinsamen Wahrscheinlichkeitsverteilungen der zu erwarteten Kosten abgebildet.

TABLE OF CONTENTS

1 INTRODUCTION 15
  1.1 Relevance 15
    1.1.1 Accident risk assessment 16
    1.1.2 Road infrastructure decision making 17
  1.2 Aim of the thesis 20
    1.2.1 Aim 1: Model development 20
    1.2.2 Aim 2: Model comparison 21
    1.2.3 Aim 3: Model implementation 21
  1.3 Scope and issues of the thesis 22
    1.3.1 Issue 1: Accident contributing factors 22
    1.3.2 Issue 2: Statistical methodologies 25
    1.3.3 Issue 3: Objects for application 27
    1.3.4 Issue 4: Model implementation 28
  1.4 Outline of the Thesis 29

2 PREDICTION OF ROAD ACCIDENTS: A BAYESIAN HIERARCHICAL APPROACH (PAPER I) 33
   Abstract 34
   2.1 Introduction 34
   2.2 Methodology 36
     2.2.1 Determination of Model Variables 37
     2.2.2 Construction of Homogeneous Segments 38
     2.2.3 Gamma-Updating of Model Response Variables 39
     2.2.4 Development of Regression Models 43
     2.2.5 Construction and Parameter Learning of BPNs 45
     2.2.6 Prediction of the Expected Number of Events 47
   2.3 Case Study 47
     2.3.1 General 47
     2.3.2 Determination of Model Variables 48
2.3.3 Construction of Homogeneous Segments 52
2.3.4 Gamma-Updating of Model Response Variables 52
2.3.5 Development of Regression Model 54
2.3.6 Construction and Parameter Learning of BPNs 63
2.3.7 Prediction of Expected Number of Events 66
2.4 Discussion 69
2.5 Summary and Conclusions 70
Acknowledgements 71

3 PREDICTION OF ROAD ACCIDENTS: COMPARISON OF TWO BAYESIAN METHODS (PAPER II) 73

ABSTRACT 74
3.1 Introduction 74
3.2 Bayesian inference 77
  3.2.1 General 77
  3.2.2 Difference between Empirical and Full Bayes 78
  3.2.3 Use of the Empirical Bayes approach to develop prediction models 78
3.3 Methods 79
  3.3.1 Definitions and data pre-processing 79
  3.3.2 Regression Analysis 81
  3.3.3 Empirical Bayes Method 82
  3.3.4 Bayesian Probabilistic Networks Method 84
3.4 Model development 87
  3.4.1 Data 87
  3.4.2 Definitions and Data Pre-Processing 88
  3.4.3 Multivariate Poisson-lognormal Regression Analysis 93
  3.4.4 Empirical Bayes Method 93
  3.4.5 Bayesian Probabilistic Networks Method 100
3.5 Model comparison 103
3.6 Discussion 106
3.7 Conclusions 108
Acknowledgements 110
INTRODUCTION

1 INTRODUCTION

1.1 Relevance

The economic strength of a nation and the quality of life of its public are reflected in its infrastructure assets. The success and progress of a nation’s society strongly depends on the capability of its physical infrastructure to provide an adequate level of distributing resources and essential services to the public (Hudson et al., 1997). Transportation infrastructure networks are an important part of a nation’s entire infrastructure asset. They substantially contribute to the economic and social development of that nation and sustainable management strategies for transportation networks are needed to achieve and obtain a required level of service and a desired level of safety in the most cost effective manner.

Accident risk is the core element of this thesis. It is considered as the product of three major dimensions: the exposure, the accident and injury occurrence probabilities and the corresponding consequences. Exposition is the magnitude and character of activity on the road networks with specific design characteristics and hence, exposure is closely connected to traffic volume and density. The occurrence probabilities reflect that road accidents are random events. It cannot absolutely be predicted where and when accidents may occur. However, there are statistical methods to assess the probabilities with which accident and injury events are expected to occur on specific road segments given a determinable amount of uncertainty. The current and the predictive occurrence probabilities of accident and injury events are assessed based on so-called accident prediction models. The consequences can be considered as a public health problem based on a distinction of the seriousness of the injury of each accident victim and the loss of health in terms of time spent in hospital or away from work. Additionally, consequences can be described in economic terms such as medical costs, rehabilitation costs, loss of production costs, property damage costs and administration costs (OECD, 1997). Consequences in terms of costs can further be attributed to the number of injured and killed road users in order to demonstrate the seriousness of road accidents as a social problem, to facilitate comparisons between road traffic accident costs and the costs of other causes of injuries and deaths, and to estimate the costs of accident injuries and fatalities to facilitate the reflection of social benefits when these costs are reduced by means of safety increasing measures and interventions. The latter additionally enables to set priorities between different risk reducing interventions and constructional refinements. In general, high numbers of injured and killed road users are connected to high social and economic consequences.

The differentiation between these three dimensions helps to deal with the complex issue of accident risk assessment and its implementation into road infrastructure decision making.
Accident risk is triggered by many different factors and the consequences are a societal concern involving different stakeholder groups. It is desired to promote one general methodology for accident risk based decision making, which could be used in different countries and societies, for which national road safety strategies are intended to be implemented or improved. In order to advance in that direction, there is certainly a need for scientific research in both directions: (1) the improvement of accident risk assessment and prediction as well as (2) the embedment of accident risk models into instruments for road infrastructure safety management and accident risk-based decision making.

1.1.1 Accident risk assessment

On a world-wide scale, of all transportation infrastructures that people have to deal with every day, the public roads transportation system is the most dangerous and the resultant injuries are a major public health issue. An road injury is understood to be a fatal or non-fatal injury incurred as a result of a collision on a public road involving at least one moving vehicle (OECD, 1997). Road accidents belong to the leading causes of death by injury and to the ten leading causes of all deaths globally. As given in Table 1.1, if present trends continue, road injuries are predicted to be one of the five leading contributors to the global burden of decease and injury by 2030 (Worley, 2006, Peden et al., 2009).

Table 1.1: Leading causes of death world-wide, comparison of 2004 observations and 2030 predictions (Peden et al., 2009)

<table>
<thead>
<tr>
<th>Rank</th>
<th>Leading Cause</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Ischaemic heart disease</td>
<td>12.2</td>
</tr>
<tr>
<td>2</td>
<td>Cerebrovascular disease</td>
<td>9.7</td>
</tr>
<tr>
<td>3</td>
<td>Lower respiratory infections</td>
<td>7.0</td>
</tr>
<tr>
<td>4</td>
<td>Chronic obstructive pulmonary disease</td>
<td>5.1</td>
</tr>
<tr>
<td>5</td>
<td>Diarrhoeal diseases</td>
<td>3.6</td>
</tr>
<tr>
<td>6</td>
<td>HIV/AIDS</td>
<td>3.5</td>
</tr>
<tr>
<td>7</td>
<td>Tuberculosis</td>
<td>2.5</td>
</tr>
<tr>
<td>8</td>
<td>Trachea, bronchus, lung cancer</td>
<td>2.3</td>
</tr>
<tr>
<td>9</td>
<td>Road traffic injuries</td>
<td>2.2</td>
</tr>
<tr>
<td>10</td>
<td>Prematurity and low birth weight</td>
<td>2.0</td>
</tr>
<tr>
<td>11</td>
<td>Neonatal infections</td>
<td>1.9</td>
</tr>
<tr>
<td>12</td>
<td>Diabetes mellitus</td>
<td>1.9</td>
</tr>
<tr>
<td>13</td>
<td>Malaria</td>
<td>1.7</td>
</tr>
<tr>
<td>14</td>
<td>Hypertensive heart disease</td>
<td>1.7</td>
</tr>
<tr>
<td>15</td>
<td>Birth asphyxia and birth trauma</td>
<td>1.5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Rank</th>
<th>Leading Cause</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Ischaemic heart disease</td>
<td>14.2</td>
</tr>
<tr>
<td>2</td>
<td>Cerebrovascular disease</td>
<td>12.1</td>
</tr>
<tr>
<td>3</td>
<td>Chronic obstructive pulmonary disease</td>
<td>8.6</td>
</tr>
<tr>
<td>4</td>
<td>Lower respiratory infections</td>
<td>3.8</td>
</tr>
<tr>
<td>5</td>
<td>Road traffic injuries</td>
<td>3.6</td>
</tr>
<tr>
<td>6</td>
<td>Trachea, bronchus, lung cancer</td>
<td>3.4</td>
</tr>
<tr>
<td>7</td>
<td>Diabetes mellitus</td>
<td>3.3</td>
</tr>
<tr>
<td>8</td>
<td>Hypertensive heart disease</td>
<td>2.1</td>
</tr>
<tr>
<td>9</td>
<td>Stomach cancer</td>
<td>1.9</td>
</tr>
<tr>
<td>10</td>
<td>HIV/AIDS</td>
<td>1.8</td>
</tr>
<tr>
<td>11</td>
<td>Nephritis and nephrosis</td>
<td>1.6</td>
</tr>
<tr>
<td>12</td>
<td>Self-inflicted injuries</td>
<td>1.5</td>
</tr>
<tr>
<td>13</td>
<td>Liver cancer</td>
<td>1.4</td>
</tr>
<tr>
<td>14</td>
<td>Colon and rectum cancer</td>
<td>1.4</td>
</tr>
<tr>
<td>15</td>
<td>Oesophagus cancer</td>
<td>1.3</td>
</tr>
</tbody>
</table>
The very first traffic injury accident recorded in a database was supposedly suffered by a cyclist in New York City on 30th May 1896. Till then, it was mainly due to the increasing accessibility of cars for the public and the increasing amount of traffic that made the problem of road accidents and injuries escalate. Today, in accordance to the reports of the World Health Organization and World Bank (Peden et al., 2004, Peden et al., 2009) worldwide the number of injured people due to road accidents is estimated up to 50 million and the number of fatalities about 1.23 million. That corresponds to a daily average of more than 130’000 people being injured and more than 3000 people being killed due to road accidents. Road accident induced fatalities are the third most important cause of the overall mortality and the main cause of death for people younger than 40 years. In addition, for each fatality, there are dozens of injured survivors who are left with short-term or permanent disabilities that may result in a reduced quality of life due to continuous restrictions on their physical functioning or psychosocial activities. Taking into account national epidemiological evidences, road injuries greatly exceed the road fatalities with a world-wide average ratio between fatalities, severe injured and light injured road users estimated to 1:15:70 (EU in 2011 (Koch, 2011): 1:12:50, Switzerland in 2011 (FEDRO, 2012): 1:13:59). In the European Union more than 30’000 people were killed and more than 1.5 million road users were injured in road accidents in 2011 (Koch, 2011, European Parliament, 2012).

According to the global status report on road safety (Peden et al., 2009) the direct economic costs of all road accidents worldwide can be estimated at US$ 518 billion. The costs for the national governments are between 1% and 3% of their gross national product. In Europe, the overall social costs of road accidents are estimated at between EUR 130 billion (about US$ 170 billion) (Koch, 2011) and EUR 180 billion (about US$ 236 billion) per year (European Commission on Mobility and Transport/Road Safety 2012).

Ambitious aims are currently set for the member states of the European Union targeting to halve the number of fatalities on the European road network by 2020, which would correspond to save the lives of 15’000 road users per year. These targets and road safety in general are a shared societal responsibility which has to be supported by efficient road infrastructure safety management procedures. These procedures should cover the entire lifecycle of a road infrastructure from planning to operation.

Relevance for the investigations of this thesis is seen in the development of a methodology that facilitates the development of accident prediction models which contribute to the establishment of national safety strategies.

1.1.2 Road infrastructure decision making

Accident risk models are relevant for road infrastructure decision making and for road infrastructure safety management (European Parliament, 2008). Only one third of the
countries being investigated in the World Report on Road Traffic Injury Prevention (Peden et al., 2004) have a national road safety strategy that is endorsed by their government with specific targets and being funded appropriately for its implementation. Nonetheless, within the last decade, road safety in general has received increasing world-wide attention and support. There has also been an increase in political will in many countries to implement national road safety strategies in order to decrease the accident risk for the road users. For such improvements, safety enhancing interventions can be performed on existing road networks. Road projects in the planning phase can already be tested in terms of their accident risk performance. For the implementation of safety enhancing measures on the road network or the implementation of national safety strategies, financial resources of the countries should be allocated in a purposeful and optimized manner. An optimal allocation of resources requires accurate estimates of the expected accident and injury events and an elaborate evaluation of effective risk reducing measures.

Infrastructure management is defined as the systematic, coordinated planning and programming of investments or expenditures, design, construction, maintenance, operation, and in-service evaluation of physical facilities (Hudson et al., 1997). It covers those actions and processes that ensure that existing infrastructure objects and networks provide an adequate level of service. The risk of failing to provide an adequate level has to be acceptable to all stakeholders. Infrastructure management requires taking into consideration the benefits and costs of infrastructure to all members of society, including economic, environmental and social benefits and costs. It also requires balancing the need for prediction accuracy with analysis effort (World Bank, 1994, Adey, 2002). Optimized scheduling and performance is required of any interventions to maintain or improve the level of service of the road infrastructure network, no matter if such interventions are carried out on structures or roadways. Optimization in this regard refers to management strategies that ensure that the total benefits for society are maximized. Maximizing the benefits of infrastructure interventions comprises taking into account the positive and/or negative consequences of the intervention on all involved stakeholders of a road infrastructure network, i.e. the public, the users and the owner (Adey, 2002, Adey et al., 2012).

Interventions on roads are executed for different reasons, e.g. to maintain the physical condition states of the road or to counteract black spots with potentially increased accident risks. The strategy for interventions, defined in the jurisdiction of the road authorities, has to be optimized in order to minimize the impacts on different stakeholders being affected by the road network under consideration. Modelling these impacts is important for road infrastructure decision making in order to evaluate different design or intervention strategies and to choose the most beneficial one. This also implies the development of impact models for road users which reflect the influence of interventions on the risk of accidents and injuries and the corresponding consequences in terms of costs for the road users. Relevance for the present thesis is seen in the development of impact models that take into account the demand
of different affected stakeholder groups and that can be implemented into road infrastructure
decision making in general and into algorithms and models for the assessment of optimal
intervention strategies in specific.

Innovative methods are required for every country to preserve and develop their road
infrastructure and to apply regular maintenance. A well-preserved road infrastructure might
contribute to reducing fatalities and injuries of road users. Furthermore, research is still
needed to support the implementation of cost-effective measures and to provide models and
tools that will enable policy-makers and road engineers to better understand how road
infrastructure should develop in order to minimize accident and injury risk and therefore to
maximize the benefits to society.
1.2 Aim of the thesis

Taking into account the above mentioned relevance for accident risk assessment and road infrastructure safety management, the aim of this thesis was to develop accurate accident prediction models which can be used as supportive instruments for road infrastructure safety management, optimized intervention planning and stakeholder impact assessment. The aim is furthermore to support an optimized allocation of available budgets e.g. for the implementation of accident risk reducing interventions. The choice of the intervention type or accident countermeasure can be evaluated based on their particular predicted consequences determined by means of the accident risk models. If the methodology presented in this thesis is used appropriately, it will contribute to improve accident risk analysis and, based on this, to a reduction in the general societal burden of injuries and fatalities incurred by road accidents.

Starting with the assumption that road accidents can be predicted by a set of risk indicating random variables, the following three aims are addressed explicitly:

1.2.1 Aim 1: Model development

The first aim of the thesis is to answer the question: Is it possible to develop a methodology which can be used to generate reliable and consistent, efficient and flexible accident prediction models?

An indicator-based generic methodology for the assessment of accident and injury occurrence probabilities on road networks is developed. The proposed methodology is designed in a way that it can be used by accident analysts and road engineers. The methodology is generic in the sense that it is formulated in terms of observable risk indicating variables. Thus, it is adaptable to road networks in different regions, given appropriate data is accessible.

Accident risk assessment and prediction is based on statistical methods that facilitate the incorporation of uncertainties related to the data used and model used. The accident prediction models comprise multivariate and hierarchical model structures that can be modified, learned or updated based on data, available either from historical observations, from observations on other (similar) road segments, or from data collections on a regular basis (e.g. yearly). The results can be used for road infrastructure decision making, e.g. to assess the conditional probability of observing less injury accidents in the following year on a selected road segment, given, that the signalized speed limit is reduced by a certain amount (e.g. 20km/h).
1.2.2 Aim 2: Model comparison

The second aim of the thesis is to answer the question: How do the models developed using the proposed methodology perform compared to those determined using a state-of-the-art methodology?

Accident prediction models developed based on the methodology addressed in Aim 1 are compared to models developed based on the empirical Bayes method (Hauer et al., 2002). The empirical Bayes method is considered to be the standard method for accident risk analysis in the American Highway Safety Manual (AASHTO, 2010) and is also recommended by the European Parliament as the state-of-the-art method for road infrastructure safety management (European Parliament, 2008). The two methods, one based on Bayesian Probabilistic Networks and one based on the empirical Bayes approach, are close in concept and outcome. However, there are significant differences in their applied updating algorithms and in their modelling concepts. A comprehensible modelling concept is essential for the employability of the model in praxis, where different analysts and experts might not be absolutely familiar with Bayesian probability calculus.

The differences of the two Bayesian methods are addressed in terms of the precision of their accident and injury predictions and their applicability as a supportive instrument in accident risk assessment and road infrastructure decision making.

1.2.3 Aim 3: Model implementation

The third aim of the thesis is to answer the question: Is it possible to use the proposed methodology to develop other types of models and other types of impacts at the same time as accidents and injuries?

The proposed methodology for the development of accident prediction models is implemented into a framework for road impact assessment. An important feature of the proposed framework for road impact assessment is that it provides decision support, e.g. on how to optimally allocate budgets into accident risk reducing interventions and evaluate the portfolio of changeable measures in terms of their effect on the accident risk before and after they are implemented. With regard to the efficient and targeted allocation of resources to reduce the accident and injury risk on road networks, this issue is rather complex due to numerous uncertainties prevailing within the field of risk and impact assessment. This aim of the thesis explicitly addresses the use of Bayesian Probabilistic Networks to be implemented into the framework of a road user impact assessment strategy. Different types of uncertainties shall be taken into account representing the uncertain model assumptions which have to be made for such an impact assessment. The model outputs are the joint posterior distribution functions of the consequences (in terms of costs) for the particular stakeholders.
INTRODUCTION

1.3 Scope and issues of the thesis

The overall scope of the thesis is to investigate the aims mentioned in section 1.2 in the contexts of road infrastructure safety management and road infrastructure decision making.

The scope for the model development and its application is set on a probabilistic investigation of the influences of road design and traffic related factors on the occurrence probabilities of injury accidents and different levels of injuries. Hierarchical multivariate Poisson-lognormal regression analysis and Bayesian Probabilistic Networks are used for model development. The developed models are applied to road segments of a rural motorway network and implemented into impact models for road users.

The overall scope can be subdivided into four specific issues addressing (1) to investigate the accident contributing factors that should be used as model variables, (2) to investigate statistical methodologies that should be used for the development of accident prediction models, (3) to investigate the objects the models should be developed and applied for and (4) to investigate the capability of the methodology to be implemented into a framework for road impact assessment.

1.3.1 Issue 1: Accident contributing factors

Accident risk analysis is performed based on defined sets of explanatory risk indicating variables and dependent response variables. The risk indicating variables are considered to influence the occurrence probabilities of the response variables (e.g. accident and injury events).

The occurrence of accidents is random in a sense that they occur as a function of a set of random and unpredictable events which are influenced by different accident contributing factors\(^1\) (Valent et al., 2002, Yau, 2004, Shankar et al., 1995). Figure 1.1 illustrates different groups of contributing factors to vehicle accidents on roads: Human factors, vehicle related factors as well as road design and traffic related factors.

**Human factors** have the largest influence on the occurrence probability of accident events and include e.g. the age of the road users, driver skills, attention and fatigue, as well as experience, use of intoxicative substances or use of cellular-telephones (Petridou and Moustaki, 2000, Odgen, 1996, Redelmeier and Tibshirani, 1997, Movig et al., 2004). The contributions of human factors are too manifold and too complex to be controlled directly by the road infrastructure decision makers. In order to control for the human factors law-enforcement actions would be needed, e.g. stop-and-search operations by the police authorities.

\(^1\) The terming accident contributing factor can be used interchangeably with the terming risk indicating variable. The latter is more frequently used in the context of the model development.
INTRODUCTION

Accident contributing factors can also be connected to the vehicle, including design, maintenance and manufacture. These **vehicle related factors** comprise e.g. the engine power or the age of the vehicles abetting technical malfunctions or failures (Bédard et al., 2002, Langley et al., 2000). Available data about such factors can be used to indicate possible occurrence probabilities of accidents. Being part of the permanent progress of technical innovations in automotive industry, there is a considerable growth in the development of new technological features which can be implemented into the technical equipment of vehicles in order to avoid accident situations and reduce the consequences of accidents (e.g. automated distance or speed control, braking assistance). Such developments also help to protect the occupants by reducing the level of injuries the occupants might suffer in case that accidents cannot have been avoided (e.g. seatbelts, airbags, automated seat-belt tensioner) (Robertson, 1996, Harvey and Durbin, 1986, Richter et al., 2005). The contributions of vehicle related factors cannot be controlled directly by the road infrastructure decision makers. In order to control for them, law-enforcement actions would be needed or technical developments have to be intensified.

The third group of accident contributing factors comprises **road design and traffic related factors** (including the signalized speed limit). If related to road design, such factors are e.g. the number and width of the driving lanes, the radius of curves or the design speed (Miaou, 1994, Aarts and van Schagen, 2006, Nilsson, 2004, Karlaftis and Golias, 2002, Amundsen and Ranes, 2000). Traffic related factors are e.g. the daily traffic volume or the percentage of heavy good vehicles (Zhou and Sisiopiku, 1997, Hauer, 1995, Martin, 2002, Qin et al., 2004, Pei et al., 2012). Only accident contributing factors related to road design and traffic volume can directly be controlled by road infrastructure decision makers, since they are in the position to modify the values of the factors. By understanding the influences of these factors on accident events, specific measures can be implemented to minimize the accident promoting effects of the particular factors. Model variables can be selected that represent these factors and models can be developed that help that traffic engineers and motorway designers can better estimate the impacts of their management and design choices on the road users’ accident risk.
A geometric road design which is optimized in terms of accident risk, merely can help to control driving speeds and to reduce the number of accidents (Ma, 2006). It is difficult to determine the relationships between individual design-related contributing factors and particular accident incidents. Some accidents, however, may have been prevented if the motorway was designed differently, meaning that changing the motorway design has considerable potential to contribute to road safety strategies in order to prevent similar accidents in future.

The investigations presented in this thesis concentrate on the reduction of accident events and injuries on such road sections where it is believed that deterministic accident contributing factors only related to road design and traffic characteristics are the relevant ones. Herewith such a set of factors is assumed to contribute to the occurrences of accident events either directly or indirectly through interactions with human factors or vehicle factors, or both. The methodology presented in this thesis was developed to deal with road design and traffic related accident contributing factors. Human factors as well as vehicle related factors are not considered in the case study of this thesis (section 2.3). The proposed methodology, however, allows including human and vehicle related factors in future investigations.

---

2 Note that neither size nor overlap of the circles used to illustrate the scope of the thesis in Figures 1 - 4 do have any meanings in terms of their general relevance or interrelations.
1.3.2 Issue 2: Statistical methodologies

Probability in general is a numerical measure of an event’s occurrence likelihood relative to a set of alternative events. Occurrence probabilities can be interpreted from different perspectives. As illustrated in Figure 1.2 the main approaches for interpreting probabilities are classical, frequentistic and Bayesian. The brief descriptions of the different interpretations in the subsequent paragraphs are based on Benjamin and Cornell (1970), Ang and Tang (2007) and Faber (2012).

The classical interpretation of probability was developed first in the games of cards and dice. It is formulated as the share of equally likely ways by which an experiment may lead to a specific event over the total number of equally likely ways in the experiment. Since the answer is known in advance, there is no need to actually carry out experiments within the classical interpretation. However, the solution for the likelihood could only be determined if all possible outcomes of the experiment could be derived analytically.

The frequentistic interpretation of probability states that probabilities are a characteristic of nature. The probabilities based on that interpretation typically correspond to the results of experiments, meaning that the probability for a specific event is similar to the relative frequency with which that event has been observed during the performance of the experiment.

The Bayesian interpretation of probability, in contrast, is formulated as a degree of believe. This degree of believe is considered to be subjective, since it is representing the state of mind in terms of experience, expertise and preferences. The classical and frequentistic approaches have no way to incorporate such subjective information into their modelling processes. The incorporation of prior believes can particularly be useful when only a small data sample is available. As an additional aspect, Bayesian probability can comprise both, frequentistic and classical approaches. The first in terms that subjective probabilities are assigned based on the outcome of an experiment, and the latter in terms of symmetry considerations (Faber, 2012). Bayesian inference is the mathematical formulation of the Bayesian interpretation of probability. This is representing the process of fitting a probability model to a set of data and summarizing the result by a probability distribution on the parameters of the model and on predictions of new observations (Gelman et al., 2004).

Argumentation based on the frequentistic interpretation assumes that there is an unknown but fixed set of model parameters. Bayesian interpretation in contrast assumes that the parameters are not fixed but random variables, the uncertainty of which is quantified by means of

---

3 For sake of unambiguity in the terminologies used in this paper, it is distinguished between statistical approaches, methods, models and tools. Statistical approaches are fundamental principles for the mathematical and theoretical formulation of probability calculus (e.g. Empirical vs. Full Bayes). Statistical methods are applying a selected statistical approach using one or more statistical tools (e.g. Bayesian Probabilistic Networks) in order to develop statistical models, e.g. for the prediction of accident events.
Bayesian probability theory. In the frequentistic approach all possible data sets are considered to be generated by the unknown fixed parameters while the Bayesian approach treats data sets as given and considers all possible values of unknown parameters (Ma, 2006, Gelman et al., 2004).

Modern accident risk analysis is mostly based on the Bayesian interpretation of probability (de Oña et al., 2013, de Oña et al., 2011, Eenink et al., 2008, El-Basyouny and Sayed, 2011, Hossain and Muromachi, 2012, Huang and Abdel-Aty, 2010, Karwa et al., 2011, Lan et al., 2009, Meng et al., 2010, Park et al., 2010, Persaud et al., 2010, Schubert et al., 2011, Xu et al., 2010). Different methods have evolved for development of accident prediction models using either Empirical or Full Bayesian approaches. The methods applying these approaches are then denoted Empirical Bayes method or Full Bayes methods, correspondingly. The difference between Empirical and Full Bayes approaches is that for the former (historical) data is used for the estimation of the prior model parameters, while the latter comprises a full prior probability model with a joint probability distribution for all observable and unobservable quantities in a system without any consideration of available data (section 3.2.2).

This thesis uses Bayesian Probabilistic Networks (BPNs) which can be based on both approaches, Empirical or Full Bayes (Pearl, 1988, Jensen and Nielsen, 2007) (Annex VI). A methodology is proposed for cross-sectional accident risk modelling based on a systems theoretic definition for elements of road networks. The application of BPNs within this methodology is discussed and evaluated. The elements within a BPN comprise the set of
accident risk indicating variables (contributing factors) and model response variables (injury accidents and different degrees of accident induced injuries) considered within the risk assessment and management problem. The joint probability distributions of the model response variables are considered to be of high value for decision makers, since they directly reflect the uncertainties involved in modelling the considered problem.

1.3.3 Issue 3: Objects for application

For the development of accident prediction models it has to be distinguished between the objects the models are intended to be used for. In accordance with the Highway Safety Manual (AASHTO, 2010) and as illustrated in Figure 1.3 a differentiation can be made for accident prediction models on urban and suburban arterials, rural two-lane/two-way roads and rural multilane motorways. The proposed methodology can be used to develop accident prediction models for all of these different road objects.

In this thesis the proposed methodology is tested during a case study in which accident prediction models are developed for rural motorways only (note that the proposed methodology is also applicable to other types of roads). Motorways are roads being specifically designed and built for motor vehicle traffic and being provided with separate carriageways for the two directions of traffic, separated from each other. Motorways additionally do not cross at any level with any other road, railway or tramway track and are especially sign-posted as a motorway being reserved for specific categories of road motor vehicles only (IRF, 2009). Additionally, in the case study of this thesis it is distinguished between open roads, intersections, exit corridors and tunnels. For each of these types of road separate accident prediction models are developed. Other civil engineering structures like e.g. bridges and their influence on the occurrence probabilities of accident events could not be taken into account due to a lack of such data.
The accident risk assessment is performed on homogeneous motorway segments. A homogenous segment is a segment of road where it can be assumed that the values of all risk indicating variables included in the model are constant, and thus also the risk is uniform. The rural motorway network is sub-divided into homogeneous segments based on available data of the risk indicating variables. A 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.

1.3.4  Issue 4: Model implementation

Optimal intervention strategies are those that result in the lowest negative impacts for stakeholders. Stakeholder for public roads can be grouped as the owner, the user, and the public (Lethanh et al., 2012, Adey et al., 2012) (Figure 1.4). The determination of cost-optimal intervention strategies on road segments is herewith the main goal. This includes making decisions about when to intervene, the type of intervention and the traffic configuration to use during the intervention (Adey et al., 2003, Chien and Tang, 2012). The largest impacts caused during interventions and between interventions on a road network are the impacts on road users (Adey et al., 2012).

The estimation of these impacts is complex due to the many different and interconnected physical relationships (Gomez et al., 2011). For instance, an increase in travel time due to a deteriorated road condition also depends on the slope, the curvature, the capacity, the signalized speed limit as well as the number of vehicles travelling on the road. At the same
time, the number of vehicles and the road condition have – amongst other factors (section 1.3.1) – a certain influence on the occurrence probabilities of accident events.

As a first step towards an overall probabilistic multiple impact model for all stakeholder groups shown in Figure 1.4, this thesis investigates the possibility of using the proposed methodology to develop models of multiple user impact types. Impacts on the stakeholder groups owner and public are not considered. Consequences incurred by injured road users are determined. Additionally, user costs due to extended travelling time (e.g. caused by congestions) and due to vehicle costs (e.g. fuel costs) are taken into account.

![Figure 1.4: Different stakeholder groups for impact assessment of road networks.](image)

### 1.4 Outline of the Thesis

The thesis is segmented into five chapters. In chapter 1 the relevance and motivations are outlined, the aims and scopes are presented and the applied methodologies discussed. A separate literature review is not included in chapter 1, since literature reviews are comprised in each of the subsequent chapters.

Chapters 2-4 represent research work which has been published, accepted or submitted for publication in two peer reviewed journal papers and one peer reviewed conference paper.
INTRODUCTION

during the PhD study\textsuperscript{4}. Each chapter focuses on one of the three issues described in section 1.2.

Chapter 2 (paper I) comprises the development of a novel methodology for the development of accident and injury prediction models. The methodology utilizes a combination of 1) Gamma-updating of the occurrence frequencies of injury accidents and injured road users as a method for pre-processing of the data, 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. Prior Bayesian Probabilistic Networks are first established by means of multivariate regression analysis. Parameter learning is performed 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. In the case study – using the Austrian motorway network and randomly selected road segments – the proposed method is used to produce a model to simultaneously predict the expected number of injury accidents and the expected number of lightly, severely and fatally injured road users.

In chapter 3 (paper II), two Bayesian methods for the development of accident prediction models are compared. The models are based on the state-of-the-art Empirical Bayes methodology and based on the proposed methodology using Bayesian Probabilistic Networks as described in Chapter 2. The principles of Bayesian probability theory are briefly described in order to highlight the differences between the two methodologies using Bayesian approaches for the development of accident prediction models. Both methodologies are used to develop models for a multivariate prediction of accident and injury occurrences. Both methods are applied on a test dataset containing road segments which have not been used before for model development to predict the expected number of accidents and injuries which are then compared to the actual observed numbers. The application of the two methods to the test dataset can be considered as an example of the use of accident prediction models as helpful instruments for road safety impact assessment during the planning phase of not yet build road links or of existing road links, for which major interventions or re-designing are intended to be performed.

\textsuperscript{4} The publications listed in this thesis (chapters 2 – 4) are partly based on other work or projects which have been conducted during the period of the PhD studies. These are in particular:

- A pilot research project has been conducted with the Austrian road safety board (Kuratorium für Verkehrssicherheit, KVF) in 2010. The data, which has been used for the investigations described in this thesis, are originating from that project and the supply and pre-processing of the data by the KVF is gratefully acknowledged.
- The comparison between the two Bayesian methods, namely the empirical Bayes method and the Bayesian Probabilistic Networks based method, was also partly described in a peer reviewed paper accepted for the annual conference of the Canadian Society for Civil Engineering CSCE in 2013. However, descriptions and investigations of the journal paper represented in chapter 2 of the thesis go far beyond of what has been described in the corresponding conference paper. Hence, in order to avoid overlap in content, the conference paper has not been included into this thesis.
INTRODUCTION

Bringing to mind the three major dimensions of accident risk (compare chapter 1.1), namely exposure, occurrence probability and consequences, the Chapters 2 and 3 comprise research mainly about the first two dimensions. They consider the exposure of the road users and the development of a methodology to estimate the occurrence probabilities of accident and injury events by means of Bayesian Probabilistic Network based accident prediction models.

In Chapter 4, the third dimension of accident risk assessment, namely the consequences, is considered. The developed accident prediction models are implemented into models for road user impact assessment within the context of road infrastructure decision making. A user cost model is established by extending and modifying the proposed Bayesian Probabilistic Network methodology. Although all models are developed based on the same methodology, the models which are developed in this chapter are different to the models developed in Chapter 2:

- The road injury prediction model was implemented and embedded into a novel framework for multivariate impact assessment for road users on rural, multilane road networks. Impacts of road links can be categorized into different stakeholder groups and focus was set on the costs for road users, differentiated into travelling time, vehicle operation and accident injury costs. The accident prediction model as described in Chapter 2 now additionally provides joint probabilities of the expected consequences resulting from the predicted probabilities for light, severe and fatal injuries; the sum of the consequences for each injury severity level results in the estimated accident user costs.

- The road injury prediction model is modified to provide a more generic character than the Bayesian Probabilistic Network developed in Chapter 2. The direct input and risk indicating variable “signalized speed limit” is substituted by an accident modification factor which is based on theoretical probabilistic models for the actual driving speed, given a particular signalized speed limit. The causal relations of driving speed on different levels of injuries are based on the power models of Nilsson (2004).

- The road injury prediction model is extended with an accident modification factor representing the influence of the pavement surface friction on the expected number of accident events. This extension enlarges the range of applied infrastructure related contributing factors and provides more interfaces between accident cost contributing factors and the remaining variables of the impact model for different stakeholders.

A Bayesian Probabilistic Network is established for the assessment of road user costs. For this the proposed methodology for the development of accident prediction models is implemented into a novel framework for multiple impact assessment. Such an application of Bayesian Probabilistic Networks is new and helpful for road infrastructure decision making. It facilitates the incorporation of probability theory and the representation of uncertainties.
appended to each cost estimation. Such uncertainties are not taken into account in the existing impact models.

In Chapter 5, conclusions about the main findings in each of the Chapters 2-4 are provided. The scientific achievements of this thesis are outlined and the limitations of the work are discussed. This chapter closes with an outlook and with recommendations for future research endeavours. In the Annex of this thesis, a brief background on the statistical methods used for the proposed methodology is provided.
2 PREDICTION OF ROAD ACCIDENTS: A BAYESIAN HIERARCHICAL APPROACH (PAPER I)

Markus Deublein
Institute of Construction and Infrastructure Management,
Swiss Federal Institute of Technology, ETHZ, Zurich, Switzerland

Matthias Schubert
Matrisk GmbH, Managing Technical Risks,
Zurich, Switzerland

Bryan T. Adey
Institute of Construction and Infrastructure Management,
Swiss Federal Institute of Technology, ETHZ, Zurich, Switzerland

Jochen Köhler
Department of Structural Engineering,
Norwegian University of Science and Technology, NTNU, Trondheim, Norway

Michael H. Faber
Department of Civil Engineering,
Technical University of Denmark, DTU, Lyngby, Denmark

published in
Accident Analysis and Prevention, Vol. 51, pp. 274-291
Abstract

This paper presents a novel methodology for the prediction of the occurrence of road accidents. The methodology utilizes a combination of three statistical methods: 1) Gamma-updating of the occurrence frequencies 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 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 black spots 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.

Keywords road safety assessment; accident prediction; injury accidents; Bayesian Probabilistic Networks; accident risk modelling; multivariate regression analysis; hierarchical Bayes;

2.1 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 (Peden et al., 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 (Song et al., 2006b, Elvik, 2011, Bijleveld, 2005). 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 well discussed phenomenon of over-dispersion (Cox, 1983, Dean and Lawless, 1989, Hauer, 2001, Karlis and Meligkotsidou, 2005, Gschlößl 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. (2006b) 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. (2006a), Maes and Dann (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 co-variance 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.

2.2 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 2.2.1), 2) subdivision of the road network into homogenous segments (section 2.2.2), 3) Gamma-updating of the response variables (section 2.2.3), 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 2.2.4), 5) the construction and parameter learning of the BPN (section 2.2.5) and 6) the prediction of the expected number of response variable events, i.e. the expected number of injury accidents (section 2.2.6).

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.

2.2.1 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 for 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 to the prediction of accident events also of any road sections, which 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.

### 2.2.2 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.

Figure 2.1 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)).

![Figure 2.1](image)

Figure 2.1: Example for segregation of a road section into three homogenous segments based on values of the observed risk indicating variables

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 2.3).
2.2.3 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.

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 \( i^{th} \) homogeneous segment is assumed to be Poisson distributed as suggested e.g. by Song et al. (2006b) and Park and Lord (2007).

\[
Y_{ik} | \mu_{ik} \sim \text{Poisson}(\mu_{ik}) \tag{2.1}
\]

with mean occurrence frequency

\[
\mu_{ik} = \nu_i \cdot \lambda_{ik} \tag{2.2}
\]

where \( \nu_i \) is the exposure (in million vehicle kilometres of travel (\( mvk \) ) per year) and \( \lambda_{ik} \) 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 \( \lambda \) itself defined by the parameters \( \alpha \) and \( \beta \) (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:
with expected value and variance

\[
E[y] = \frac{\alpha}{\beta} \quad \text{and} \quad \text{VAR}[y] = \frac{\alpha}{\beta^2} \left(\beta + 1\right) = \frac{\alpha}{\beta} \cdot \frac{\beta + 1}{\beta}
\]  

with shape (dispersion) parameter \( \alpha > 0 \) and inverse scale parameter \( \beta > 0 \). \( E[.] \) is the expectation operator and \( \text{VAR}[.] \) the variance operator. According to equation \( \{2.4\} \) the variance is always greater than the expected value. As \( \beta \) 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 \( \lambda_{ik}' \) are first calculated based on averaged information over all homogeneous segments of the network. The prior distribution parameters of \( \lambda_{ik}' \) are then updated on the second level of hierarchy using observations of the response variables for the individual homogeneous segments.

**Second Level of Hierarchy**

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

\[
\lambda_{ik}' \sim \text{Gamma}(\alpha_{ik}', \beta_i')
\]  

with probability density

\[
p(\lambda_{ik}') = \frac{\beta_i'^{\alpha_{ik}'} \cdot \lambda_{ik}'^{\alpha_{ik}' - 1} \cdot e^{-\beta_i' \cdot \lambda_{ik}'}}{\Gamma(\alpha_{ik}')}
\]  

and expected value of \( \lambda_{ik}' \) as
The prior shape parameter $\alpha_{ik}$ represents the expected number of accidents and is calculated for the $i^{th}$ homogeneous segment and the $k^{th}$ response variable as $\alpha_{ik} = \lambda_k \cdot \beta'_i$. $\beta'_i$ is the prior inverse scale parameter of the Gamma distribution representing the weighted exposure as given by $\beta'_i = v_i \cdot \omega_i = v_i \cdot \frac{\psi}{l_i}$. 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 $v_i$ is multiplied by the prior weight $\omega_i$ being the fraction of the weighting factor $\psi$ and the individual homogeneous segment length $l_i$. The length is included in the estimation of the prior weight since short sections are less likely to experience events than longer segments. $\psi$ is introduced in a Bayesian sense to give weight to the prior parameter $\beta'_i$ 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 $\lambda_k$ are used for the assessment of the prior Gamma parameters $\alpha_{ik}$ 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 $\lambda_k$ for the $k=1,...,z$ different response variables and the optimal value for the weighting factor $\psi$, which is the same for all response variables and all homogeneous segments. The objective is to minimize the difference between the observed ($\hat{y}_{ik}$) and the NB distributed ($\tilde{y}_{ik}$) number of events of the response variables, the latter being assessed using equations {2.2} and {2.7} as $\tilde{y}_{ik} = v_i \cdot \frac{\alpha_{ik}}{\beta'_i}$. The optimization problem may be formulated as

$$E[\lambda'_i] = \frac{1}{\beta'_i} \quad \text{(2.7)}$$
arg min \{ f(\psi, \hat{\lambda}_k) \} \quad \{2.8\}

with

\[
f(\psi, \hat{\lambda}_k) = \sum_{i=1}^{n} \tilde{y}_{ik} - \sum_{i=1}^{n} \hat{y}_{ik}\quad \{2.9\}
\]

subject to \( 0 < \psi \leq 1 \) and \( \hat{\lambda}_k > 0 \).

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

\[
\lambda_{ik}'' | \tilde{y}_{ik}, \tilde{v}_i \sim Gamma(\alpha_{ik}''', \beta_{ik}''). \quad \{2.10\}
\]

with

\[
\alpha_{ik}''' = \alpha_{ik}' + \tilde{y}_{ik} \quad \text{and} \quad \beta_{ik}''' = \beta_{ik}' + \tilde{v}_i \quad \{2.11\}
\]

Where \( \tilde{y}_{ik,j} \) represents the observed sum of response variable counts and \( \tilde{v}_i \) represents the exposure in the \( i^{th} \) 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.
2.2.4 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 $\varepsilon_k$ 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 $\lambda_{ik}^\prime\prime$ for every $i^{th}$ 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

$$
\ln \left( E \left[ \Lambda | X \right] \right) = BX + \Xi \quad \triangleq \quad E \left[ \Lambda | X \right] = \exp (BX + \Xi) \quad \{2.12\}
$$

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

$$
\Lambda = \begin{pmatrix}
\lambda_1^\prime & \ldots & \lambda_z^\prime \\
\vdots & \ddots & \vdots \\
\lambda_{u1}^\prime & \ldots & \lambda_{uz}^\prime \\
\end{pmatrix}, \quad X = \begin{pmatrix}
x_{11} & \ldots & x_{1u} \\
\vdots & \ddots & \vdots \\
x_{u1} & \ldots & x_{zu} \\
\end{pmatrix} \quad \text{and} \quad \Xi = \begin{pmatrix}
\varepsilon_{11} & \ldots & \varepsilon_{1z} \\
\vdots & \ddots & \vdots \\
\varepsilon_{u1} & \ldots & \varepsilon_{uz} \\
\end{pmatrix} \quad \{2.13\}
$$
\( \mathbf{\varepsilon}_{\mathbf{e}} \) represents the vector of the normal distributed random variation for each homogeneous segment and for the \( k^{th} \) different response variable with \( \mathbf{M}_{\mathbf{e}} \) as the mean and \( \mathbf{\Sigma}_{\mathbf{e}} \) as the covariance matrix. For normally distributed response variables (logarithmically transformed), the regression coefficients \( \mathbf{\hat{B}} \) estimated by means of the Maximum Likelihood Method correspond to the least squares estimates assessed as

\[
\mathbf{\hat{B}} = \left( \mathbf{X}^T \mathbf{X} \right)^{-1} \mathbf{X}^T \mathbf{\Lambda}.
\]  \{2.14\}

With a matrix of residuals \( \mathbf{\hat{R}} \) being computed as the difference between the observations and the model predictions of the event occurrence rates, \( \mathbf{\Lambda} \) and \( \mathbf{\hat{\Lambda}} \), respectively:

\[ \mathbf{\hat{R}} = \mathbf{\Lambda} - \mathbf{X} \mathbf{\hat{B}} = \mathbf{\Lambda} - \mathbf{\hat{\Lambda}}. \]

The predicted covariance matrix of the error term is assessed based on the residuals as

\[
\mathbf{\hat{\Sigma}} = \frac{1}{n - p - 1} \mathbf{\hat{R}}^T \mathbf{\hat{R}}
\]  \{2.15\}

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. (2006b), 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\(^5\), 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 (Cheng and Washington, 2005, Elvik, 2002). 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.

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 & 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 \( \lambda_k \) given observations of the different risk indicating variables. The joint probability distribution represented by the BPN can be formulated as

\[
p(A) = \prod_{k=1}^{z} p(A_k | X) \tag{2.16}
\]

\( p(A_k | X) \) represents the conditional probability of \( A_k \) given \( X \).

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 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.
2.2.6 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 homogeneous segment as given in equation \(2.2\) 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 (equation \(2.1\)).

2.3 Case Study

2.3.1 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 1,821 km. Since data was provided for both driving directions separately, the total length of the investigated road network is 3,642 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 Figure 2.2. 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.
2.2 Use of Data

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 Figure 2.2. The data of this road link was also excluded from the initial dataset and was then used for testing the model predictions.

2.3.2 Determination of Model Variables

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 2.1. The discretization intervals are determined based on the rates of the response variables $\lambda_x$ 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).

Table 2.1: Definition of response variables

<table>
<thead>
<tr>
<th>acronym</th>
<th>description</th>
<th>discretization intervals for BPN</th>
</tr>
</thead>
<tbody>
<tr>
<td>IAC</td>
<td>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.</td>
<td>interval size of $\lambda_{iac}$ for $0 \leq \lambda_{iac} &lt; 0.01$ $0.01$ for $0.01 \leq \lambda_{iac} &lt; 0.2$ $0.1$ for $0.2 \leq \lambda_{iac} &lt; 2$</td>
</tr>
<tr>
<td>LINJ</td>
<td>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.</td>
<td>interval size of $\lambda_{linj}$ for $0 \leq \lambda_{linj} &lt; 0.01$ $0.01$ for $0.01 \leq \lambda_{linj} &lt; 0.2$</td>
</tr>
<tr>
<td>SINJ</td>
<td>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 severely 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.</td>
<td>interval size of $\lambda_{sijn}$ for $0 \leq \lambda_{sijn} &lt; 0.01$ $0.01$ for $0.01 \leq \lambda_{sijn} &lt; 0.2$</td>
</tr>
<tr>
<td>FAT</td>
<td>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.</td>
<td>interval size of $\lambda_{fat}$ for $0 \leq \lambda_{fat} &lt; 0.001$ $0.001$ for $0.001 \leq \lambda_{fat} &lt; 0.02$</td>
</tr>
</tbody>
</table>

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 2.2 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.

Table 2.2: Length of investigated road networks and counts of
response variables between 2004 and 2010

<table>
<thead>
<tr>
<th>length</th>
<th>IAC</th>
<th>LINJ</th>
<th>SINJ</th>
<th>FAT</th>
</tr>
</thead>
<tbody>
<tr>
<td>[km]</td>
<td>/-</td>
<td>/-</td>
<td>/-</td>
<td>/-</td>
</tr>
<tr>
<td>entire dataset</td>
<td>3'642</td>
<td>12'892</td>
<td>14'482</td>
<td>5'861</td>
</tr>
<tr>
<td>development dataset</td>
<td>2'952</td>
<td>10'282</td>
<td>11'460</td>
<td>4'677</td>
</tr>
<tr>
<td>test dataset</td>
<td>690</td>
<td>2'610</td>
<td>3'022</td>
<td>1'184</td>
</tr>
</tbody>
</table>

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 2.3.

Table 2.3: Risk indicating variables for accident risk modelling

<table>
<thead>
<tr>
<th>acronym</th>
<th>[units]</th>
<th>[intervals]</th>
<th>description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHAR</td>
<td>[exit corridors, intersections, tunnels, open roads]</td>
<td>[1, 2, 3, 4]</td>
<td>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.</td>
</tr>
<tr>
<td>AADT</td>
<td>[vehicles/day]</td>
<td>[0, 10, ..., 100] x10³</td>
<td>The variable AADT represents the annual average daily traffic per 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 [ v = \frac{AADT}{365} \cdot \text{length}, ] communicated in terms of million vehicle kilometre travelled (mvk).</td>
</tr>
<tr>
<td>HGV</td>
<td>[%]</td>
<td>[0, 5, ..., 30]</td>
<td>The variable HGV represents the fraction of heavy good vehicles travelling on the road section with respect to the AADT.</td>
</tr>
</tbody>
</table>
\[
HGV = \frac{AAHG\text{V}}{A\text{ADT}} \times 100\%
\]

with \(AAHG\text{V}\) being the annual average number of heavy good vehicles.

**BEND \([-\text{\%}]\) \([0, 2, ..., 10]\)**

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 50m grid as illustrated in Figure 2.3 using the following equation

\[
bend = \left( \frac{\sum_{i=1}^{10} x_i}{z} - 1 \right) \times 100
\]

\(z\) 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 50m.

**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 \([-\text{\%}]\) \([1, 2, 3, 4]\)**

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 \([\text{km/h}]\) \([80, 90, ..., 130]\)**

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 \([\text{yes/no}] = [1/0]\)**

The binary variable EML represents the existence of road emergency lanes having the values 0 or 1 for “no” and “yes”, respectively.
The sample mean \((m)\), sample standard deviation \((s)\), minimum \((\text{min})\) and maximum \((\text{max})\) values of the eight investigated risk indicating variables for the development and test datasets are given in Table 2.4.

<table>
<thead>
<tr>
<th>CHAR</th>
<th>AADT</th>
<th>HGV</th>
<th>BEND</th>
<th>SLP</th>
<th>LAN</th>
<th>SPD</th>
<th>EML</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>[-]</td>
<td>[vehicle/day]</td>
<td>[%]</td>
<td>[-]</td>
<td>[%]</td>
<td>[-]</td>
<td>[km/h]</td>
</tr>
<tr>
<td>n=5'546 development dataset</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(m)</td>
<td>3.08</td>
<td>17'796</td>
<td>12.5</td>
<td>4.2</td>
<td>-0.01</td>
<td>2.2</td>
<td>120</td>
</tr>
<tr>
<td>(s)</td>
<td>1.26</td>
<td>11'357</td>
<td>4.9</td>
<td>5.74</td>
<td>1.67</td>
<td>0.5</td>
<td>16</td>
</tr>
<tr>
<td>(\text{min})</td>
<td>1</td>
<td>2'800</td>
<td>3</td>
<td>0</td>
<td>-6</td>
<td>1</td>
<td>80</td>
</tr>
<tr>
<td>(\text{max})</td>
<td>4</td>
<td>99'950</td>
<td>29</td>
<td>76.3</td>
<td>6</td>
<td>4</td>
<td>130</td>
</tr>
<tr>
<td>n=1'386 test dataset</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(m)</td>
<td>3.14</td>
<td>17'521</td>
<td>12.6</td>
<td>4.49</td>
<td>0.01</td>
<td>2.1</td>
<td>120</td>
</tr>
<tr>
<td>(s)</td>
<td>1.22</td>
<td>10'996</td>
<td>4.9</td>
<td>6.24</td>
<td>1.64</td>
<td>0.5</td>
<td>16</td>
</tr>
<tr>
<td>(\text{min})</td>
<td>1</td>
<td>2'800</td>
<td>3</td>
<td>0</td>
<td>-6</td>
<td>1</td>
<td>80</td>
</tr>
<tr>
<td>(\text{max})</td>
<td>4</td>
<td>99'950</td>
<td>29</td>
<td>71</td>
<td>6</td>
<td>4</td>
<td>130</td>
</tr>
</tbody>
</table>

### 2.3.3 Construction of Homogeneous Segments

The homogeneous segments were determined using the values of the risk indicating variables shown in Table 2.3. The entire road network was sectioned into \(n_{\text{HS}} = 6'932\) homogeneous segments, which were randomly apportioned into the development dataset \((n_{\text{dev}} = 5'546, 75\%\) of data) and the test datasets \((n_{\text{test}} = 1'386, 25\%\) 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.

### 2.3.4 Gamma-Updating of Model Response Variables

The occurrence frequencies of the response variables were assessed as described in paragraph 2.2.3. The parameters of the Gamma distribution were quantified and updated for each homogeneous segment based on the background rates given in Table 2.5.
Table 2.5: Weighting factor and background rates assessed for model development dataset

<table>
<thead>
<tr>
<th>$\psi$</th>
<th>$\lambda_{IAC}$</th>
<th>$\lambda_{LINJ}$</th>
<th>$\lambda_{SINJ}$</th>
<th>$\lambda_{FAT}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>[-]</td>
<td>[IAC/mvk]</td>
<td>[LINJ/mvk]</td>
<td>[SINJ/mvk]</td>
<td>[FAT/mvk]</td>
</tr>
<tr>
<td>0.3</td>
<td>0.08764</td>
<td>0.09910</td>
<td>0.03705</td>
<td>0.00315</td>
</tr>
</tbody>
</table>

Both, the weighting factor $\psi$ and the background rates $\lambda_k$ 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 equation (2.8). The posterior rates $\lambda_k$ were assessed based on the Gamma-updating procedure taking into account the background rates $\lambda_k$ and the observed counts of the response variables $\tilde{y}_{ik}$ (likelihoods) for every $i^{th}$ homogeneous segment, as well as for every of the $k=1,...,z$ 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 equation (2.11). The influence of different $\tilde{y}_{ik}$ values on the expected number of injury accidents (based on posterior rates) is shown in Figure 2.4.

![Figure 2.4: Updating of Gamma probability density function (pdf) with different values of $\tilde{y}_{ik}$ with constant exposure](image-url)
The Gamma shaped prior probability density function for the expected number of injury accidents is illustrated in Figure 2.4 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 \( y_{ik} = 1 \) 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.

### 2.3.5 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 2.1). The seven risk indicating variables (excluding CHAR) from Table 2.3 were used as independent variables. The regression equation used has the multiplicative form as

\[
Y_{ik} | CHAR = v_{ik} \cdot \exp \left( \beta_{0,i,k} + \beta_{1,i,k} \cdot \ln(AADT_i) + \beta_{2,i,k} \cdot \ln(HGV_i) + \beta_{3,i,k} \cdot BEND_i + \ldots + \beta_{4,i,k} \cdot SLP_i^2 + \beta_{5,i,k} \cdot LAN_i + \beta_{6,i,k} \cdot VEL_i^2 + \beta_{7,i,k} \cdot EML_i + \epsilon_i \right)
\]

where \( Y_{ik} \) 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 \( \beta \)s 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), Malyskina and Manering (2008), Aljanahi et al. (1999), Aarts and van Schagen (2006), Haglund and Aberg (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 \( n_{\text{dev},w} = 73'389 \). The statistical significance of the results was tested using a Students t-test for the individual regression coefficients, where the Null-
hypothesis of the t-test with \( n-u-1 \) (\( u=\text{number of estimated regression coefficients} \)) degrees of freedom was rejected at the significance level of \( \alpha = 0.05 \).

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 2.6. Values of the regression coefficients considered to be statistically not significant are marked with asterisks (*). It has to be noted that in Table 2.6 the estimated regression coefficients are shown for the risk indicating variables and response variables being transformed according to equation \( \{2.20\} \) of the multivariate regression analysis (paragraph 2.3.5).
Table 2.6: Results of regression analysis. Multivariate normal maximum likelihood estimates of regression coefficients.

<table>
<thead>
<tr>
<th>columns:</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \ln (\hat{\beta}_1) )</td>
<td>( \ln (\hat{\beta}_2) )</td>
<td>( \ln (\hat{\beta}_3) )</td>
<td>( \ln (\hat{\beta}_4) )</td>
<td>( \ln (\hat{\beta}_5) )</td>
<td>( \ln (\hat{\beta}_6) )</td>
<td>( \ln (\hat{\beta}_7) )</td>
<td>( \ln (\hat{\beta}_8) )</td>
</tr>
<tr>
<td>Exits</td>
<td>( \ln (\hat{\beta}_{\text{inc}}) )</td>
<td>-11.319</td>
<td>0.801</td>
<td>0.120</td>
<td>0.014</td>
<td>-4.23E-4</td>
<td>-0.066</td>
<td>2.57E-5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(-43.37)</td>
<td>(30.81)</td>
<td>(4.80)</td>
<td>(7.00)</td>
<td>(-0.14)</td>
<td>(-3.00)</td>
<td>(9.38)</td>
</tr>
<tr>
<td></td>
<td>( \ln (\hat{\beta}_{\text{dly}}) )</td>
<td>-13.462</td>
<td>1.027</td>
<td>0.263</td>
<td>0.013</td>
<td>5.83E-3</td>
<td>-0.152</td>
<td>1.94E-6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(-46.26)</td>
<td>(35.41)</td>
<td>(9.39)</td>
<td>(6.50)</td>
<td>(1.94)</td>
<td>(-6.08)</td>
<td>(0.64)</td>
</tr>
<tr>
<td></td>
<td>( \ln (\hat{\beta}_{\text{si}}) )</td>
<td>-11.937</td>
<td>0.573</td>
<td>-0.051*</td>
<td>0.013</td>
<td>7.86E-3*</td>
<td>0.240</td>
<td>9.39E-5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(-35.53)</td>
<td>(17.36)</td>
<td>(-1.59)</td>
<td>(6.50)</td>
<td>(1.97)</td>
<td>(8.57)</td>
<td>(26.60)</td>
</tr>
<tr>
<td></td>
<td>( \ln (\hat{\beta}_{\text{v}}) )</td>
<td>-11.708</td>
<td>0.291</td>
<td>-0.045*</td>
<td>0.016</td>
<td>-2.19E-3*</td>
<td>0.133</td>
<td>4.21E-5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(-41.96)</td>
<td>(10.39)</td>
<td>(-1.67)</td>
<td>(8.00)</td>
<td>(-0.73)</td>
<td>(5.54)</td>
<td>(14.42)</td>
</tr>
<tr>
<td>Intersections</td>
<td>( \ln (\hat{\beta}_{\text{inc}}) )</td>
<td>-13.877</td>
<td>1.023</td>
<td>-0.088</td>
<td>0.023</td>
<td>-0.014</td>
<td>0.012*</td>
<td>5.23E-5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(-46.26)</td>
<td>(33.00)</td>
<td>(-2.58)</td>
<td>(23.60)</td>
<td>(-3.50)</td>
<td>(0.46)</td>
<td>(15.12)</td>
</tr>
<tr>
<td></td>
<td>( \ln (\hat{\beta}_{\text{dly}}) )</td>
<td>-22.619</td>
<td>1.845</td>
<td>0.250</td>
<td>0.031</td>
<td>-0.045</td>
<td>-0.310</td>
<td>8.00E-5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(-28.82)</td>
<td>(12.22)</td>
<td>(-4.24)</td>
<td>(11.00)</td>
<td>(-3.50)</td>
<td>(-6.0)</td>
<td>(16.25)</td>
</tr>
<tr>
<td></td>
<td>( \ln (\hat{\beta}_{\text{si}}) )</td>
<td>-10.404</td>
<td>0.452</td>
<td>-0.174</td>
<td>0.022</td>
<td>-0.004*</td>
<td>0.342</td>
<td>6.76E-5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(-61.63)</td>
<td>(49.86)</td>
<td>(5.95)</td>
<td>(15.50)</td>
<td>(-9.00)</td>
<td>(-6.99)</td>
<td>(18.91)</td>
</tr>
<tr>
<td></td>
<td>( \ln (\hat{\beta}_{\text{v}}) )</td>
<td>-20.412</td>
<td>1.056</td>
<td>0.295</td>
<td>0.005</td>
<td>-0.054</td>
<td>-0.258</td>
<td>1.23E-4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(-61.30)</td>
<td>(31.06)</td>
<td>(7.76)</td>
<td>(2.50)</td>
<td>(-13.50)</td>
<td>(-8.90)</td>
<td>(32.03)</td>
</tr>
<tr>
<td>Tunnels</td>
<td>( \ln (\hat{\beta}_{\text{inc}}) )</td>
<td>-18.811</td>
<td>1.482</td>
<td>0.947</td>
<td>0.013</td>
<td>0.041</td>
<td>-0.316</td>
<td>-3.57E-5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(-28.63)</td>
<td>(26.00)</td>
<td>(11.84)</td>
<td>(3.25)</td>
<td>(5.86)</td>
<td>(-6.20)</td>
<td>(-2.98)</td>
</tr>
<tr>
<td></td>
<td>( \ln (\hat{\beta}_{\text{dly}}) )</td>
<td>-21.314</td>
<td>1.698</td>
<td>1.048</td>
<td>0.017</td>
<td>0.023</td>
<td>-0.309</td>
<td>-1.53E-5*</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(-27.33)</td>
<td>(25.34)</td>
<td>(11.05)</td>
<td>(3.40)</td>
<td>(2.88)</td>
<td>(-5.15)</td>
<td>(-1.08)</td>
</tr>
<tr>
<td></td>
<td>( \ln (\hat{\beta}_{\text{si}}) )</td>
<td>-13.696</td>
<td>1.015</td>
<td>0.590</td>
<td>-0.001*</td>
<td>0.060</td>
<td>-0.697</td>
<td>-8.81E-5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(-16.21)</td>
<td>(13.90)</td>
<td>(5.73)</td>
<td>(-0.20)</td>
<td>(6.67)</td>
<td>(-10.72)</td>
<td>(-5.72)</td>
</tr>
<tr>
<td></td>
<td>( \ln (\hat{\beta}_{\text{v}}) )</td>
<td>-19.655</td>
<td>1.408</td>
<td>-0.184*</td>
<td>0.001*</td>
<td>0.075</td>
<td>-1.938</td>
<td>2.46E-4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(-22.36)</td>
<td>(18.53)</td>
<td>(-1.72)</td>
<td>(0.20)</td>
<td>(7.50)</td>
<td>(-28.30)</td>
<td>(15.28)</td>
</tr>
<tr>
<td>Open Roads</td>
<td>( \ln (\hat{\beta}_{\text{inc}}) )</td>
<td>-7.609</td>
<td>0.445</td>
<td>-0.088</td>
<td>0.011</td>
<td>-0.002*</td>
<td>0.062</td>
<td>1.59E-5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(-59.91)</td>
<td>(34.23)</td>
<td>(-8.00)</td>
<td>(11.00)</td>
<td>(-2.00)</td>
<td>(5.64)</td>
<td>(8.83)</td>
</tr>
<tr>
<td></td>
<td>( \ln (\hat{\beta}_{\text{dly}}) )</td>
<td>-7.073</td>
<td>0.402</td>
<td>-0.166</td>
<td>0.013</td>
<td>-0.002*</td>
<td>0.118</td>
<td>4.37E-6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(-48.12)</td>
<td>(26.80)</td>
<td>(-12.77)</td>
<td>(13.00)</td>
<td>(-2.00)</td>
<td>(9.08)</td>
<td>(2.10)</td>
</tr>
<tr>
<td></td>
<td>( \ln (\hat{\beta}_{\text{si}}) )</td>
<td>-12.214</td>
<td>0.676</td>
<td>0.087</td>
<td>-0.010</td>
<td>-0.005</td>
<td>0.113</td>
<td>5.33E-5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(-64.28)</td>
<td>(35.58)</td>
<td>(5.44)</td>
<td>(-3.50)</td>
<td>(-5.00)</td>
<td>(6.65)</td>
<td>(19.74)</td>
</tr>
<tr>
<td></td>
<td>( \ln (\hat{\beta}_{\text{v}}) )</td>
<td>-8.762</td>
<td>-0.030*</td>
<td>-0.071</td>
<td>0.008</td>
<td>-0.007</td>
<td>0.446</td>
<td>1.04E-5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(-38.10)</td>
<td>(-1.30)</td>
<td>(-3.55)</td>
<td>(4.00)</td>
<td>(-3.50)</td>
<td>(21.24)</td>
<td>(3.18)</td>
</tr>
</tbody>
</table>

Even though some regression coefficients in Table 2.6 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 2.6, 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 2.6 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 2.6, 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 2.7, “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 $\beta$-values mostly smaller than 1. The strongest impact of the AADT on the response variables is observed in tunnels and on intersections with $\beta$-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 $\beta$-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 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 Manering (1998), Anastasopoulos and Manering (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 2.6, 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 2.6, 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 2.6, 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.* (Shankar et al., 1995), Milton and Mannering (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 2.6, 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 2.6, 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 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

**LAN:** On open roads the variable LAN has positive impact on all response variables (Table 2.6, column 6). A higher number of lanes implies more traffic and hence, more lane changing behaviour and overtaking manoeuvres 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 (Noland, 2003) as well as Noland and Oh (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 2.6, 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 (Figure 2.5 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 2.6, 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 2.7. 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.

Table 2.7: Covariance matrices for the error terms, showing the covariances (C), the variances (VAR) and the correlation coefficients ($r$) for the logarithmic response variables.

<table>
<thead>
<tr>
<th>$\Sigma_{1,k}$</th>
<th>Exit Corridors</th>
<th>$\ln(\lambda_{AC}^{*})$</th>
<th>$\ln(\lambda_{LINJ}^{*})$</th>
<th>$\ln(\lambda_{SINJ}^{*})$</th>
<th>$\ln(\lambda_{FAT}^{*})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\ln(\lambda_{AC}^{*})$</td>
<td>VAR=1.215</td>
<td>r=0.855</td>
<td>r=0.581</td>
<td>r=0.218</td>
<td></td>
</tr>
<tr>
<td>$\ln(\lambda_{LINJ}^{*})$</td>
<td>C=1.157</td>
<td>VAR=1.508</td>
<td>r=0.291</td>
<td>r=0.194</td>
<td></td>
</tr>
<tr>
<td>$\ln(\lambda_{SINJ}^{*})$</td>
<td>C=0.909</td>
<td>C=0.507</td>
<td>VAR=2.012</td>
<td>r=0.236</td>
<td></td>
</tr>
<tr>
<td>$\ln(\lambda_{FAT}^{*})$</td>
<td>C=0.283</td>
<td>C=0.280</td>
<td>C=0.394</td>
<td>VAR=1.381</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$\Sigma_{2,k}$</th>
<th>Intersections</th>
<th>$\ln(\lambda_{AC}^{*})$</th>
<th>$\ln(\lambda_{LINJ}^{*})$</th>
<th>$\ln(\lambda_{SINJ}^{*})$</th>
<th>$\ln(\lambda_{FAT}^{*})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\ln(\lambda_{AC}^{*})$</td>
<td>VAR=0.960</td>
<td>r=0.877</td>
<td>r=0.511</td>
<td>r=0.249</td>
<td></td>
</tr>
<tr>
<td>$\ln(\lambda_{LINJ}^{*})$</td>
<td>C=1.030</td>
<td>VAR=1.436</td>
<td>r=0.241</td>
<td>r=0.271</td>
<td></td>
</tr>
<tr>
<td>$\ln(\lambda_{SINJ}^{*})$</td>
<td>C=0.590</td>
<td>C=0.340</td>
<td>VAR=1.390</td>
<td>r=0.195</td>
<td></td>
</tr>
<tr>
<td>$\ln(\lambda_{FAT}^{*})$</td>
<td>C=0.265</td>
<td>C=0.353</td>
<td>C=0.250</td>
<td>VAR=1.183</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$\Sigma_{3,k}$</th>
<th>Tunnels</th>
<th>$\ln(\lambda_{AC}^{*})$</th>
<th>$\ln(\lambda_{LINJ}^{*})$</th>
<th>$\ln(\lambda_{SINJ}^{*})$</th>
<th>$\ln(\lambda_{FAT}^{*})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\ln(\lambda_{AC}^{*})$</td>
<td>VAR=1.290</td>
<td>r=0.800</td>
<td>r=0.593</td>
<td>r=0.192</td>
<td></td>
</tr>
<tr>
<td>$\ln(\lambda_{LINJ}^{*})$</td>
<td>C=1.225</td>
<td>VAR=1.817</td>
<td>r=0.301</td>
<td>r=0.044</td>
<td></td>
</tr>
<tr>
<td>$\ln(\lambda_{SINJ}^{*})$</td>
<td>C=0.983</td>
<td>C=0.592</td>
<td>VAR=2.133</td>
<td>r=0.100</td>
<td></td>
</tr>
<tr>
<td>$\ln(\lambda_{FAT}^{*})$</td>
<td>C=0.332</td>
<td>C=0.091</td>
<td>C=0.223</td>
<td>VAR=2.311</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$\Sigma_{4,k}$</th>
<th>Open Roads</th>
<th>$\ln(\lambda_{AC}^{*})$</th>
<th>$\ln(\lambda_{LINJ}^{*})$</th>
<th>$\ln(\lambda_{SINJ}^{*})$</th>
<th>$\ln(\lambda_{FAT}^{*})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\ln(\lambda_{AC}^{*})$</td>
<td>VAR 1.007</td>
<td>r=0.822</td>
<td>r=0.569</td>
<td>r=0.178</td>
<td></td>
</tr>
<tr>
<td>$\ln(\lambda_{LINJ}^{*})$</td>
<td>C=0.960</td>
<td>VAR 1.355</td>
<td>r=0.283</td>
<td>r=0.106</td>
<td></td>
</tr>
<tr>
<td>$\ln(\lambda_{SINJ}^{*})$</td>
<td>C=0.861</td>
<td>C=0.496</td>
<td>VAR 2.270</td>
<td>r=0.151</td>
<td></td>
</tr>
<tr>
<td>$\ln(\lambda_{FAT}^{*})$</td>
<td>C=0.326</td>
<td>C=0.225</td>
<td>C=0.415</td>
<td>VAR 3.338</td>
<td></td>
</tr>
</tbody>
</table>

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 2.3. When the value of one risk indicating variable was changed all other variables were kept constant to their most probable value (Table 2.8), i.e. the interval with the highest relative occurrence frequency of one risk indicating variable determined using the entire dataset.

Table 2.8: Most probable values of the risk indicating variables according to the relative frequencies in the defined intervals.

<table>
<thead>
<tr>
<th>CHAR</th>
<th>AADT</th>
<th>HGV</th>
<th>BEND</th>
<th>SLP</th>
<th>LAN</th>
<th>SPD</th>
<th>EML</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>[vehicles/day]</td>
<td>[%]</td>
<td>[km/h]</td>
<td>[no/yes]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>open road</td>
<td>20'000</td>
<td>10</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>130</td>
<td>yes</td>
</tr>
</tbody>
</table>
The results of the sensitivity analysis are shown in Figure 2.5. Accident modification factors (AMF) were calculated. The AMF of the $k^{th}$ response variable was assessed as

$$AMF_k = \frac{\tilde{\lambda}_k}{\lambda_k}.$$

Figure 2.5: Sensitivity analysis by means of the AMF when the different risk indicating variables were varied.
The results of the sensitivity analysis show for the different road characteristics in Figure 2.5 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 (Figure 2.5 b), c), 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 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. Figure 2.5 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 (Figure 2.5 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 (Figure 2.5 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 at first glance to be contradicting the results presented in Table 2.6, 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 2.6, column 8, it can be observed in Figure 2.5 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.
2.3.6 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.

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 Figure 2.6.

![Figure 2.6: Structure of the developed BPN](image_url)

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:

\[
q = 48 \cdot 4 \cdot 10 \cdot 6 \cdot 5 \cdot 7 \cdot 4 \cdot 6 \cdot 2 = 19,353,600
\]  

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.

**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 (paragraph 2.2.5). 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 $\hat{\lambda}_k$ being multiplied with the observed exposure $\bar{v}_i$ (equation {2.2}) to calculate the expected number of injury accidents and injured road users according to equation {2.1}.

The observed and predicted numbers of injury accidents and light injuries are plotted in Figure 2.7 separately for the prior and posterior BPN.

![Figure 2.7: Comparison of predicted and observed number of injury accidents and light injuries for the prior and posterior BPN modelling results.](image)

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 Figure 2.7. 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 $r = 26$ 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.

### 2.3.7 Prediction of Expected Number of Events

**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 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 Figure 2.8.

![Figure 2.8: 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.](image)

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 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 difficulty for defining the causal relationships in the model structure. Related investigations and discussions have effectively been done in previous studies (Milton et al., 2008, Milton and Mannering, 1998).

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).

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 to 2010.

The results are shown in Figure 2.9, separate for the injury accidents and the lightly injured road users. Figure 2.9 is sectioned into 4 sub-graphics, each of which contains two coloured lines. The shape and position of the upper line in Figure 2.9 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 to 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.
Figure 2.9: 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.

2.4 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 outweigh 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.

2.5 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.
3 PREDICTION OF ROAD ACCIDENTS: COMPARISON OF TWO BAYESIAN METHODS (PAPER II)

Markus Deublein
Institute of Construction and Infrastructure Management,
Swiss Federal Institute of Technology, ETHZ, Zurich, Switzerland

Matthias Schubert
Matrisk GmbH, Managing Technical Risks,
Zurich, Switzerland

Bryan T. Adey
Institute of Construction and Infrastructure Management,
Swiss Federal Institute of Technology, ETHZ, Zurich, Switzerland

Journal paper, accepted by
Structure and Infrastructure Engineering
ABSTRACT

In this paper two Bayesian methods for the development of accident prediction models are compared: the well acknowledged Empirical Bayes method and a recently developed method based on Bayesian Probabilistic Networks. Brief descriptions of the two methods are provided and their commonalities, differences, advantages and disadvantages are discussed. Both methods can be used to develop models for a multivariate prediction of accident events and can be included into road infrastructure safety management systems, e.g. road safety impact assessments or road safety audits. On a comprehensive dataset taken from the Austrian rural motorway network it is shown that the predictions of both models are in good accordance with the data. It is observed that the Bayesian Probabilistic Network models show a better correlation with the data than the models developed using the Empirical Bayes method; measured through the higher values of the correlation coefficients (approximately 5-10%).

Keywords: Bayesian networks, Empirical Bayes, roads & highways, prediction models, road accidents, statistical analysis

3.1 Introduction

In recent years, road safety is significantly increased due to developments in the automotive industry and due to the implementation of national road safety strategies. One aim of such safety strategies is to reduce the consequences due to traffic injuries by applying target-oriented and efficient countermeasures. The knowledge about the influences of road design and traffic related factors on the occurrence probabilities of accident and injury events as well as the application of road accident prediction models are therefore important aspects of road infrastructure safety management. Numerous studies have been performed to identify the most relevant of these accident contributing factors as so-called risk indicating variables and to describe the relationships between them and the probabilities of occurrence of the road accidents events. Comprehensive overviews of the different methods used for predicting road accident events can be found in Hauer (2009), Lord and Mannering (2010) and Savolainen et al. (2011).

Since road accidents are relatively rare events when compared to the amount of vehicles and people using the roads, considerable epistemic uncertainty is associated with the expected number of accident events predicted based on observed data (Der Kiureghian and Ditlevsen, 2009). The capability of a method to deal with this uncertainty in the development of accident prediction models is important in order to gain accurate model results. In Hauer et al. (2002) it is stated that methods that are based only on the observed numbers of accident events may lead to inaccurate modelling results, either due to a large variance in the observations (over-
dispersion (Cox, 1983, Dean and Lawless, 1989, Hauer, 2001, Karlis and Meligkotsidou, 2005, Gschlöessl and Czado, 2006, Berk and MacDonald, 2008)) or due to a systematic bias in the predictions (regression-to-the-mean (Hauer, 1986, Hauer et al., 1991a, Barnett et al., 2005, Davis, 1976, Brüde and Larsson, 1988)). These weaknesses can be overcome by using methods that make use of a combination of theoretical prediction models and real observations. In this paper, two such methods are compared, the so-called Empirical Bayes method (EB method) and a recently developed Bayesian Probabilistic Network method (BPN method) (Deublein et al., 2013).

For sake of unambiguity in the terminologies used, in this paper it is distinguished between statistical approaches, methods, models and tools. Statistical approaches are fundamental principles for the mathematical and theoretical formulation of probability calculus (e.g. Empirical vs. Full Bayes). Statistical methods are applying a selected statistical approach using one or more statistical tools (e.g. Bayesian Probabilistic Networks) in order to develop statistical models, e.g. for the prediction of accident events.

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 Hauer (1986, 1992), Hauer et al. (1991b, 2002), Persaud and Dzbik (1993), Carlin and Louis (1997), Persaud et al. (1999), Tunaru (2002), Cheng and Washington (2005) and Elvik (2008). Guidelines for accident risk analysis in the American Highway Safety Manual (AASHTO, 2010) are based on the EB method and it is also recommended by the European Parliament to be included into concepts for road infrastructure safety management (European Parliament, 2008).

The EB method has already been compared to methods based on the Full Bayes approaches e.g. by Schlüter et al. (1997), Heydecker and Wu (2001), MacNab (2003), (Miaou and Lord, 2003), Ying (2004), Carriquiry and Pawlovich (2005), Miaou and Song (2005), Qin et al. (2005), Song et al. (2006a), Maes and Dann (2007), Persaud et al. (2010), Park et al. (2010) and Huang and Abdel-Aty (2010). The results of these investigations commonly indicate that the Empirical and Full Bayes methods result in similar predictions of the expected numbers of accident events. 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 are different between the approaches and indicate that estimates based on methods which use the Full Bayesian approach are more precise. According to Carriquiry and Pawlovich (2005) the use of Empirical Bayes approaches may result in unrealistic low standard errors.

BPNs were developed in the late 1970s and emerged as a general instrument for representing uncertain knowledge. In the mid-1980s, different algorithms were developed for more efficient computation of probabilities (Pearl, 1988, Pearl, 1986), and due to increasing computing capacities, Bayesian inference and updating algorithms have gradually become
more relevant in the field of accident risk assessment. Since then, the application of BPNs has increased enormously; Aguilera et al. (2011) observed an exponential increase. Today, BPNs are applied the most in the scientific areas of computer sciences, mathematics, health sciences as well as environmental and civil engineering. However, the application of BPNs for the analysis and prediction of accidents and accident related injury severities is still rather scarce.

Recent research on modelling the expected number of accident/incident related events by means of BPNs includes the work done by Davis and Pei (2003), Marsh and Bearfield (2004), Ozbay and Noyan (2006) and Simoncic (2004), Maes and Dann (Maes and Dann, 2007), de Oña et al. (2011, 2013), Karwa et al. (2011), Schubert et al. (2011), Hossain and Muromachi (2012) and Deublein et al. (2013). None of these studies, however, compared the predictions of the BPN models with the predictions of models developed using other methods.

Similarly, none of the recently conducted reviews of methods for the development of accident prediction models has investigated the applicability and performance of BPNs regarding the prediction of accident events, e.g. by Sorensen and Elvik (2007), Elvik (2007), Eenink et al. (2008), Lord and Mannering (2010) and Savolainen et al. (2011). The work presented in this paper is intended to fill this gap, i.e. to compare the performance of multivariate prediction models developed using BPNs with those developed using the EB method.

As indicated by their names, both methods are based on Bayesian probability and inference theory. For a consistent comparison of the methods and the theories beyond each of these methods, the paper is structured in the following way.

In section 3.2, the differentiations between the Bayesian approaches and the EB method and BPN methods are explained.

In section 3.3, the major steps for the development and learning of accident prediction models based on both methods are briefly outlined. Their commonalities and differences are emphasized. The hierarchical multivariate Poisson-regression analysis is explained in this section as it is used in the EB method and in the BPN method for the assessment of prior probabilities and causalities.

In section 3.4, prediction models are developed based on the two methods to predict different accident events on road sections in the Austrian rural motorway network. For both methods the same dataset is used.

In section 3.5, the predictive performance of both methods is evaluated by comparing the expected numbers of accident events with the observed ones. For this comparison a test dataset is used which has not been used before for the model development.

In section 3.6, the results of the model comparison and the main advantages and disadvantages of the methods in regard to their technical performance and practical implementation are discussed.
Section 3.7 includes a summary of the investigation and the conclusions about the main findings.

3.2 Bayesian inference

3.2.1 General

Probability theory and statistics form the basis for the assessment of probabilities of the occurrence of uncertain events and thus constitute a cornerstone in the accident analysis on road segments. Probability can basically be interpreted on three different ways: frequentistic, classical and Bayesian (Benjamin and Cornell, 1970, Faber, 2012). In the context of this paper, the assessment of probabilities for the occurrence of accident events is based on the Bayesian interpretation. For general concepts of Bayes theory and Bayesian inference calculations the reader is referred to Benjamin and Cornell (1970), Pearl (1988), Congdon (2006) and Ang and Tang (2007).

Bayesian inference is defined by the Bayes theorem which states that conditional on the observed values of the data \( y \) the posterior density of the vector of unknown parameters \( \theta \) can be assessed. In some statistical applications, it is appropriate to model the investigated problem hierarchically. That means that the observable outcomes are modelled conditionally on parameters, which themselves have a given probabilistic distribution defined by their own parameters, known as hyperparameters.

The posterior density of the vector of unknown parameters is given as

\[
p(\theta|y, \eta) = \frac{p(y, \eta|\theta)}{p(y|\eta)} \tag{3.1}
\]

where \( \eta \) is the vector of hyperparameters defining the distributions of the unknown parameters \( \theta \). The posterior predictive distribution of the unknown but observable data conditional on observed values of \( y \) is then given as

\[
p(\theta|y) = \int p(\theta|y, \eta) \cdot p(\eta|y)d\eta. \tag{3.2}
\]
3.2.2 Difference between Empirical and Full Bayes

In accident risk analysis, it is commonly distinguished between Empirical and Full Bayes approaches; both are used to combine prior information with current observations. According to Carlin and Louis (2000) the term Empirical Bayes is chosen because observed data is used to estimate the hyperparameters $\eta$ by means of the maximum likelihood method, meaning that the posterior probability distribution functions are maximized instead of integrated, which is a considerable simplification in terms of computing. Statistical methods based on the Empirical Bayes approach can be viewed as approximations of methods based on the Full Bayes approach. Uncertainties related to the estimation of the hyperparameters are not taken into account in the Empirical Bayes approach.

In the Empirical Bayes approach, the prior information comes from road segments similar to those under evaluation to estimate the parameters of the model based on historical data, which is not the case in Full Bayes methods. The point estimates of the parameters are then combined with the site specific observed parameters to obtain improved estimates of the long-term expected numbers of accident events (Persaud et al., 2010). In the Full Bayes approach, instead of the prior point estimates of the parameters, prior probability distribution functions are defined to represent the hyperparameters. This is again combined with the observed site-specific accident data, but results in a more accurate estimate of the variance of the expected numbers of accident events.

3.2.3 Use of the Empirical Bayes approach to develop prediction models

In this paper, two methods that can be used to implement Bayes inference for model development are analysed: 1) the EB method and 2) the BPN method. The EB method can only be used to implement an Empirical Bayes approach, whereas the BPN method can be used to implement either the Empirical or the Full Bayes approach or both. Since the target of the current investigation is to develop purely data-based accident models (extracting the predictive potential of historical data) and in order to better facilitate the model comparisons, for the current investigations both methods are applied using the Empirical Bayes approach. That means that for both methods the prior distributions of the random variables and the regression coefficients are established and estimated based on available historical accident data of similar road segments.
3.3 Methods

3.3.1 Definitions and data pre-processing

The EB method and the BPN method are both based on defined sets of risk indicating variables (model input) and model response variables (model output). The risk indicating variables are observable road and traffic characteristics that are considered to influence the conditional occurrence probability of the model response variables. The model response variables are the occurrence rates $\lambda_{ki}$ of the $k=(1,..,u)$ different, simultaneously investigated accident events on the $i=(1,...,n)$ different investigated road segments. The sets of risk indicating variables and model response variables are defined for model development, from case to case, taking into consideration the problem to be investigated and the availability of data.

With both methods the investigated road network is sub-divided into homogeneous segments, for each of which it can be assumed that the values of all risk indicating variables to be included in the accident prediction models are constant. Therefore, also the occurrence probabilities of the model response variables over the length of the homogeneous segment can be considered to be constant for each homogeneous segment. Generic accident risk models are first developed based on data for the entire network and are only made specific when the values of the risk indicating variables for the homogeneous segments are used as model input.

In the method comparison that is made in this paper, the values of the model response variables are not used directly for the regression analysis. Instead, a gamma-updating of the model response variables is performed as described in Deublein et al. (2013) with the assumption that the accident counts can be represented by a negative-binomial distribution. This is done (a) to deal with accident counts that are characterised by over-dispersion, (b) to dilute the effects of regression-to-the-mean and (c) to avoid the preponderance of zero values. The latter of these arguments is something that better reflects the fact that accident frequencies are larger than zero, even if there are no observations on one homogeneous segment over the time of observation, which is often triggered by short observation periods or short segment lengths.

The core element of the gamma-updating process is the assumption that accident counts are negative binomial distributed, which is a mixture of a Poisson distribution and the natural conjugate gamma distribution. The first with parameter $\mu_{ki}$ describing the probability of having a defined number of accident events on one particular homogeneous segment over a defined period of time and the latter describing the distribution of the parameter $\mu_{ki}$ by means

---

$^6$ It is distinguished between occurrence frequencies and occurrence rates. The first represent the number of accident events per road segment and year, the latter represent the number of accident events per million vehicle kilometer and year.
of the hyperparameters \((\alpha, \beta)\). \(\mu_{ki}\) in this case is the mean frequency of accident events, given by:

\[
\mu_{ki} = \left( \lambda_{ki}^* \right)_{\alpha_{ki}^*, \beta_{ki}^*} \cdot \nu_i
\]

where:

- \(\lambda_{ki}^*\) is the gamma-updated accident rates (accidents per million vehicle kilometres per year).
- \(\nu_i\) is the exposure (number of vehicles per kilometre and year) of the homogenous segments.
- \(k\) indicates the type of accident event (e.g. light injury, fatalities, etc.).
- \(i\) indicates the homogeneous segment.

The gamma updating of the accident rates is done by updating the prior parameters of the gamma distribution according to Gelman et al. (2004) as:

\[
\alpha_{ki}^* = \alpha_{ki}' + \tilde{y}_{ki} \quad \text{and} \quad \beta_{ki}^* = \beta_{ki}' + \tilde{v}_i
\]

where

- \(\alpha_{ki}'\) is the posterior shape (dispersion) parameter of the gamma distribution and
- \(\beta_{ki}'\) is the posterior inverse scale parameter of the gamma distribution.
- \(\tilde{y}_{ki}\) and \(\tilde{v}_i\) are the observed accident counts and observed values of exposure, respectively.
- \(\alpha_{ki}'\) and \(\beta_{ki}'\) are the prior parameters of the gamma distribution, assessed as

\[
\alpha_{ki}' = \hat{\lambda}_k \cdot \beta_i' \quad \text{and} \quad \beta_{ki}' = v_i \cdot \omega_i = v_i \cdot \frac{\nu_i}{l_i}
\]

where
\( \omega_i \) is the weight calculated as the fraction of the weighting factor \( \psi \) and the individual homogeneous segment length \( l_i \),

\( \psi \) is a weighting factor attributed to information about the prior gamma parameter \( \beta_i' \). It is used to take into account the time period of observations based on which the prior information has been gathered, experts experience and appraisal of the quality of the prior information and

\( \hat{\lambda}_x \) are the averaged background accident rates of the model response variables, determined based on analysis of available historical data.

### 3.3.2 Regression Analysis

Regression analysis is performed to explore the systematic components in the observed accident events based on additional observations of the risk indicating variables. The EB method and the BPN method are both using hierarchical multivariate Poisson-lognormal regression analysis (subsequently only referred to as regression analysis) to establish prior accident prediction models, i.e. to describe the linear causal relationships between the risk indicating and the model response variables. These causal relationships are often referred to as safety performance functions but can also be used directly as accident prediction models (subsequently referred to as regression models). The multivariate Poisson-lognormal regression model, also used e.g. in Park and Lord (2007), Ma et al. (2008) and El-Basyouny and Sayed (2009b) is different from the more frequently applied multivariate Poisson-gamma regression model (Kim et al., 2002, Lord et al., 2005, Lord and Miranda-Moreno, 2008, Lord and Park, 2008, Miaou, 1994, Abdel-Aty and Radwan, 2000, Zhou et al., 2012) mainly due to different non-negative prior distributions for the error term of the regression equation. For the Poisson-lognormal model, the error term is assumed to be lognormal distributed, for the Poisson-gamma model the error term is assumed to be gamma distributed. Good arguments can be found for using either of the two model assumptions. The multivariate Poisson-gamma model is considered as the model that is easier to implement and hence, it is applied more frequently, however, the multivariate Poisson-lognormal model may fit the data better (Winkelmann, 2008). Additionally, the latter is considered to be more general, more flexible to handle over-dispersion, less restrictive for the assessment of the correlation-structure of the model response variables (the Poisson-gamma model requires the co-variances to be non-negative) and more consistent with probability theory (according to the central limit theorem, the multiplicative structure of the multivariate regression model indicates the response variables and the corresponding error terms to be approximately lognormal distributed).

The explanatory variables in the regression analysis are the risk indicating variables and the dependent variables are the multidimensional model response variables representing the expected accident event rates. The structural component of the proposed regression model is:
\[
\ln \left( E \left[ \Lambda | X \right] \right) = BX + \Xi \ \overset{\Delta}{=} \ E \left[ \Lambda | X \right] = \exp \left( BX + \Xi \right) \quad \{3.6\}
\]

where

- \( X \) is the design-matrix of explanatory risk indicating variables (road design parameters and traffic characteristics),
- \( \Lambda \) is the matrix of the model response variables (accident events),
- \( B \) the matrix of regression coefficients for different model response variables and types of roads \( (\beta_0, \beta_1, \ldots) \) and
- \( \Xi \) the matrix of the error terms for the different model response variables \( (\varepsilon_k) \).

The regression coefficients indicate the change in the expected number of accident events on the \( i^{th} \) homogeneous segment when the risk indicating variables are changing. The regression equation \{3.6\} represents a two-level hierarchical modelling approach in which the regression coefficients (parameters) are given a probability distribution described by hyperparameters estimated based on available data. From this point, methods that use the Empirical Bayes approach continue only with the point estimates of the hyperparameters. Methods that use the Full Bayes approach continue with the complete probability distribution of the hyperparameters.

### 3.3.3 Empirical Bayes Method

The EB method is used to combine theoretical accident prediction models assessed on historical data of similar homogeneous segments with observed accident data on the specific road segments under investigation. Since the EB method has been described many times before (references are provided in section 3.1) only the major steps are highlighted in the subsequent paragraphs. The steps are i.e. the determination of (Step 1) the safety performance functions, (Step 2) the over-dispersion parameters, (Step 3) the relative weights and (Step 4) the expected number of accident events.

#### Step 1: Determination of the safety performance functions

The first step of the EB method is to determine the safety performance functions which can equally be considered as accident prediction models. The determination of the safety performance functions is based on the regression analysis (as described in section 3.3.2) and is a mathematical model that describes the relationship between accident risk predicting
variables and the model response variables. According to Hauer (2001) the occurrence of accident events is best predicted by using multivariate statistical models, by means of which the safety performance functions can be used to predict the number of expected accident events on different road sites. Each type of road section (i.e. tunnels, intersections, open roads, etc.) for which the relationship between risk indicating variables and model response variables is assumed to be different from the others, have to have their own safety performance functions.

**Step 2: Determination of the over-dispersion parameters**

The second step of the EB method is to determine the over-dispersion parameters for every of the model response variables. Data is considered to be over-dispersed when the sample variance of the model response variables exceeds the sample mean values. For estimating the over-dispersion parameters of the model response variables, a negative-binomial distribution is fitted to the observed numbers of accident events on the homogeneous segments (Hauer, 2001). The over-dispersion parameter is equivalent to the reciprocal value of the negative binomial shape parameter \( \alpha \). Using the EB method there are two types of over-dispersion parameters, the individual over-dispersion parameters \( \phi_i \) of the particular homogeneous segments, and the overall over-dispersion parameter \( \phi_k \) for the entire investigated network. Since the homogeneous segments vary in length and in the values of their risk indicating variables, an individual over-dispersion parameter is adjusted to represent that variation. According to Hauer (2001), the individual over-dispersion parameter is related to the overall over-dispersion parameter as:

\[
\phi_i = \phi_k \cdot \tau^\tau.
\]

where \( \tau \) is the weighting exponent and has a value between zero and one. Its value is selected to ensure that homogeneous segments are weighted as a function of their number of observed accident events. The weighting exponent takes into account the differences between the groups of homogeneous segments defined by different risk indicating variable values. Only if each road segment was completely different from the other segments, the weighting exponent would be \( \tau = 0 \) and the two over-dispersion parameters (individual and overall) would be the same. The other way round, only if all road segments would be completely similar then \( \tau = 1 \) and the overall over-dispersion parameter would be adjusted only by the segment length. The latter can be assumed to provide the most conservative estimates of expected accident events.
Step 3: Determination of the relative weights

The third step of the EB method is to determine the relative weight \( w_{ki} \) which is applied to every homogeneous segment for the different model response variables in order to adjust for varying degrees of the individual over-dispersion and given as

\[
\frac{1}{1 + \left( \frac{\hat{\lambda}_{ki}}{\alpha_{ki}} \right)} \quad \{3.8\}
\]

where \( w_{ki} \) is the relative weight given to the number of expected accident events (predicted based on regression analysis) taking into consideration the number of observed accidents and the individual over-dispersion parameters through \( \alpha_{ki} = 1/\phi_{ki} \).

Step 4: Determination of the expected number of accident events

The fourth step is to estimate the expected number of accident events. Based on the estimates of the safety performance functions \( \hat{\lambda}_{ki} \), the over-dispersion parameters \( \phi_{ki} \) and the observed counts of accident events \( \hat{y}_{ki} \) on one specific homogeneous segment \( i \) the expected numbers \( E[\cdot] \) of the accident events \( \hat{\mu}_{ki} \) are calculated as

\[
E\left[ \hat{\mu}_{ki} \mid \hat{y}_{ki}, \hat{\lambda}_{ki} \right] = w_{ki} \cdot \hat{\lambda}_{ki} + (1 - w_{ki}) \cdot \hat{y}_{ki} \quad \{3.9\}
\]

with \( w_{ki} \) being the relative weight as described in step 3. More information about the EB method can be found e.g. in Elvik (2008), Hauer (1992, 2001), Hauer et al. (2002) and Persaud et al. (1999, 2010).

3.3.4 Bayesian Probabilistic Networks Method

A Bayesian Probabilistic Network or simply Bayesian network, is a form of artificial intelligence which could be described as a probability-based expert system incorporating uncertainty by means of probability theory and conditional dependencies (McCabe et al., 1998). In the engineering sector BPNs are used due to their flexibility and efficiency in regard to systems representation (Faber et al., 2002) and in the presented work BPNs are used in accordance with the definitions in Faber (2012) as normative expert systems, meaning that the domain of uncertainty is modelled instead of the expert himself, they are based on classical probability calculus and decision theory instead of using variations of probability calculations.

The major steps for the development of a BPN based model are (Step 1) the pre-processing of the data to be used for BPN construction and learning, (Step 2) model development and learning and (Step 3) the validation of the model.

Step 1: Pre-processing of data

The first step of the BPN method is to pre-process the data which shall be used for BPN construction and learning. That means that (beyond the data pre-processing which has already been done in the precedent process of the regression analysis in section 3.3.1) the data is prepared that it can directly be used for BPN development and learning. For efficiency reasons it often becomes necessary to discretize continuous state variables into discrete state variables. In this case, data pre-processing for BPNs comprises mainly the discretization of the random variables and the allocation of the available observed data into intervals which are defined in accordance with the states of the discretised variables of each node. The discretisation of a continuous random variable is an approximation to the actual data implying a loss of information, which becomes smaller the narrower the discretisation intervals are. Hence, the intervals of the BPN described in this paper are the result of a trade-off between a maximized number of intervals per discrete state variable, given the constraints of commonly applied computer equipment and reasonable computing time. According to Aguilera et al. (2011) discretising continuous state random variables is currently the most common approach for BPN data pre-processing.

Step 2: Model development and learning

The second step of the BPN method is to construct the directed acyclic graph and to learn the parameters of the BPN. Given a sufficiently large dataset the BPN can be learned by both, structure learning and parameter learning, the former meaning the definition of the conditional dependencies and independencies through the determination of an optimized directed acyclic graph structure, and the latter meaning the updating of the conditional probability tables based on additionally available data. However, in most engineering problems the structure is modelled on the basis of causal relations. Structural learning is only based on linear correlations in the data. Since dependencies might not always be represented appropriately through linear correlations and correlations might not always represent causality, pure data based structural learning is not recommended in engineering problems for which the
causalities are in principle known. Thus, in this paper the structure of the BPN is determined by causal relations.

Parameter learning of the BPN is done by constructing so called contingency tables which provide input information for the BPN based on observations of the risk indicating and model response variables. Different algorithms can be used for the parameter learning of BPNs. Most of the classical statistical techniques are not able to deal with missing data, meaning that if only one value was missing for a homogeneous road segment of the contingency table, the entire segment would have to be excluded. Such a loss of information can be overcome by using one of the following algorithms which are able to deal with missing data: the Expectation-Maximization algorithm (EM algorithm) (Dempster et al., 1977, Lauritzen, 1995, Cox, 1983, Carlin and Louis, 2000, Fahrmeir and Osuna, 2003, Karlis, 2003), the Gibbs Sampler (Gilks and Wild, 1992) or the Metropolis-Hastings algorithm (Hastings, 1970). Due to its simplicity and efficiency the EM algorithm is used in the present investigations. It exploits the partial information of the incomplete cases, finding the maximum likelihood parameter estimates basically based on alternates between an expectation step and a maximization step until a certain convergence criteria is fulfilled (Carlin and Louis, 2000). In the expectation step the dataset is completed by calculating expectations for the missing values based on current parameter estimates. In the maximization step the (now) completed dataset is used to find new maximum likelihood estimates of the parameters. In that way, the internal causal interrelationships and dependencies in the BPN are iteratively updated based on the additional data. Experience factors are applied to weight the information of the prior BPN. The values of the experience factors are to be determined for every investigated problem individually, according to the experts’ experience on how much weight should be given to the available prior information and the informative value of the available data. The results of the parameter learning are updated conditional probability tables for all variables of the BPN.

As during the parameter learning process only the cells of the conditional probability table are updated, for which new data are available, the remaining cells persist unchanged. For the updated cells, purely empirical regression model based probabilities and linear relationships are replaced by observation based posterior probabilities and non-linear relationships. When new data is collected regularly it can be directly used to update the conditional probability tables and, therefore, the model performance can continuously be improved. The contingency tables can also be used to evaluate the future demand of specific data.

Step 3: Model validation and testing

The third step of the BPN method is to validate and test the developed and learned BPN. The representativeness and the accuracy of the inference results are checked and the influences of the particular input variables on the output variables are determined, e.g. by means of
sensitivity analyses being the most basic tool for investigating the epistemic model uncertainty. When conducting sensitivity analysis, reasonable modifications can be made to the assumption of any node of the BPN, regardless of whether the variables are input or output variables. The posterior conditional probabilities of the nodes of interest are then recomputed and compared to the conditional distributions before the modifications.

3.4 Model development

3.4.1 Data

In order to demonstrate the differences between the two investigated methods, both methods were used to develop accident prediction models for the Austrian rural motorway network comprising only roads with multiple lanes being physically separated based on the driving direction. Nearly 40,000 geographical coordinates were accessible to represent the total length of 3,642km (considering that data was provided for the two driving directions separately). The same data set was used in (Deublein et al., 2013). The investigated motorway network was sub-divided into $n=6,932$ homogeneous segments. Since the accident prediction models are supposed to be learned from an available set of cases, but are at the same time intended to perform properly for a wider range of data (e.g. on a different motorway network), the dataset was split into two different subsets, the development dataset and the testing dataset. The former was exclusively used for model development and learning. The latter was exclusively used for comparing the performance of the models. The division was made randomly, selecting about 75% of the entire data for the development dataset and 25% of the entire data for the testing dataset. For the period of observation between the years 2004-2010, 12,892 injury accidents were considered being distinguished into light injuries, severe injuries and fatalities as defined in Table 3.

All accident events together with the corresponding numbers of injured road users have been recorded by the Austrian police authorities and were allocable to the motorway network via GPS coordinates. Each accident record is thereby considered as a case, containing observed data of the model response variables and observed data of the risk indicating variables. Variables of both types are defined in section 3.3.1. Information about drivers being under the influence of alcohol was available and was used to exclude accidents which were verifiably caused by drunk drivers in order to avoid biased modelling results. The datasets additionally did not comprise accidents which took place on ramps, within work zones or intervention areas, or at tollbooths, since information about such road sections were not available. The datasets were purged of double entries and cases were excluded from the dataset, where either no information about the exact location of the accident was available or when the accident
was verifiably part of a multiple collision with more than ten vehicles involved. For such multiple collisions it was assumed that the origin for the accident causation has been beyond the set of road design and traffic related factors and variables taken into account for the present model development.

The cumulative length of the investigated homogeneous segments, together with the observed numbers of the different accident events for the entire observation period, are given separately in Table 3.1 for the entire dataset and the two sub-sets.

<table>
<thead>
<tr>
<th>length</th>
<th>IAC</th>
<th>LIN</th>
<th>SIN</th>
<th>FAT</th>
</tr>
</thead>
<tbody>
<tr>
<td>[km]</td>
<td>[-]</td>
<td>[-]</td>
<td>[-]</td>
<td>[-]</td>
</tr>
<tr>
<td>entire dataset</td>
<td>3’642</td>
<td>12’892</td>
<td>14’482</td>
<td>5’861</td>
</tr>
<tr>
<td>development dataset</td>
<td>2’952</td>
<td>10’282</td>
<td>11’460</td>
<td>4’677</td>
</tr>
<tr>
<td>test dataset</td>
<td>690</td>
<td>2’610</td>
<td>3’022</td>
<td>1’184</td>
</tr>
</tbody>
</table>

### 3.4.2 Definitions and Data Pre-Processing

The input variables used in this comparison were referred to as a set of observable road- and traffic specific random risk indicating variables. The investigated risk indicating variables’ acronyms, states and units are given in Table 3.2 and described in more detail in Deublein *et al.* (2013).

<table>
<thead>
<tr>
<th>RIV</th>
<th>Definition</th>
<th>States</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHAR</td>
<td>Different types of road sections, namely 1) exit corridors, 2) intersections, 3) tunnels and 4) normal/open roads. Exit corridors and intersections are defined over a range of one kilometre including a 500 m section before and after their centroid. It is assumed that there are different relationships between the risk indicating variables and model response variables on the different types of road sections.</td>
<td>(1,2,3,4)</td>
<td>[-]</td>
</tr>
<tr>
<td>AADT</td>
<td>Annual average daily traffic per driving direction. It is assumed that the traffic volume remained the same over the time of observation (2004-2010).</td>
<td>(1,2,...,10)-10⁴</td>
<td>[veh/d]</td>
</tr>
<tr>
<td>HGV</td>
<td>Percentage of heavy good vehicles with respect to the AADT.</td>
<td>(5,10,...,30)</td>
<td>[%]</td>
</tr>
<tr>
<td>BEND</td>
<td>Horizontal curvature; integer variable having values between zero (straight road) and ten (very high curvature), determined as the fraction of the sum of the lengths of ten subsequent 50m road sections divided by the length of the straight distance between the starting point of the first and end point of the tenth section. Data of a geographical information system (GIS) was used for assessment of the bend factor.</td>
<td>(0,2,...,10)</td>
<td>[-]</td>
</tr>
</tbody>
</table>
SLP  Percentage of the upwards or downwards gradient (slope).  
\[ (-6, -4, ..., 0, ..., 6) \]  
[%]

LAN  Number of driving lanes per direction.  
\[ (1, 2, 3, 4) \]  
[-]

SPD  Signalized speed limit.  
\[ (80, 90, ..., 130) \]  
[km/h]

EML  Existence of emergency lanes, binary variable  
\[ (0 = no, 1 = yes) \]  
[-]

Variables which are merely related to the characteristics of the individual road users (i.e. gender, age, etc.) are presently not taken into account. Four different model response variables were modelled simultaneously – injury accidents, light injuries, severe injuries and fatalities – taking into consideration the definitions of injury levels in accordance with §84 of the Austrian Penal Code (Bundesrepublik Österreich, 2012) (Table 3.3). The model response variables are the occurrence rates of the accident events \( \lambda_{ki} \) being the observed numbers of accident events \( \tilde{y}_{ki} \) on the \( i^{th} \) homogeneous segment divided by its exposure \( v_i \) with the unit accident events per million vehicle kilometres and year (\( mvk \)).

<table>
<thead>
<tr>
<th>MRV</th>
<th>Definition</th>
<th>Intervals</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>IAC</td>
<td>Occurrence rates of all injury accidents, i.e. accident events where at least one vehicle is involved and at least one occupant becomes at least lightly injured.</td>
<td>[ 0.001 \text{ for } 0 \leq \lambda_{IAC} &lt; 0.01 ] [ 0.01 \text{ for } 0.01 \leq \lambda_{IAC} &lt; 0.2 ] [ 0.1 \text{ for } 0.2 \leq \lambda_{IAC} &lt; 2 ]</td>
<td>IAC/( mvk )</td>
</tr>
<tr>
<td>LIN</td>
<td>Occurrence rates of light injured road users. A road user is considered to be lightly injured if the damage to his well-being lasts less than 25 days following the accident event.</td>
<td>[ 0.001 \text{ for } 0 \leq \lambda_{LIN} &lt; 0.01 ] [ 0.01 \text{ for } 0.01 \leq \lambda_{LIN} &lt; 0.2 ]</td>
<td>LIN/( mvk )</td>
</tr>
<tr>
<td>SIN</td>
<td>Occurrence rates of severely injured road users. A road user is considered to be severely injured if the damage to his well-being lasts more than 24 days following the accident.</td>
<td>[ 0.001 \text{ for } 0 \leq \lambda_{SIN} &lt; 0.01 ] [ 0.01 \text{ for } 0.01 \leq \lambda_{SIN} &lt; 0.2 ]</td>
<td>SIN/( mvk )</td>
</tr>
<tr>
<td>FAT</td>
<td>Occurrence rates of fatally injured road users. A road user is fatally injured when he has died within 30 days following the accident event as a consequence of accident induced injuries.</td>
<td>[ 0.0001 \text{ for } 0 \leq \lambda_{FAT} &lt; 0.001 ] [ 0.001 \text{ for } 0.001 \leq \lambda_{FAT} &lt; 0.02 ]</td>
<td>FAT/( mvk )</td>
</tr>
</tbody>
</table>

The dependent variables used for the regression analysis (performed for prior models of both compared methods) were the values of the gamma-updated model response variables as described in section 3.3.1. The parameters of the gamma distribution were quantified and updated for each homogeneous segment based on the background accident rates \( \hat{\lambda}_{k} \) and the weighting factor \( \psi \) as given in Table 3.4. The values of the background accident rates and \( \psi \) were derived based on an optimization algorithm objecting to minimize the difference.
between the sum of the regression model based accident predictions and the sum of the observed average accident events in the development dataset (Deublein et al., 2013).

Table 3.4: Weighting factor and background rates for model development

<table>
<thead>
<tr>
<th></th>
<th>( \psi )</th>
<th>( \hat{\lambda}_{IAC} )</th>
<th>( \hat{\lambda}_{LIN} )</th>
<th>( \hat{\lambda}_{SIN} )</th>
<th>( \hat{\lambda}_{FAT} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>[-]</td>
<td>([IAC/mvk])</td>
<td>([LIN/mvk])</td>
<td>([SIN/mvk])</td>
<td>([FAT/mvk])</td>
<td></td>
</tr>
<tr>
<td>0.3</td>
<td>0.08764</td>
<td>0.09910</td>
<td>0.03705</td>
<td>0.00315</td>
<td></td>
</tr>
</tbody>
</table>

As an example, in accordance with equations \{3.4\} and \{3.5\}, and under the assumption that for a specific homogeneous road segment with length \( l = 0.25 \) km, exposure \( v = 1.2962 \) mvk and observation time \( t = 7 \) years, the updating of the gamma distribution parameters to represent the Poisson parameter \( \lambda \) is determined as

\[
\beta^* = \beta' / 0.3 + 1.2962 / 7 = 5.3699 \quad \{3.10\}
\]

with

\[
\beta' = \frac{0.3}{0.25} \cdot 1.2962 = 1.5554. \quad \{3.11\}
\]

While the values of the gamma parameter \( \beta \) (related to exposure) remain the same for the different model response variables, the values of the gamma shape parameter \( \alpha \) (related to accident observations) change as shown in Table 3.5. Numbers of observed accident events on a fictive homogeneous segment (\( \tilde{y}_k \)) were estimated for the different model response variables.

Table 3.5: Values of observed number of accident events, prior and posterior values of Gamma shape parameter

<table>
<thead>
<tr>
<th></th>
<th>( \tilde{y}_k )</th>
<th>( \alpha' )</th>
<th>( \alpha^* )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( IAC )</td>
<td>1</td>
<td>0.1363</td>
<td>0.2792</td>
</tr>
<tr>
<td>( LIN )</td>
<td>2</td>
<td>0.1541</td>
<td>0.4399</td>
</tr>
<tr>
<td>( SIN )</td>
<td>0</td>
<td>0.0577</td>
<td>0.0577</td>
</tr>
<tr>
<td>( FAT )</td>
<td>0</td>
<td>0.0049</td>
<td>0.0049</td>
</tr>
</tbody>
</table>
Since no severe injuries and no fatalities were recorded, the shape parameter remains unchanged during the updating procedure. The corresponding graphical illustrations are shown in Figure 3.1, separately for the particular model response variables. The solid lines are the prior gamma probability density functions (pdf) and the dashed lines are the posterior gamma pdfs of the Poisson distribution parameter $\lambda_k$.

![Figure 3.1: Example for gamma updating for the probability density functions (pdf) of the model response variables; solid lines are the prior pdfs and dashed lines the updated posterior pdfs.](image)

A shift of the pdf to higher values can be observed in Figure 3.1 (a) and (b) for the model response variables IAC and LIN due to $\tilde{y}_{IAC}$ and $\tilde{y}_{LIN}$ values inducing mean observed accident rates $\tilde{\lambda}_{IAC}$ and $\tilde{\lambda}_{LIN}$ larger than their background rates, $\hat{\lambda}_{IAC}$ and $\hat{\lambda}_{LIN}$ (Table 3.4). On the contrary, zero observations as observed for the model response variables SIN and FAT result in pdfs shifted slightly to lower values as illustrated in Figure 3.1 (c) and (d). However, these shifts are very small since the background rates for SIN and FAT are also low. As given in Table 3.5, the updated values of the gamma shape parameter $\alpha_k^*$ remain unchanged when neither severe injuries nor fatalities are observed and the small change in the functions of Figure 3.1 are induced only by the updated values of the gamma scale parameter $\beta^*$. 

-91-
The selection of the weighting factor $\psi$ for the updating of the background rates is found to have no significant influence on the predictive performance of the regression models. This can be seen in Figure 3.2, where the differences in the coefficient of correlation (r-values, representing the quality of the relationship between the regression model predictions and the real observations) are illustrated over increasing values of the weighting factor $\psi$. The difference is assessed by subtracting the r-value given $\psi = 0.1$ from the r-values given increasing values of $\psi = (0.2, 0.3, ..., 1)$. The overall maximum difference in the r-values is $\Delta r_{LIN} = 0.034$ considered to be an insignificant difference and too small to award any influence to $\psi$ on the coefficients of correlations of the regression model results. However, a systematically increasing pattern can be observed for the r-values of the model response variables IAC, LIN and SIN, but not for the model response variable FAT. In contrast to the remaining model response variables, the r-values of FAT are decreasing with increasing values of $\psi$. The decrease is, however, of such negligible magnitude that it is not further discussed and considered to be a random effect due to statistical uncertainty for the rarely observed fatal injury events.

Figure 3.2: Coefficients of correlation between predictions of regression analysis and observed values for IAC, LIN, SIN and FAT when values of the weighting factor $\psi = (0.1, 0.2, ..., 1)$. 
3.4.3 Multivariate Poisson-lognormal Regression Analysis

For the regression analysis a hierarchical multivariate Poisson-Lognormal model was used on the development dataset with the data of the homogeneous segments being weighted according to their individual exposure values in order to account for homogeneous segments with a large exposure having a large influence on the results of the regression analysis. The structural component of the regression analysis was determined to have the multiplicative form

\[
Y_{ik} | \text{CHAR} = \nu_{ik} \cdot \exp \left( \beta_{0,ik} + \beta_{1,ik} \cdot \ln(AADT_i) + \beta_{2,ik} \cdot \ln(HGV_i) + \beta_{3,ik} \cdot \text{BEND}_i + \ldots \right)
\]

The values of the variable SLP were used in quadratic form, where the same accident promoting effects were assigned to upwards and downwards slope. Additionally, the signalized speed limit (SPD) was used in the regression model in a squared format as done in other recent investigations (e.g. Hauer (2009), Malyshtina and Mannering (2008), Aljanahi et al. (1999), Aarts and van Schagen (2006), Haglund and Aberg (2000) and Nilsson (2004)). The regression coefficients \( \beta_{ki} \), as well as the standard deviations of the standard errors \( \varepsilon_k \), were assessed by means of maximum likelihood estimation, simultaneously for all model response variables (due to the multivariate model structure) per road section type (CHAR). The road section types are handled differently from the other risk indicating variables as they are discrete, and fundamentally different, categories. The statistical significance of the results was tested using a Students t-test for the individual regression coefficients at a significance level of \( \alpha = 0.05 \). The regression model was without consideration of time trends in the variables and with non-time dependent random effects for each site. Time dependencies, however, can principally be implemented into the regression model and into the BPNs (then denoted as dynamic BPNs), if desired.

3.4.4 Empirical Bayes Method

Step 1: Determination of the safety performance functions

The results of the regression analysis can be expressed in the form of safety performance functions. The results of the regression analysis can be expressed in the form of safety performance functions. Examples for the different model response variables for the different types of road sections are shown in Figure 3.3 when only the risk indicating variable AADT was chosen as

\[7\] The term safety performance function is herewith equivalent to the term prediction model, since the outcomes, i.e. the estimates of the expected number of accident events, of both are identical.
explanatory variable for the regression analysis as given in section 3.4.3. The variables AADT was selected for this example since a covariance analysis for all risk indicating variables showed the strongest relationship to the accident events for the AADT. The safety performance functions are shown one-dimensionally for illustrative reasons. Sensitivity analyses and parameter studies for particular variables can be applied instead for showing such multi-dimensional relationships as it is done e.g. in Deublein et al. (2013).

The comparison of the AADT based safety performance functions in Figure 3.3 shows that the safety performance functions of the injury accidents and light injuries for all types of road are rather close in shape and magnitude while the safety performance functions of the severe injuries and the fatalities are deviating in a way of lower values and different shapes. For example, in terms of AADT the probability of injury accidents, light injuries and severe injuries in tunnels is considerably higher than for open roads but the probability of fatalities is considerably lower (maximum expected number of fatalities $\hat{y}_{Fat} = 0.0003$ per $mvk$); something which may be attributed to safety enhancing measures such as reduced speed limits, increased lane width, or clear signalization and illumination.
Step 2: Determination of the over-dispersion parameters

The overall over-dispersion parameters were assessed as described in (Elvik, 2008) based on the development dataset by fitting negative binomial distributions to the numbers of homogeneous segments on which $\bar{y}_k = 0,1,\ldots,q$ events of the different model response variables have been observed. In Figure 3.4, for injury accidents a negative binomial pmf is shown which has been fitted to the relative frequencies of $\bar{y}_{IAC}$ over the observation period of seven years (2004-2010). The original observation period is kept for the model fit since the negative binomial distribution is only representative for integer values.
Figure 3.4: Relative frequencies of the observed injury accidents and the negative binomial distribution fitted to the observations of injury accidents

The values of the assessed overall over-dispersion parameters are given in Table 3.6 and are found to be quite similar for all model response variables on exit corridors and open roads, except for fatalities. The much higher values for fatalities for exit corridors and intersections may be due to the very small number of fatalities observed there and hence, a higher statistical uncertainty associated with these values. Compared to the overall over-dispersion on open roads, lower values were observed for IAC and LIN on intersections and higher ones for SIN and FAT as well as for all model response variables in tunnels.

<table>
<thead>
<tr>
<th></th>
<th>exit corridors</th>
<th>intersections</th>
<th>tunnels</th>
<th>open roads</th>
</tr>
</thead>
<tbody>
<tr>
<td>injury accidents</td>
<td>2.4476</td>
<td>1.9379</td>
<td>4.5054</td>
<td>2.5312</td>
</tr>
<tr>
<td>light injuries</td>
<td>2.2328</td>
<td>1.8623</td>
<td>4.1462</td>
<td>2.3217</td>
</tr>
<tr>
<td>severe injuries</td>
<td>2.9240</td>
<td>3.2878</td>
<td>5.2546</td>
<td>2.8254</td>
</tr>
<tr>
<td>fatalities</td>
<td>11.1127</td>
<td>16.2621</td>
<td>9.5581</td>
<td>5.1938</td>
</tr>
</tbody>
</table>

The different overall over-dispersion parameters were taken into consideration in the assessment of the individual over-dispersion parameters of the homogeneous segments in the test dataset according to equation {3.7}, where the exponent $\tau$ can have values between zero (no influence of individual section length on over-dispersion parameter) and one (over-dispersion parameter is multiplied by the length of segment). The effect of varying the values of the weighting exponent $\tau$ on the coefficients of correlation between the predictions and the
actual observed values of the model response variables for the test dataset are shown in Figure 3.5.

![Figure 3.5: Relationship between coefficient of correlation and the exponent $\tau$ for the determination of the individual over-dispersion](image)

The differences are assessed by subtracting the $r$-value given $\tau = 0$ from the $r$-values given the increasing values of $\tau = (0.1, 0.2, ..., 1)$. The overall maximum difference in the $r$-values is $\Delta r_{LIN} = -0.0035$. Such a difference is considered to be insignificant and from this perspective, there is no influence of $\tau$ on the coefficients of correlation for the regression model results. However, systematic trends can be observed for IAC, LIN and SIN: The more weight is given to the segment length for the assessment of the individual over-dispersion parameter the less precise the regression model predictions become in terms of the coefficient of correlation. Only for fatalities, the $r$-values are slightly increasing when the value of $\tau$ is increasing. The different trends between injuries and fatalities are attributed to the large statistical uncertainties in the datasets.

The simultaneous influence of the length of one homogeneous segment and of the choice of the weighting exponent $\tau$ on the values of the individual over-dispersion parameters is illustrated in the mesh plot of Figure 3.6. For this example, the overall over-dispersion parameter is assumed to be $\phi = 2.5$ for a homogeneous segment with length $l = 1$ km. The length of this fictive segment is varied between 0.1 and 5 km. If $\tau = 0$ the length of the homogenous segment has no influence and the individual over-dispersion parameter remains
the same as the overall over-dispersion parameter over the entire range of the length. The influence of the segment length increases with increasing values of $\tau$ yielding higher values for the individual over-dispersion when the segment length is greater than 1km and lower values when the segment length is lower than 1km.

![Figure 3.6: Influence of segment length and weighting exponent on the value of the individual over-dispersion parameter](image)

The choice of the value for the weighting exponent depends on the circumstances of the investigation. For the current comparison a value of $\tau = 0.8$ was chosen to give considerable, but not full, influence of the segment length on the over-dispersion parameter.

**Step 3: Determination of the relative weights**

Based on individual over-dispersion parameters, the relative weights for the updating process were determined in accordance with equation (3.8). The relative weight defines the influence of the safety performance functions (assessed on similar road segments) on the prediction of the expected number of accident events on a specific road segment. The higher the relative weight, the more influence is given to the predictions of the safety performance function. An increasing over-dispersion in the available data of the specific road segments leads to a higher weight given to the safety performance functions assessed based on historical data of similar road segments.
In order to demonstrate the effect of transforming the overall over-dispersion parameter for the entire network into an individual over-dispersion parameter for the individual homogeneous segments, the relative weight was determined based on both, the overall and individual over-dispersion parameter. In Figure 3.7 (a), for injury accidents (IAC), the relative weight given the values of the individual over-dispersion parameter ($\phi_{\text{INDIVIDUAL}}$) is plotted over the relative weight given the values of the overall over-dispersion parameter ($\phi_{\text{OVERALL}}$). It can be observed that with increasing values of the relative weight, the difference between the relative weights assessed based on the individual and overall over-dispersion parameters become smaller. The maximum value of both is equal to one, which means that full weight is given to the predictions of the safety performance functions.

Figure 3.7: Comparison of individual weights and expected numbers of accidents assessed based on individual and overall over-dispersion parameters.

In Figure 3.7 (b), the expected values of injury accidents for the individual homogeneous segments of the test dataset are plotted when these are assessed given either the individual over-dispersion parameter ($E[IAC]|\phi_{\text{INDIVIDUAL}}$, y-axis) or the overall over-dispersion parameter ($E[IAC]|\phi_{\text{OVERALL}}$, x-axis). It is apparent that the difference in the predictions is very small and nearly the same expected values of the model response variable occurrences are determined regardless of how the overall over-dispersion parameter is transformed into individual over-dispersion parameters.

Step 4: Determination of the expected number of accident events

The basic assumption for the model evaluation is that no information is available about accident events for the road segments in the test dataset. The updating within the EB method
was, therefore, done using segments similar to those of the development dataset, i.e. the specific homogeneous segments in the test data set to be evaluated had the same constant values of the risk indicating variables as the homogeneous segments in the development dataset. The results are given in section 3.5.

3.4.5 Bayesian Probabilistic Networks Method

Step 1: Discretization of random state variables

The values of the model response variables assumed to be continuously lognormal distributed random variables were discretized to facilitate more efficient model development and parameter learning procedures (intervals and states for discretization are provided in Table 3.2 and Table 3.3).

Step 2: Model learning

Based on the results of the regression analysis the prior predictive distributions of the model response variables were determined and the structure of the BPN was established. Using the estimated distributions of the regression coefficients and the covariance matrices of the error terms the predictive probability density functions of the model response variables were estimated, and the information of the observed data was extrapolated into the entire modelling space in order to provide distributions of the model response variables also in those domains of the model where no observations were available. The open-source inference engine of the program GeNie® 2.0 (Decision-Systems-Laboratory-Pittsburgh, 2006) was used to construct the directed acyclic graph and to calculate the marginal probability distribution functions. In Figure 3.8 the structure of the developed BPN is shown.
Eight risk indicating variables were chosen as the input nodes of the BPN. The values shown in the bar charts correspond to the relative frequencies with which the values of the risk indicating variables were observed in the development dataset. If nothing was known about the considered homogeneous segments, default probabilities of the values of the risk indicating variables would be used as given in the bar charts of the parent nodes in Figure 3.8. For the comparison, the same homogeneous segments were used as for the EB method evaluation (test dataset). Each homogeneous segment was described by putting evidence in the nodes representing the risk indicating variables and appropriate states were selected (e.g. CHAR=4, AADT=40’000, HGV=12%, etc.). The four model response variables were taken as the output nodes of the BPN. In the directed acyclic graph, all input nodes were connected to all output nodes by directed edges.

Parameter learning was then performed using the EM-algorithm and assuming a small value for the experience factor of $e=0.1$. This value of the experience factor means that almost no influence is attributed to the prior information and almost full weight is attributed to the posterior distribution when the model is updated. The role of the experience factor in the updating process of the BPN method is comparable to the role of the overall over-dispersion parameter in the updating process of the EB method. Both are responsible to attribute appropriate weight to the prior information.

However, in case of the updating process of the BPN method, the linear relationships of the
regression analysis are released and, the more data becomes available the more the model is representing the observed data. That means that in long term the regression model and the associated epistemic uncertainty with regard to the regression model are diminishing. A change in the value of the experience factor $e$ results in a changing intensity with which the linear regression based relationships are transformed into data based relationships. Thus, the following coherence is assumed to hold: increasing amount of available data leads to a decreasing value of $e$, which leads to an increasing transformation of the BPN into non-linear model structures.

In contrast, within the updating process of the EB method the linear relationships remain unchanged for the updated models and only the relative weights attributed to the prior information are changed, if the values of the individual over-dispersion parameter are changing, i.e. the epistemic uncertainty in regard to the regression model will remain, independent of the available amount of data. For the EB method the following coherence is assumed to hold: increasing amount of available data, decreasing values of the over-dispersion parameter, decreasing weight given to the prior information, increasing representation of the data in the results.

Both, the experience factor $e$ and the weighting exponent $\tau$ cannot be assessed analytically but have to be selected based on reasonable assumptions (sections 3.3.3 and 3.3.4). They are however, independent.

During the parameter learning process, only the domains of the prior BPN were updated for which there were observations in the development dataset. In the creation of the posterior BPN, the purely empirical regression model based probabilities and linear relationships in the prior BPN were replaced by observation based posterior probabilities and non-linear relationships.

**Step 3: Model validation and testing**

For the model validation, the posterior predictive probability density functions of the model response variables $\lambda_k^*$ were determined, given the evidences for the risk indicating variables on each homogeneous segment of the test dataset. Their mean values were multiplied with the exposures of the homogeneous segments $\nu_i$ to obtain the parameter of the Poisson distribution $\mu_{ki}$ which is used to estimate the expected number of road accident events $\hat{y}_{ki}$ over a defined period of time on the individual homogeneous segments. Model validation and testing was done by comparing the predicted numbers of the model response variables with the observed number of accident events by means of scatter plots and correlation analysis. The results are given in section 3.5.
3.5 Model comparison

The numbers of predicted accident events using the models developed from the two investigated methods, i.e. the models resulting from the EB and the BPN methods as well as the number of observed accident events on the homogeneous segments of the test dataset are shown in Figure 3.9. Additionally, predictions of the regression models, resulting from the regression analysis, are shown to emphasise the possible improvement to the predictions by using the EB or the BPN method. The numbers of accident events are shown separately for all model response variables and the different methods (EB and BPN). The results of the regression analysis are also shown, since they are the basis for the EB method and the BPN method.

The regression lines, which graphically represent the quality of the relationship between model predictions and actual observations, are indicated with solid lines and the regression equations, as well as the coefficients of correlation (r-values), are provided for each scatter plot. The dashed lines indicate (on average) perfect accordance between the model predictions and real observations ($E[y_{ij}] = 0 + 1 \cdot \mu_{ij}$).

Regression lines above and below the dashed reference line indicate that the model predictions on average underestimate and overestimate the real number of accident events, respectively. For injury accidents (Figure 3.9, (a)-(c)) the regression lines indicate that the EM models, the BPN models and the regression models tend to overestimate the occurrence frequencies, which can be considered as rather conservative modelling results. However, in the remaining plots (except of regression predictions in (d) and BPN predictions in (i)), all models tend to underestimate the expected number of injured road users.
Figure 3.9: Comparison of model results based on regression analysis, the Empirical Bayes method and the Bayesian probabilistic network method for the different model response variables.

<table>
<thead>
<tr>
<th></th>
<th>REGRESSION</th>
<th>EMPIRICAL BAYES</th>
<th>BAYESIAN NETWORKS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>μ_{IAC,REG}</td>
<td>y = 0.0815 + 0.5501μ</td>
<td>y = 0.0227 + 0.8187μ</td>
<td>y = 0.0532 + 0.7542μ</td>
</tr>
<tr>
<td>μ_{LIN,REG}</td>
<td>y = 0.0719 + 0.8868μ</td>
<td>y = 0.0390 + 1.1066μ</td>
<td>y = 0.0387 + 1.1366μ</td>
</tr>
<tr>
<td>μ_{SIN,REG}</td>
<td>y = 0.0333 + 1.0017μ</td>
<td>y = 0.0285 + 1.0566μ</td>
<td>y = 0.0287 + 0.9121μ</td>
</tr>
<tr>
<td>μ_{FAT,REG}</td>
<td>y = 0.0046 + 1.3957μ</td>
<td>y = 0.0045 + 1.4288μ</td>
<td>y = 0.0039 + 1.681μ</td>
</tr>
</tbody>
</table>

expected occurrence frequencies from prediction models [per year]
The scatter plots in Figure 3.9 show in terms of the coefficient of correlation and the parameters of the regression equations that the predictive capabilities of the regression models generated using regression analysis are significantly improved by applying the updating algorithms either of the EB or the BPN methods. For all model response variables the coefficients of correlations become increasingly larger from the left column to the right column in the same order as the models are based on regression analysis, EB and BPN methods are applied, the differences are rather small for some model response variables, though.

As can be seen in Figure 3.9, (j)-(l), for all models the capability to predict fatal accident events is very limited. The correlations are weak with $r_{SPF} = 0.238$ and $r_{BPN} = 0.292$, and hence, do not allow meaningful prediction of the expected number of fatalities. Due to the small numbers of observed fatalities (per mvk and year), the data points in the scatter plots show strictly ordered patterns for which it obviously becomes difficult to fit any regression model representing the interrelationship between the predictions and the observations. The poor correlation is assumed to be due the low occurrence frequencies of fatalities and hence, only a very limited amount of data to be used to determine models that sufficiently represent the covariance structure.

For the prediction of severe injuries (Figure 3.9, (g)-(i)), the coefficients of correlation are higher, with $r_{SPF} = 0.477$ and $r_{BPN} = 0.503$. They are still, however, not high enough to result in meaningful predictions. The values of the parameters of the regression equations show good accordance with the dashed reference lines, which signifies good accordance between average model predictions and average accident event observations.

The models developed using the EB and the BPN methods that are meaningful in terms of the prediction of the occurrence of injury accidents and light injured road users. The BPN models achieve the highest correlations with $r_{BPN} = 0.668$ for the prediction of light injured road users and $r_{BPN} = 0.731$ for the prediction of injury accidents. The coefficients of correlation for the EB method based models are in average 5-10% lower for all model response variables, but still allow for almost as accurate predictions as the BPN models. In the scatter plot (e) of Figure 3.9 it can be observed, that the BPN models over-estimate, and the EB models under-estimate, the expected number of light injured road users.

The direct comparison of the predictions for injury accidents on the homogeneous segments of the test dataset resulting from the EB and BPN models are illustrated in Figure 3.10 in which it is shown that BPN models tend to provide higher estimates than EB models and hence, are more conservative. The value of the correlation coefficients between the predictions made using the EB and the BPN models, however, is high for all model response variables with $r_{FAT} = 0.803$ and $r_{SID} = 0.937$. The high coefficients of
correlations indicate that the predictions of both methods are very close and the predictive capability of both is comparably accurate.

![Figure 3.10](image)

Figure 3.10: Direct comparison between BPN and EB predictions for the expected counts of model response variables

### 3.6 Discussion

Both the EB and the BPN methods are used empirically (purely data based) using Bayes and classical probability theory and using the same prior prediction models. The methodological difference between the two methods, hence, reduces to the different procedures used to updating the prediction models that were determined by means of the multivariate Poisson-lognormal regression analysis as described in section 3.3.2. In comparison with the EB method, the following advantages are attributed to the BPN method in terms of its statistical features and practical application.

- The nodes of a BPN are modelled by means of conditional probability distributions which allow a more accurate representation of the uncertainties connected to the random variables than in models which are only based on mean values neglecting such uncertainties.
The proposed BPN method can straightforwardly be modified to use the Full Bayesian approach. Full Bayesian approaches, when compared to Empirical Bayes approaches, make it easier to consistently take into consideration 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, Schubert and Faber, 2012). They also allow consideration of the complete joint posterior probability distributions of the parameters up to a reasonable level within the hierarchy, whereas Empirical Bayes approaches do not. Full Bayes approaches also enable more detailed inference, adding more flexibility in selecting prior distributions (El-Basyouny and Sayed, 2011), in incorporating different theoretical models, e.g. models for lane changing behaviour (Ahmed, 1999). Such theoretical models are needed when no data is available for risk indicating variables, the accident influencing effect of which is intended to be covered by the accident prediction model. There is an increasing demand to use Full Bayes approaches in accident analysis (Carriquiry and Pawlovich, 2005, Miaou and Lord, 2003, Persaud et al., 2010, El-Basyouny and Sayed, 2009b) and hence, tools which allow the use of Full Bayes approaches are increasingly being provided.

- Depending on the structure of the BPN and the nodes where evidence is entered, the flow of information in BPNs can be reversed, which allows analysing both, the effects given evidence for the states of the risk indicating variable nodes and the causes given evidence for the states of the model response variable nodes. Once developed and learned, Bayesian Probabilistic Networks show the impacts of any particular action on the values of all model variables linked to it, with all corresponding uncertainties. The final choice of action is then left to the road planner or infrastructure manager. This characteristic of BPNs, which means that the probabilities of the states can be adjusted to new knowledge without redesigning the whole system, is not as straightforward to be implemented in other methods; e.g. in the EB method or in rule-based systems.

- The initially defined multidimensional linear model structure of the regression analysis remains unchanged during the updating procedures of the EB method. It cannot be modified when new data becomes available and when such new data indicates different types of functional relationships. In contrast, the updating procedures of the BPN method facilitates both, the updating of the model parameters of the hierarchical regression analysis, and the modification and adaption of the model structure to the information obtained through new data; e.g. non-linearity in the interrelationships between risk indicating and model response variables can be accounted for.

- The directed acyclic graph of the BPN makes it possible, that complex systems can be represented in a compact and understandable manner due to the clear structure of nodes and connections in the network. The learning process of the BPN can be based on a conceptualisation of the system and the outcome of discussions with stakeholder groups and experts. Stakeholder participation (recommended already for the definition of the modelling aims) is essential for an integrated management approach (Bromley et al., 2005). In engineering sciences, it is often required to include experts experiences and opinions from the very beginning of the systems modelling process. Such a participatory
modelling procedure can guide the model to focus on the most relevant problems (Aguilera et al., 2011, Henriksen and Barlebo, 2008) and avoid misleading assumptions.

Due to their illustrative nature the application of BPNs for the development of accident prediction models might be appealing. However, this method is very sensitive to erroneous assumptions and thus, the main disadvantage of the BPN method might be hidden in its suspected simplicity. When the directed acyclic graph of the BPN model is to be defined based on experts and stakeholder participation, a sufficient large amount of experience is needed to ensure that the resulting structure of the BPN is consistent with the causalities of the system under consideration. Once, the structural part of the BPN is defined by the directed acyclic graph it is relatively easily communicated to stakeholders. However, the quantitative part of the models to be applied and the parameter learning, with the CPTs and the conditional relationships, is the modelling part, where discussions between experts and stakeholders involved are expected to rise (Henriksen and Barlebo, 2008). Additionally, incorporating experts’ opinion sometimes suffers from a general lack of understanding of classical and Bayes probability theory. Research has shown that significant errors in the application of BPNs might result from the perception of risk depending on the risk-aversion characteristics of the individual (McCabe et al., 1998). In this view, the EB method as described comprehensibly e.g. in the American Highway Safety Manual (AASHTO, 2010) or in Hauer (2002) might be the more straightforward and robust choice for some road engineers.

3.7 Conclusions

In this paper, a brief introduction in the general Bayes probability and inference theory is given and the differences between Empirical and Full Bayes approaches are summarized to provide a basis for the distinction and comparison between methods developed to implement these approaches.

The differences between the Empirical Bayes method and the Bayesian Probabilistic Networks method are outlined and their capabilities with regard to develop models that result in accurate prediction of accident events were compared. In both methods Bayesian inference and updating algorithms are used and for both methods a multivariate Poisson-lognormal regression analysis is performed for the assessment of prior inferences and the determination of the safety performance functions. The theoretical predictions of the safety performance functions are combined with real observations to predict the site specific expected number of the model response variables. The model response variables used for the comparison were the numbers of injury accident events and the numbers of injured road users having no more than
light injuries, severe injuries and fatal injuries. The risk indicating variables were selected taking into consideration both, road design and traffic parameters.

A comparison of the two methods was done by predicting the expected values of the model response variables. Prediction models were developed using the Empirical Bayes and the Bayesian Probabilistic Network method. The predictions were compared with the real observed numbers of accident events on road sections which were not used for model development. It was found that:

- the updating algorithms of both methods significantly improved the coefficients of correlations between model predictions and actual observations when compared with the predictions of the regression analysis alone.

- the models developed using both methods showed good agreement between the predicted and observed numbers of injury accidents and light injuries, i.e. both sets of models resulted in a good correlations between the predicted and observed numbers of the values of these model response variables. They, however, do both not allow drawing meaningful conclusions about their capabilities to predict severe and fatal injuries. This is most likely due to the rareness of these events in the data sets which have been available for the current investigations. One way to overcome this problem could be in the implementation of theoretical severe and fatal injury prediction models being available in literature as results from other investigations on different datasets. In that context, it would be straightforward to use the BPN method to combine data-based interrelationships between the risk indicating variables and the model response variables and literature-based theoretical models. The implementation of theoretical prediction models was, however, beyond the scope of this work. Furthermore, severe injuries and fatalities could be merged to one category of injury severity and would then be represented by more observations in the dataset.

- the models developed using the Bayesian Probabilistic Network method gave slightly more accurate estimates of the number of accident related events, i.e. the values of the correlation coefficients of the Bayesian Probabilistic Network models were approximately 5-10% higher than those of the Empirical Bayes models.

Additionally the predicted values of the model response variables using both methods were graphically compared. It was found that the direct comparison results in high coefficients of correlation indicating that the predictions of both methods are very close and that the predictive capability of both is comparably precise.

Accident prediction models are usually developed as components in a system for first identifying road segments which have atypically high accident rates and second to evaluate the effects of countermeasures on these accident rates. They can also be used to predict the accident rates on not yet built roads or to determine the optimum design and route.

The models resulting from the Bayesian Probabilistic Network method can be used in all situations that the models from the Empirical Bayes method can be used. However, the BPN
models have the additional advantage that a Full Bayes approach can be implemented seamlessly, when desired.

**Acknowledgements**

The work presented in this paper was conducted with financial support by the Austrian Road Safety Board (KFV Austria) in 2011. The KFV Austria also provided and pre-processed the data.
4 USER COST ESTIMATION ON ROAD NETWORKS BY MEANS OF BAYESIAN PROBABILISTIC NETWORKS (PAPER III)

Markus Deublein
Institute of Construction and Infrastructure Management,
Swiss Federal Institute of Technology, ETHZ, Zurich, Switzerland

Bryan T. Adey
Institute of Construction and Infrastructure Management,
Swiss Federal Institute of Technology, ETHZ, Zurich, Switzerland

peer reviewed conference paper, accepted for publication
in proceedings of the 16th International Conference on Road Safety on Four Continents,
15-17 May 2013, Beijing, China
ABSTRACT

In this paper a methodology for the development of multiple impact models for road networks is described. Road impacts can be categorized into impacts for different stakeholder groups, namely the public, the owner and the road users. The current investigations address multiple impacts only for road users, but the developed methodology can be extended to estimate the impacts of all stakeholder groups. Impacts for road users are differentiated into costs due to travelling time, vehicle operation and accident injuries. The accident and injury costs are assessed based on multivariate regression analysis and Bayesian Probabilistic Networks. The proposed methodology for the assessment of road user impacts is different from existing impact models since uncertainties are incorporated into the model. Accordingly, all variables of the model are represented probabilistically.

To demonstrate the usefulness of the methodology, models were developed for the prediction of multiple road user impacts on three road segments that were different in terms of traffic configurations, road designs and surface conditions. Based on the assumptions made for the model development, the results show that accident and injury costs represent only a small share of the total user impacts.

An important feature of the introduced methodology for road impact assessment is that it provides decision support, e.g. on how to optimally allocate budgets into accident risk reducing interventions and evaluate the portfolio of changeable measures in terms of their effect on the accident risk before and after they are implemented.

4.1 Introduction

Optimal intervention strategies are those that result in the lowest negative impacts for stakeholders which for public roads can be grouped as the owner, the user, and the public (Figure 4.1) (Adey et al., 2012). The determination of optimal intervention strategies on road segments, including when to intervene, the type of intervention to execute and the traffic configuration to use during the intervention is the main goal of road owners with the largest impacts on road users. Although it is acknowledged that impacts on the stakeholder groups “owner” and “public” including e.g. environmental impacts are very important contributors to the entire impacts caused by a road network, this paper – as a preliminary investigation – solely focusses on the estimation of road user impacts, namely impacts due to travel time costs, vehicle operation costs and accident costs.

The estimation of these impacts is tremendously difficult due to the many different and interconnecting physical relationships. For instance, the increase in travel time due to deteriorating road condition depends on the slope, the curvature, the capacity, and the speed.
limit of the road as well as the number of vehicles travelling on the road. At the same time, the number of vehicles and the road condition have amongst other variables an influence on the occurrence probabilities of accident events.

Estimations of impacts and assumptions of physical relationships are at least partly subject to uncertainties. These uncertainties can broadly be categorized into aleatory (natural variability of the phenomenon itself) and epistemic uncertainties (modelling uncertainty, statistical uncertainty) (Faber and Stewart, 2003) and are so far not considered in the existing studies for evaluation of road impacts (Adey et al., 2012).

In this paper a recently proposed methodology for the development of accident prediction models (Deublein et al., 2013) is implemented into the development of impact models for road users. For this purpose, the accident prediction model is modified as described in section 4.3.4. The proposed methodology is different from many that are currently used because uncertainties are incorporated into the modelling development and the output variables of the model are represented probabilistically. The use of the methodology is demonstrated by developing models to predict vehicle operation, travel time and accident costs due to multiple types of road segments.

![Figure 4.1: Public road stakeholder groups and cost relevant impacts (Adey et al., 2012)](image)

### 4.2 Methodology

#### 4.2.1 General

The described methodology uses accident prediction models based on multivariate Poisson-Lognormal regression analysis and Bayesian Probabilistic Networks (BPNs).
BPNs are directed acyclic graphs containing chance nodes which represent variables either as continuous random variables or as sets of discrete states, where each state represents an interval. The nodes are connected through directed arrows representing the causal dependencies between variables. BPNs were originally developed in the field of artificial intelligence approximately 30 years ago, but are now widely used in the field of engineering (Faber, 2003). They are used to compactly represent the joint probability density function of all variables in the model. The outcomes of the compilation of the BPN are the marginal probability distributions of the model response variables, i.e. the distributions of the different impact types.

The structure of a BPN is described by using family relations. If node A is directly linked to node B then the node A is denoted as the parent node of the child node B. The joint probability distributions of all nodes are calculated by multiplying the conditional probabilities of the child nodes with their parent nodes. The marginal probabilities of the child nodes are calculated by summing the joint probabilities of all possible states. When the value of one or more variables is observed, referred to as evidence in BPN terminology, the prior probability distributions of the remaining variables can easily be updated. The Bayesian approach necessitates that a prior (conditional) probability mass is assigned for each node. The prior probability represents the existing knowledge before any (new) evidence in terms of data is available (Box and Tiao, 1992). The adequate choice of the prior distributions is very much dependent on the investigated problem and the availability of information (Gelman et al., 2004). A more detailed description of BPNs is given in Jensen & Nielsen (2007), Bayraktarli (2009) and Schubert (2011).

4.2.2 Modelling Steps

The modeling approach can be sub-divided into five consecutive steps. The first three steps are to determine the indicator variables (Step 1), the intermediate variables (Step 2) and the response variables (Step 3) to be included in the BPN and the corresponding prior probabilities of each node. The indicator variables, e.g. road design parameters, surface conditions and traffic characteristics, are the input variables (parent nodes) of the BPN. The intermediate variables (intermediate nodes) are defined to establish a network structure which a) is capable of representing the functional relationships to be used and b) ensures at the same time that the conditional probability tables of the child nodes do not become too large and, therefore, computational expensive. The intermediate nodes are child nodes of the indicator variables and become parent nodes of the response variables. The response variables are the variables, with which it is desired to have information, e.g. accident costs.

Step 4 is to construct the BPN and Step 5 is to enter the values of the indicator variables (evidences) for the road segments to be investigated and compile.
4.3 Model development

4.3.1 Step 1: Determine indicator variables

The definitions, prior probability assumptions and the states of the BPN nodes for the indicator variables are given in Table 4.1.

<table>
<thead>
<tr>
<th>VARIABLE</th>
<th>UNIT</th>
<th>STATES</th>
<th>DEFINITION</th>
<th>PRIOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>AADT</td>
<td>[veh/day]</td>
<td>[1,2,...,10] x 10^4</td>
<td>The variable AADT represents the annual average daily traffic per driving direction.</td>
<td>data-based (Figure 4.2)</td>
</tr>
<tr>
<td>HGV</td>
<td>[%]</td>
<td>[5,10,...,30]</td>
<td>The variable HGV represents the heavy good vehicles (trucks) in percentage of AADT.</td>
<td>data-based (Figure 4.2)</td>
</tr>
<tr>
<td>HOUR</td>
<td>[h]</td>
<td>[0,1,...,23]</td>
<td>The variable HOUR represents the time of the day. All estimates of the user costs are made for one specific hour of the day.</td>
<td>assumption: uniform (0.04)</td>
</tr>
<tr>
<td>TVC</td>
<td>[-]</td>
<td>[A,B,C,D,E,F]</td>
<td>The variable TVC represents different traffic variation curves in percentages of AADT with given HOUR. Six different TVCs are distinguished having different traffic configurations and quantities. E.g. TVC (A) in Figure 4.3 has a clear peak of traffic volume (rush hour) at 7am with a percentage of 12.5% of the AADT.</td>
<td>(Pinkofsky, 2005) (Figure 4.3)</td>
</tr>
<tr>
<td>LAN</td>
<td>[-]</td>
<td>[1,2,3,4]</td>
<td>The variable LAN represents the number of lanes per driving direction.</td>
<td>data-based (Figure 4.2)</td>
</tr>
<tr>
<td>SPDsig</td>
<td>[km/h]</td>
<td>[60,70,...,130]</td>
<td>The variable SPDsig represents the signalized speed limit.</td>
<td>data-based (Figure 4.2)</td>
</tr>
<tr>
<td>CUR</td>
<td>[-]</td>
<td>[0,2,...,10]</td>
<td>The variable CUR represents the horizontal curvature of the road. It is defined as an integer variable having values between zero (straight road) and ten (very high curvature). CUR is determined as the fraction of the sum of the length of ten subsequent 50m road sections, divided by the length of the straight connection between the starting point of the first section and the end point of the tenth section.</td>
<td>data-based (Figure 4.2)</td>
</tr>
<tr>
<td>GRD</td>
<td>[%]</td>
<td>[-6,-4,-2,0,2,4,6]</td>
<td>The variable GRD represents the percentage of upwards or downwards longitudinal gradient.</td>
<td>data-based (Figure 4.2)</td>
</tr>
<tr>
<td>WZ</td>
<td>[-]</td>
<td>[1=no,2=yes]</td>
<td>The variable WZ represents the presence of Work Zones. A road section at which interventions are currently performed is considered as a work zone. Within work zones, specific accident risk increasing variables might be</td>
<td>assumption: ( p(WZ = 1) = 0.9 ) ( p(WZ = 2) = 0.1 )</td>
</tr>
</tbody>
</table>
present, which are currently not included in the model, e.g. reduced lane width, distraction, soiled pavement, work zone length and duration. Different studies show that the crash rates increase when work zones are introduced (Rophail et al., 1988, Wang et al., 1996, Ha and Nemeth, 1995, Khattak et al., 2002). However, the amount of the increase varies across these studies and between the levels of injuries. Taking into account these studies, for the current investigations the accident costs within work zones are assumed to be increased by 25%.

I2 and I4

The variables I2 and I4 represent the road pavement conditions. In Switzerland the pavement conditions are measured using six indices according to VSS SN 640 925b. Increasing values of the indices indicate decreasing road conditions. For the assessment of road user impacts the following two indices are relevant:

\[ I_2 = \text{index for longitudinal unevenness}; \]
\[ I_4 = \text{index for skid resistance}; \]

For both indices the prior probabilities are determined to follow the shape of a Gamma distribution functions with parameters which were assessed based on the assumption that the road conditions have in average \( I_2 \)- and \( I_4 \)-values of \( \mu = 3 \) with a standard deviation of \( \sigma = 2 \) (Figure 4.4). Please note that in reality the observed values of the indices are likely to be different. Only if nothing is known about the condition states of the road they are assumed to have similar prior probabilities.

\( \Gamma(\alpha, \beta) \)abbreviates the Gamma distribution with parameters \( \alpha \) (shape) and \( \beta \) (scale).

In Figure 4.2 the (prior) relative frequencies of the indicator variables are shown for which data was available. The prior probabilities of the indicator variables for which no data was available are modelled either based on assumptions or based on literature information.

Figure 4.2: Prior probabilities based on data used in Deublein et al. (2013)
Figure 4.3: Traffic variation curves (Pinkofsky, 2005)

Figure 4.4: Prior probabilities for the states of I2 and I4 (*assumption of Gamma distribution*)

### 4.3.2 Step 2: Determine Intermediate Variables

The intermediate variables of the BPN are illustrated by the grey shaded nodes in Figure 4.10. The intermediate variables are introduced into the BPN to represent the complex functional relationships between the model input (risk indicating variables) and the model output (model response variables). The data for all injury accidents together with the corresponding numbers
of injured road users (LIN, SIN, FAT, Table 4.2) is the same as in the case study described in (Deublein et al., 2013). 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. In Table 4.2 definitions, prior model assumptions and states of the intermediate nodes are listed.

**Table 4.2: Intermediate variables of the BPN**

<table>
<thead>
<tr>
<th>VARIABLE</th>
<th>UNIT</th>
<th>STATES</th>
<th>DEFINITION</th>
<th>PRIOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>LOS</td>
<td>[-]</td>
<td>[A,B,C,D,E,F]</td>
<td>The variable LOS represents the level of service. It indicates the quality of the traffic flow. LOS depends on the number of lanes and boundary values for the maximum number of vehicles per hour on a particular road segment. The boundary values used to define the level of service are taken from FGSV (2001). Polynomial functions were adapted to the values for LAN=1 to LAN=3 in order to extrapolate the boundaries to roads with LAN=4 per driving direction (Figure 4.5).</td>
<td>deterministic</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>A = free flow</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>B = reasonably free flow</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>C = stable flow</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>D = pending unstable flow</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>E = unstable flow</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>F = breakdown flow</td>
<td></td>
</tr>
<tr>
<td>VEHph</td>
<td>[veh/h]</td>
<td>[0.5,1,...,10]x10³</td>
<td>The variable VEHph represents the vehicles per hour on a considered road segment. The traffic volume per hour during the day is assessed as the fraction of the AADT given the type of TVC for a specific hour of the day (HOUR).</td>
<td>assumption: ( \text{transN}(\mu_{\text{ADT}},\sigma_{\text{ADT}}) ) bounds: (0,10'000)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>where ( \mu_{\text{ADT}} = AADT \cdot TVC \cdot \text{HOUR} )</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>( \sigma_{\text{ADT}} = 0.05 \cdot \mu_{\text{ADT}} )</td>
<td></td>
</tr>
<tr>
<td>SPDact</td>
<td>[km/h]</td>
<td>[5,10,...,150]</td>
<td>The variable SPDact represents the actual driving speed. Given information about SPDsig and LOS, the SPDact is assumed to be normal distributed with mean value derived by establishing polynomial functions under two constraints: (a) if LOS=A then ( \mu_{\text{SPDsig}} = \text{SPDsig} ) and (b) if LOS=F then ( \mu_{\text{SPDsig}} = 5 \text{km} / \text{h} ). The results of the polynomial fit are given in Figure 4.6. The standard deviation of SPDact is assessed by fitting a polynomial function to observed values of standard deviation given SPDsig as illustrated in Figure 4.7.</td>
<td>assumption: ( \text{transN}(\mu_{\text{SPDact}},\sigma_{\text{SPDact}}) ) bounds (0,150)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>where ( \mu_{\text{SPDact}} = f(\text{LOS}) )</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>( \sigma_{\text{SPDact}} = f(\text{SPDsig}) )</td>
<td></td>
</tr>
<tr>
<td>AMF14</td>
<td>[-]</td>
<td>[1,1.02,...,2]</td>
<td>The variable AMF14 represents the accident modification factor related to surface friction index 14. The functional relationship between AMF14 and 14 is given in Figure 4.8 assessed based on the equations provided in Adey et al. (2012). The AMF14 is</td>
<td>assumption: ( N(\mu_{\text{AMF14}},\sigma_{\text{AMF14}}) )</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>where ( \mu_{\text{AMF14}} = f(14) )</td>
<td></td>
</tr>
</tbody>
</table>
representing the relative change in the expected number of injury events given changes in the surface friction \(I_4\) compared to the assumed base condition of \(I_4=0\) (new constructed/replaced pavement).

The functional relationship between AMF_{I4} and the expected number of accident events is assumed to be the same for all investigated accident events.

\[ \sigma_{AMF_{I4}} = 0.05 \]

The variable AMF_{S} represents the crash modification factor related to SPDact. The functional relationships between the SPDact and the AMF_{S} are based on the power-functions given in Nilsson (2004).

The AMF_{S} is representing the relative change in the expected number of injury events given changes in the driving speed (SPDact) with respect to the base condition of the driving speed which is assumed to be SPDbase=100 km/h (assumed average driving speed on Swiss rural motorways). The functional relationship between the values of AMF_{S} and the expected number of accident events is different for the different levels of injury.

\[ \mu_{AMF_{S}} = \frac{SPDact}{SPDbase} \]

\[ \sigma_{AMF_{S}} = 0.1 \]

The variable LIN_{r} represents the occurrence rate of lightly injured road users per million vehicle kilometre \((mvk)\) and year. A road user is considered to be lightly injured if the damage to his well-being lasts less than 25 days following the accident.

The rates of lightly injured road users are determined based on the accident prediction methodology given in (Deublein et al., 2013).

\[ \text{assumption: } N(\mu_{LIN_{r}}, \sigma_{LIN_{r}}) \]

\[ \mu_{LIN_{r}} = LIN_{r} \cdot \text{length}_\text{HS} \cdot VEHph \cdot 10^6 \cdot \text{AMF}_{I4} \]

\[ \sigma_{LIN_{r}} = 0.1 \cdot \mu_{LIN_{r}} \]

The variable SIN_{r} represents the occurrence rate of severely injured road users per \(mvk\) and year. A road user is considered to be severely injured if the damage to his well-being lasts more than 24 days following the accident.

The rates of severely injured road users are determined based on the accident prediction methodology given in (Deublein et al., 2013).

\[ \text{assumption: } N(\mu_{SIN_{r}}, \sigma_{SIN_{r}}) \]

\[ \mu_{SIN_{r}} = SIN_{r} \cdot \text{length}_\text{HS} \cdot VEHph \cdot 10^6 \cdot \text{AMF}_{I4} \]

\[ \sigma_{SIN_{r}} = 0.1 \cdot \mu_{SIN_{r}} \]

The variable FAT_{r} represents the rate of fatally injured road users per \(mvk\) and year. A road user is fatally injured when he has died within 30 days following the accident event as a consequence of accident induced injuries.

The rates of fatally injured road users are determined based on the accident prediction methodology given in (Deublein et al., 2013).

\[ \text{assumption: } \text{trunc}N(\mu_{FAT_{r}}, \sigma_{FAT_{r}}) \]

\[ \mu_{FAT_{r}} = FAT_{r} \cdot \text{length}_\text{HS} \cdot VEHph \cdot 10^6 \cdot \text{AMF}_{I4} \]

\[ \sigma_{FAT_{r}} = 0.1 \cdot \mu_{FAT_{r}} \]
The variable LIN represents the costs for lightly injured road users. Taking into account the information provided by the Highway Safety Manual (AASHTO, 2010), the costs in Swiss Francs (CHF) per lightly injured road user are assumed to correspond to $58'400 [LINUC CHF] = 0.1 \cdot \mu_{LIN}$.

The variable SIN represents the costs for severely injured road users. Taking into account the information provided by the Highway Safety Manual (AASHTO, 2010), the costs per severely injured road user are assumed to correspond to $203'700 [SINUC CHF] = 0.1 \cdot \mu_{SIN}$.

The variable FAT represents costs for fatally injured road users on the homogeneous segment. Taking into account the information provided by the Highway Safety Manual (AASHTO, 2010), the costs pro fatally injured road user are assumed to correspond to $3'780'400 [FATUC CHF] = 0.1 \cdot \mu_{FAT}$.

The variable TIM represents the travelling time per km given information about SPDact and I2. TT is assessed in accordance with Adey et al. (2012) where a penalty factor increases TT on a road segment as a function of the longitudinal unevenness (I2). Values needed for the assessment of TT are assumed to be a $= 2.3$, b $= 2$ and c $= 0.6$ (Adey et al., 2012).

The variable HR represents the vehicle costs per hour driven. It is assessed according to Adey et al. (2012) and it is distinguished between trucks and passenger cars.

The variable KM represents the vehicle costs per kilometre driven. It is assessed according to Adey et al. (2012) and it is distinguished between trucks and passenger cars.

The variable FUEL represents the fuel costs per kilometre driven. It is assessed according to Adey et al. (2012) and it is distinguished between trucks and passenger cars.
A matrix \( (T_v) \) containing the fuel consumptions of passenger vehicles and trucks given SPDact is used. The assumed fuel consumptions per 100km for trucks, diesel and gasoline passenger cars are given in (Figure 4.9).

\[ U_{Cf\text{,corr}} = 0.53 \text{ [CHF/km]} \quad \sigma_{Cf} = 0.1 \cdot \mu_{Cf} \]

\( N(\mu, \sigma) \) abbreviates the normal distribution with parameters \( \mu \) (mean) and \( \sigma \) (standard deviation). \( \text{truncN}(\mu, \sigma) \) abbreviates the truncated normal distribution with lower and upper bounds \((l, u)\).

Figure 4.5: Maximum road capacities (VEHph) for the different numbers of lanes and for the different levels of service based on FGSV (2001)

Figure 4.6: Actual driving speed for different signalized speed limits (SPDsig=60-130) as a function of the level of service (LOS) (assumption)
Figure 4.7: Standard deviation of the actual driving speed given signalized speed limits (based on data described in Deublein et al. (2013))

Figure 4.8: Increase of accident modification factor (AMFI4) given road surface friction index (I4) based on Adey et al. (2012)
4.3.3 Step 3: Determine Response Variables

The response variables are the three different user cost variables, namely the travelling time costs (TT), the vehicle maintenance and operation costs (VO) and the accident costs (AC). The definitions, modelling assumptions and states of the response variables are given in Table 4.3.

Table 4.3: Model response variables

<table>
<thead>
<tr>
<th>VARIABLE</th>
<th>UNIT</th>
<th>DEFINITION</th>
<th>PRIOR</th>
</tr>
</thead>
</table>
| TT | CHF/[h km] | The variable TT represents the travelling time costs. They are due to time which is lost during travelling on the road section. TT varies indirectly as a function of pavement condition represented by the longitudinal unevenness index I2. The unit costs for TT are \( UC_{TT} = 30.93 \times [mu/(veh \cdot h)] \), assuming an occupancy rate of 1.57 persons per car (Adey et al., 2012). | assumption: \( \text{truncN}(\mu_{TT}, \sigma_{TT}) \) bounds (0,90'000) \[
\mu_{TT} = 11.37 \times \text{VEHph} \times UC_{TT} \\
\sigma_{TT} = 0.15 \times \mu_{TT}
\]|
| VO | CHF/[h km] | The variable VO represents the costs of maintenance and operation of a vehicle. Vehicle maintenance comprises man-hours needed for maintenance work and material costs. Vehicle operation costs include additionally costs for fuel consumption. | assumption: \( \text{truncN}(\mu_{VO}, \sigma_{VO}) \) bounds (0,24'000) \[
\mu_{VO} = Ch + Ckm + Cf \\
\sigma_{VO} = 0.1 \times \mu_{VO}
\]|
| AC |  | The variable AC represents the costs of road users suffering different injury | |
levels which are caused by accidents on the road section. The total amount of user costs due to injuries is represented by the sum of the costs resulting from the different injury severities. Accidents with property damage only (PDOs) are neglected since the expected costs for such damages are very low compared to the costs arising due to injuries.

\[
\text{truncN}(\mu_{ac}, \sigma_{ac}) = \begin{cases} 
(0, 2'000) & \text{if } \mu_{ac} < 2'000 \\
(\mu_{ac}, \sigma_{ac}) & \text{otherwise}
\end{cases}
\]

where

\[
\mu_{ac} = \text{LIN} + \text{SIN} + \text{FAT} \\
\sigma_{ac} = 0.1 \cdot \mu_{ac}
\]

\(N(\mu, \sigma)\) abbreviates the normal distribution with parameters \(\mu\) (mean) and \(\sigma\) (standard deviation).

\(\text{truncN}(\mu, \sigma)\) abbreviates the truncated normal distribution with lower and upper bounds \((l, u)\).

### 4.3.4 Step 4: Develop model

The development of the impact model is based on observable risk indicating variables (model input) and predicted model response variables (model output). Additionally, intermediate variables are introduced to simplify the structure and the computation processes of the model. The different variables are represented in the BPN as nodes. The nodes are connected by means of arrows which represent the functional relationships between the variables as given in the Table 4.1 - Table 4.3. The functional relationships are determined empirically based on available data or are defined based on theoretical assumptions.

The entire BPN can be described as two sub-models (Figure 4.10). The first sub-model consists of the model response variables travel time costs (TT) and vehicle operation costs (VO) and the corresponding indicator and intermediate variables. The second sub-model consists of the response variable accident costs (AC) and represents the user costs due to injuries caused by accidents on the road segments. It contains an accident prediction model, that is developed based on the methodology described in Deublein et al. (2013) and uses data-based multivariate Poisson-Lognormal regression analysis and Bayesian updating algorithms to establish the internal relationships between its nodes and the predictive distribution functions of the injury occurrence rates (LINr, SINr and FATr). The variables which were used as input variables for the accident prediction model are indicated by the small stars in the nodes (Figure 4.10). The difference between the prediction model developed here and that proposed in Deublein et al. (2013), is that the signalized speed limit is only used to determine the crash modification factor AMF_S. The AMF_S is used because previous case studies have revealed that the signalized speed limit on rural motorways shows only very small variation. Hence, using the signalized speed limit directly as indicator variable for the injury occurrence rates would result in a rather weak relationship between the signalized speed limit and the different injury occurrence rates. An additional AMF is introduced to represent the influence of the indicator for skid resistance (I4) on the occurrence probabilities of injury events caused by road accidents, AMF_{I4}. The skid resistance of the road surface is considered to have influence on the braking distance of vehicles which can be a decisive factor for the emergence of accidents.
The two sub-models are connected by indicator variables and intermediate variables that are used as input variables for both sub-models. These ‘connecting’ variables are indicated by the dashed lines of the nodes LAN, AADT, SPDsig, SPDact, LOS, VEHph, TVC and HOUR. For example, as illustrated in Figure 4.10, when evidence is given for the node SPDsig the change in the signalized speed limit will change the probabilities in the states of the child node SPDact (being additionally dependent on LOS). The change in the node SPDact in turn influences all sub-models for accident costs, travel time costs and vehicle operation costs.

Each node in the network contains a probability mass function over the defined range of states. A conditional probability table is associated with each node that contains the conditional probabilities for each state of the node given the states of the other nodes. The values of all variables were represented as being in one of a reasonable number of discrete states. The total impact for road users on a road network is obtained by summing the costs predicted for the three response variables using the two sub-models.
Figure 4.10: Developed Bayesian Probabilistic Network for estimation of the total road user costs including the sub-models for travelling time (TT), vehicle operation (VO) and accident injuries (AC).
4.3.5 Step 5: Enter values of indicator variables and compile

In order to evaluate the methodology, it is used to develop models to predict vehicle operating costs, travel time costs and accident costs for three different types of road segments. The three fictive road segments (HS1, HS2, HS3) are independent and represent three different road designs and traffic densities, i.e. the traffic volume or traffic flow on one HS has no influence on the other HSs and the individual HSs can be spatially separated. The values of all investigated variables are considered to be the same within each HS. The road segments are assumed to be representative of typical highway segments on the Austrian highway network for which substantial data is available (Deublein et al., 2013).

<table>
<thead>
<tr>
<th>Variables</th>
<th>[Unit]</th>
<th>HS1</th>
<th>HS2</th>
<th>HS3</th>
</tr>
</thead>
<tbody>
<tr>
<td>length</td>
<td>[km]</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>time</td>
<td>[h]</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>AADT</td>
<td>[veh/day]</td>
<td>30’000</td>
<td>70’000</td>
<td>70’000</td>
</tr>
<tr>
<td>HGV</td>
<td>[%]</td>
<td>20</td>
<td>30</td>
<td>30</td>
</tr>
<tr>
<td>LAN</td>
<td>[-]</td>
<td>2</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>CUR</td>
<td>[-]</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>GRD</td>
<td>[%]</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>TVC</td>
<td>[-]</td>
<td>C</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>HOUR</td>
<td>[h]</td>
<td>11am</td>
<td>11am</td>
<td>8am</td>
</tr>
<tr>
<td>WZ</td>
<td>[-]</td>
<td>no</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>SPDsig</td>
<td>[km/h]</td>
<td>130</td>
<td>100</td>
<td>60</td>
</tr>
<tr>
<td>I2</td>
<td>[-]</td>
<td>1</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>I4</td>
<td>[-]</td>
<td>1</td>
<td>3</td>
<td>4</td>
</tr>
</tbody>
</table>

The values of the indicator variables for the three different HS to be entered into the BPN are given in Table 4.4.

HS1 is a 2-laned motorway section in good condition (I2 = 1, I4 = 1 (definitions according to Table 4.1)). The TVC=C represents a relative constant traffic flow during daytime (Figure 4.3). The average daily traffic is moderate with 30’000 vehicles per day of which 20% are heavy good vehicles. The signalized speed limit is 130km/h. The user cost models are developed for the hour between 10am and 11am.

HS2 is a 3-laned motorway section in medium condition (I2 = 3, I4 = 3). The traffic is assumed to vary on this road segment significantly over the course of the day, e.g. due to additional traffic going into a city in the morning (TVC=A). The average daily traffic is 70’000 vehicles per day of which 30% are heavy good vehicles. The signalized speed limit is 100 km/h. The user cost models are developed for the hour between 10am and 11am.
HS3 is essentially the same motorway section as HS2, but during the execution of an intervention, where the number of lanes has been reduced from 3 to 1, and the signalized speed limit is reduced from 100 to 60km/h. The road condition state on to which the vehicles are being deviated is poor (I2 = 4, I4 = 4). The user cost models are developed for the hour between 7am and 8am being the hour over the course of the day with the highest percentage of the traffic volume. The hour between 7am and 8am was chosen for this segment as the lowest level of service (LOS = F – breakdown flow / congestion) occurs then as opposed to the hour between 10am and 11am.

The values of the indicator variables (Table 4.4) are entered in the BPN and it is compiled. The marginal probability distribution functions are calculated using an inference engine as implemented in the software packages of e.g. Hugin (Hugin, 2008) or GeNie (Decision-Systems-Laboratory-Pittsburgh, 2006).

4.4 Results

The expected values $E[.]$, the standard deviations $STD[.]$ and the coefficients of variation ($COV$) of the three response variables (i.e. the expected impacts) are given in Table 4.5 in Swiss Francs (CHF) per km and hour. The response variables of the developed impact models facilitate the representation of the uncertainties which are connected to the data used for model development and to the model itself. The results provided in Table 4.5 indicate that the traffic volume, the number of lanes and the signalized speed limit have considerable influence on the values of the impact types. When the road condition is in relatively good condition (best three of five condition states), the vehicle operation costs comprise the largest share of the impacts. When the road condition is in poor condition (worst two of five condition states) the travel time costs comprise the largest share of the impacts. The extremely high values for travel time costs when the road is in the worst condition state is due to the fact that the traffic flowing on the road segment is close to the maximum capacity of the road segment, and congestion is occurring. The accident impacts represent only a very small share of the total user costs. The highest values for injury costs are obtained on HS1 triggered by the high signalized speed limit due to the applied power functions (Nilsson, 2004). In case the traffic flow is congested (as it is on HS3) only very few injury causing accidents are expected to occur.
Table 4.5: Values of the response variables - expected user costs in \([CHF/(km\cdot h)]\)

<table>
<thead>
<tr>
<th></th>
<th>HS1</th>
<th></th>
<th></th>
<th>HS2</th>
<th></th>
<th></th>
<th>HS3</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(E)</td>
<td>(STD)</td>
<td>(COV)</td>
<td>(E)</td>
<td>(STD)</td>
<td>(COV)</td>
<td>(E)</td>
<td>(STD)</td>
</tr>
<tr>
<td>TT</td>
<td>540.57</td>
<td>279.64</td>
<td>0.52</td>
<td>1'380.59</td>
<td>1'091.13</td>
<td>0.79</td>
<td>41'948.21</td>
<td>7'762.63</td>
</tr>
<tr>
<td>VO</td>
<td>1'053.39</td>
<td>176.70</td>
<td>0.17</td>
<td>1'524.64</td>
<td>231.97</td>
<td>0.15</td>
<td>6'341.95</td>
<td>982.47</td>
</tr>
<tr>
<td>AC</td>
<td>37.44</td>
<td>26.91</td>
<td>0.72</td>
<td>19.28</td>
<td>10.44</td>
<td>0.54</td>
<td>28.56</td>
<td>6.07</td>
</tr>
<tr>
<td>total</td>
<td>1'631.49</td>
<td>244.08</td>
<td>0.15</td>
<td>2'924.14</td>
<td>558.55</td>
<td>0.19</td>
<td>48'318.72</td>
<td>8'921.19</td>
</tr>
</tbody>
</table>

The coefficient of variation (\(COV\)) varies considerably between different road segments and response variables. The lowest values of the \(COV\) (and hence, the lowest dispersion in the estimates of the user costs) are observed on the road section HS3. For vehicle operation costs the \(COV\) is almost constant on the three considered road segments and for injury costs it decreases with decreasing speed limits and decreasing levels of service.

The discretized marginal probability distributions of the TT, VO and AC response variables are shown for the three different road segments (Figure 4.11).
4.5 Discussion

It is repeatedly discussed that road impact modelling in general requires numerous assumptions to be made for the implemented model variables. The more variables are implemented, the more assumptions have to be made. In the current investigations, assumptions were made in regard to the distribution families and distribution parameters of the random variables and in regard to their causal relationships. These assumptions can be based either on empirical investigations, theoretical suppositions or on experts’ experiences. In this paper, all assumptions are made explicitly for all variables of the models. If fewer variables would be used, fewer assumptions would have to be made, but the influences of the omitted variables on the modelling results would remain. These influences would not be described explicitly any longer but just be represented implicitly as non-causal random noise.
However, the more variables and dependencies are considered in a Bayesian Probabilistic Network and the more states are attributed to each node, the more data is needed to estimate the probability distribution function and to perform the parameter learning.

### 4.6 Conclusions

A methodology based on Bayesian Probabilistic Networks was presented to develop models to be used to predict multiple impacts of road segments, namely travel time costs, vehicle operation costs and accident injury costs. The methodology allows taking into account all uncertainties which are connected to the assumptions of physical models or functional relationships used for model development. The developed models can be updated when additional data for the same problem becomes available, or when it is desired to use the models for a new but similar problem (e.g. to predict impacts in different countries with different road conditions, driving behaviours, etc.). To demonstrate the methodology, models were developed for the prediction of impacts on three road segments that were different in terms of traffic configurations, road designs and surface conditions.

Through this example it can be seen that the proposed methodology can be used to simultaneously estimate multiple impacts on road segments taking into consideration the complex relationships between them. An important feature of the proposed methodology for multiple road impact assessment is that it provides decision support, e.g. on how to optimally allocate budgets into accident risk reducing interventions and evaluate the portfolio of changeable measures in terms of their effect on the accident risk before and after they are implemented. It is believed that this ability will yield decision makers substantial benefits as they try to determine optimal intervention strategies for highway networks in terms of saved time and increased accuracy with respect to how estimations are currently made.
5 CONCLUSIONS AND OUTLOOK

5.1 Conclusions

Accident risk on road infrastructure networks remains a major problem for society. A methodology is proposed for the development of probabilistic multi-level and multivariate accident and injury prediction models based on Poisson-lognormal regression analysis and Bayesian Probabilistic Networks. The methodology facilitates the simultaneous prediction of the expected number of injury accidents as well as different severities of injured road users. In a case study, accident prediction models were developed for rural motorways and design and traffic related risk indicating variables were used. Models developed based on the proposed methodology were compared to models developed based on a state-of-the-art methodology. Furthermore, the developed accident prediction models were modified to be implemented into a framework for multiple road user impact assessment. Knowing the impacts of road networks on the road users is important for road infrastructure decision making, e.g. for the determination of cost-optimal intervention strategies.

Chapter 2 considers the development of a methodology that allows the development of reliable and consistent, efficient and flexible accident prediction models. Such models are applied for the prediction of the expected numbers 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 risk models developed are formulated in terms of model response variables (model output) and risk indicating variables (model input) using Bayesian Probabilistic Networks for homogeneous road segments.

The precision of the model predictions and their usefulness are verified in a case study using the Austrian road network. In that case study, the model response variables considered are gamma-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 expected values of the model response variables are predicted for randomly selected road segments, the data of which has been excluded from the dataset used for initial model development. Satisfying compliance is found between the predicted and observed numbers of response variables with correlation coefficient of $r \leq 0.73$. It is also shown that the model could be used in a geo-referenced manner for the application to identify and predict the expected number of injury accidents and injury severities.
Chapter 3 presents the comparison of two Bayesian methods for the prediction of road traffic accidents and injuries. The methods applied are the proposed methodology being based on Bayesian Probabilistic Networks (Chapter 2) and the well acknowledged Empirical Bayes method.

A basic repetition of the fundamentals of Bayes probability and inference theory is given. The differences between Empirical (data based) and the Full Bayes (prior assumptions based) approaches are described. These two approaches have to be distinguished, since the two compared methods apply them in different ways for the development of accident and injury prediction models.

It is found, that the investigated methodologies are close in terms that both use Bayesian inference and updating algorithms. In the work of this thesis, the same hierarchical multivariate Poisson-lognormal regression analysis is used for both methods to assess the prior inferences and to determine the regression models. Both methods use algorithms for the combination of the theoretical predictions of the regression models with real observations of accident data to predict the site specific expected number of the model response variables, namely light injuries, severe injuries and fatalities. The risk indicating variables were selected taking into consideration accident contributing factors related to road design and traffic characteristics.

The performances of the two methods with regard to develop accident prediction models that result in accurate prediction of accident events were compared by using data of road segments which have not been used for their model development steps. Additionally, the predictive capability of the regression models is compared to the models based on the two Bayesian methodologies.

For the predictions of injury accidents and light injured road users, it is found that the predictions are in good correlation with the observed numbers of accident events when only the models of the hierarchical multivariate Poisson-lognormal regression analysis are used. The updating algorithms of both methods could additionally improve the coefficients of correlations significantly. For all methods the predictions of severe injuries and fatalities are rather weak. This phenomenon can be explained by the natural rareness with which these events have been observed on road segments used for the case study. Related to the size of the entire modelling domain, only a very small amount of observations were available for severe injuries and fatalities. The amount of observed data was insufficient to provide satisfying inference between observed risk indicators and these two model response variables.

The models developed using the proposed methodology gave slightly more accurate estimates of the number of accident related events than those developed by the Empirical Bayes methodology, i.e. the values of the correlation coefficients of the were approximately 5-10% higher than those of the Empirical Bayes models.
Accident prediction models based on Bayesian Probabilistic Networks can be applied at least for the same tasks as the models based on the Empirical Bayes method are applied. For practical application one major advantage of the Bayesian Probabilistic Network method can be seen that it allows representing and communicating the investigated problem in an understandable and comprehensible manner. That makes it possible to establish a participatory modelling procedure, where experts and stakeholders are supposed to contribute to the structural model learning process by defining, adding or removing variables and links in the network, since they can contribute to the building of the model to focus on the most relevant problems and help to avoid bias in the model output due to imprecise assumptions.

Chapter 4 addresses the implementation of the proposed methodology into the context of road infrastructure decision making by developing a Bayesian Probabilistic Network for the assessment of multiple road user impacts.

The work in chapter 4 shows the adaption of the existing accident model and the general implementation of Bayesian probabilistic into the framework of road user impact assessment. In existing impact models the accident costs are often estimated based on very simplified assumptions only conditional on the pavement surface roughness (e.g. Keller and Zbinden (2004) and Adey et al. (2012)). For the impact model development in this thesis, the accident costs are estimated by using the results of the implemented BPN based accident prediction model. This allows for more realistic estimates of the accident occurrence frequencies and hence more realistic estimates of their consequences in terms of costs. For the user impact model different user costs are taken into account related to travel time, vehicle operation and accident injuries. In order to evaluate the user impact model, the total user costs on three road segments are predicted. The three road segments are different in terms of traffic configurations, road designs and surface conditions.

This chapter shows that the proposed methodology can be used to simultaneously estimate impacts on road segments and that it facilitates taking into consideration the complex relationships between the different accident and cost contributing factors and representing all types of uncertainties. If extended to all stakeholder groups in future research, it is believed that this would increase the ability of decision makers to determine optimal intervention strategies for highway networks.
5.2 Scientific achievements and limitations

The scientific achievements of the present thesis are:

- The development of a new methodology to determine advanced multivariate probabilistic models for the simultaneous prediction of road injury accidents and injuries of different levels of severity. The application of Bayesian Probabilistic Networks in the field of accident risk analysis is scarce and little research has been done on this subject. There is a considerable need for accurate, robust, reliable and applicable methods for the development of accident prediction models. The accident prediction models developed using the proposed methodology fulfil these criteria.

- The comparison of the proposed methodology with the state-of-the-art Empirical Bayes methodology. The so-called Empirical Bayes method is presently considered as the state-of-the-art method for the development of accident prediction models. It has been compared to various different methods. However, it has never been compared to methods based on Bayesian Probabilistic Networks. The work of this thesis fills this gap and discusses advantages and disadvantages of both methods. For road accident analysts and road engineers a method is provided, which could be used and recommended for application alternatively to the established Empirical Bayes approach.

- A preliminary investigation of the ability to use the proposed methodology to develop more general multiple impact models. A Bayesian Probabilistic Network is established for the impact assessment of road users in the context of road infrastructure decision making. In contrast to the existing models for road user impact assessment, the developed Bayesian Probabilistic Network represents all variables, which are included in the model probabilistically and provides estimates of the joint probability functions of user costs. Decision makers benefit from this feature since uncertainties connected to these impact estimates can directly be evaluated and considered within the process of decision making. The probabilistic framework for such impact assessment models is a new scientific achievement.

Limitations of the investigations in this thesis are provided in the subsequent paragraphs. Limitations related to the proposed methodology are:

- The discretization of the random variables. For the development of the Bayesian Probabilistic Network, all model variables were discretized into defined intervals, into which the observed values have been allocated. In general, it would be possible for the methodology to treat the risk indicating variables as continuous random variables providing a more consistent probabilistic procedure. However, discrete values facilitate to remarkably increasing computing speed. This advantage is considered to outweigh the small loss of information caused by using discretized data. The methodology allows every
time to modify the Bayesian probabilistic network nodes for continuous distributions of the risk indicating variables as soon as this is required and the effect on the predictive precision can then be evaluated.

- Spatial correlations are not considered. Although it is acknowledged that there may be unobserved confounding variables, which result in spatial correlations between the consecutive road-sections that also influence the accident rates, the applied hierarchical multivariate Poisson-lognormal regression analysis does not consider such spatial correlation. Additional accuracy might be obtained in future investigations by implementing algorithms that are capable to represent such spatial correlations (Meliker et al., 2004, Song et al., 2006b).

- Temporal effects are not considered. The developed accident prediction model in Chapter 2 does not consider temporal effects, e.g. an increase in the traffic volume over time. If appropriate data was available, time trends could be considered in the contingency tables for parameter learning of the Bayesian Probabilistic Networks. Alternatively, dynamic Bayesian Probabilistic Networks can be developed and adapted to observations or assumption of temporal and demographic trends.

- Dependencies between the risk indicating variables are not considered. The risk indicating variables in the multivariate Poisson-lognormal regression model and in the Bayesian Probabilistic Network are treated as independent random variables. However, it might be the case that interactions between the risk indicating variables do exist, e.g. combinations of number of lanes and lane width or horizontal and vertical curves within one homogeneous segment. The possible dependencies between the risk indicating variables should be investigated, described and considered within future modelling approaches.

- A large number of model assumptions are made for the development of multiple impact models in Chapter 4. For every variable in the model, specific assumptions have to be made to represent the complex interrelationships of cost contributing factors. Distribution families were assigned to the particular model variables and the parameters have been estimated. The model results might be sensitive to the assumptions which are made and the sensitivity should be illustrated in future research.

Other limitations are merely related to the developed accident prediction models and the basic conditions of the case study (i.e. the availability of data). These do not affect the proposed methodology for the development of the accident prediction models. In most cases such limitations can be overcome by more information in the available datasets. Limitations related to the case study are:

- The selection of risk indicating variables are not thoroughly investigated. The risk indicating variables were selected only taking into account road design and traffic related accident contributing factors, since these were considered to be of major interest to road
infrastructure decision makers. The selection of risk indicating variables was also influenced by the availability of observed data. Hence, different accident contributing factors which might also play an important role for accident and injury risk prediction might not be represented. For instance lane width, shoulder width or weather conditions.

- The effects of construction sites are not considered. Construction sites are a potentially significant factor in prediction of accident rates on a road segment. Construction sites are temporary interruptions of the original road design and traffic flow and hence are assumed to have an influence on the occurrence probabilities of accident events. They could not be considered in the current investigations due to lack of information with respect to their locations and timing. A (supposedly small) confounding effect of construction sites on the observed accident events might be present in the presently used datasets.

5.3 Outlook

Based on the above mentioned limitations of the proposed methodology, future research endeavours are intended to contribute to useful methodological model modifications and developments as well as more advanced applications of the proposed methodology in the context of road infrastructure decision making and safety management.

Suggestions for future research on methodologies for the development of models to predict accidents and other impacts are:

- Methodologies for the development of accident prediction models should be based on the Full Bayes approach as discussed in section 3.2.2. The trend in accident prediction modelling goes into the direction of substituting the Empirical Bayes approach by methodologies, which are based on the Full Bayes approach. Herewith, the Full Bayes approach can already be implemented seamlessly into the proposed Bayesian Probabilistic Network methodology.

- Accident modelling approaches should account for the spatial and temporal correlations in the accident data obtained over several years, across the investigated road infrastructure network based on approaches described e.g. in Maes et al. (2007).

- The Expectation Maximization algorithm in the parameter learning procedure of the Bayesian Probabilistic Networks might be replaced by Markov Chain Monte Carlo integration methods. The former is considered to be straightforward but rather slow and the latter are faster and provide more complete information about the posterior conditional distributions. Such algorithms are for instance the Metropolis-Hastings algorithm (Hastings, 1970) and the Gipps sampler (Gilks and Wild, 1992).

- Ways have to be found to handle the rare observation frequencies and the scarce data of severe injured and fatally injured road users in order to overcome the vague prediction
results with large uncertainties. 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. Information about such confounding variables is seldom available, and not easy to be considered and incorporated into an accident risk model structure, which is intended to have controllable input parameters. The same time, the model complexity would be increased and so the difficultnness for defining the causal relationships in the model structure. The problem has also been recognized in other studies (Milton et al., 2008, Milton and Mannering, 1998).

- The consideration of construction site data would be of great benefit for the development of accident prediction models, since it is assumed that interventions do influence the occurrence probabilities of accident events. However, such data is rarely recorded since work zones are very temporary phenomena. However, with the developments in infrastructure management systems it is expected that it will not be long until this data is regularly available.

- Understanding and application of the methodology to develop sight, link or network specific accident prediction models might require a lot of coding work for practitioners and engineers. Hence the development of software tools and the implementation of the developed method in such tools would be very beneficial for practitioners. The generic methodology could then be transformed straightforwardly towards different objects of investigation.

Future research on the application and implementation of the proposed methodology shall include:

- Implementation into instruments for road infrastructure safety management. Implementation in Geographic Information Systems (GIS) for geo-referenced road safety assessment and accident risk based decision making. Graphical interfaces can be developed to directly illustrate the accident risk on planned or existing road networks when evidences for different risk indicators are entered into the programs. By now, the proposed methodology can already be used to develop models for the prediction of accidents and injuries linked to the coordinates of the considered road network. On this way, local black spots of noticeable high numbers of injury accidents and injured road users can be identified in an illustrative manner and the effect of countermeasures can directly be evaluated by changing the evidences in the input nodes of the network. In Figure 5.1 an example for the prediction of the expected number of injury accidents on a road section of the Austrian motorway network is illustrated. For each homogeneous segment of that road section the expected number of injury accidents is predicted and plotted over the geo-referenced track of the investigated road section.
CONCLUSIONS AND OUTLOOK

Figure 5.1: Example for geo-referenced application of a BPN based injury accident prediction model (IAC) on a 50km-section on the Austrian motorway A1

- Implementation of the developed accident and injury prediction models into the framework of road infrastructure safety management. Since accident prediction models are usually not developed for their own sake, but rather as components in a system for first, identifying road segments which have atypically high accident risk and second, to evaluate the accident modification effects of countermeasures. On planned but not yet built road links the models are supposed to predict the accident risk and to find the optimum design and track of the planned road. For instance, the methodology can be implemented into road safety impact assessment (RIA) and road safety audits (RSA) which are applied already in the planning phase of road infrastructure projects, or for the instruments of black spot management (BSM) and network safety management (NSM) for existing roads (European Parliament, 2008, RipCORD-iSEREST-Project, 2010). These are helpful instrument for road infrastructure decision making in the future. The prediction of accidents and injuries and their corresponding consequences helps the infrastructure managers to evaluate different design alternatives by comparing the estimated accident and injury risks. They can then chose appropriate measures to reduce accident and injury risk even before these accidents and injuries happen. Furthermore, a re-designing of deficits on the planned road link before traffic opening can avoid expensive and time-consuming changes to existing road infrastructure.

- Adaption of the developed accident and injury prediction models to different countries and different systems. As an example, the scope of investigation could be moved from rural motorways to rural two-lanes and two-way highways (Figure 1.3).

- Show the ability to use the proposed methodology to develop accident and injury prediction models to take into consideration motorway construction sites and work zones. This could be the basis for the model implementation into concepts for optimized intervention planning.
• Extension of the multiple impact assessment models to cover also the impacts on the public and the owners. Part of such investigations should contain a careful treatment and discussions about the assumptions which are made for the different stakeholder impacts.
ANNEX – BACKGROUND OF THE METHODOLOGY

I. General

Accident risk assessment within the framework of road infrastructure safety management can help to decide on the optimal intervention strategies in terms of cost-benefit and reduced numbers of accidents and injuries. However, a prerequisite for the risk assessment and consequence modelling is that appropriate means for assessing the probabilities of accidents and injuries are established. This in turn requires that accurate probabilistic models for the prediction of future accident and injury events are available.

The difference of the proposed methodology for the development of accident prediction models to models developed earlier based on different methodologies, is the combined use of two statistical tools: the hierarchical multivariate Poisson-lognormal regression analysis and the Bayesian Probabilistic Networks (BPNs). The first is used for structure and prior parameter learning. The latter is used to update the results of the regression analysis to take into account aleatory and epistemic uncertainties as well as non-linear dependencies. This novel combination makes it possible to deal with a very general dependency structure (low individual correlations between risk indicating and model response variables) in the data as well as non-linear causal relationships.

This section comprises a brief background on the statistical methods applied in this thesis. Further it is shown how these methods can be used in the context of road accident risk analysis. More detailed descriptions of the particular methods are provided in sections 2-4.

II. Uncertainties

In the developed methodology different types of uncertainties are taken into account. When engineering designs are developed (e.g. for motorways), decisions have to be made irrespective of the state of completeness and quality of available information. Thus decisions are formulated under conditions of uncertainty since the consequences of the respective decision cannot be determined with complete confidence (Ang and Tang, 2007). The effects of such uncertainty on design and planning are important to be considered, however, they are also difficult to be quantified and evaluated in terms of their current and future effects on the performance and design. Uncertainties can be interpreted and differentiated with regard to their type and origin into two commonly used categories: aleatory and epistemic uncertainties (Der Kiureghian and Ditlevsen, 2009, Faber and Maes, 2005, Faber and Stewart, 2003, Steward and Melchers, 1997). The first (type 1 uncertainties) refers to inherent natural variability which cannot be reduced by means of collection of additional information. The
latter (type 2 uncertainties) comprises model uncertainties and statistical uncertainties. The choice of the model for instance introduces model uncertainty into the system (one never knows if the chosen model currently represents the natural phenomenon to 100 percent). The estimation of the model parameters for instance introduces statistical uncertainty, which could be reduced by collection of additional information. When uncertainties are considered explicitly in the statistical analyses, the models which are to be derived are probabilistic and subject to analysis based on probability theory.

III. Bayesian inference

In the developed methodology Bayesian inference approaches are used. In the Bayesian interpretation the probability \( P(A) \) of the event \( A \) is formulated as a degree of belief that \( A \) will occur. This degree of belief is subsequently referred to as prior probability which can be based on subjective experience (theoretical assumptions) or based on data of experiments or observations (frequentistic, empirical). Bayesian inference combines prior information and current information to derive an estimate for the expected numbers of the investigated uncertain events. This can be considered as a process of fitting a probabilistic model to a set of data and summarizing the result by a probability distribution on the parameters of the model and on unobserved quantities such as predictions for new observations (Gelman et al., 2004).

Bayesian inference is based on the Bayes theorem which states that the posterior density of the vector of unknown parameters \( \theta \) can be assessed conditional on the observed values of the data \( y \).

\[
p(\theta|y) = \frac{p(y|\theta) \cdot p(\theta)}{p(y)} \tag{A.1}
\]

where

\( p(\theta) \) is the prior distribution of the unknown parameters \( \theta \),

\( p(y|\theta) \) is the likelihood function, being a function of \( \theta \) with fixed values of \( y \) and

\( p(y) = \sum_{\theta} p(y|\theta) \cdot p(\theta) \) is the sum over all possible values of \( \theta \).

The prior predictive distribution for unknown but observable data is not conditional on any previous observations and is assessed as
The posterior predictive distribution for unknown but observable data can be assessed conditional on the observed values of $y$ formulated as

$$p(\tilde{y}|y) = \int p(\tilde{y}|\theta) \cdot p(\theta|y) d\theta \quad \{A.3\}$$

where $\tilde{y}$ are the posterior predictive values of the observable variable $y$ assuming that these are conditional independent.

In some statistical applications it is appropriate to model the investigated problem hierarchically. That means that the observable outcomes are modelled conditionally on certain parameters, which themselves have a given probabilistic distribution defined by their own parameters, known as hyperparameters.

The posterior density of the vector of unknown parameters is then extended to

$$p(\theta|y, \eta) = \frac{p(y, \eta|\theta)}{p(y|\eta)} \quad \{A.4\}$$

where $\eta$ is the vector of hyperparameters defining the distributions of the unknown parameters $\theta$. The posterior predictive distribution of the unknown but observable data conditional on observed values of $y$ is then given as

$$p(\theta|y) = \int p(\theta|y, \eta) \cdot p(\eta|y) d\eta \quad \{A.5\}$$


When performing probability calculations based on the Bayesian interpretation of probability, there are two related approaches combining prior information with current observations, the Empirical and the Full Bayesian approach.
In the context of accident risk analysis, the most important difference between Empirical and Full Bayes is that in the Empirical Bayes approach, the prior information comes from road segments similar to those under evaluation to estimate the parameters of the model based on historical data. The point estimates of the parameters are then combined with the site specific observed parameters to obtain improved estimates of the long-term expected numbers of accident events (Persaud et al., 2010). In the Full Bayes approach, instead of the prior point estimates of the parameters, prior probability distribution functions are assumed to represent the hyperparameters. This is again combined with the observed site-specific accident data. Now, the variance of the expected numbers of accident events can be assessed more accurately. The Full Bayes approaches are currently gaining more and more weight in the context of accident analysis (El-Basyouny and Sayed, 2011, Miaou and Lord, 2003, Persaud et al., 2010) and hence, methods and tools which allow to implement the Full Bayes theory have to be provided. The method for the development of accident prediction models, which is introduced in this thesis, can be based on both, the Empirical and the Full Bayes approach.

IV. Multivariate model structure

In the proposed methodology a multivariate model structure is chosen. Road accidents are rare and random events. Different approaches and methodologies exist for establishing models to predict such rare events. However, many of them are connected with considerable limitations since they do not take into account e.g. different categories of uncertainties or do not consider multivariate covariance structures which might be present in the accident and injury data.

The multivariate modelling approach is especially appropriate for accident analysis since it is based on both, a set of more than one explanatory variables and a set of more than one model response variables (e.g. simultaneous modelling of the expected numbers of accidents, light injuries, severe injuries and fatalities). In multiple regression models, accident and injury observations are analysed separately. Such models may result in a substantial decrease in efficiency for parameter estimation. Unobserved confounding variables may exist which are not considered as model variables but which still create correlations amongst the model response variables. Taking into account such correlations can improve the prediction and estimation efficiencies. Multivariate accident prediction models based on the frequentistic interpretation of probability suffer from their difficulties in parameter estimation. Multidimensional integrals have to be solved, which makes them too complicated to be efficiently used in practice. Therefore, a Bayesian multivariate model structure is chosen.
V. Hierarchical multivariate Poisson-lognormal regression analysis

In the proposed methodology a hierarchical multivariate Poisson-lognormal regression analysis is used to establish prior accident prediction models. It is used to probabilistically describe the linear causal relationships between the risk indicating and the model response variables. In the context of accident risk modelling, the outcomes of such regression analyses are often referred to as safety performance functions but can also be used directly as accident prediction models.

The regression model is termed hierarchical, because the parameters of the model (regression coefficients, first level of hierarchy) themselves, are described by so-called hyperparameters representing the second level of hierarchy. The hyper-parameters can be determined based on an Empirical or Full Bayes approach and subsequently updated when new data becomes available. Hierarchical models have a great degree of flexibility since mixtures of distribution assumptions can be applied with different assumptions on the different levels of hierarchy. Additionally, criteria like homoscedasticity can be relaxed, e.g. when Monte Carlo simulations are applied to estimate the posterior parameter distributions.

The regression model is termed multivariate, because not only a set of multiple explanatory random variables are used as model input parameters, but also a set of multiple model response variables is used as model output parameters. For the latter, joint probability distributions are assessed simultaneously.

The regression model is termed Poisson-lognormal, because a Poisson-lognormal structural component was chosen for regression analysis. Such a Poisson-lognormal regression model was also used e.g. in Park and Lord (2007), Ma et al. (2008) and El-Basyouny and Sayed (2009b). It is different from the more frequently applied multivariate Poisson-gamma regression model (Kim et al., 2002, Lord et al., 2005, Lord and Miranda-Moreno, 2008, Lord and Park, 2008, Miaou, 1994, Abdel-Aty and Radwan, 2000) mainly due to different non-negative prior distributions for the error term of the regression equation. For the Poisson-lognormal model, the error term is assumed to be lognormal distributed (Ma et al., 2008). For the Poisson-gamma model the error term is assumed to be gamma distributed (Zhou et al., 2012). Good arguments can be found for using either of the two model assumptions. The multivariate Poisson-gamma model is considered as the model which is easier to be applied. Hence, it is used more frequently. However, the multivariate Poisson-lognormal model may fit the data better (Winkelmann, 2008). Additionally, the latter is considered to be more general, more flexible to handle over-dispersion, less restrictive for the assessment of the correlation-structure of the model response variables (the Poisson-gamma model requires the co-variances to be non-negative) and more consistent with probability theory (according to the central limit theorem, the multiplicative structure of the multivariate regression model indicates the response variables and the corresponding error terms to be approximately lognormal distributed).
VI. Bayesian Probabilistic Networks

In the proposed methodology Bayesian Probabilistic Networks (BPN) are used since they constitute a very efficient way of representing the joint probability distribution by taking into account conditional dependencies as well as aleatory and epistemic uncertainties. A BPN is a form of artificial intelligence, which can be described as a probability-based expert system incorporating uncertainty by means of probability theory and conditional dependencies (McCabe et al., 1998, Pearl, 1985, Cowell et al., 1999, Jensen and Nielsen, 2007, Kjaerulff and Madsen, 2008, Korb and Nicholson, 2004, Lauritzen, 1996, Pelikan, 2005). For the application in the field of artificial intelligence Bayesian Probabilistic Networks were mostly used as rule-based systems, which may lead to problems when dealing with uncertainties, especially when new knowledge or data shall be introduced into the system (Cooper, 1990). In contrast to the rule-based systems the Bayesian Probabilistic Networks applied in this thesis are rather normative expert systems, meaning that the domain of uncertainty is modelled instead of the expert himself, they are based on classical probability calculus and decision theory instead of using probability calculations tailored for rules, and they are intended for expert’s support instead of replacement (Faber, 2012). In the engineering sector Bayesian Probabilistic Networks are used due to their flexibility and efficiency in regard to systems representation. A comprehensive overview of the use of Bayesian Probabilistic Networks in today’s research endeavours is provided in Aguilera (2011).

The structural component

The structural component (or qualitative component) of Bayesian Probabilistic Networks is encoded by a directed acyclic graph containing chance nodes which are graphically representing a set of random variables. The random variables are represented either as continuous variables or as variables with discrete states or intervals. The nodes are connected through directed edges representing informal or causal dependencies among the random variables. Hence, the structure of the associated graph determines the dependence relationships among the variables. This feature makes it possible to evaluate the relevance of individual variables to be included or excluded in the structure of the Bayesian Probabilistic Network with no need for retraining, adaption or new numerical calculations. The structure of a Bayesian Probabilistic Network is often described by using family relations. E.g. when node A is directly connected with node B then node A is denoted as the parent node (\(\text{pa}\)) of the child node B (Figure A.1).
Figure A.1: Example of a simplified Bayesian Probabilistic Network

The arcs between the nodes in a directed acyclic graph can be defined as serial, converging and diverging connections, explaining the flow of information through the network structure as soon as evidence is given to one of the variables. Considering Figure A.1, the different connections are defined as

- **serial connections**: The node A (parent of B) has a causal influence on node B (child of A and parent of C) which in turn has a causal influence on C (child of B). If evidence is introduced about the state of A this information flows from A to C and vice versa, since knowledge about one of the variables provides information about the other. If the state of B is known with certainty the variables A and C become conditional independent.

- **diverging connections**: The node D (parent of B and E) has a causal influence on both child nodes, B and E. Information about any of the states of the child nodes can influence the other child nodes as long as the state of the parent node D is not known. If D is observed any information about B is irrelevant to E and vice versa.

- **converging connections**: The node B is causally influenced by both parent nodes, A and D. As long as no evidence is given for B, the parent nodes remain independent and no information is transferred through the child node B. However, as soon as knowledge about the state of the child node B or the parents A and D, all variables will become dependent. This means, that the extent, to which child node B is influenced by parent A also provides information on how much parent D is influencing B.

**The parameter component**

The parameter component (or quantitative component) defines the strength of the dependencies between the variables of the Bayesian Probabilistic Network by conditional probabilities that are determined for each cluster of parent and child nodes in the network. The parameter component of a Bayesian Probabilistic Network is represented by using
multidimensional conditional probability tables. In a conditional probability table, all probabilities for a variable are assessed conditional on the probabilities of all variables the considered one depends on. To fully specify a Bayesian Probabilistic Network with a previously developed structure, it would be necessary to establish conditional probability tables (CPTs) for each variable given any information about the states of the parent variables. Bayesian inference and updating algorithms are used to assess the joint probability density functions of all random variables of the model. A joint probability distribution function is equal to the product of the conditional distributions attached to each node in the Bayesian Probabilistic Network and given as

\[
P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i | pa_i) \quad \{A.6\}
\]

where

- \( P(x_i | pa_i) \) is the conditional probability of \( x_i \) given information of the parents \( pa_i \).
- \( X = (X_1, \ldots, X_n) \) is a vector of all the variables in the system;
- \( x = (x_1, \ldots, x_n) \) is a vector of specific states of the variables in \( X \) and
- \( pa_i \) is the set of predecessors of \( X_i \).

Evidence can be introduced into any of the nodes of the Bayesian Probabilistic Network in terms of measured observations. The inference calculation of the Bayesian Probabilistic Network 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 distributions. Different algorithms which can be used to compute and update the conditional probabilities in a Bayesian Probabilistic Network have been proposed e.g. by Pearl (1988), Lauritzen and Spiegelhalter (1988) and Jensen et al. (1990). 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. According to Faber and Maes (2005), Li et al. (2008) and Der Kiureghian and Ditlevsen (2009), the consideration of uncertainties is necessary in the estimation of accident risk, since it allows for more realistic standard errors of the resulting model than would otherwise be determined.
Learning the Bayesian Probabilistic Network

The updating of a Bayesian Probabilistic Network when new data becomes available is denoted as learning the Bayesian Probabilistic Network. Both, the structure component and the parameter component of the Bayesian Probabilistic Network can be learned based on available data.

Structure learning is an optimization problem to construct a directed acyclic graph of the Bayesian Probabilistic Network, which best represents the covariance of the available dataset. Optimization of the structure can purely be based on data exploiting algorithms and statistical measures, e.g. the hill climbing algorithm (de Oña et al., 2011, 2013, Tsamardinos et al., 2006, 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. However, learning the structure directly from data may require a great amount of data and the result may be a Bayesian Probabilistic Network structure which is very much fitted to that data. If the amount of data is not sufficient, experts’ knowledge can be incorporated into the development of the network structure: The directed acyclic graph of the Bayesian Probabilistic Network is then either fully defined by the experts or just a particular part while the rest of it is learned with data.

In most engineering problems, the structure is modelled on the basis of causal relations. Pure data based structure learning is only representing the linear correlations in the data. Linear relationships do often not represent engineering problems appropriately. Therefore, pure data based structural learning is not recommended for model development in engineering systems, where the causalities are in principle known. Thus, in this thesis the structure of the directed acyclic graph of the Bayesian Probabilistic Network is determined by causal relations and experts judgment.

Parameter learning of the Bayesian Probabilistic Network is done by constructing so called contingency tables which contain observations of the risk indicating variables and the response variables for each road segment in the investigated time period. Different algorithms for estimating and learning the parameters can be used for the parameter learning process of Bayesian Probabilistic Networks. For complete datasets the maximum likelihood parameter estimation is commonly chosen. This estimation method however, might have some drawbacks in case the datasets are small. It is based on the fraction of positive counts over the total counts and hence, can have (unrealistic) zero values (Carlin and Louis, 2000). Alternatively the Bayesian estimation can be used for complete datasets which starts with a prior distribution of the parameters being updated by new experience to maximize the posterior distributions (Jensen and Nielsen, 2007). In contrast, when the contingency tables for parameter learning are incomplete (a so-called missing data situation), the Expectation-
Maximization algorithm (EM algorithm) is used as described e. g. in Dempster et al. (Dempster et al., 1977), Lauritzen (Lauritzen, 1995), Cox (1983), Carlin and Louis (2000), Fahrmeir and Osuna (2003) and Karlis (2003). Unlike the EM algorithm, most of the classical statistical techniques are not able to deal with missing data, meaning that if only one value is missing for a homogeneous road segment of the database, the entire segment is excluded from data analysis. This loss of information can become large and can be overcome by the EM algorithm capable to still using the partial information of the incomplete cases (Carlin and Louis, 2000). The implementation of algorithms which can handle missing data makes Bayesian Probabilistic Networks useful instruments to deal with expensive or incomplete data. The EM algorithm is very efficient and widely featured in todays’ Bayesian Probabilistic Network software (e.g. HUGIN™ researcher (Kjaerulff, 1992) or GeNie™ (Decision-Systems-Laboratory-Pittsburgh, 2006)). It is also used in the proposed methodology to find the maximized estimates of the distribution parameters based on alternates between an expectation step and a maximization step until a certain convergence criteria is fulfilled. In the expectation step, the dataset is completed by calculating expectations for the missing values based on current parameter estimates. In the maximization step, the completed dataset is used to find new maximum likelihood estimates of the parameters. On that way, the internal causal interrelationships and dependencies in the BPN are iteratively updated based on the additional data. Alternatively to the EM-algorithm, Markov Chain Monte Carlo (MCMC) simulation techniques can be applied to assess the updated posterior distribution function in the Bayesian Probabilistic Network. MCMC algorithms are used for accident risk analysis for instance in Qin et al. (2005), Song et al. (2006b), Park et al. (2010) and Persaud et al. (2010). The EM-algorithm was chosen for the proposed methodology due to its simplicity to be implemented into the Bayesian Probabilistic Network learning process.

The results of the parameter learning are the updated conditional probability tables for all variables of the Bayesian Probabilistic Network. The joint probability distribution function can be assessed for all nodes of the Bayesian Probabilistic Network as the product of the conditional distributions (equation {A.6}). So-called evidence can be introduced into any of the nodes of the Bayesian Probabilistic Networks in terms of measured observations and the joint probability distributions are updated accordingly. When new data are collected regularly, these data can directly be used to learn the parameters of the Bayesian Probabilistic Network and to improve the model performance.
VII. Example

Based on very simplified assumptions, this example illustrates the calculations for the six different steps of the proposed methodology as described in Section 2.2.

STEP 1: Determination of Model Variables

The risk indicating variables and model response variables used for this example are listed in Table A.1 and Table A.2 and their different states (a1, a2, b1, b2, c1 and c2) are defined.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Definition</th>
<th>States</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>AADT</td>
<td>Daily traffic volume per driving direction&lt;sup&gt;8&lt;/sup&gt;.</td>
<td>a1: $AADT &lt; 25'000$ a2: $AADT \geq 25'000$</td>
<td>[veh/day]</td>
</tr>
<tr>
<td>SLP</td>
<td>Percentage of absolute upwards or downwards gradient.</td>
<td>b1: $0 \leq SLP &lt; 2$ b2: $SLP \geq 2$</td>
<td>[%]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variable</th>
<th>Definition</th>
<th>States</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIN</td>
<td>The response variable LIN represents the occurrence rate of light injured road users being involved in injury accidents. Light injuries are bodily harms with less than 24 days of damage to health or incapacity to work.</td>
<td>c1: $0 &lt; LIN \leq 0.05$ c2: $LIN \geq 0.05$</td>
<td>[LIN/mvk]</td>
</tr>
</tbody>
</table>

STEP 2: Construction of Homogeneous Segments

The values for the model variables as given in Tables A.1 and A.2 are defined for one homogeneous road segment and are listed in Table A.3. Additionally, the average occurrence rate of light injuries of the entire Austrian rural motorway network is defined in accordance with the results of the case study in section 2.3 and used as background rate for the occurrence of light injuries.

<sup>8</sup> Commonly in traffic research, the AADT is provided for the entire cross section of a roadway. In order to develop accident prediction models for the separate driving directions, the values of the variable AADT refer to one specific driving direction.
Table A.3: Risk indicating values, background rate and weighting factor for the selected homogeneous road segment

<table>
<thead>
<tr>
<th>homogeneous Segment Nr.</th>
<th>$l$ [km]</th>
<th>AADT [veh/day]</th>
<th>SLP [%]</th>
<th>$\bar{y}_{LIN}$ [LIN]</th>
<th>$\lambda_{LIN}$ [LIN/mvk]</th>
<th>$\psi$ [-]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.75</td>
<td>20'000</td>
<td>1.2</td>
<td>2</td>
<td>0.01</td>
<td>0.3</td>
</tr>
</tbody>
</table>

where

- $l$ is the length of the homogeneous segment,
- $\bar{y}_{LIN}$ the number of observed light injured road users between (and including) the years 2004-2010,
- $\lambda_{LIN}$ is the background occurrence rate for the number of light injured road users assessed as average value over the entire road network (section 2.3.4), and
- $\psi$ is the weighting factor.

**STEP 3: Gamma-Updating of Model Response Variables**

The exposure ($v$) of one road segment is defined as the product of the traffic volume ($AADT$) per year ($year = 365$ days) and its length ($l$). For the selected homogeneous road segment the exposure is determined as

$$v = AADT \cdot l \cdot year = 20'000 \left(\frac{veh}{day}\right) \cdot 0.75 \left(\frac{km}{day}\right) \cdot 365 \left(\frac{day}{year}\right) = 5.475 \left(\frac{10^6 veh}{km \cdot year} \right) = mvk \{A.7\}$$

The exposure is commonly expressed in terms of million vehicle kilometre ($mvk$).

On the first level of hierarchy, the product of the exposure and the occurrence rate ($\lambda$) is the average occurrence frequency ($\mu$) being used as the parameter of the assumed Poisson process (section 2.2.3).

On the second level of hierarchy, the occurrence rate $\lambda$ can be assumed to be Gamma distributed with prior hyper-parameters $\alpha'$ and $\beta'$ assessed as
\[ \beta' = \frac{\psi}{l} AADT \left( \frac{\text{veh}}{\text{day}} \right) \cdot l \left( \text{km} \right) \cdot \text{year} \cdot \frac{0.3}{0.75} \left( \text{km} \right) = \]
\[ 20'000 \cdot 0.75 \cdot 365 \cdot 0.3 \left( \frac{\text{veh}}{\text{km} \cdot \text{year}} \right) = 2.190 \left( \frac{10^6 \text{veh}}{\text{km} \cdot \text{year}} \right) \]

\[ \alpha'_{\text{LIN}} = \beta_{\text{LIN}} \cdot \beta' = 0.01 \left( \frac{\text{LIN}}{\text{mvk} \cdot \text{year}} \right) \cdot 2.190 \left( \text{mvk} \right) = 0.0219 \left( \frac{\text{LIN}}{\text{year}} \right) \]

The posterior hyper-parameters are determined as

\[ \alpha^*_{\text{LIN}} = \alpha'_{\text{LIN}} + \frac{\lambda_{\text{LIN}}}{t} = 0.0219 \left( \frac{\text{LIN}}{\text{year}} \right) + \frac{2}{7} \left( \frac{\text{LIN}}{\text{year}} \right) = 0.3076 \left( \frac{\text{LIN}}{\text{year}} \right) \]

\[ \beta^* = \frac{\beta'}{\psi} + \frac{\psi}{l} = \frac{2.190}{0.3} + \frac{5.475}{7} \left( \text{mvk} \right) = 8.0821 \left( \text{mvk} \right) \]

with \( t \) being the observation period of the available dataset (\( t = 7 \) years).

**STEP 4: Development of Regression Model**

The expectation operator \( E[\cdot] \) of the Gamma-updated model response variable is used as dependent variable for the regression analysis. The expectation operator is assessed as

\[ E[\hat{\lambda}^*_{\text{LIN}}] = \alpha^*_{\text{LIN}} \cdot \frac{1}{\beta^*_{\text{LIN}}} = 0.3076 \cdot \frac{1}{8.0821} \left( \frac{\text{LIN}}{\text{mvk}} \right) = 0.0381 \left( \frac{\text{LIN}}{\text{mvk}} \right) \]

As a comparison, the Gamma-updated occurrence rate for the number of light injured road users is considerably higher than the background occurrence rate (\( \hat{\lambda}_{\text{LIN}} \)) which was determined based on the observations of the entire network. This is because two light injuries
have been observed on the selected road section in the considered time period. They are increasing the values of the updated Gamma parameters.

For this example a simplified multiple Poisson-lognormal regression analysis is performed. The regression equation is:

\[
E[Y_{aLin}] = v_i \cdot \exp(\beta_0 + \beta_1 \cdot AADT_i + \beta_2 \cdot SLP_i + \epsilon_i) \tag{A.13}
\]

This regression equation is transformed into the logarithmic domain as

\[
\ln(\lambda) = \beta_0 + \beta_1 \cdot AADT_i + \beta_2 \cdot SLP_i + \epsilon_i \tag{A.14}
\]

This transformation leads to a linearized structure of the regression equation simplifying the estimation of the expected values of the regression coefficients and allowing treating the error term as normal distributed random variable (result: \(\epsilon \sim N(0,0.738)\)).

Regression analysis is performed based on equation A.14 using the observations of the 5'546 homogeneous road segments of section 2.3.3. The expected values and standard deviations of the regression coefficients are given in Table A.4.

<table>
<thead>
<tr>
<th>(\ln(\lambda_{aLin}^*))</th>
<th>(E[\beta_1])</th>
<th>(STD[\beta_1])</th>
<th>(E[\beta_2])</th>
<th>(STD[\beta_2])</th>
</tr>
</thead>
<tbody>
<tr>
<td>-3.1038</td>
<td>0.0257</td>
<td>1.581 (10^3)</td>
<td>1.029 (10^6)</td>
<td>0.0172</td>
</tr>
</tbody>
</table>

**STEP 5: Construction and Parameter Learning of the Bayesian Probabilistic Network**

STEP 5 contains the construction and parameter learning of the BPN. In Figure A.2 a very simplified structure of a BPN is illustrated. The following sub-sections contain four stages of the BPN development and application based on the assumptions being made for the current example.
Stage 1: BPN without prior information

Assuming nothing is known about the nodes in the network, in this example, the prior probability distributions for the parent nodes AADT and SLP are assumed to be uniform distributed over their different states. The prior conditional probability table of the output node LIN and the joint probability table of the BPN contain uniform probability distributions, accordingly (Figure A.2).

The joint probabilities are assessed as

\[ P(LIN, AADT, SLP) = P(LIN | AADT, SLP) \cdot P(AADT) \cdot P(SLP) \]  \hspace{1cm} \{A.15\}

and the marginal distribution of the node LIN is determined as

\[ P(LIN) = \sum_{AADT} \sum_{SLP} P(AADT, SLP, LIN) \]  \hspace{1cm} \{A.16\}

The resulting probabilities are given in Figure A.2.

Stage 2: BPN parameters learned based on regression results

The prior probability distributions for the parent nodes AADT and SLP are still assumed to be uniform distributed over the states of these nodes. The prior conditional probability table of

---

Figure A.2: BPN with uniformly distributed prior probabilities in all nodes

---
the output node LIN is now assessed based on the dependencies identified and quantified by means of the regression analysis. In order to make this example reproducible, Monte Carlo simulations are made for the distributions of the input nodes of the BPN, the regression coefficients and the error term. The normal distributed model parameters for the regression coefficients are provided in Table A.4. The assumptions for the input variables are $AADT \sim LN(9.9729,0.5545)$ and $SLP \sim N(0,2)$ (Figure A.3). Based on the regression analysis, the error term was assessed to be $\varepsilon \sim N(0,0.738)$.

![Figure A.3: Assumed probability density functions for the input variables AADT and SLP.](image)

The expected values for the light injuries (LIN) are assessed by using the regression equation \{A.14\} and the conditional probabilities are determined. The joint probability table of the BPN is now assessed based on the new values of the conditional probability table by using equation \{A.15\}.

The results for the conditional probability table and the joint probabilities are given in Figure A.4.
Stage 3: **BPN updated based on observed data**

When new data becomes available, such datasets can be used in terms of contingency tables to update the probability distributions of the nodes in the BPN. The observed probabilities are then termed as observed likelihoods.

The posterior distributions in the nodes of the BPN are calculated by using the maximum likelihood estimator $\theta$

$$\theta = \hat{P}(X_i|pa_i) = \frac{n(X_i, pa_i)}{n(pa_i)}.$$  \hspace{1cm} \{A.17\}

where $pa_i$ are the parent nodes of node $X_i$ having $u$ different states $X_i = x_{i1}, x_{i2}, ..., x_{iu}$.

The contingency table (Table A.5) is created for this example containing observations of the selected homogeneous road segment over ten subsequent years. As a comparison, the contingency table of the case study in section 2.3.6 contains 5’546 homogeneous road segments over an observation period of 7 years.

In the updating process, using the EM algorithm, the numbers of observations in the contingency table are counted for all different possible state combinations of the nodes in the BPN.

As an example when using the observations recorded in the contingency table given in Table A.5, the maximum likelihood estimator for the conditional probability of the occurrence rate of light injuries being smaller than 0.05 light injuries per $mvk$, given that the AADT is larger (or equal) than 25’000 vehicles per day and the SLP is less than 2% is assessed by
\[ \hat{P}(\text{LIN} = c_1 | \text{AADT} = a_2, \text{SLP} = b_1) = \]
\[ \frac{\hat{P}(\text{AADT} = a_2, \text{SLP} = b_1)}{n(\text{LIN} = c_1, \text{AADT} = a_2, \text{SLP} = b_1)} \]
\[ \frac{n(\text{AADT} = a_2, \text{SLP} = b_1)}{N} = \frac{5}{10} = 0.5 \]  \{A.18\}

where \( n(\text{LIN} = c_1, \text{AADT} = a_2, \text{SLP} = b_1) \) represents the number of observations in the dataset of the contingency table where the variables LIN, AADT and SLP are equal to c1, a2 and b1 respectively (Figures A.2-6).

Table A.5: Contingency table for parameter learning of the Bayesian Probabilistic Network for one homogeneous road segment

<table>
<thead>
<tr>
<th>obs. count</th>
<th>segm nr [-]</th>
<th>obs time [year]</th>
<th>( l ) [( \text{km} )]</th>
<th>AADT [veh/day]</th>
<th>( v ) [v/km]</th>
<th>SLP [%]</th>
<th>( y_{\text{LIN}} ) [LIN]</th>
<th>( y_{\text{LIN}} ) [LIN/mvk]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.75</td>
<td>37'000</td>
<td>10.13</td>
<td>1.2</td>
<td>0</td>
<td>0.0000</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>2</td>
<td>0.75</td>
<td>38'000</td>
<td>10.40</td>
<td>1.2</td>
<td>1</td>
<td>0.0961</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>3</td>
<td>0.75</td>
<td>39'000</td>
<td>10.68</td>
<td>1.2</td>
<td>0</td>
<td>0.0000</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>4</td>
<td>0.75</td>
<td>40'000</td>
<td>10.95</td>
<td>1.2</td>
<td>1</td>
<td>0.0913</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>5</td>
<td>0.75</td>
<td>41'000</td>
<td>11.22</td>
<td>1.2</td>
<td>0</td>
<td>0.0000</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>6</td>
<td>0.75</td>
<td>42'000</td>
<td>11.50</td>
<td>1.2</td>
<td>0</td>
<td>0.0000</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>7</td>
<td>0.75</td>
<td>43'000</td>
<td>11.77</td>
<td>1.2</td>
<td>1</td>
<td>0.0850</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>8</td>
<td>0.75</td>
<td>44'000</td>
<td>12.05</td>
<td>1.2</td>
<td>0</td>
<td>0.0000</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>9</td>
<td>0.75</td>
<td>45'000</td>
<td>12.32</td>
<td>1.2</td>
<td>1</td>
<td>0.0812</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>10</td>
<td>0.75</td>
<td>46'000</td>
<td>12.59</td>
<td>1.2</td>
<td>2</td>
<td>0.1588</td>
</tr>
</tbody>
</table>

In large BPNs, not all states of the nodes can be observed, which does not imply that the probability for the un-observed states is equal to zero. In such cases, so-called experience can be introduced into the updating process. The experience corresponds to the number of observations being available in the contingency table for the different state combinations. The experience is then used to weight the observed information. E.g. if an old and a new dataset are assumed to have the same value of information, each observation in the corresponding datasets are attributed with experience factor \( e=1 \). In case the new information is considered to be more valuable, the information of the old dataset is weighted with less high experience factors. For the current example and also for the investigations in 2.2.5, the prior information based on the regression analysis is assumed to have only very little informative value and the observed information in the contingency table is supposed to almost completely replace the prior information. For this reason, an experience factor of \( e=0.1 \) is attributed to the prior
probabilities in the nodes of the BPN and the observed probabilities of the contingency table are weighted with an experience factor $e=1$.

The posterior probabilities are assessed as:

$$P^*( LIN = c1 | AADT = a2, SLP = b1 ) =$$

$$\frac{0.1 \cdot P( LIN = c1 | AADT = a2, SLP = b1 ) + n( LIN = c1 | AADT = a2, SLP = b1 )}{0.1 + 10} = \frac{0.1 \cdot 0.05 + 5}{10.1} = 0.5005$$

From equation \{A.19\} it can be seen that due to the low weight given to the prior probabilities the determined posterior probabilities mainly represent the observed likelihoods.

The posterior probabilities of the node LIN and the posterior joint probabilities are given in Figure A.5.

![Figure A.5: BPN with updated posterior probabilities in all nodes based on observations](image)

**Figure A.5: BPN with updated posterior probabilities in all nodes based on observations**

**Stage 4: BPN applied for particular homogeneous segment**

When evidence is introduced into the input nodes of the risk indicating variables of the Bayesian Probabilistic Network, the state e.g. of the node AADT will become known with certainty. The probability changes to a value of e.g. $P(a1)=1$ for the observed state. The
conditional probability table as well as the joint probability table are updated accordingly based on equations \{A.15\} and \{A.16\} and are given in Figure A.6.

Figure A.6: BPN with evidence introduced into the node AADT

<table>
<thead>
<tr>
<th>P(AADT)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>a1</td>
<td>0.00</td>
</tr>
<tr>
<td>a2</td>
<td>1.00</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>P(SLP)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>b1</td>
<td>0.70</td>
</tr>
<tr>
<td>b2</td>
<td>0.30</td>
</tr>
</tbody>
</table>

| P(LIN|AADT,SLP) |  |  |
|---|---|---|
| b1 | b2 |  |
| a1 | a2 | a1 | a2 |  |
| c1 | 0.00 | 0.05 | 0.00 | 0.01 |  |
| c2 | 0.00 | 0.95 | 0.00 | 0.99 |  |

<table>
<thead>
<tr>
<th>P(LIN,AADT,SLP)</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>b1</td>
<td>b2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>a1</td>
<td>a2</td>
<td>a1</td>
<td>a2</td>
</tr>
<tr>
<td>c1</td>
<td>0.00</td>
<td>0.04</td>
<td>0.00</td>
</tr>
<tr>
<td>c2</td>
<td>0.00</td>
<td>0.66</td>
<td>0.00</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>P(SLP)</th>
<th>0.70</th>
<th>0.30</th>
</tr>
</thead>
<tbody>
<tr>
<td>P(AADT)</td>
<td>0.00</td>
<td>0.70</td>
</tr>
</tbody>
</table>

| P(LIN) | 0.043 | 0.957 |

**STEP 6: Prediction of the Expected Number of Events**

For the prediction of the expected number of light injury events based on the predictions of the BPN, the marginal probability distribution of the node LIN is used and its mean value is assessed. The mean value is then multiplied with the exposure as given in equation \{2.2\} of section 2.2.3 to be used as parameter for the Poisson distribution to predict the expected number of light injured road users.

**VIII. Data**

The data used for the investigations in this thesis has been provided by the Austrian Road Safety Board (KFV). The release of the data is not permitted due to confidentiality agreement.
REFERENCES


ADEY, B. T. 2002. A supply and demand system approach to the development of bridge management strategies. PhD, EPFL.


ADEY, B. T., HERRMANN, T., TSAFATINOS, K., LÜKING, J., SCHINDELE, N. & HAJDIN, R. 2012. Methodology and base cost models to determine the total benefits of preservation interventions on road sections in Switzerland. Structure and Infrastructure Engineering, 8, 639-654.


AHMED, K. I. 1999. Modeling Drivers' Acceleration and Lane Changing Behavior. PhD, Massachusetts Institute of Technology MIT.


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, 591-600.

REFERENCES


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


EL-BASYOUNY, K. & SAYED, T. 2009a. Accident prediction models with random corridor parameters. Accident Analysis & Prevention, 41, 1118-1123.


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


REFERENCES


MA, J. 2006. *Bayesian multivariate Poisson-lognormal regression for crash prediction on rural two-lane highways*. PhD, University of Texas at Austin.


REFERENCES


MARSH, W. & BEARFIELD, G. Using Bayesian Networks to model Accident Causation in the UK Railway Industry. 7th International Conference on Probabilistic Safety Assessment and Management, PSAM7, June 14-18 2004 Berlin, Germany.


REFERENCES


RIPCORD-ISEREST-PROJECT 2010. Road Infrastructure Safety Management, Results from the RIPCORD-iSEREST Project. BAST, German Federal Highway Research Institute


REFERENCES


CURRICULUM VITAE

PERSONAL DETAILS

Name Markus Klaus Deublein
Address Mühlezelgstrasse 24, 8047 Zürich, Switzerland
Telephone +41 44 633 71 31
E-mail deublein@ibi.baug.ethz.ch
Date of birth 15 March 1979
Citizenship German

EDUCATION

2011- ETH Zurich, Switzerland, Institute of Construction and Infrastructure Management (IBI), D-BAUG; Ph.D. student and research assistant in the Group of Infrastructure Management
2007-2011 ETH Zurich, Switzerland, Institute of Structural Engineering (IBK), D-BAUG; Ph.D. student and lecture assistant in the Group of Risk and Safety
2006-2007 TU Munich, Germany, Research Department Ecology and Ecosystem Management, Woodscience Munich; Research Assistant in the Group of Timber Testing and Grading
2004-2006 TU Munich, Germany, Research Department Ecology and Ecosystem Management; Master of Science in Forest and Wood Science
2001-2004 TU Munich, Germany, Research Department Ecology and Ecosystem Management; Bachelor of Forest Science and Resource Management
2000-2001 University of Rostock, Germany, Research Department Environmental Sciences; Bachelor Student in Landscape Management and Environmental Protection
1998-2000 Jung Gartenbau, Nürnberg, Germany; Apprenticeship and Diploma as Landscape Gardener