TRYSE
TRANSPORTE Y SEGURIDAD

Universidad de Granada

E.T.S. Ingenieros de Caminos, Canales y Puertos

Departamento de Ingeniería Civil

Área de Ingeniería e Infraestructuras de los Transportes

Tesis Doctoral

APPLICATION OF BAYESIAN NETWORKS FOR THE ANALYSIS OF TRAFFIC ACCIDENTS INJURY SEVERITY ON RURAL HIGHWAYS

Randa Oqab Mujalli

Granada, Julio de 2011

Universidad de Granada

E.T.S. Ingenieros de Caminos, Canales y Puertos

Departamento de Ingeniería Civil

Área de Ingeniería e Infraestructuras de los Transportes

Tesis Doctoral

APPLICATION OF BAYESIAN NETWORKS FOR THE ANALYSIS OF TRAFFIC ACCIDENTS INJURY SEVERITY ON RURAL HIGHWAYS

Randa Oqab Mujalli

Granada, Julio de 2011

Editor: Editorial de la Universidad de Granada
Autor: Randa Oqab Mujalli
D.L.: GR 896-2012
ISBN: 978-84-694-9408-0

TESIS DOCTORAL

APPLICATION OF BAYESIAN NETWORKS FOR THE ANALYSIS OF TRAFFIC ACCIDENTS INJURY SEVERITY ON RURAL HIGHWAYS

Por

Randa Oqab Mujalli

Ingeniera Civil

Presentada en el

Departamento de Ingeniería Civil

de la

Universidad de Granada

Director de Tesis:

D. Juan de Oña López

Dr. Ingeniero de Caminos, Canales y Puertos

Granada, Julio de 2011

TESIS DOCTORAL

APPLICATION OF BAYESIAN NETWORKS FOR THE ANALYSIS OF TRAFFIC ACCIDENTS INJURY SEVERITY ON RURAL HIGHWAYS

Por: Randa Oqab Mujalli

Ingeniera Civil

Director de Tesis:

D. JUAN DE OÑA LÓPEZ

Dr. Ingeniero de Caminos, Canales y Puertos

TRIBUNAL CALIFUCADOR

Presidente: Dr. D.

Vocales: Dr. D.

Dr. D.

Dr. D.

Secretario: Dr. D.

Acuerda otorgarle la calificación de,

Granada, Julio de 2011

To my parents, brothers and sisters
who supported me during all times.

ACKNOWLEDGEMENTS

First and before all, I would like to express my gratitude to Prof. Juan de Oña López, his continuous assistance and guidance throughout the period of my research work were the motive behind my dedication to overcome all the obstacles that encountered me.

Also, I would like to thank Erasmus Mundus External Cooperation Window Lot 4 (JOSYLEEN) for the financial support.

I would like to thank all the members of my research group for their support and encouragement, Laura, Griselda, Rocio and Pachi.

Thus, none of what I have achieved would have become true without my family, Mom and Dad you were the beacon that kept guiding me throughout this tough road. Thanks Rana, Mohammad, Basel and Ayat for being the lovely brothers and sisters whom always encouraged me to keep following my goal!

Randa,

Granada, July 2011

RESUMEN

En esta Tesis doctoral, se proponen modelos de redes bayesianas para analizar la severidad de los accidentes de tráfico. Estos modelos son capaces de hacer predicciones, sin necesidad de pre-supuestos y se utilizan para hacer representaciones gráficas de los sistemas complejos con componentes relacionados entre sí.

En esta investigación se analizaron los accidentes de tráfico en carreteras rurales en España en función de la severidad de los mismos. Se construyeron tres Redes Bayesianas (Bayesian networks - BN) usando 18 variables que representan las características del conductor, de la carretera, del vehículo, de los accidentes, y las condiciones meteorológicas. Las BN construidas se utilizaron para clasificar la gravedad de los accidentes en heridos leves y muertos o heridos graves. Los resultados indicaron que utilizando las 18 variables para construir las BN, las variables que se encuentran más relacionada con un accidente con un muerto o con heridos graves en los accidentes de tráfico son: tipo de accidente, edad del conductor, iluminación y número de heridos.

A continuación, mediante diferentes algoritmos se seleccionaron las variables más significativas, se construyeron Bayesian networks con un menor número de variables (entre 4 y 16 variables) y se compararon los resultados con los del modelo base, que utilizaba las 18 variables originales. Los resultados indicaron que si se utilizaba un grupo seleccionado de variables (p.e. tipo de accidente, edad, factores atmosféricos, género, iluminación, número de heridos y número de ocupantes) para construir una BN, los indicadores de rendimiento (p.e. precisión) mejoraban con respecto a los de la BN original en la mayoría de los casos. Por lo tanto, se puede decir que mediante BN es posible reducir el número de variables que se utilizan para analizar la gravedad de un accidente de tráfico sin reducir la precisión de los modelos.

ABSTRACT

In this Ph.D. Thesis, Bayesian networks' models are proposed to study traffic accident severities. These models are capable of making predictions without the need for pre-assumptions and are used to make graphic representations of complex systems with interrelated components.

In this research work accidents on rural highways in Spain were analyzed for injury severity. Three different Bayesian networks (BN) were built using 18 variables representing driver characteristics, highway characteristics, vehicle characteristics, accidents characteristics, and atmospheric factors. The BN built were used to classify the injury severity of accidents into slightly injured and killed or severely injured. The results indicated that using the 18 variables to build BN, the variables that are mostly associated with a killed or severely injured in traffic accidents were: accident type, driver age, lighting and number of injuries.

Following, using different algorithms the most significant variables were selected and were used to build BN with a smaller number of variables (from 4 to 16 variables). Results were compared with the base model which used the 18 original variables. The results indicated that using a selected group of variables (accident type, age, atmospheric factors, gender, lighting, number of injured and occupant involved) to build a BN, the performance indicators (i.e. accuracy) improved with respect to the original BN in most of the cases. Thus, it is possible to reduce the number of variables used to model traffic accidents injury severity through BNs without reducing the performance of the model.

Table of Contents

LIST OF TABLES

LIST OF TABLES

LIST OF FIGURES

LIST OF FIGURES

CHAPTER 1

Introduction

1.1. Statement of the Problem

The presented doctorate thesis is focused on the analysis and modeling of injury severity of traffic accidents. Due to the fact that road traffic injuries are a growing public health and development problem, special attention is dedicated in order to study and analyze this problem in traffic safety and traffic accidents' research.

The World Health Organization (WHO) report on the global trends and prediction (WHO, 2004) indicated that the number of road traffic injuries has continued to rise in the world as a whole, but there has been an overall decline in high-income countries since the 1970s, and an increase in many of the low-income and middle income countries. With respect to road traffic injuries the results indicated that they are predicted to rise from 10th place in 2002 to 8th place by 2030 as a contributor to the global burden of diseases. On the other hand, road traffic deaths are predicted to increase by 83% in low-income and middle-income countries (if no major action is taken), and to decrease by 27% in high-income countries. If an appropriate action is not taken, the overall global increase is predicted to be 67% by 2020 (Table 1.1).

As illustrated in Table 1.1 the fatalities that traffic accidents produce could be considered one of the major problems that large parts of the world are and will be facing in the future. Thus, the economic burden that accidents produce is one of the major problems that might encounter low and medium income countries.

Thus, understanding and defining the causes that are responsible for traffic accidents' injuries, more specifically traffic accidents fatalities, are one of the main reasons that could lead to find countermeasures in order to control or mitigate this problem.

Many research studies have analyzed the injury severity of traffic accidents; in general the main purpose of these studies was to find the variables that most significantly contribute to the occurrence of a specific injury severity level when a traffic accident occurs.

Table 1. 1: Predicted road traffic fatalities by region (in thousands), adjusted for underreporting, 1990-2020

World Bank Region*	Number of countries surveyed	1990	2000	2010	2020	Change (%) 2000-2020	Fatality rate (deaths per 100 000 person)	
							2000	2020
East Asia and Pacific	15	112	188	278	337	79	10.9	16.8
East Europe and Central Asia	9	30	32	36	38	19	19.0	21.2
Latin America and Caribbean	31	90	122	154	180	48	26.1	31.0
Middle East and North Africa	13	41	56	73	94	68	19.2	22.3
South Asia	7	87	135	212	330	144	10.2	18.9
Sub-Sahara Africa	46	59	80	109	144	80	12.3	14.9
Sub-total	121	419	613	862	1124	83	13.3	19.0
High-income countries	35	123	110	95	80	-27	11.8	7.8
Total	156	542	723	957	1204	67	13.0	17.4

* Data are displayed according to the regional classifications of the World Bank.

Source: reproduced from The World Bank (2003).

In general, injury severity data is represented by discrete categories such as fatal injury, incapacitating, possible injury, and property damage only which is usually referred to as KABCO scale (Morgan, 2009). The first step in the process of developing and applying an appropriate statistical methodology is to define the characteristics of the data used to explain the injury severity of a traffic accident. However, data used for analyzing injury severity of traffic accidents were found to share a number of characteristics, some of which will be briefly mentioned below (Savolainen, 2011).

The first characteristic is related to assuming that statistical data are drawn from a random sample. This, however might not always be guaranteed, not all accidents are always reported. Individuals involved in no injury or minor injury accidents are less likely to have their accidents reported to police.

The second common characteristic is that adjacent injury severity levels might share some unobserved effects because sometimes these levels might be closely related. Failing to account for such correlation in certain types of model estimation methods might result in problems related to biased parameter estimates and incorrect inferences.

In addition, sometimes relevant explanatory variables might be omitted from the model due to a limited amount of data available for the researcher. This can result in inconsistent parameter estimates if the omitted variable is correlated or has different variances among severity levels.

Finally, volume of accidents data available to the researcher is an important factor in the selection of an appropriate model. The size of the sample can guide the methodological selection process in which small samples may necessitate more simplistic models.

To that end, several approaches have been developed and refined to analyze the factors that contribute to the occurrence of a specific injury severity level conditioned on the occurrence of a traffic accident and to predict the severity level when certain conditions apply. These approaches included historical accident data analysis, predictions based on statistical models, results from before-and-after studies, and expert judgments made by experienced engineers.

Most of the studies found to analyze injury severity of traffic accident used discrete outcome models; recently new methods that use data mining and soft computing techniques were applied, these methods will be discussed in detail in chapter two.

Thus, this study is focused on using Bayesian networks to the analysis of injury severity of traffic accidents, in which its utilization to this field was limited so far into only one study by Simoncic (2004).

However, Bayesian Networks were introduced in the 1980's as formalism for representing and reasoning with models of problems involving uncertainty, adopting probability theory as a basic framework (Lucas, 2001). Over the last decade, the Bayesian network has become a popular representation for encoding uncertain expert knowledge in expert systems. The field of Bayesian networks has grown enormously over the last few years, with theoretical and computational developments in many areas.

Bayesian networks are used for modeling knowledge in bioinformatics, medicine, document classification, information retrieval, image processing, data fusion, decision support systems, engineering, gaming, and law.

In addition, using Bayesian networks to analyze injury severity of traffic accidents is supported by the fact that traffic accidents do not follow a determined distribution, and they are not distributed normally, and since Bayesian networks does not suppose these assumptions, it could be applied to such a problem.

1.2. Thesis Organization

This thesis consists of seven chapters. The first chapter includes an introduction to thesis. Chapter two presents a brief overview of the injury severity models used in previous studies, discussion of these studies, a brief introduction about Bayesian networks, their applications, advantages, and finally a conclusion is presented. Chapter three presents the objectives and hypotheses to be fulfilled in this research work. Chapter four presents the dataset to be used in the models, the methodology followed, methods of learning Bayesian networks, evaluation indicators of Bayesian networks and a brief description of the variables' selection algorithms. In chapter five the results of modeling traffic accidents injury severity using Bayesian networks and the results of using variables' selection algorithms are also presented. Finally, chapter six, presents the major conclusions for the injury severity constructed Bayesian networks' models; Bayesian networks built using selected variables and the future work.

CHAPTER 2

State of the Art

This chapter includes a brief review of existing literature for injury severity of traffic accidents. Generally, in the research of roadway safety, two measures are commonly used:

1. Accidents frequency: this measure evaluates the frequency of accidents on roadway segments. Accidents frequencies on different roadway segments are used to estimate count data models, such as; Poisson, negative binomial, and their zero inflated. In which the independent variables used are the roadway segment characteristics, e.g. roadway segment length, curvature, slope, type, pavement quality, etc.) (Malyshkina, 2008).

2. Injury severity: this measure evaluates the injury severity of accidents as determined by the most severely injured individual involved in the accidents. In this case discrete outcome models (e.g. ordered probit and multinomial logit models) are commonly used to do the evaluation using data available on individual accidents. The independent variables used in these models are the individual accident characteristics (e.g. time and location of an accident, weather conditions, roadway characteristics at the location of the accident, vehicles characteristics, and driver characteristics) (Malyshkina, 2008).

This chapter presents an overview and a discussion of the used models in the analysis of traffic accidents' injury severity studies.

2.1. Injury severity models

Several modeling techniques were found to be used by researchers analyzing injury severity of traffic accidents. The most commonly used ones are briefly discussed below. The modeling techniques used to analyze injury severity of traffic accidents were classified into 4 groups: discrete outcome models, data mining techniques, soft computing techniques and other techniques.

The review performed found that since 1996, 45 studies analyzed injury severity of traffic accidents. The review carried out was limited to those studies that analyzed injury severity from an engineering perspective. Also, studies that only analyzed accidents of pedestrians, cyclists, motorcycles were not considered.

2.1.1. Discrete outcome models (DOM)

DOM are used to represent probabilities of having an outcome based on certain factors or characteristics. The discrete nature of the accident outcome are considered as a physical event, in which modeling accident outcomes is derived from simple probabilistic theory (Washington et al. 2003). It should be noticed that some discrete data might have an ordered nature while others not, in the following subsection a description of each type of DOM is presented.

2.1.1.1. Logit Models (LM)

A special case of general linear regression is logistic regression or logit model, which assumes that the response variable follows the logit-function. Logistic model is an approach that is used in mathematical modeling in order to describe the relationship of single or several independent variables to a binary outcome variable. This modeling approach is usually preferred by researchers, since the logistic function must lie in the range between zero and one, and this is not usually the case with other possible functions. Also the logistic function has an S-shaped description and this shape can be easily interpreted (Kleinbaum and Klein, 2002).

The simplest form of the LM is the binary form, where the outcome variable is one of two outcomes. Binary logit model (BLM) and some extensions of them are the models which were found to be mostly used in the literature of accidents' injury severity analysis.

A brief description of these extensions is the following (Ortúzar and Willumsen, 2001; Kleinbaum and Klein, 2002; Train, 2009; and Keele and Park, 2006; Washington et al., 2003):

1. Multinomial logit model (MNL) is used when the outcome variable has more than two unordered categories.
2. Hierarchical logit, nested logit or multi-level logit model (HL) is used when certain assumptions valid for the MNL are violated, e.g., when either the outcome are not independent, or when there are variations among individuals.
3. Mixed logit model (MXL) is a generalized extreme value (GEV) model where, this distribution allows for correlations over outcomes and it is a generalization of the univariate extreme value distribution that is used for standard LM. This model alleviates the three limitations for the standard LM by allowing for random variation, unrestricted substitution patterns and correlation in unobserved factors over time. MXL are actually the integrals of the standard logit probabilities over density parameters.
4. Ordered logit model (OLM), also known as proportional odds model, this model has an observed ordinal variable (Y), where Y is a function of another latent continuous unmeasured variable (Y*). Values of Y* determine the values of the resulting Y. Y* has various threshold or cut points, where the value of Y depends on these thresholds. The random disturbance or the error term here follows a logistic distribution (Washington et al., 2003).
5. Heteroskedastic logit model (HKL) is also a GEV model; however, instead of capturing correlations among outcomes, it allows the variance of unobserved factors to differ over outcomes.
6. Heterogeneous model (HM) is used when dealing with categorical dependent variables. If the variants of the error term are non-constant, the standard error will be incorrect and the parameters will be biased and inconsistent. In order to deal with unequal error variances the HM is used.
7. Generalized estimating equations (GEE) are an extension of the logistic model to handle outcome variables that have binary correlated outcomes. GEE takes into account the correlated nature of the outcome.

Table 2.1: Studies that analyze injury severity of traffic accidents using Logit Models

Study Authors (publication year)	Model type	No. records used	No. variables	Area type	Features	Injury level	No. injury levels	Model fit test
Shanker et al., 1996	HL	1,505	21	Rural	SEG	KABCO	5	ρ^2 = 0.52 (Probably McFadden's)
Donelson et al., 1999	BLM	55,000	11	mixed	SEG	K	1	Kendall's tau (τ) coefficient c statistic=[0.882-0.916] for all models , concordance) = [88.4%-85.5%] for all models, discordance =[6.4%-9.8%]for all models, tied pair of a fatal crash=[4%-5.6%] for all models
Krull et al., 2000	BLM	59,743	16	mixed	SEG	K+A, B+C+O	2	ρ^2 = 0.147 (McFadden's)
Abdelwahab and Abdel-Aty, 2001	OLM	1,168	14	n.a.	INT	A, C+B, O	3	accuracy =58.9%
Dissanayake and Lu, 2002	BLM	8,382	16	mixed	SEG	KABCO	5	accuracy =89.2%
Bédard et al., 2002	BLM	109,837	12	n.a.	n.a.	K	1	n.a.
Srinivasan, 2002	OMXL, OLM	3,492	6	mixed	BOTH	KACO	4	OMXL against OLM: $\chi^2 > \chi^2$ critical (60.86>28.86 at 0.05) for the observed data, $\chi^2 > \chi^2$ critical (31.92>28.86 at 0.05) for the predictive data
Ouyang et al., 2002	BLM	2,986	24	mixed	BOTH	A+K, O+C	2	ρ^2= 0.172 (Probably McFadden's)
Khattak and Rocha, 2003	OLM	4,552	5	n.a.	n.a.	AIS	7	ρ^2 = 0.1040 (McFadden's)
Dissanayake, 2004	BLM	n.a.	15	mixed	SEG	KABCO	5	n.a.

Wang and Kockelman, 2005	HKL	n.a.	25	n.a.	SEG	KABCO	5	ρ^2 (McFadden's)
								for one vehicle: HOP= 0.237, OP= 0.235
								for two vehicle: HOP= 0.257, OP= 0.251

Lenuguerrand et al., 2006	GEE, HL, BLM	12,030	9	n.a.	INT	K, not K	2	n.a.
Awadzi et al., 2008	MNL	n.a.	18	mixed	BOTH	KBO	3	n.a.
Milton et. al., 2008	MXL	22,568	26	mixed	BOTH	K+A, BO	2	ρ^2 = 0.1145 (Calculated McFadden's)
Malyshkina and Mannering, 2009	MNL	81,172	16	mixed	SEG	KBO	3	*p*-value for χ^2= 0.20-0.50
Schneider et al., 2009	MNL	10,029	24	Rural	SEG	ABCO	4	ρ^2 (Probably McFadden's) for small radius model: ρ^2=0.258 for medium radius model: ρ^2=0.230 for large radius model: ρ^2=0.253
Jung et al., 2010	OLM, BLM	255	30	n.a.	n.a.	K+A, B+C, O	3	accuracy=88% for the KA, accuracy=68% for the B+C
Haleem and Abdel-Aty, 2010	HL	2,043	21	mixed	INT	K+A, B+ C+ O	2	AIC= 34040
Jin et al., 2010	OLM	13,218	7	n.a.	INT	KABCO	5	ρ^2 = 0.0542

Continues Table 2.1

Study Authors (publication year)	Model type	No. records used	No. variables	Area type	Features	Injury level	No. injury levels	Model fit test
Daniels et al., 2010	HL	1,491	7	n.a.	INT	K+ A, K	2	χ^2 for K+A= 10.88 (DF=8, *p*-value= 0.21), for K = 4.86 (DF= 6, *p*-value= 0.56)
Paleti et al., 2010	HKL	6,950	15	n.a.	n.a.	KABCO	5	ρ^2 =0.1188 (Calculated McFadden's)
Quddus et al., 2010	OLM, HM	39,98	17	n.a.	SEG	KAC	3	ρ^2 (McFadden's) ρ^2 = 0.096 for the OLM ρ^2 = 0.099 for the HCM
Dupont et al., 2010	BLM	1,296	14	n.a.	BOTH	K	1	n.a.
Peek-Asa et al., 2010	BLM	87,185	12	mixed	BOTH	KA	2	n.a.
Kononen et al., 2011	BLM	n.a.	7	n.a.	n.a.	A	1	sensitivity= 40% specificity= 98% ROC area= 0.84

- n.a.: not available data; KABCO (K=killed, A=incapacitating, B=non-incapacitating, C=possible injury, O=no injury); AIS (0=no injury, 1=minor, 2=moderate, 3=serious, 4=severe; 5=critical; 6=unsurvivable); SEG=basic segment only, INT=intersection only, BOTH=intersection + segments

  Probably McFadden's: ρ^2 value was given in the study, and was found to apply to McFadden's formula; Calculated McFadden's: ρ^2 value was not given, but was calculated using log-likelihoods given in the study
- For more information about each study please refer to Appendix I.

2.1.1.2. Probit models (PM)

PM deal with the three limitations that LM presents: they can handle random variation, they allow any pattern of substitution, and they are applicable to panel data with temporally correlated errors (Train, 2009).

Extensions of PM have been used in the field of accident severity. The most used were the ordered probit models (OPM). OPM is a generalization of the PM to the case of more than two outcomes of an ordinal dependent variable. In the case the model cannot be estimated using the ordinary least square, it is usually estimated using the maximum likelihood (Train, 2009).

Other PM used are the following:

- Heteroskedastic probit model (HOP): it is used when the error terms are not homoskedastic and their variance may be parametrized as a function of covariates. HOP offers more flexibility than OPM, since they capture the effect of the independent variables on the variance or uncertainty in the outcome (Lemp, et al., 2011).
- Bayesian ordered probit model (BOP): it is an extension of the Bayesian inference into the OPM, in which the parameters to be estimated are assumed to follow certain prior distributions. Based on the data, the likelihood function is used to update the prior distribution and obtain the posterior distribution (Xie et al., 2009).

Table 2.2: Studies that analyze injury severity of traffic accidents using Probit Models

Study Authors (publication year)	Model type	No. records	No. variables	Area type	Features	Injury level	No. injury levels	Model fit test
Renksi et al. 1999	OPM	27,29	7	mixed	SEG	KABCO	5	ρ^2 = 0.116 (Probably McFadden's)
Khattak, 2001	OPM	3,912	36	mixed	SEG	KABCO	5	ρ^2 (McFadden's) ρ^2= 0.0319, 0.0671, 0.0660 for drivers 1, 2, 3 respectively
Kockelman and Kweon, 2002	OPM	n.a.	13	n.a.	n.a.	KACO	4	ρ^2= (0.0451-0.0868) (McFadden's)
Khattak et al., 2002	OPM	17,045	16	mixed	BOTH	KABC	4	ρ^2= 0.057 (Probably McFadden's)
Abdel-Aty and Abdelwahab, 2004	OPM	7,891	12	mixed	SEG	K+A, BCO	2	accuracy=61.7%
Abdel-Aty and Keller, 2005	OPM	21,371	34	n.a.	INT	KABCO	5	ρ^2= 0.24 (Calculated McFadden's)
Oh, 2006	OPM	4,49	16	Rural	INT	KACO	4	ρ^2= 0.176 for all crashes model ρ^2= 0.480 for 3 or more vehicles ρ^2= 0.197 for 2 vehicles ρ^2=0.378 for single vehicle
Gårder, 2006	OPM	3,136	7	Rural	SEG	KABCO	5	n.a.
Gray et al., 2008	OPM	622,431	13	mixed	BOTH	KACO	4	LL =-33665.05 for London model LL =-267706.85 for UK
Xie et al., 2009	OPM, BOP	76,994	14	n.a.	INT	KABCO	5	accuracy for BOP for small data size= [55%-68%], for OP for small data size= [58%-68%], for BOP for predicted rest of the data= [61.8%-65.4%], for OP for predicted rest of the data =[59.9%-62.9%]
Wang et al., 2009	OPM	10,946	17	n.a.	INT	KABCO	5	ρ^2 = 0.0273
Haleem and Abdel-Aty, 2010	OPM, PM	2,043	21	mixed	INT	for the OPM: KABCO for the PM: K+A, BCO	5, 2	AIC= 17091 (OPM + 3-legged) AIC= 9423 (OPM + 4-legged) AIC= 3804 (PM + 3-legged) AIC= 2100 (PM + 4-legged)

Lemp et al., 2011	OPM, HOP	1,849	27	mixed	n.a.	KABCO	5	LL
								for OPM= -1993
								for HOP= -1896

Zhu and Srinivasan, 2011	OPM	953	28	mixed	BOTH	KA, B+C	2	ρ^2
								for PAR model= 0.1780
								for RES model= 0.1827

- n.a.: not available data; KABCO (K=killed, A=incapacitating, B=non-incapacitating, C=possible injury, O=no injury); AIS (0=no injury, 1=minor, 2=moderate, 3=serious, 4=severe; 5=critical; 6=unsurvivable); SEG=basic segment only, INT=intersection only, BOTH=intersection + segments
  Probably McFadden's: ρ^2 value was given in the study, and was found to apply to McFadden's formula; Calculated McFadden's: ρ^2 value was not given, but was calculated using log-likelihoods given in the study
- For more information about each study please refer to Appendix I.

2.1.2. Other models

Other models found to be used by researchers in the field of traffic accidents' injury severity included those that belong to the following techniques: Data mining, soft computing and others.

Data mining is defined as the process of discovering patterns in data. This process may be automatic or (more usually) semiautomatic. The patterns discovered must be meaningful in that they lead to some advantages (Witten and Frank, 2005). Many data mining techniques are being in use in different fields of science, economy, engineering, etc. Some of the data mining techniques that were found to be used in the analysis of injury severity of traffic accidents include: Decision trees (DT) and Bayesian networks (BN).

Soft computing is a mix of distinct methods which in a way or another cooperate in their fundamentals. The principal objective of soft computing is to exploit the tolerance for imprecision and uncertainty in order to achieve manageability, robustness and solutions at low cost (Zadeh, 1994). Models belonging to soft computing techniques that were used to analyze injury severity of traffic accidents include: Artificial neural networks (ANN) and Evolutionary algorithms (EA).

Other models which were found to be used in the literature include the log-linear model (LLM).

The following sections present a brief description of the models belonging to the aforementioned modeling techniques.

2.1.2.1. Decision trees (DT)

Decision trees are a kind of nonlinear predictive models that use the tree to represent the recursive partition. Each of the terminal nodes, or leaves, of the tree represents a cell of the partition, and has attached to it a simple model which applies in that cell only.

Within the literature of injury severity for traffic accidents, two types have been found to be used (see Table 2.3).

1. Classification and regression trees (CART): it constructs binary trees, in which each internal node has exactly two outgoing edges. CART can consider misclassification costs in the tree induction. It also enables users to provide prior probability distribution. An important feature of CART is its ability to generate regression trees, where the leaf predicts a real number and not a class (Rokach and Maimon, 2008).
2. Chi squared automatic interaction detection (CHAID): it is a procedure used to generate decision trees, and it was originally designed to handle nominal variables only. For each input variable, CHAID finds the pair of values that is least significantly different with respect to the target variable. The significant difference is measured by the *p*-value obtained by a statistical test. If this is positive, CHAID merges the values and searches for an additional potential pair to be merged, this process is repeated until no significant pairs are found (Rokach and Maimon, 2008).

2.1.2.2. Bayesian networks (BNs)

BNs are graphical models of interactions among a set of variables, where the variables are represented as nodes (also known as vertices) of a graph and the interactions (direct dependences) as directed links (also known as arcs and edges) between the nodes. Any pair of unconnected/nonadjacent nodes of such a graph indicates (conditional) independence between the variables represented by these nodes under particular circumstances that can easily be read from the graph. Each node contains the states of the random variable and it represents a conditional probability table. The conditional probability table of a node contains probabilities of the node being in a specific state given the states of its parents (Mittal and Kassim, 2007). More information about BNs can be obtained in section 2.2.

2.1.2.3. Artificial Neural Networks (ANN)

A neural network (NN) is an interconnected assembly of simple processing elements, units or nodes, whose functionality is based on the animal neuron. The processing ability of the network is stored in the inter-unit connection strengths, or weights, obtained by a process of adaptation to, or learning from, a set of training patterns (Gurney, 1997). NN are composed of neurons which in turn are composed of a number of inputs, and each input comes with a connection that has a weight and a threshold value.

Many types of ANN exist. A number of these types have been used by the researchers of accident severity analysis (see Table 2.3):

1. Multi-layer perceptron ANN (MLP): they usually consist of three layers: input layer, hidden layer, and output layer. The connections in MLP are feed-forward type in which they are allowed from a certain index to layers of a higher index. In order to train MLP, the back propagation algorithm is used (Rumelhart, et al., 1986).
2. Fuzzy ART MAP ANN (ARTMAP): they are based on adaptive resonance theory. It is a clustering algorithm that maps a set of input vectors to a set of clusters. Models built by fuzzy ARTMAP have a fast, stable learning in response to binary input patterns (Carpenter et al., 1992).

2.1.2.4. Evolutionary algorithms (EA)

EA mimic the natural evolution in order to optimize a solution to a problem (Brameier and Banzhaf, 2007). These algorithms exploit differential fitness advantages in a population of solutions to gradually improve the state of that population.

General EA is summarized as follows:

1. Randomly initialize a population of individual solutions
2. Select individuals from the population that are fitter than others using selection methods
3. Generate new variants by applying genetic operators.
4. If termination criterion is not met, then the best individual represents the best solution found

GP representation is usually executable and of variable size and shape. It is defined as any direct evolution or breeding of computer programs for the purpose of inductive learning. It might be considered as prediction models that approximate an objective function. Unlike other EAs, GPs can complete missing parts of existing model.

Linear genetic programming (LGP) is a GP variant that evolves sequences of instructions from an imperative programming language or from a machine language. Linear refers to the structure of the imperative program representation, where the nodes do not have to be linearly listed nor the method itself needs to be linear.

2.1.2.5. Others

The log-linear model (LLM) is one of the specialized cases of generalized linear models for Poisson-distributed data. It is an extension of the two-way contingency table where the conditional relationship between two or more discrete, categorical variables is analyzed by taking the natural logarithm of the cell frequencies within a contingency table. They are more commonly used to evaluate multi-way contingency tables that involve three or more variables. In log-linear models, there is no distinction between independent and dependent variables, all variables are treated as response variables. Therefore, log-linear models only demonstrate association between variables, where if one or more variables are treated as explicitly dependent and others as independent, or if the variables being investigated are continuous and cannot be broken down into discrete categories then LM should be used instead.

Table 2.3: Studies that analyze injury severity of traffic accidents using other techniques

Study Authors (publication year)	Model type	No. records	No. variables	Area type	Features	Injury level	No. injury levels	Model fit test
TREES								
Council and Stewart, 1996	CART	n.a.	7	mixed	BOTH	KBO	3	n.a.
Chen and Jovanis, 2000	CHAID	408	24	Rural	BOTH	KB	2	n.a.

Chang and Wang, 2006	CART	26,831	14	mixed	BOTH	KBO	3	accuracy
								for fatal (0%),
								for injury (94%),
								for no injury (68%)

BAYESIAN NETWORKS								
Simoncic, 2004	BN	17,558	12	mixed	n.a.	K+A, other	2	n.a.
NEURAL NETWORKS								
Abdelwahab and Abdel-Aty, 2001	MLP, Fuzzy ARTMAP	1,168	14	n.a.	INT	A, B+C, O	3	accuracy =65.6%

Abdel-Aty and Abdelwahab, 2004	MLP, Fuzzy ARTMAP	7,891	12	mixed	SEG	K+A, BCO	2	accuracy
								for MLP= 73.5%
								for fuzzy ARTMAP 40.6%

Delen et. al., 2006	MLP	30,358	13	n.a.	n.a.	KABCO	5	accuracy= 40.73%
LINEAR GENETIC PROGRAMMING								
Das and Abdel-Aty, 2010	LGP	104,952	58	mixed	BOTH	B+O, A+C	2	accuracy= 60.4%
OTHERS								
Chen and Jovanis, 2000	LLM	408	24	Rural	BOTH	KB	2	R^2= 0.95

- n.a.: not available data; KABCO (K=killed, A=incapacitating, B=non-incapacitating, C=possible injury, O=no injury); AIS (0=no injury, 1=minor, 2=moderate, 3=serious, 4=severe; 5=critical; 6=unsurvivable); SEG=basic segment only, INT=intersection only, BOTH=intersection + segments
  Probably McFadden's: ρ^2 value was given in the study, and was found to apply to McFadden's formula; Calculated McFadden's: ρ^2 value was not given, but was calculated using log-likelihoods given in the study
- For more information about each study please refer to Appendix I.

2.1.3. Discussion

The usage of the modeling techniques varied with time, thus DOM (logit and probit models) continued to be dominant along the years. Since 2001 new methods started to be applied in the analysis of injury severity, such as Neural Networks, Bayesian Networks, and most recently Genetic Algorithms.

In this research work 19 modeling techniques used to model injury severity of traffic accidents, applied in 57 case-studies, have been analyzed. The most used techniques are the DOM (46 cases), highlighting the models BLM, OLM and OPM over all the others. These three models were used in more than 54% of the cases.

As illustrated in the previous section, many studies analyzed and modeled injury severity of traffic accidents. Figure 2.1 shows the frequency of usage of the different models which have been applied to the analysis of traffic accidents' injury severity where it is illustrated that the most used are the logit and probit, followed by other models. However, Tables 2.1, 2.2 and 2.3 show that there is a large dispersion in the principal magnitudes used in these models (number of records considered in the analysis, number of variables for analyzing severity, type of the analyzed roadway segment, area type in which the roadway exist, type and number of injury levels, model fit parameters, and type of model).

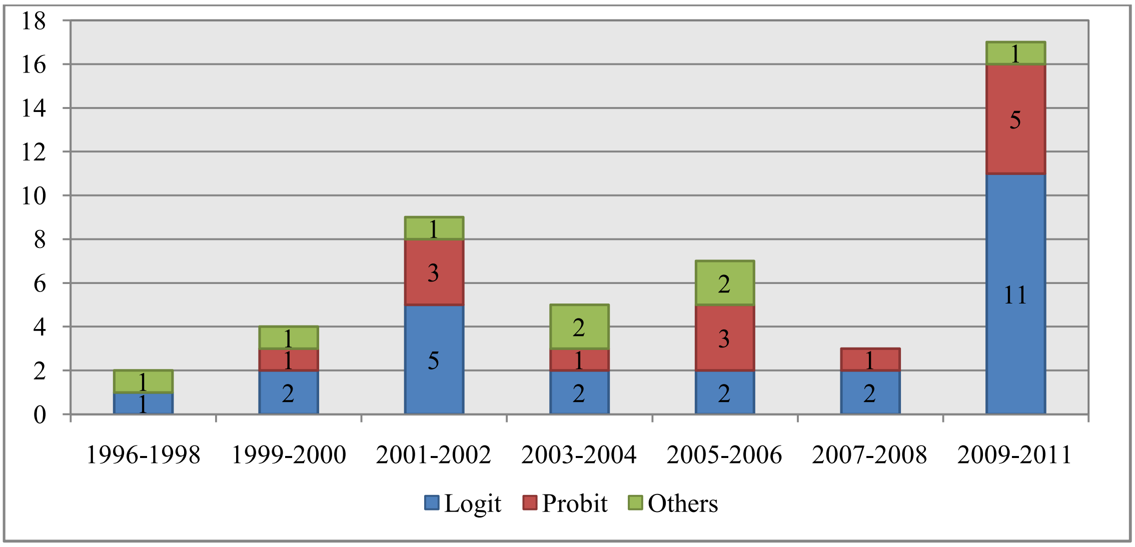


Figure 2.1: number of studies performed (1996-2011)

To have a further insight about the modeling techniques used in the literature to analyze injury severity, a statistical analysis is done, using the Mann-Whitney test, to test the significance of using a certain number of variables, injury severity levels, or number of accidents (Table 2.4).

Table 2.4: Maximum, minimum and median for number of accident records, number of variables and number of injury levels

	No. of accidents' records	No. of Variables	No. of Injury Levels
All studies			
Max	81172	36	7
Min	255	5	1
Median	3,998	15	3
Logit studies			
Max	81,172	30	7
Min	255	5	1
Median	4,552	15	3 a
Probit studies			
Max	76,994	36	5
Min	449	7	2
Median	3,136	16	5 b
Other studies			
Max	30,358	24	5
Min	408	7	2
Median	12,088	13.5	2a

a, b: denotes differences statistically significant ($p < 0.05$), based on Mann-Whitney U test

2.1.3.1. Number of records

The number of records considered in the analysis ranges between 255 and 622,432. However, four extreme outliers were identified. Extreme outliers are defined as any data values which lie more than 3 times the interquartile range below the first quartile or above the third quartile (Montgomery and Runger, 2003). Table 2.4 shows the statistical analysis results without extreme outliers.

Table 2.4 shows that the number of records (without extreme outliers) for all the studies ranges between 255 and 81,172, with a median value of 3,998. PM present the lowest median, with 3,136 crashes, followed by LM with (4,552) records. The other models present the highest median (12,088) for the number of records considered in the analysis.

All the values are very similar and no significant statistical difference was observed between the three groups (logit, probit and others) based on the Mann-Whitney U test.

Thus, it is quite important to mention that there is no objective criterion for the choice of the optimum subset size. Intuitively, one would choose the optimum subset size to be the minimum size that does not cause a meaningful decrease in the statistic describing the predictive quality of the model, e.g. R^2, ρ^2, etc. (Freund et al. 2006).

2.1.3.2. Number of variables

The number of variables used in the modeling of injury severity range between 5 and 58. However, one extreme outlier was identified, which was not considered for the statistical analysis.

Table 2.4 shows that the number of variables (without extreme outliers) for all the studies ranges between 5 and 36, with a median value of 15. PM present the highest median, with 16 variables, followed by LM and the other models with 15 variables.

All the values are very similar and no significant statistical difference was observed between the three groups (logit, probit and others) based on the Mann-Whitney U test.

No correlation was found to exist between the number of variables for analyzing severity and the number of records considered in the analysis.

Thus, it is rather important to mention that leaving out variables that should be in the model, results in inflation of the error variance and such a model is said to be underspecified. Since this is a rather well-known result, one common practice for avoiding such results is to put into an initial model all conceivably relevant variables, with the number of variables often restricted only by the availability of data. When this procedure is followed, it usually happens that the initial model contains too many variables; that is, some of these variables are not needed in the sense that they do not contribute to the fit of the model. Such a model is said to be overspecified. Therefore, if the researcher does not know which variables are not needed, any selection of variables must be based on the data (Freund et al. 2006).

2.1.3.3. Type and number of categories for injury levels

The definition of the injury severity might refer to the emphasis of the study (Krull et al., 2000), for convenience (Ouyang et al., 2002), or because of the small counts of certain category with respect to other categories (Peek-Asa et al., 2010).

Most of the studies used the KABCO scale, which is the scale used in police observed accident records (Morgan, 2009). Others combined one or more categories into one, in which the more severe cases were put into one category and the least severe were put into another.

Table 2.4 shows that the number of injury levels for all the studies ranges between 1 and 7, with a median value of 3. In this case no extreme outliers were identified. PM present the highest median, with 5 levels, followed by LM with 3 levels. The other models present the lowest median (2) for the number of injury levels.

In this case, significant statistical differences were observed ($p<0.05$), based on the Mann-Whitney U test, between the number of injury levels considered in the probit models with respect to the other two types of models (logit and others).

2.1.3.4. Focus of the study

There is a relatively high dispersion in the type of analyzed roadway segment. Figure 2.2 shows that 13 studies analyzed only basic segments, 9 studies analyzed only intersections, and 13 studies analyzed both intersections and roadways segments in the same study. So, the number of studies found to use mixed features is relatively high. However, Moore et al. (2010) recommended that intersections and road segments should not be analyzed together, since the factors related to accidents occurring on intersections are different from those related to roadway segments.

Something similar occurs with the area type in which the roadway or the intersection exists. Only five studies analyze rural areas, while 22 studies mixed the data for rural, urban and/or suburban highways, keeping in mind that the characteristics of the roadways and intersections differ significantly (mostly in the number of access points and the traffic volume) upon if they were in an urban or in a rural area.

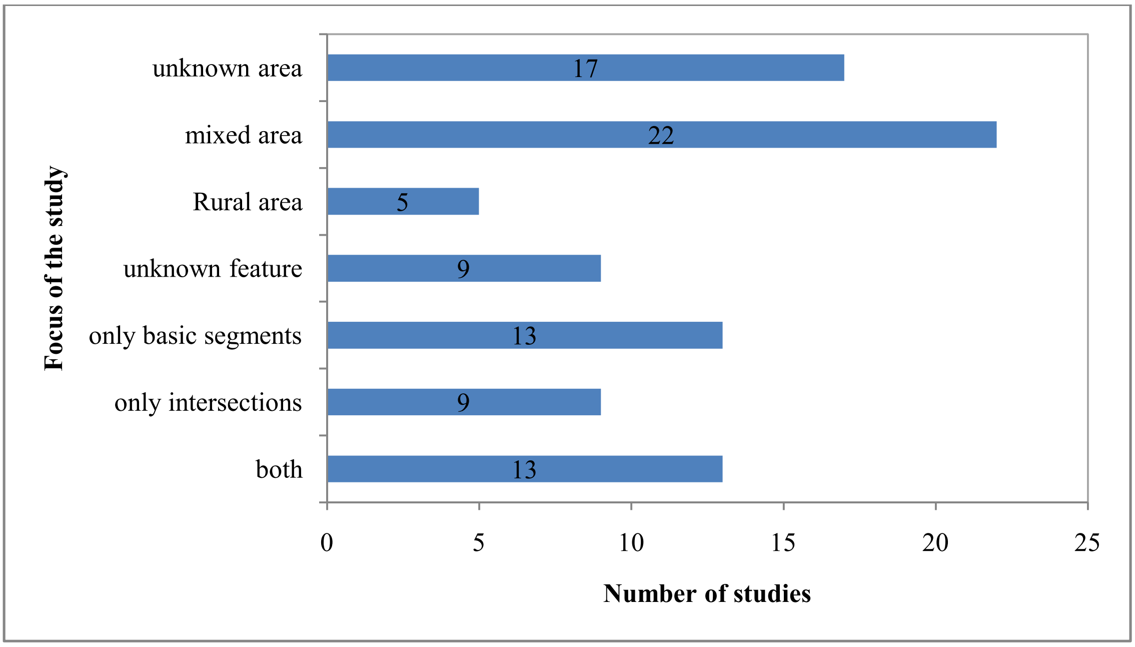


Figure 2.2: number of studies performed according to the focus of the study

2.1.3.5. Model fit

In general, the statistical tests used to validate the performance of the model vary with the study. Each modeling technique uses a specific type of model fit test or parameter. These tests permit indicating if the model fits the data adequately or not. But the number of different tests and their characteristics does not permit comparing the results of one study with another.

In the following subsections, descriptions of the fit parameters most commonly used for the analyzed models are presented:

2.1.3.5.1. R-squared

R-squared (R^2) is a statistic that is generated in ordinary least squares (OLS) regression that is often used as a goodness-of-fit measure. This statistic is used to measure the level of explanation of the dependent variable by the independent variable(s) that the model provides. Its value ranges between 0 and 1, where 1 indicates a high level of explanation of the variance as explained by the regression model and zero as a low level of explanation (Bruin, 2006).

Chen and Jovanis (2000) used R^2 to test LLM fit (see Table 2.3). The results indicated that the log-linear model fitted the data very well (R^2=0.95).

2.1.3.5.1. Pseudo R-square

When analyzing data with a logistic regression, an equivalent statistic to R-squared does not exist. The model estimates from a logistic regression are maximum likelihood estimates arrived at through an iterative process. They are not calculated to minimize variance, so the OLS approach to goodness-of-fit does not apply. However, to evaluate the goodness-of-fit of logistic models, several pseudo R-squareds (ρ^2) have been developed (McFadden, adjusted McFadden, Efron's, Cox & Snell, Negelkerke/ Cragg & Uhler's, McKelvey & Zavoina, Count, adjusted count, etc.). They look like R^2 in the sense that they are on a similar scale, ranging from 0 to 1 (though some ρ^2 never achieve 0 or 1) with higher values indicating better model fit (Bruin, 2006).

In this thesis (see Table 2.1, 2.2 and 2.3) several studies used a ρ^2 to test the fit of their models. Other studies supplied the log-likelihood (LL) of the model for both the model with no parameters and LL calculated at convergence. Thus, when information about LLs was available, we calculated the McFadden ρ^2 in order to homogenize the model fit results.

Anyway, ρ^2 cannot be interpreted as one would interpret a R^2. Bruin (2006) indicated that ρ^2 cannot be interpreted independently or compared across datasets; a ρ^2 only has meaning when compared to another ρ^2 of the same type, on the same data, predicting the same outcome. In this situation, the higher ρ^2 indicates which model better predicts the outcome. So, it is only possible to make comparisons within models in the same study. However, it is possible to indicate if the results of a study are among the satisfactory range for such parameter for model fit. As indicated by McFadden (1979), the satisfactory range for the McFadden's ρ^2 lies between 0.2 and 0.4. Results above 0.4 are considered to be superior.

Most of the models in this paper that use McFadden's ρ^2 present values below 0.2. However, there are five studies that present values over 0.2: Shanker et al. (1996) (ρ^2=0.52), Wang and Kockelman (2005) (ρ^2=0.235-0.257), Abdel-Aty and Keller (2005) (ρ^2=0.24), Oh (2006) (ρ^2=0.378-0.480) and Schneider et al. (2009) (ρ^2= 0.23-0.258).

2.1.3.5.3. Accuracy

Accuracy measures the percentage of cases in the accident data correctly predicted by each injury severity expression in each model. Therefore, accuracy is obtained at the case-specific level, that is, cases that are correctly classified as e.g. fatal or nonfatal according to their observed injury experience (Saccomanno et al., 1996).

Most of the studies used this parameter to test the capability of their models to correctly classify the injury severity into a specific level (see Table 2.1, 2.2 and 2.3). Global accuracy range lies between 0.41 and 0.89. The highest global accuracy achieved was for a BLM model built by Dissanayake and Lu (2002) and the lowest global accuracy was obtained by Abdel-Aty and Abdelwahab (2004) when they constructed a Fuzzy ARTMAP ANN model.

Having a look on the study with the highest accuracy (Dissanayake and Lu, 2002) indicates that the number of accident classified under each severity level was homogeneous along all the levels. On the other hand, the lowest accuracy obtained for a specific level (fatal accidents) with a CART model was practically zero (Chang and Wang, 2006). The authors referred this result to the fact that their dataset was imbalanced, where the fatal accidents accounted only for about 0.4% of the whole sample used to build the model.

Delen et al. (2006) also obtained relatively low accuracy results (40.7%) for their model (MLP ANN). They explained their results by a multi-class classification problem in which the cases in a data that is highly skewed (unbalanced among the class labels) and the problem domain is complex. A possible solution would be in such cases by reducing the multi-class problem into series of two-class (binary) classification problems. Applying such a solution, the complete dataset was separated into eight subsets with binary output variables. In which a top-down (more serious injury versus the less serious injuries) and a bottom-up (less serious injury versus more serious injuries) were built. The method applied by Delen et al. (2006) could be compared to that applied by Dissanayake and Lu (2002) where, after developing a series of binary logistic regression models, they chose to interpret the coefficients of the “best fit” model as opposed to looking at the problem from a progression of severity perspective (disaggregating the model containing all the injury severity levels into various models, each representing a binary injury severity level). The average accuracy of the eight models improved to 77%.

2.1.3.5.4. Other Measures

Other measures used to test the model fit are: Akaike Information Criterion (AIC), log-likelihood (LL), chi-squared (χ^2), and Kendall rank correlation coefficient (Kendall's tau (τ) coefficient).

The likelihood is the probability of the data given the parameter estimates. The goal of a model is to find values for the parameters (coefficients) that maximize value of the likelihood function, that is, to find the set of parameter estimates that make the data most likely. Many procedures use the log-likelihood, rather than the likelihood itself, because it is easier to work with, where, higher values of the log likelihood indicate a better fitting model (Bruin, 2006). Two studies used the LL as the only fit test. However, only Lemp et al. (2011) used it to compare two models (OPM against HOP).

The Akaike Information Criterion (AIC) is used only once (Haleem and Abdel-Aty, 2010) to select a model from a set of models. The chosen model is the one that minimizes the Kullback-Leibler distance between the model and the truth (low AIC and high LL indicates good fit). It's based on information theory, but a heuristic way to think about it is as a criterion that seeks a model that has a good fit to the truth but few parameters (Burham and Anderson, 2002).

χ^2 test is used to verify if a sample of data came from a population with a specific distribution. χ^2 is applied to binned data (i.e., data put into classes). However, the value of χ^2 statistic is dependent on how the data is binned. Another disadvantage of χ^2 is that it requires a sufficient sample size in order for χ^2 approximation to be valid (NIST/SEMATECH, 2003). Three studies (Srinivasan, 2002; Malyshkina and Mannering, 2009; and Daniels et al., 2010) were found to use χ^2 as the model goodness of fit test.

Only Donelson, et al. (1999) used Kendall's tau (τ) coefficient, which is a statistic used to measure the association between two quantities. A τ test is a non-parametric hypothesis test which uses the coefficient to test for statistical dependence Kruskal (1958). The results indicated that the BLM fitted the data used.

2.1.3.6. Modeling techniques

The usage of the modeling techniques varied with time, thus DCM (logit and probit models) continued to be dominant along the years. Since 2001 new methods started to be applied in the analysis of injury severity, such as Neural Networks, Bayesian Networks, and most recently Genetic Algorithms.

Table 2.5 shows the frequency of usage of each model. In this paper 19 modeling techniques used to model injury severity of traffic accidents, applied in 57 case-studies, have been analyzed. The most used techniques are the DOM (46 cases), highlighting the models BLM, OLM and OPM over all the others. These three models were used in more than 54% of the cases.

Table 2.5: Frequency of usage of each model

	Family of models	Model	Frequency
Discrete Outcome Model	Logit Models	BLM OLM HL MNL HKL MXL GEE OMXL	11 6 4 3 2 1 1 1
	Probit Models	OPM BOP HOP PM	14 1 1 1
Other Models	Decision Trees	CART CHAID	2 1
	Bayesian Networks	BN	1
	Artificial Neural Networks	MLP Fuzzy ARTMAP	3 2
	Evolutionary algorithms	LGP	1
	Log-linear models	LLM	1

2.1.3.6.1. Logit Models

Table 2.5 shows that the most used logit models are BLM followed by OLM, while the least used were MXL, GEE and OMXL.

The frequent usage of BLM to analyze accident severity might refer to the fact that most of the studies used the response variable as binary. Even when there was more than one category for the response variable, researchers still have divided these categories into distinct models each representing the specific category (Dissanayake and

Lu, 2002; Delen et al., 2006; Jung et al., 2010). This refers to the fact that BLM is easily interpretable.

On the other hand, OLM uses the logistic distribution on ordered alternatives, where the probabilities in OLM incorporate the BLM formula. The OLM has one utility with multiple alternatives to represent the level of that utility, while the BLM has two alternatives with utility for each one (Train, 2009). Consequently, using the OLM enables the researcher to make comparisons among the different alternatives available, while using the BLM does not offer this advantage.

An OLM restriction is that regression parameters have to be the same for different accident severity levels, called proportional odds. However, it is not always clear if the distance between accident severity levels is equal, and hence it is arbitrary to assume that all coefficients of ordered probability models are the same (Jung et al., 2010).

Moreover, Srinivasan (2002) stated that the primary restriction of the ordered models comes from the assumption of deterministic thresholds that are often identical across all observations for each ordinal response level or category. Also, it is assumed that the response is homogeneous and independent from exogenous variables. In addition, these models disregard possible correlations across the thresholds of different alternatives.

Consequently, the aforementioned assumptions could lead to significant bias and inconsistency in ordered response models. Therefore, and in order to overcome these restrictions, Srinivasan (2002) used an OMXL, where she compared OMXL to OLM using a χ^2 test. The results indicated that the χ^2 test rejected the restrictive OLM.

Lenuguerrand et al. (2006) compared different models (BLM, HL and GEE) to determine which deals best with the analysis of accident severity. HL was found to be more adequate for problems with correlated data than BLM and GEE, clusters and sub-clusters, since BLM and GEE both underestimate parameters and confidence intervals. Thus, they recommended the use of HL when the number of vehicles per accident or the number of occupants per accident is high.

2.1.3.6.2. Probit Models

The most frequently used probit model is the OPM (see Table 2.5). OPM have been used to model injury severity of accidents on roadways and intersections. Some researchers have used models that combined the accident occurring at intersections with accidents off intersections (Gray et al., 2008; Xie et al., 2009; Zhu and Srinivasan, 2011).

OPM proved to be a good choice for modeling injury severity of accidents, and even when compared with other models such as the BOP, the OPM still performed as well (Xie et al., 2009). BOP and OPM produced similar results for large data size, where they recommended using BOP for smaller data sizes, as it can produce more reasonable parameter estimation and better prediction performance. On the contrary, when comparing OPM with HOP, the HOP was preferred over the OPM in terms of log-likelihoods (Lemp et al., 2011).

Haleem and Abdel-Aty (2010) used OPM, PM and HL to analyze accident injury severity at intersections. The results indicated that the PM fits the data better than the OPM (lower AIC and higher log-likelihood). The model built using HL had fewer significant variables, the goodness-of-fit criterion gets worsened, and unexpected signs of variables were found.

2.1.3.6.3. Other modeling techniques

Classification and regression trees (CART) were used by Council and Stewart (1996) and Chang and Wang (2006) to model injury severity of accident. CART suits the injury severity analysis due to the fact that there is no need to a pre-defined relationship between the response variable and the predictors. CART results are simply interpreted and they rapidly classify new observations. CART can also consider misclassification costs in the tree induction; it is also capable to generate regression trees (Rokach and Maimon, 2008). The results presented by Chang and Wang (2006) indicated that CART can effectively handle multi-collinearity problems, and they could handle the outliers that exist in the data by isolating them into a node.

However Chang and Wang (2006) indicated that one of the problems with applying the CART is that they do not provide confidence intervals for the risk factors (splitters) and predictions. Also, they have a difficulty in applying the sensitivity analysis which does not permit examining the marginal effects of the predictors on the response variable. In addition the CART models are unstable, the structure and the accuracy alter if different strategies are followed to create learning and test sets. Moreover, the results presented by Chang and Wang (2006) showed that CART correct classification of the fatal injuries was 0%. The authors explained this result by the fact that their dataset was imbalanced.

BNs were used by Simoncic (2004) to model injury severity of accident. The work presented by Simoncic (2004) based the conclusion upon one single network, which was not validated using a test set, neither being compared to other possible configurations. However, the method presented highlighted the advantages that BNs offer to analyze injury severity in which the results of modeling could be viewed in a graphical manner. In addition, the correlations that exist between different variables could be determined graphically, as well as, the conditional probabilities of each variable.

MLP and Fuzzy ARTMAP ANNs have been compared twice to analyze injury severity on road segments and on intersections (Abdelwahab and Abdel-Aty, 2001; Abdel-Aty and Abdelwahab, 2004). Both studies indicated that MLP ANN performance is superior to Fuzzy ARTMAP ANN. Their performance on roadway segments was better than that on signalized intersections. Delen et al. (2006) also used MLP ANN to model injury severity on roadways. They used more injury levels and their results (in terms of accuracy) were worse than those of previous studies (Abdelwahab and Abdel-Aty, 2001; and Abdel-Aty and Abdelwahab, 2004).

Abdelwahab and Abdel-Aty (2001) compared the performance of MLP ANN and fuzzy ARTMAP ANN with the performance of OLM. Their results indicated that the best in terms of accuracy was the MLP ANN followed by OLM, and finally by fuzzy ARTMAP. Thus, OLM was superior in performance with respect to certain types of ANNs. Abdel-Aty and Abdelwahab (2004) used MLP ANN, fuzzy ARTMAP and OPM. The results showed once again the superiority of MLP ANN over all the other techniques, however, this time OPM did not perform better than the fuzzy ARTMAP.

2.1.4. Summary

The studies included in this research work are only those that have analyzed and modeled the injury severity of traffic accidents from an engineering perspective. Those related to medical research were excluded, as well as those related to analyzing pedestrian, cyclists, and traffic accidents that occurred in urban areas only.

Many modeling techniques have been in use to analyze the injury severity of accidents. The most used models are the logit and probit (discrete outcome models). However, in recent years, methods based on data mining techniques (decision trees and Bayesian networks), as well as other models based on soft computing techniques (artificial neural networks and linear genetic programming), have started to gain acceptance in their application to the analysis of traffic accidents.

Within the discrete outcome models, the most used are OPM, BLM and OLM. BLM are commonly used when the study uses a binary variable for severity. OPM and OLM are used when the number of injury severity levels increases.

There is a large diversity in the number of accidents' records and the number of variables used in the models for analysis. The number of records range from 255 to 622,432 accidents, with a median value of 3,998 records (without extreme outliers). The number of variables range from 5 to 58 variables, with a median value of 15 (excluding an extreme outlier). However, no significant statistical difference was found to exist in both cases between logit, probit and other models. No correlation was found to exist between these two variables. The number of records and the number of variables are found to be mostly dependent upon the availability of data.

Information provided by the accident injury reports has a great influence in the number of injury levels considered in the study. Most of the studies use the KABCO scale or a modification. The number of injury levels for all studies range between 1 and 7, with a median value of 3. However, the probit models use a higher number of injury levels (5) than the logit models (3 levels) or the rest of the models (2 levels). In this case, significant statistical differences were observed ($p<0.05$) between the probit models and the other types of models.

The model fit results are satisfactory in most of the cases (e.g. global accuracy in the range of 0.41 and 0.89; McFadden's pseudo R-square values between 0.2 and 0.4), although it can also be observed some exceptional results (e.g. Chen and Jovanis (2000) obtained $R^2=0.95$), while others were not so satisfactory (e.g. many studies with McFadden's pseudo R-square below 0.2).

Different factors affect the accuracy obtained by a model, such as the balance of cases among the different categories that lie under the injury severity levels. If the number of observed cases classified among the different levels is not relatively different, this identifies a balanced dataset; and accuracy would improve since the classification will not be biased towards a specific injury severity level.

In general, when the number of injury levels increases the results of the model fit gets worse, since increasing the number of alternatives increases the variance in the data so the statistical fit will necessarily get worse. This could explain why many studies, aggregate the levels of injury severity in order to reduce them into 2 or 3 levels after initial estimations due to convergence problems resulting from insufficient variation among discrete-injury outcomes. (Krull. et al., 2000; Abdelwahab and Abdel-Aty, 2001; Ouyang et al., 2002; Abdel-Aty and Abdelwahab, 2004; Simoncic, 2004; Milton et al., 2008; Das and Abdel-Aty, 2010; Haleem and Abdel-Aty, 2010; Jung et al., 2010; Zhu and Srinivasan, 2011).

In general, it is not possible to identify which is the best method to be used. Using a certain model might be suitable under certain circumstances, while not under others. Many examples are available in the literature (Lenuguerrand et. al., 2006; Xie et al., 2009). Probably this is one of the main reasons why, in recent years, the number of studies that analyze injury severity of traffic accidents has greatly increased.

Finally, the literature indicated that there exists a large diversity on the possible modeling techniques to be used, in which each technique has its advantages and disadvantages. No accordance on the best technique to be used was found, researchers can choose a modeling technique that applies to their specific conditions, based on the dataset characteristics and size, number of variables used, etc. Thus, experimenting new modeling techniques is still an open field for researchers.

In addition, special caution should be made when selecting relevant variables; it is advised to use statistical methods to do the selection process in order to avoid problems related to relevance of selected variables.

2.2. Bayesian networks

2.2.1. What is a Bayesian network?

Bayesian networks (BN) belong to the family of probabilistic graphical models; they are used to represent knowledge about uncertain domain. In general graphical models' structures are composed of nodes that represent random variables, and edges between these nodes that represent probabilistic dependencies among the corresponding random variables.

BNs correspond to a structure of the graphical models called directed acyclic graph (DAG). They enable an effective representation and computation of the joint probability distribution over a set of random variables.

The structure of a DAG is defined by two sets: the set of nodes (vertices) and the set of directed edges. The nodes represent random variables and are drawn as circles labeled by the variable names. The edges represent causality, relevance or direct dependencies between variables and are drawn by arrows between nodes. In particular, an edge from node X_i to node X_j represents a statistical dependence between the corresponding variables. Thus, the arrow indicates that a value taken by variable X_j depends on the value taken by variable X_i. Node X_i is then referred to as a parent of X_j and, similarly, X_j is referred to as the child of X_i. The DAG structure of the BN represents the qualitative component of the BN. However, even though the arrows represent direct dependence connection between the variables, the reasoning process can operate on BNs by propagating information in any direction. Figure 2.3 illustrates the qualitative component of a BN (Ruggeri, et al. 2007).

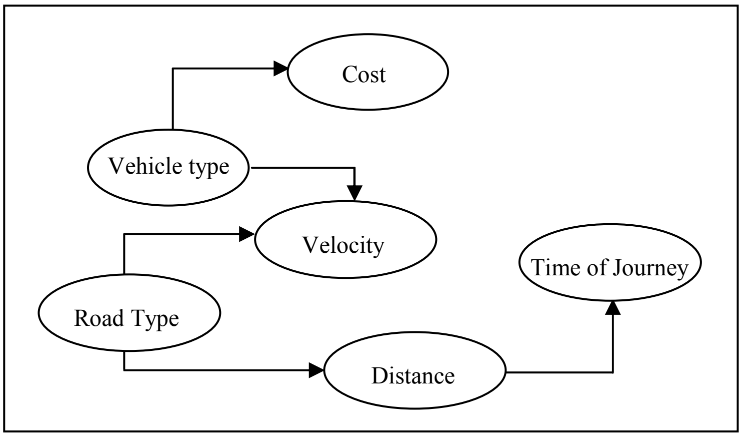
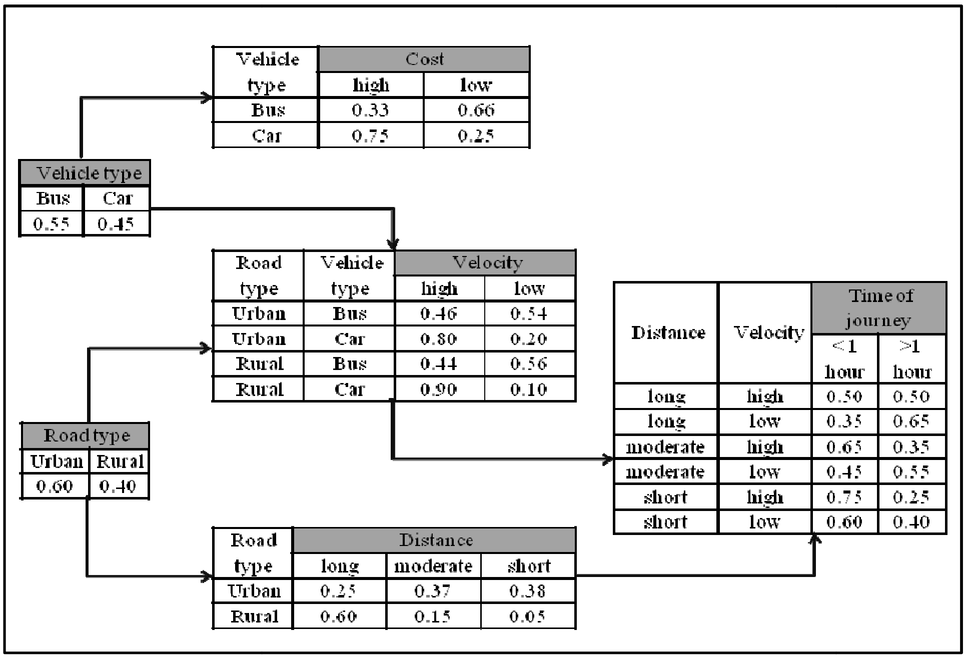


Figure 2. 3: An example of a BN graph structure

In addition to the qualitative component, BNs have a quantative component. The quantative component represents the conditional probability distribution (CBD) at each node, which measures the uncertainty of existing relations between variables, in which the CBD at each node depends only on its parents. For discrete random variables, the conditional probability is represented by a table that lists the local probability that a child node takes on each of the possible values for each combination of values of its parents. Thus, the joint distribution of a collection of variables can be determined by these local conditional probability tables (CPTs). Figure 2.4 illustrates CPT for the DAG given in Figure 2.3.



Vehicle type	
Bus	Car
0.55	0.45

Vehicle type	Cost	
	high	low
Bus	0.33	0.66
Car	0.75	0.25

Road type	
Urban	Rural
0.60	0.40

Road type	Vehicle type	Velocity	
		high	low
Urban	Bus	0.46	0.54
Urban	Car	0.80	0.20
Rural	Bus	0.44	0.56
Rural	Car	0.90	0.10

Road type	Distance		
	long	moderate	short
Urban	0.25	0.37	0.38
Rural	0.60	0.15	0.05

Distance	Velocity	Time of journey	
		<1 hour	>1 hour
long	high	0.50	0.50
long	low	0.35	0.65
moderate	high	0.65	0.35
moderate	low	0.45	0.55
short	high	0.75	0.25
short	low	0.60	0.40

Figure 2. 4: An example of a CPT for a BN

2.2.1.1. Previous concepts

The mathematical foundation for all applications of probability is considered the one that was developed by Kolmogorov (1950).

Probability theory is related to experiments that have a set of distinct outcomes, in which the experiment is not considered to be well defined until a set of outcomes is identified, and that the same act can be associated with many different experiments, depending on what is considered to be a distinct outcome. The sample space of an experiment could be set as soon as the experiment be well defined. From a mathematical point of view, a sample space is a set and the outcomes are the elements of the set (Neapolitan, 2003).

One of the most important concepts in probability theory is the *conditional probability* which is defined as follows:

$$P(E \middle| F) = \frac{P(E \cap F)}{P(F)} \tag{4.1}$$

Where,

E and F are events such that $P(F) \neq 0$.

If there exists n events $E_1, E_2, \ldots E_n$ such that $E_i \cap E_j = \Phi$ for $i \neq j$ and $E_1 \cup E_2 \cup \ldots \cup E_n = \Omega$. Such events are called *mutually exclusive and exhaustive*. The *law of total probability* says for any other event F,

$$P(F) = \sum_{i=1}^{n} P(F \cap E_i) \tag{4.2}$$

However, conditional probabilities of events of interest are usually computed with known probabilities using Bayes' theorem (Neapolitan, 2003). The Bayes theorem states the following:

Given two events E and F such that $P(E) \neq 0$ and $P(F) \neq 0$, then,

$$P(E \mid F) = \frac{P(F \mid E)P(E)}{P(F)} \tag{4.3}$$

Two random variables induce a probability function on the Cartesian product of their spaces. In which, P (X = x, Y = y) is called the *joint probability distribution* on X and Y.

Given a joint probability distribution, the law of total probability (Equation 2.2) implies the probability distribution of any one of the random variables can be obtained by summing over all values of other variables. In which, if exists a joint probability P (X = x, Y = y) then,

$$P(X = x) = \sum_{y} P(X = x, Y = y) \tag{4.4}$$

Where,

$\sum_{y}$: Means the sum as y goes through all values of Y.

P(X =x): is called the *marginal probability distribution* of X, because it is obtained using a process similar to adding across a row or column in a table of numbers.

2.2.1.2. The Markov Condition

A *directed graph*, is a pair (V, E) where V is a finite, non empty set, and its elements are called *nodes* (or vertices), and E is a set of ordered pairs of distinct elements of V. Elements of E are called *edges* (or arcs), and if (X, Y) ∈ E, then it is said that there is an edge from X to Y and that X and Y are each *incident* to the edge.

A directed graph $\mathbb{G}$ is called a directed acyclic graph (DAG) if it contains no directed cycles. Figure 2.5 illustrates an example of a DAG a non-DAG.

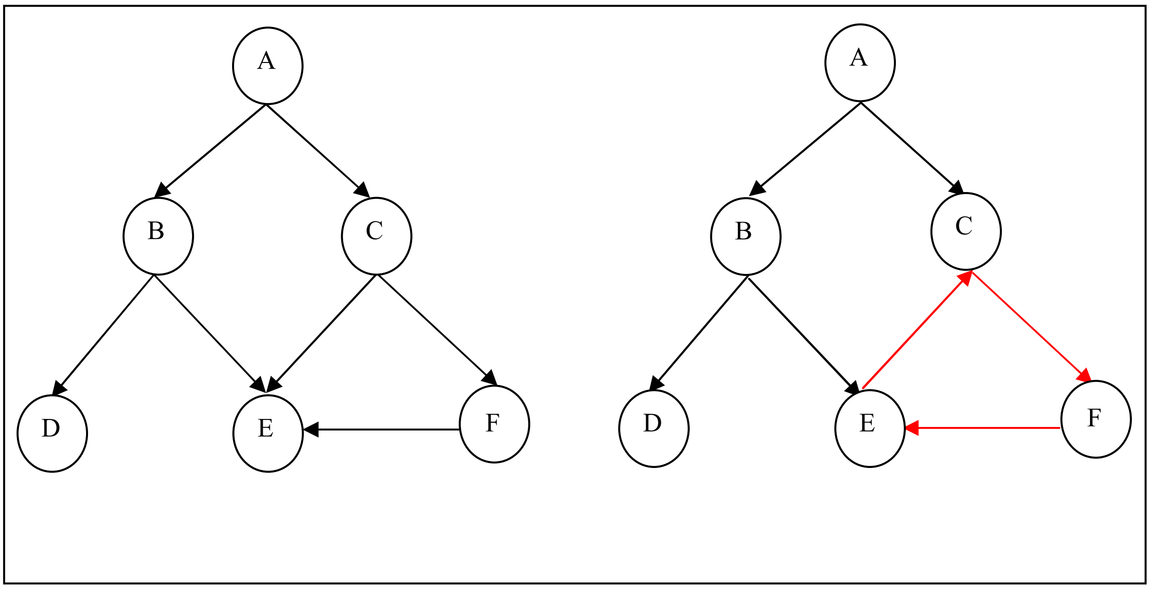


Figure 2. 5: An example illustrating the difference between DAG and a non-DAG

Give a DAG $\mathbb{G}$ = (V, E) and nodes X and Y in V, Y is called a *parent* of X (PA_x). Y is called a parent of X (PA_x), if there is an edge from Y to X. Supposedly there is a joint probability distribution P of the random variables in some set V and a DAG $\mathbb{G}$ = (V, E). It is said that ($\mathbb{G}$, P) satisfies the *Markov condition* if for each variable X ∈ V, {X} is conditionally independent of the set of all its nondescendents given the set of all its parents.

The number of terms in a joint probability distribution is exponential in terms of the number of variables. Therefore, in the case of a large number of instances, the joint distribution could not be fully described by determining each of its values directly; this is when Markov condition function powerfully.

2.2.2. Types of BNs

There are two types of BNs, a causal BN and a non-causal BN; in the following sub-sections a brief description of each is presented.

2.2.2.1. Causal Bayesian Networks

Given a set of random variables V, if for every X, Y ∈ V an edge is drawn from X to Y if and only if X is a direct cause of Y relative to V, the resultant DAG is called a *causal DAG*.

Causality is an operational method for identifying causal relationships. That is, if the action of making variable X take some value sometimes changes the value taken by variable Y, then it is assumed that X is responsible for sometimes changing Y's value, and it is concluded that X is a cause of Y. Therefore, it is assumed that causes and their effects are statistically correlated. However, variables can be correlated without one be the cause of the other.

2.2.2.2. Non-Causal Bayesian Networks

Causal edges are one way to develop a DAG and a probability distribution that satisfies the Markov condition. The Markov condition is simply a property of the probabilistic relationships among the variables.

For instance, as Figure 2.6 shows that the joint distribution of Value, Shape, and Color satisfied the Markov condition, however, it would not be acceptable to say that the color of an object has a causal influence on its shape. The Markov condition is simply a property of the probabilistic relationships among the variables.

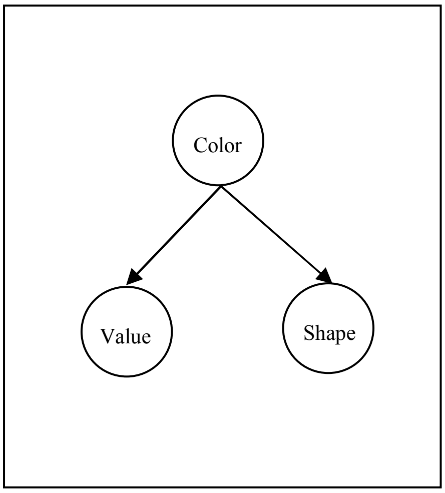


Figure 2. 6 : A non-causal BN

2.2.3. Learning Bayesian Networks

There are many techniques for automatic learning such as, rote learning, learning by instructions, learning by deduction, learning by analogy, discovery learning, and inductive learning (learning from examples).

Constructing Bayesian networks could be done by experts, learned from data, or using a combination of both techniques, however, constructing Bayesian networks from experts can be slow, subjective, expensive, and even difficult in a case of large networks or huge amount of data.

Network learning involves structure (DAG) learning and parameter learning (conditional probability estimation). The structure learning is to find a graphical relationship of the variables and the parameter learning is to determine quantitatively how they relate to each other. Generally the structure learning is much more difficult than parameter learning.

When there are masses of data available and it is necessary to interpret them and to provide a model for predicting the behavior of unobserved cases, the learning of both structure and parameters is used (Cooper & Herskovits, 1992). There are two main approaches to structure learning in BNs:

• Constraint based: Perform tests of conditional independence on the data, and search for a network that is consistent with the observed dependencies and independencies.

• Score based: Define a score that evaluates how well the dependencies or independencies in a structure match the data and search for a structure that maximizes the score.

The advantage of score-based methods over the constraint based methods is that they are less sensitive to errors in individual tests; compromises can be made between the extent to which variables are dependent in the data and the cost of adding the edge. Detailed description of the BN learning approaches is presented in section 4.2.2.

2.2.4. Bayesian Network Inference

The specification of the random variables and their values must be precise enough to satisfy the requirements of the particular situation being modeled, and must be sufficiently precise to pass the *clarity test*. The test was developed by Howard (1988), it states the following: Imagine a clairvoyant who knows precisely the current state of the world (or future state if the model concerns events in the future). Would the clairvoyant be able to determine unequivocally the value of the random variable?

This test is used to test the goodness of the defeintion of the model's elements, whether elements such as variables, events, outcomes, and alternatives are sufficiently defined to make the decision needed.

The probabilities of the random variables having their values are judged after defining the possible values of the random variables (i.e. their spaces). But, prior probabilities are not always determined, neither the values in a joint probability distribution of the random variables are determined. Instead probabilities concerning relationships among other variables which are accessible are ascertained.

It should be noticed that even if the probabilities used to do Bayesian inference (as will be seen later on) are obtained from frequency data, they are only estimates of the

actual relative frequencies. So they are subjective probabilities obtained from estimates of relative frequencies, they are not relative frequencies. Even when manipulated using Bayes' theorem, the resultant probability is also a subjective probabiliy.
Once the probabilities are judged for a given application, values in a joint probability distribution of the random variables can often be obtained.

When Bayesian inference involves only two related variables, it is fairly simple to compute. However, it becomes more complex when the inference is needed to be done with many related variables.

Difficulties exist when representing large instances and in doing inference when there are large number of variables. Thus, sometimes, it is needed to do probabilistic inference involving variables that are not related via a direct influence. In such a case of requiring the determination of many values and doing many calculations, Bayesian network is the alternative.

Bayesian networks address the problems of representiing the joint probability distribution of a large number of random variables, and doing Bayesian inference with these variables.

Inference in Bayesian networks consists of computing the conditional probability of some variables, given that other variables are set to evidence. Inference may be done for a specified state or value of a variable, given evidence on the state of other variable(s). Thus, using the conditional probability table for the BN built, their values can be easily inferred. Figure 2.4 shows an example of a conditional probability table, where it could be seen that given evidence for the distance to be "short" and the velocity to be "high", the probability that the time of journey will be less than 1 h is 0.75. Thus, other inferences could be extracted using this Figure, where the example presented here is used to explain how inference in BNs works.

2.2.5. Applications of Bayesian Networks

Figure 2.7 shows the number of studies performed using BNs, as it could be seen that the first study found to apply BNs was performed in 1990 in physical sciences. The real boost in the usage of BN began in the year 2003. Gradually more fields of different sciences started to apply BNs in their researches.

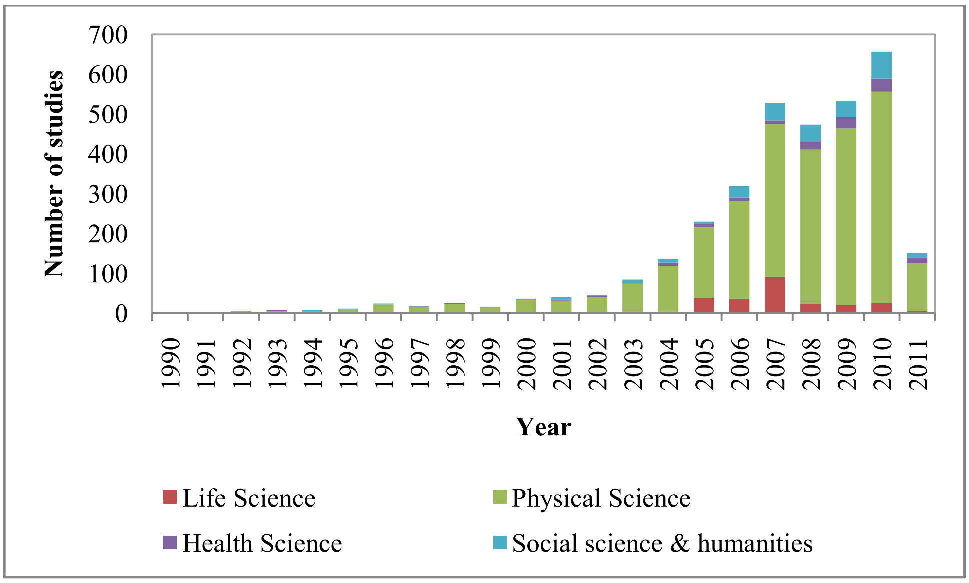


Figure 2. 7: number of studies performed in different fields of science that used BNs. Source: SCOPUS search engine

Thus, the vast applications of BNs to different fields apparently attracted the attention of researchers in transportation engineering. To our knowledge, it was not until the year 2003 that the first application of BNs found its way to Transportation Engineering (see Figure 2.8).

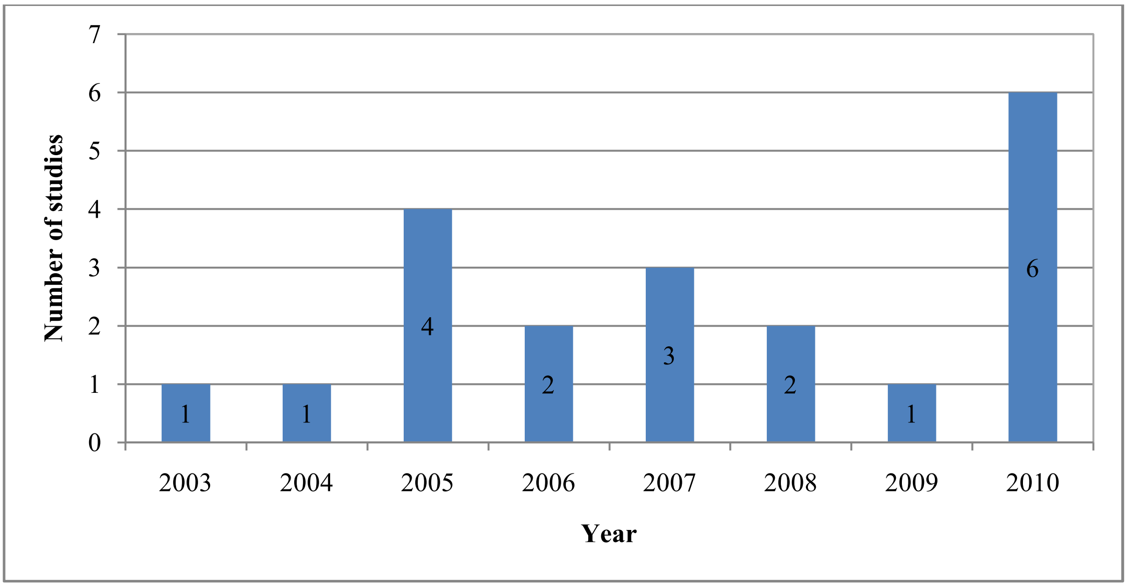


Figure 2. 8: Number of Studies found to apply BNs into the fields of Transportation Engineering

The applications varied from those in safety research, traffic flow…. ,etc. In the next sections a brief summary of these studies is presented.

2.2.5.1. Application in safety

In 2003 Davis and Pei was the first (to our knowledge) to apply BNs in order to illustrate how BNs can be used to support inductive reasoning about road accidents. In their study they used three actual accidents situations, two of which dealt with a pedestrian/vehicle accident and the third of two vehicles being involved in an accident in an attempt to reconstruct the three accidents.

Their results indicated that while it is often possible to identify an underlying causal structure for a reconstruction problem, deductive certainty is not possible because the initial conditions of the accident cannot be measured, and are usually underdetermined by the available evidence. In addition, there is no comprehensive or commonly accepted method for rationally accounting for this uncertainty, then an accident reconstruction problem can often be formulated as an example of processing information in a Bayesian network.

One year after Simoncic (2004) used BNs to model road accidents and accordingly made inferences for accident analysis. In his work, a two car accident injury severity model was constructed using BNs. A BN was built using several variables, and the Most Probable Explanation (MPE) was calculated for the most probable configuration of values for all the variables in the BN, in order to serve as an indication of the quality of the estimated BN. The results pointed out that BNs could be applied in road accident modeling. Thus, some improvements, such as using more variables and larger datasets, were recommended. Although this study highlighted the possibility of using BNs to model traffic accidents' injury severity, the results were based on building only one possible network, without measuring the performance of the Bayesian classifier. Also, his work concentrated on the development of the model and not how to use it in order to improve safety.

Gregoriades (2007) integrated microscopic road network simulation with BN technology for improved prediction of road accident risk. His work described a method along with the current state of the development of an accident prediction system. The

BN model developed assesses the likelihood of an accident occurring based on real time observations from the simulation. The BN model is developed on multi disciplinary principles ranging from human factors to road network design.

Based on the first work presented by Gregoriades (2007), in 2010 Gregoriades et al. applied BNs and road traffic simulation for validating the safety requirements of prospective road designs. The theoretical platform of the method is the concepts of human performance and mental workload and how these affect accident likelihood. Their paper presented a novel method and a tool that integrates these two mature technologies, for assessing the safety performance of road designs before they are developed. A case study was included that illustrated the application of the method and tool.

Based on data survey and statistical analysis, a BN for traffic accident causality analysis was developed by Xu et al. (2010). The structure and parameter of the BN was learnt with K2 algorithm and Bayesian parameter estimation respectively. With the Junction Tree algorithm, the effect of road cross-section on the accident casualties was inferred. The results showed that the BN can express the complicated relationship between the traffic accident and the causes, as well the correlations among the factors of causes. Also, the results of analysis provided valuable information on how to reveal the traffic accident causality mechanisms and how to take effective measures to improve the traffic safety situations.

In order to have a comprehensive and clear description of incident clearance patterns, Ozbay and Noyan (2005) used BNs to represent these patterns. They employed a unique database created using incident data collected in Northern Virginia. This database was then used to demonstrate the advantages of employing BNs as a powerful modeling and analysis tool especially due to their ability to consider the stochastic variations of the data and to allow bi-directional induction in decision-making. Their results indicated that the prediction methodology employed by BNs is shown to be fully capable of representing the stochastic nature of incidents. Moreover, this methodology enables the decision makers to make real-time decisions by giving them the flexibility of employing the probabilistic inference capabilities of BNs.

Timely and accurate incident detection is an essential part of any successful advanced traffic management system. Zhang et al. (2006) presented a new arterial road incident detection algorithm TSC_ar in which in this algorithm, BNs are used to quantitatively model the causal dependencies between traffic events (e.g. incident) and traffic parameters. Using real time traffic data as evidence, the BNs update the incident probability at each detection interval through two-way inference. An incident alarm is issued when the estimated incident probability exceeds the predefined decision threshold. A total of 40 different types of arterial road incidents were simulated to test the performance of the algorithm, in which a high accuracy of 88% was obtained. The results indicated that BNs allow to subjectively building existing traffic knowledge into their conditional probability tables, which makes the knowledge base for incident detection robust and dynamic. Where, BN approach was found to be advanced in enabling effective arterial road incident detection.

Based on analysis of safety factors of major highway infrastructure, a combined method coupling fuzzy algorithm and Bayesian network was presented by Guo et al. (2010). The works was performed in order to calculate operation risk of major highway infrastructure under various events, in which fuzzy algorithm was used to calculate the operation risk of management units, and BN to deal with the risk relationship between management units. The method can comprehensively deal with various complicated relationship between safety factors of major infrastructure, and rationally reflect its subjective and objective safety. Furthermore, it can conduct forward reasoning in risk prediction and backward reasoning in cause or cause-effect network diagnosis, thus it is suitable for constructing safety management system for major highway infrastructures.

2.2.5.2. Application in traffic flow

Traffic flow forecasting is an important issue in the field of Intelligent Transportation Systems. Due to practical limitations, traffic flows recorded can be partially missing or unavailable. In this case few methods can deal with forecasting successfully.

A BN model and a two step BN model were constructed respectively to describe the causal relationship among traffic flows, and then the joint probability distribution between the cause and effect nodes with its dimension reduced by Principal Component

Analysis was approximated through a Gaussian Mixture Model (Sun et al. 2004). Their results indicated that the essence of traffic flow is consistent with the ideology of BN. Besides, as BN encodes dependencies among all variables, it readily handles situations where some data entries are incomplete. Therefore, constructing BNs for traffic flow forecasting is reasonable. Experiments with real-world data also show that BN is applicable and effective for short-term traffic flow forecasting.

A novel predictor for traffic flow forecasting called spatiotemporal BN predictor was proposed by Sun et al. (2005). Their approach incorporates all the spatial and temporal information available in a transportation network to carry the traffic flow forecasting of a current site. Their experimental results with the urban vehicular flow data of Beijing demonstrated the effectiveness of the presented spatio-temporal BN predictor.

In 2006, Sun et al. proposed another new approach based on BNs for traffic flow forecasting. In their paper, traffic flows among adjacent road links in a transportation network were modeled as a BN. The joint probability distribution between the cause nodes (data utilized for forecasting) and the effect node (data to be forecasted) in a constructed BN was described as a Gaussian mixture model (GMM) whose parameters were estimated via the competitive expectation maximization (CEM) algorithm. Traffic flow forecasting was then performed under the criterion of minimum mean square error (mmse). The approach departs from many existing traffic flow forecasting models in that it explicitly includes information from adjacent road links to analyze the trends of the current link statistically. In addition, it also encompasses the issue of traffic flow forecasting when incomplete data exist. Comprehensive experiments on urban vehicular traffic flow data of Beijing and comparisons with several other methods showed that the BN is a very promising and effective approach for traffic flow modeling and forecasting, both for complete data and incomplete data.

The problem of estimating and updating the origin-destination matrix and link flows from traffic counts and its optimal location was studied by Castillo (2008). A combination (bi-level) of an OD-pair matrix estimation model based on Bayesian networks, and a Wardrop-minimum variance model, which identifies origins and destinations of link flows, was used to estimate OD-pair and unobserved link flows based on some observations of links and/or OD-pair flows. The BN model was also used to select the optimal number and locations of the links counters based on maximum correlation. The proposed methods were applied to two networks, where the results indicated that BNs enables determining which information (link flows, for example) is relevant to other given variables (OD matrix). In addition, using the variance–covariance information and its updating when new information is available given by BNs, it is possible to build an effective procedure to obtain the number and position of traffic counts for OD-matrix estimation, based on the correlation matrix.

Another study was performed by Castillo et al. (2008) in order to deal with the problem of predicting traffic flows and updating these predictions when information about OD pairs and/or link flows becomes available. A BN was built which is able to take into account the random character of the level of total mean flow and the variability of OD pair flows, together with the random violation of the balance equations for OD pairs and link flows due to extra incoming or exiting traffic at links or to measurement errors. BN provide the joint density of all unobserved variables and in particular the corresponding conditional and marginal densities, which allow not only joint predictions, but also probability intervals. The influence of congested traffic can also be taken into consideration by combination of the traffic assignment rule. Their results indicated that the parameters in the BN can be easily learned from the topology of the network and other data. The proposed methods allow obtaining the full distribution of unobserved conditional variables accounting for all the information available (evidences).

A new hybrid approach, based on a coupling of a continuum model and a knowledge-based model (BN), was introduced by McCrea and Moutari (2010). Some experiments have been carried out on some parts of the south-east of England road network. Contrary to existing hybrid models, the specificity of this approach is not only its ability to describe effectively traffic dynamics in road networks, but also its simplicity to implement and to operate. Furthermore, the model is appropriate to operate in real-time, as required by any traffic management system.

In order to deal with the problem of predicting route flows and updating these predictions when plate scanned information becomes available. A BN is built which is able to deal with the joint distribution of route and link flows and the flows associated with all possible combinations of scanned link flows and associated random errors

(Sánchez-Cambronero et al. 2010). The BN provides the joint density of route flows conditioned on the observations, which allows not only the independent or joint predictions of route flows, but also probability intervals or regions to be obtained. A procedure was also given to select the subset of links to be observed in an optimal way. An example of application illustrated the proposed methodology and showed that it is practically applicable.

Real-time and accurate traffic speed is important for a successful traffic management system. However, the most common form of the single-loop detector is incapable of providing speed measurements. Jin et al. (2010) presented a method of speed estimation from single-loop detector data using BN method. After analyzing the causal relationship between volume, occupancy, and speed, a BN model of speed estimation was proposed using volume and occupancy from single-loop outputs. The Gaussian mixture model (GMM) and the expectation-maximization (EM) algorithm were used to represent the model and train model parameters, respectively. The proposed method was implemented and evaluated using the field data from urban expressways in Beijing. Estimated speeds were compared with the observed speed data and also with results from conventional algorithm. The results showed that the proposed method is robust for every kind of sampling intervals, lanes, and traffic condition. The mean absolute error holds more than 2 km/h decrease. This method can be efficiently applied in traffic management system.

2.2.5.3. Other Application in transportation

Applications of BNs into the field of transportation engineering are not limited for the aforementioned applications, many other applications in data imputation, travel behavior and localization measurements were found. The following is a brief description of these applications.

One of the limitations of the data usage in intelligent transportation systems (ITS) is missing data. Many imputation methods have been proposed in the past decade. Ni and Leonard (2005) proposed an advanced imputation method based on BNs to learn from the raw data and a Markov chain Monte Carlo technique to sample from the probability distributions learned by the BN. The method imputes the missing data multiple times and makes statistical inferences about the result. In addition, the method incorporates a time series model so that it allows data missing in entire rows. Empirical study shows that the proposed method is robust and accurate. It is ideal for use as a high-quality imputation method for off-line application.

A new method was proposed by Janssens et al. (2006) to combine BNs and decision trees using data that measure the travel behavior of individuals (nature of the activity, the day, start and end time, the location where the activity took place, the transport mode, the travel time, accompanying individuals and whether the activity was planned or not). For this reason, this study was designed to examine whether a decision tree (which is implicitly always less complex) that uses the structure of a BN called (BNT) to select its decision nodes can achieve simultaneously accuracy results comparable to BNs with an easier and less complex model structure, comparable to traditional decision trees. This study indicated that the new way of integrating decision trees and BNs may produce a decision tree that is structurally more stable and less vulnerable to the variable masking problem. Additionally, the results at the activity level and trip level suggested a trade-off between model accuracy and model complexity. When the main issue is the interpretation and the general understanding of the decision rules, the integrated BNT approach may be favored above CHAID decision trees when decisions need to be made at pattern level. At a more detailed level, one may benefit from the use of the CHAID approach. However, when the main issue is model accuracy, BNs should be favored.

The global positions provided by a GPS receiver are used to select the most likely segment(s) from a set of segments close to the estimation of the vehicle position. Nowadays, since the geometry of roadmaps is more and more detailed, the number of segments representing roads is increasing. The road managing module is an important stage in the vehicle localization process because the robustness of the localization depends mainly on this stage.

GPS suffers from satellite outages occurring in urban environments, under bridges, tunnels or in forests. GPS can thus be seen as an intermittently-available positioning system that needs to be backed up by a dead-reckoning system. Smaili et al. (2007) proposed a low-cost odometric method based on the use of encoders attached to the rear wheels. A dead-reckoned estimated pose is obtained by integrating the elementary rotations of the wheels starting from a known pose. The multisensor fusion of GPS and odometry is performed by a Switching Kalman Filter (a subclass of Dynamic BNs). This kind of formalism is also useful in quantifying the imprecision associated with each estimated pose. The experiments indicated that the GPS measurements are not necessary available all the time, since the merging of odometry and roadmap data can provide a good estimation of the position over a substantial period. The strategy presented in this paper doesn't keep only the most likely segment. When approaching an intersection, several roads can be good candidates for this reason by managing several hypotheses until the situation becomes unambiguous.

To that end, many applications of BNs into transportation engineering were found to be efficient. However, the only study found to analyze the injury severity of traffic accidents was that done by Simoncic (2004). This study was a preliminary trial to show the possibility of applying BNs into this field. Thus, the analysis of injury severity of traffic accidents is an important problem that still needs further analysis and new technologies to be applied in order to better define the problem and the factors that are related the most with the consequences of a traffic accident.

2.2.6. Advantages of using Bayesian networks

Using Bayesian networks has many advantages, of which are the following (de Campos Ibáñez, 2011):

1. They allow explicitly handling uncertain knowledge (most of the human knowledge has some types of uncertainty)
2. Utilizing the probability theory as a formalism to handle uncertainty, offer a clear semantic and a solid theoretical foundation, in contrast to other ad hoc techniques for reasoning uncertainty employed in artificial intelligence and expert systems.
3. They enable a graphical form of representation of knowledge, which is very intuitive and very similar to some patterns of human reasoning.
4. They allow the specification of the global model (a distribution of joint probability) through local models (marginal and conditional distributions involving subsets of variables), exploiting the conditional independence relations. This modularity:
    - Facilitate maintenance
    - Drastically reduces the number of required parameters that are needed to specify the model
    - Make the tasks of elicitation or estimation of parameters simpler.
    - At the same time, reduces the storage need
    - Contribute decisively to efficiency of reasoning

5. Allow a combination of predictive and diagnostic reasoning, which is more difficult to model using other types of systems (like rules-based systems)
6. The inference tasks (reasoning) can be carried-out relatively efficient, using algorithms existing for propagation of evidence (exact or approximate). Supporting various types of inference:

    - Probability of any event given any evidence
    - Calculating the most probable explanation (the best scenario that explains the available data)
    - Decision making (using influence diagrams, a generalization of the Bayesian networks incorporating decision nodes and utility)

2.3. Conclusions

This chapter presented a review of the existing literature for the analysis of injury severity of traffic accidents, as well as, a presentation of BNs. Based on that the following concluding remarks could be extracted:

As mentioned before, there is no such a consensus on the best model to be used in the analysis of injury severity, many models have been used and are in use. The only guidance that could be given to researchers is to use the model that best fits the characteristics of the data being used.

1. Recently, Data mining and soft computing techniques have started to gain acceptance for applications in different fields including in the analysis of traffic accidents. More specifically, BNs have proven to be efficient when applied in many different fields (life sciences, physical sciences, health sciences and social sciences). Thus, given the characteristics of the data used to analyze injury severity of traffic accidents, and the advantages of BNs that are beneficial for such characteristics, a utilization of BNs to the field of injury severity analysis could find its way.
2. An important attention should be paid towards sample size and the number of variables used in the model. One should keep an optimum sample size that maintains an optimum goodness of fit statistic, as well as, an optimum number of variables to be used without having the model being neither underspecified nor overspecified.

3. Most of the traffic agencies that are in charge of collecting traffic accidents' data around the world have an availability of data for many years. One should not be tempted to use time series for more than three years. The reason for such a recommendation would be the fact that characteristics of data collected for the same location in different time periods might change. These changes could be due to improvements to the roadway design, regular maintenance work, social and demographic changes, etc. Thus, if the researcher is interested in studying traffic accidents for larger time periods, he/she should pay attention to such changes, and if a drastic change has taken place and the researcher is still interested in the whole time period, most probably a before/ after study could be

CHAPTER 3

Objectives

This research work applies Bayesian networks to the field of traffic accidents injury severity modeling, though that Bayesian networks have been proven to apply for dealing with complicated problems, their application in the field of traffic accidents is still new. Keeping in mind that many previous studies tried to employ different statistical techniques to analyze traffic accidents and specifically accidents severity levels (as will be seen later on).

The first objective of this thesis is to validate utilizing Bayesian networks for the analysis of the injury severity of the traffic accidents.

The second objective of this thesis is validating the possibility of obtaining similar results to those obtained previously using reduced number of variables.

To that end, two hypotheses are supposed:

1. It is possible to utilize BNs for the treatment of traffic accidents' data
2. It is possible to reduce the number of explanatory variables used without losing the precision of the model.

CHAPTER 4

Materials and methods

4.1. Data

The data used in this research work was obtained from the General Directorate of Traffic (DGT). The data obtained were for traffic accidents that occurred on rural two-lane highways for the province of Granada (South of Spain) for three years (2003-2005). The total number of accidents obtained for this period was 3,302. The data was first checked out for questionable data, and those which were found to be unrealistic were screened out. Only rural highways were considered in this study; data related to intersections were not included. Finally, the database used to conduct the study contained 1,536 records. Table 4.1 provides information on the data used for this study.

Table 4. 1: Variables, values and actual classification by severity.

Variables	Values	SEV				Total
		SI		KSI		
ACT: accident type	AS: angle or side collision	381	61.45%	239	38.55%	620
	CF: fixed objects	99	52.94%	88	47.06%	187
	HO: head on	84	40.58%	123	59.42%	207
	O: other	75	59.06%	52	40.94%	127
	PU: pile up	33	78.57%	9	21.43%	42
	R: rollover	163	49.39%	167	50.61%	330
	SP: straight path	17	73.91%	6	26.09%	23
AGE: age	[18-25]	225	50.34%	222	49.66%	447
	(25-64]	586	57.73%	429	42.27%	1015
	>64	41	55.41%	33	44.59%	74
ATF: atmospheric factors	GW: good weather	730	54.23%	616	45.77%	1346
	HR: heavy rain	23	71.88%	9	28.13%	32
	LR: light rain	84	61.76%	52	38.24%	136
	O: other	15	68.18%	7	31.81%	22
CAU: cause	DC: driver characteristics	791	54.93%	649	45.07%	1440
	OF: other factors	50	66.67%	25	33.33%	75
	RC: road characteristics	3	75.00%	1	25.00%	4
	VC: vehicle charactersitics	8	47.06%	9	52.94%	17
DAY: day	BW: beginning of week	123	60.29%	81	39.71%	204
	EW: end of week	132	57.14%	99	42.86%	231
	F: festive	29	61.70%	18	38.30%	47
	WD: week day	325	55.65%	259	44.35%	584
	WE: week end	243	51.70%	227	48.30%	470
GEN : gender	F: female	148	63.79%	84	36.21%	232
	M: male	704	53.99%	600	46.01%	1304

Continues Table 4.1

Variables	Values	SEV				Total
		SI		KSI		
LAW: lane width	THI: thin: <3.25m	19	67.86%	9	32.14%	28
	MED: medium: 3.25m<=L<=3.75m	176	51.16%	168	48.84%	344
	WID: wide: >3.75m	657	56.44%	507	43.56%	1164
LIG: lighting	D: dusk	52	61.18%	33	38.82%	85
	DL: daylight	573	58.65%	404	41.35%	977
	I: insufficient	27	54.00%	23	46.00%	50
	S: sufficient	36	59.02%	25	40.98%	61
	W: without lighting	164	45.18%	199	54.82%	363
MON: month	AUT: autumn	218	54.23%	184	45.77%	402
	SPR: spring	206	59.03%	143	40.97%	349
	SUM: summer	246	56.55%	189	43.45%	435
	WIN: winter	182	52.00%	168	48.00%	350
NOI: number of injuries	1	539	49.95%	540	50.05%	1079
	>1	313	68.49%	144	31.51%	457
OI: occupants involved	1	229	51.58%	215	48.42%	444
	2	374	55.99%	294	44.01%	668
	>2	249	58.73%	175	41.27%	424
PAS: paved shoulder	missing values	66	51.56%	62	48.44%	128
	N: no	253	57.11%	190	42.89%	443
	Y: yes	533	55.23%	432	44.77%	965
PAW: pavement width	THI: thin: <6m	95	53.98%	81	46.02%	176
	MED: medium: 6 m<=law<=7m	209	54.29%	176	45.71%	385
	WID: wide: >7m	548	56.21%	427	43.79%	975
ROM: pavement markings	DME: does not exist or was deleted	60	58.25%	43	41.75%	103
	DMR: define margins of roadway	60	57.69%	44	42.31%	104
	SLD: separate lanes and defined road margins	714	55.26%	578	44.74%	1292
	SLO: separate lanes only	18	48.65%	19	51.35%	37
SHT: Shoulder type	NOS: does not exist	311	55.24%	252	44.76%	563
	THI: thin:<1.5m	402	54.47%	336	45.53%	738
	MED: medium: 1.5m<=sht<2.50m	133	58.85%	93	41.15%	226
	WID: wide >= 2.50 m	6	66.67%	3	33.33%	9
SID: sight distance	A: atmospheric	26	81.25%	6	18.75%	32
	B: building	10	55.56%	8	44.44%	18
	O: other	6	66.67%	3	33.34%	9
	T: topological	187	55.49%	150	44.51%	337
	V: vegetation	6	54.55%	5	45.45%	11
	WR: without restriction	617	54.65%	512	45.35%	1129
TIM: time	[0-6]	99	46.26%	115	53.74%	214
	(6-12]	236	57.99%	171	42.01%	407
	(12-18]	314	57.72%	230	42.28%	544
	(18-24)	203	54.72%	168	45.28%	371
VI: vehicles involved	1	316	52.06%	291	47.94%	607
	2	468	56.73%	357	43.27%	825
	>2	68	65.38%	36	34.62%	104
Total		852	55.47%	684	44.53%	1536

Eighteen variables were used with the class variable of injury severity (SEV) in an attempt to identify the important patterns of an accident that usually require an explanation.

Thus, it should be taken into account that both the choice of the eighteen variables and the categorization of these variables were mainly guided by previous studies. Also, an analysis of the data using preliminary models was performed in order to define an optimum and near optimum models that best represent the data used. In which, the number of variables and categories that obtain the best possible fit to the data used were chosen.

Following previous studies (Chang and Wang, 2006; Milton et al., 2008) the injury severity of an accident is determined according to the level of injury to the worst injured occupant. On the other hand, the injury severity level was categorized into two categories mainly based on initial models that we have built using three categories of severity (killed, severe and slight). However, the performance of the resulting models was not acceptable. This possibly was due to the convergence problems resulting from insufficient variation among discrete-injury outcomes.

In addition, only accidents that occurred on rural roadways segments and not intersections were considered. This was based on a recommendation given by Moore et al. (2010) in which they recommended that intersections and road segments should not be analyzed together, since the factors related to accidents occurring on intersections are different from those related to roadway segments.

The data included variables describing the conditions that contributed to the accident and injury severity.

- Injury severity variables: number of injuries (e.g., passengers, drivers and pedestrians), severity level of injuries (e.g., fatal, severe, slight).
- Roadway information: characteristics of the roadway on which the accidents occurred (e.g., grade, pavement width, lane width, shoulder type, pavement markings, sight distance, if the shoulder was paved or not, etc.).
- Weather information: weather conditions when the accident occurred (e.g., good weather, rain, fog, snow and windy).

- Accident information: contributing circumstances (e.g., type of accident, time of accident (hour, day, month and year), and vehicles involved in the accident).
- Driver data: characteristics of the driver, such as age or gender.

As it is shown in Table 4.1 from the 1,536 accidents 852 (55%) were slightly injured and 684 (45%) were killed or severely injured. Table 4.1 also provides the information about injury severity distribution by the variables studied.

As it can be seen from Table 4.1, a relatively larger proportion of angle or side impact accidents were involved in more KSI accidents than other types of accidents. Rollover accidents comes in the second place in producing KSI accidents, within the total number of accidents produced by rollover accident, more KSI were produced. The proportion of KSI accidents produced by rollover accidents is about two times less than those produced by angle or side impact accident. Head-on accidents produced more KSI accidents than SI accidents within the total number of accidents produced by this type of accidents, in which about 1.5 times the accidents were KSI.

With respect to age it can be noticed that the age group (25-64) had the highest number of KSI accidents, it produced more than double the accidents produced by the age group (18-25).

Good weather mainly dominated the other atmospheric factors both in the number of accidents produced and in their severity. The KSI accidents produced by good weather were more than 10 times higher than those produced by the next group which is light rain.

The pattern presented by the good weather is also consistent with the results corresponding to accidents by month. Accidents during the summer are the most, and they also produce more KSI accidents.

According to the cause of the accident as stated in the police reports, driver characteristics were responsible for the highest number of accidents and they were also responsible for the highest number of KSI accidents.

With respect to the day of an accident, accidents which occurred during the week were the most; however one should know that the category week day included three working days during the week (Tuesday, Wednesday and Thursday). While the category week end (weekend) had the second highest proportion of accidents, and it included two days (Saturday and Sunday). This means that in reality the accidents which occurred during the weekend had the highest percentage since they occurred during only two days. Roughly speaking, about 194 accidents occurred in each of the three working days; where about 235 accidents occurred in each one of the two weekend days.

Males were responsible for more accidents and also for more severe accidents. Males were found to be responsible for seven times higher KSI accidents than female drivers.

Increased lane width increases the accidents and also severity; wider lanes seem not to be related to safer highways. However, having a look on medium lanes it is seen that although the total number of accidents produced by medium lane widths were about 3 times less than those produced by wider lanes, the proportion of the KSI accidents and SI is almost the same. This means that even though the medium lanes have fewer accidents, the accidents produced on such lanes have almost an equal chance of being either a KSI or a SI.

Pavement widths are related to lane widths, wider pavements were also associated with more accidents probably due to the same explanation presented earlier for the lane width.

More accidents as well as more severe accidents occur during the daylight. Roadways without lighting were the next with regards to the total number of accidents and the resulting severity. However, KSI accidents were more than SI cases in no-light condition.

Paved shoulders were associated with higher number of accidents. Pavement markings that are well defined were found to be associated with higher number of accidents with higher severities.

Roadways without any restriction of the sight distance were associated with the highest number of accidents. Restriction due to topology was associated with the second highest number of accidents.

Accidents that occurs during the time period (12 p.m.-18 p.m.) were associated with more accidents. Accidents that involve two vehicles were the most, and those that involved two occupants as well. However, most of the accidents involved one injury, and the proportion of SI and KSI in each were almost equal.

In order to find out the associations between pairs of variables, a chi-square test was used. One of the uses of the chi-square test is to test the statistical significance that exists in the association between two variables. To test the strength of significant associations, other measures are used such as the Phi coefficient and the Cramer’s V.

The chi-square test serves both as goodness-of fit test, where the data is categorized along one dimension, and as a test for the contingency table, in which categorization is across two or more dimensions (Wuensch, 2011) .

The standard chi-square statistic is defined as:

$$\chi^2 = \sum_{i}^{n} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \tag{4.1}$$

Where,

O: is the observed frequency

E: is the expected frequency

i and j: index the rows and columns of the table

n: the number of cells in the table

Equation 4.1 results in a statistic that is approximately distributed as χ^2 on (r-1) (c-1) degrees of freedom, where r represents the number of rows in the contingency table, and c represents the number of columns in the contingency table. The null hypothesis is that the two variables are independent, so that the two variables are considered to be associated if the p-value of the χ^2 test is <0.05.

In order to measure the strength of the association between a two significantly associated variables, two other coefficients could be used the Phi-coefficient and the Cramer’s V.

Phi coefficient is used only for 2×2 contingency tables, where a value of 0 signifies no association and 1 a perfect association (Mehta and Patel, 1996).

$$\text{Phi}=\sqrt{\frac{\chi^2}{n}} \tag{4.2}$$

Where,

χ^2: the Chi-square value

n: is the sample size

Cramer's V coefficient is interpreted as a measure of relative strength of association between two variables (Crewson, 2006). Cramer's V coefficient ranges between 0 and 1, with 0 signifying no association and 1 signifying perfect association (Mehta and Patel, 1996), it is given by:

$$V = \sqrt{\frac{\chi^2}{n(q-1)}} \tag{4.3}$$

Where,

q= is smaller number of rows or columns

A guidance to describe the strength of association is given in Table 4.2 (Crewson, 2006).

Table 4. 2: Strength of association description

Value of the coefficient	Characterization
> 0.5	high association
0.3 – 0.5	moderate association
0.1 – 0.3	low association
0.0 – 0.1	little if any association

Certain restriction exist for the chi-square test, if the expected numbers on some classes are small, the chi-square test will give inaccurate results. Chi-square test could be used whenever no more than 20% of the expected counts are less than 5 and that all individual counts are 1 or greater (Yates et al., 1998). With small expected frequencies Fisher's Exact test is used (Wuensch, 2011).

If a general a 2×2 contingency table has cell entries n_{ij} (I, j=1, 2) and row totals n_{i+} and column total n_{+j} and grand total of all 4 cell entries is n, then the hyper-geometric distribution for the observed cell values has an associated probability:

$$\frac{(n_{1+})!\,(n_{2+})!\,(n_{+1})!\,(n_{+2})!}{(n_{11})!\,(n_{12})!\,(n_{21})!\,(n_{22})!\,n!} \tag{4.4}$$

To perform the test one calculates these probabilities for all possible n_{11} consistent with the fixed marginal totals and computes the p-value as the sum of all such probabilities that are less than or equal to that associated with the observed configuration (Fisher, 1934).

For the research work herein, the associations between variables were calculated using the measures described. Table 1 in the Appendix II shows the association's coefficients and their corresponding p-values.

The results illustrated in Table 1 in the Appendix II indicates that there exist associations between variables; however, the strength of that association differs amongst them. Table 4.3 shows the variables that were found to be significantly associated and their corresponding strength.

As illustrated from Table 4.3 ACT was the variable which had the most associated variables. Also, it was the variable that had the most highly associated variables. Of the interesting findings that can be extracted is the high association that exists between SEV and DAY.

Table 4. 3: Association between variables and its corresponding strength

variable	Associated variables and their strength of association			
	Little	Low	Moderate	High
SEV	AGE, ATF, GEN, LAW, OI, SID, TIM, VI	ACT, LIG, NOI	-	-
ACT	AGE,CAU	LAW, LIG, NOI, PAS, ROM, SHT, SID, TIM	-	DAY, OI, PAW, VI
AGE	GEN, LAW, OI, SHT, VI	LIG, TIM	-	-
ATF	CAU, DAY, NOI, OI	MON, SID	-	-
CAU	GEN, LIG, ROM, SID	LAW	-	-
DAY	GEN, MON, NOI, OI, SID	TIM, VI	-	-
GEN	LIG, NOI, OI, TIM	-	-	-
LAW	VI	-	PAS, PAW, ROM, SHT	-
LIG	SID	MON, OI, VI	TIM	-
MON	SID	PAS	-	-
NOI	-	VI	-	OI
OI	PAS, PAW, ROM	SID, TIM	-	VI
PAS	SID, VI	-	ROM	PAW, SHT

Continues Table 4.3

variable	Associated variables and their strength of association			
	Little	Low	Moderate	High
PAW	SID, VI	-	ROM, SHT	-
ROM	-	SHT, SID, VI	-	-
SHT	-	SID, VI	-	-
SID	-	VI	-	-
TIM	-	VI	-	-

To have a deeper insight towards this association Figure 4.1 shows that the highest number of AS occur during the WD however it should be noticed that the WD variable is composed of 3 days in which that approximately 89 accidents occur during a working day of the week of the AS type. Moreover, approximately 74 accidents of the AS type occur in a day of the WE.

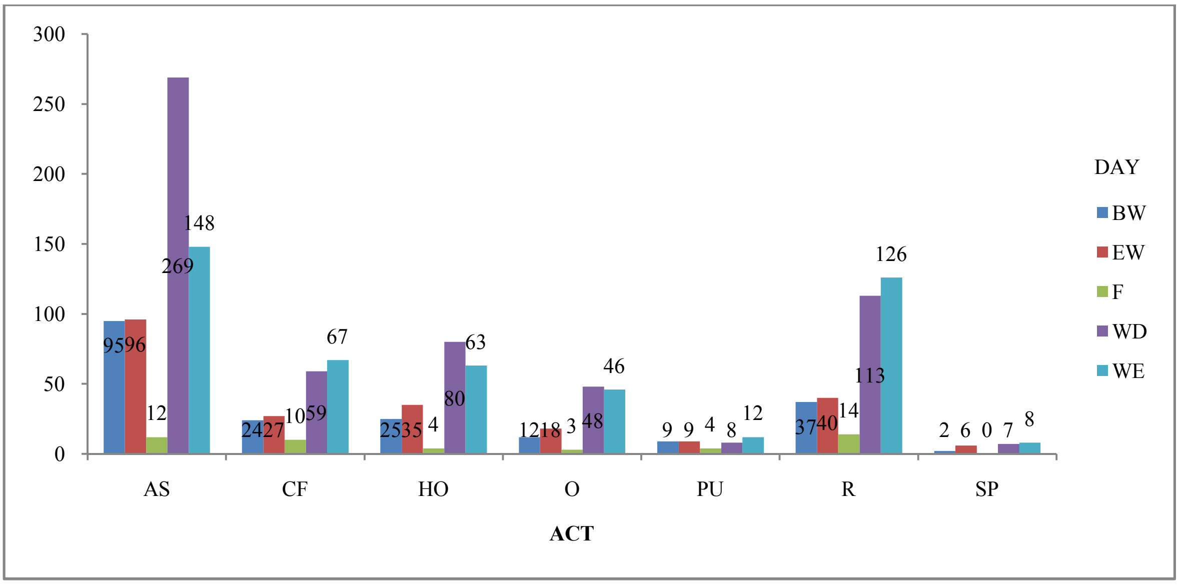


Figure 4. 1: Distribution of DAY by ACT

Another high association for ACT was with OI. Figure 4.2 shows that AS type of accidents involve the highest number of occupants. In which at least two occupant are found to be involved in an AS accident.

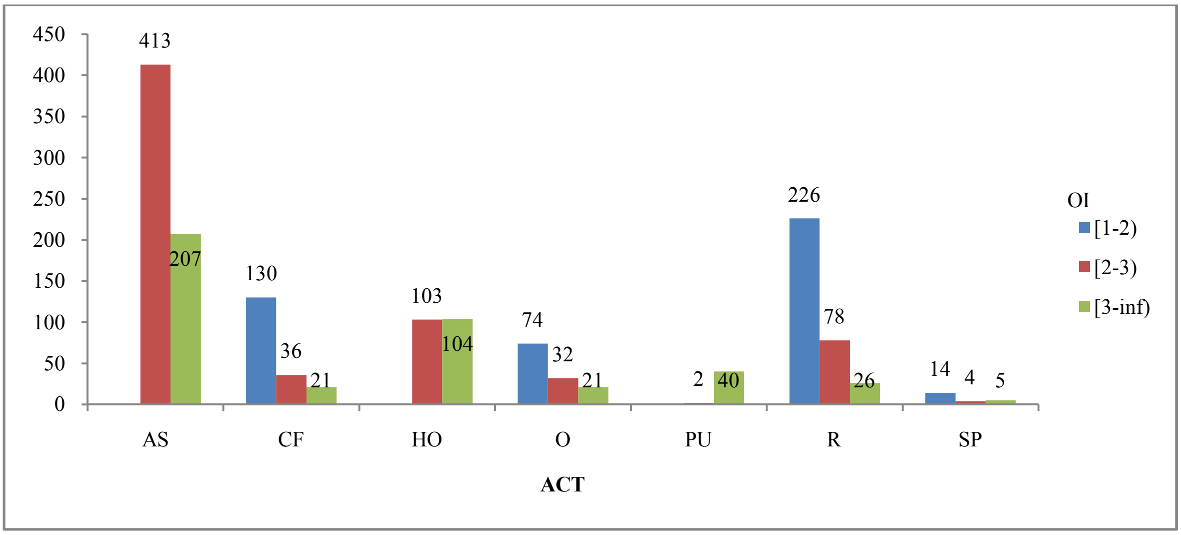


Figure 4. 2: Distribution of OI by ACT

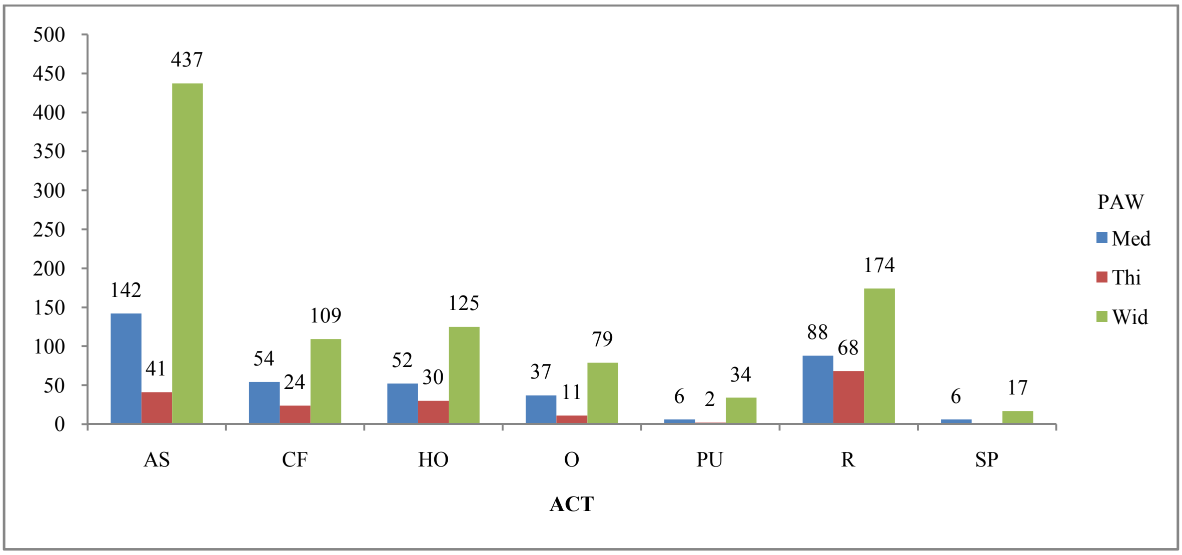


Figure 4. 3: Distribution of PAW by ACT

PAW was found to be highly associated with ACT, as shown in Figure 4.3 that Wid and Med were the ones with most AS accidents.

VI was found to be highly associated with ACT, as shown in Figure 4.4 that accidents between two vehicles were mostly of AS type.

On the other hand, SEV was the second variable with the most associated variables; however, the associations of variables with SEV were mostly of the low strength associations (8 associations). Neither high nor moderate associations were found for SEV.

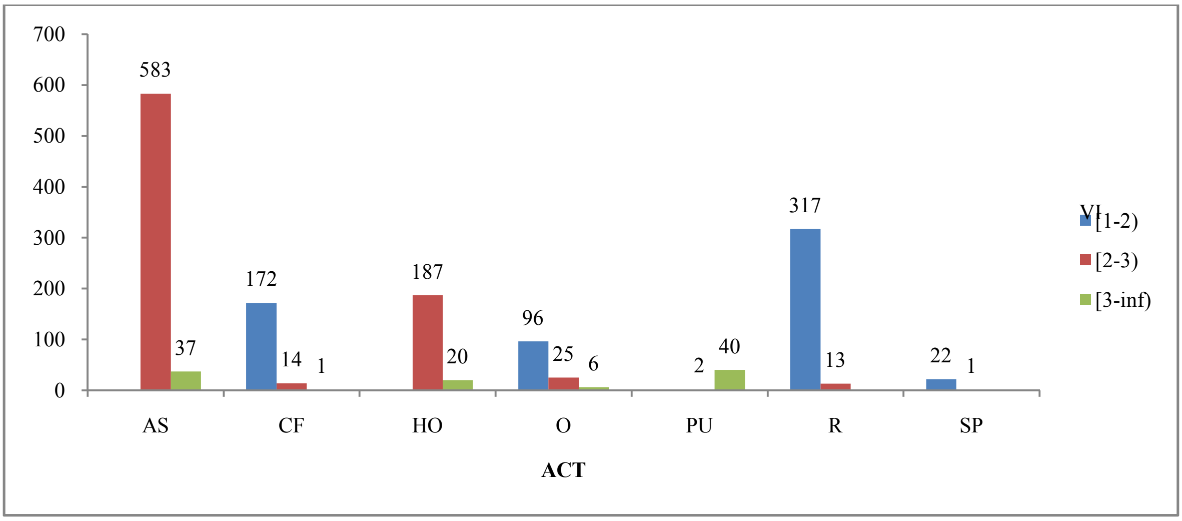


Figure 4. 4: Distribution of VI by ACT

Other high and moderate associations were found for other variables however, not as those found for ACT. For example, moderate associations were found to exist for LAW with [PAS, PAW, ROM, SHT], LIG with TIM, PAS with ROM, PAW with [ROM, SHT].

In order to better represent moderate associations between variables; Tables 4.4-4.7 show contingency tables for moderately associated variables.

Table 4. 4: Contingency table for LAW against PAS, PAW, ROM, and SHT

		LAW			Total
		Med	Thi	Wid	
PAS	N	210	14	219	443
	Y	134	14	945	1093
PAW	Med	144	5	236	385
	Thi	154	10	12	176
	Wid	46	13	916	975
ROM	Dme	69	9	25	103
	Dmr	73	5	26	104
	Sld	180	14	1098	1292
	Slo	22	0	15	37
SHT	Med	6	2	218	226
	Nos	254	14	295	563
	Thi	78	12	648	738
	Wid	6	0	3	9
Total		344	28	1164	1536

As seen from Table 4.4 the following combinations were found to have the highest number of accidents: PAS=Y with LAW= Wid, PAW=Wid and LAW= Wid, ROM= Sld with LAW= Wid, and finally SHT=Thi with LAW= Wid. LIG was found to have a moderate association with TIM, in which (as shown in Table 4.5) accidents were the most if they occur in LIG=DL and in the time periods TIM= (12-18) and TIM= (6-12).

Table 4. 5: Contingency table for LIG against TIM

		LIG					
		D	DL	I	S	W	Total
TIM	(12-18]	18	520	0	1	5	544
	(18-24)	31	117	24	34	165	371
	(6-12]	31	336	4	5	31	407
	[0-6]	5	4	22	21	162	214
Total		85	977	50	61	363	1536

PAS was found to have a moderate association with ROM, in which (as shown in Table 4.6) accidents were the most if they occur on PAS= Y and if ROM= Sld.

Table 4. 6: Contingency table for PAS against ROM

		PAS		
		N	Y	Total
ROM	Dme	76	27	103
	Dmr	87	17	104
	Sld	253	1039	1292
	Slo	27	10	37
Total		443	1093	1536

PAW was found to have a moderate association with both ROM and SHT, in which (as shown in Table 4.7) accidents were the most if they occur on PAW= Wid and if the ROM= Sld and/or SHT= Thi.

Other highly associated variables were found for other variables however, not as the high associations found for ACT. For example, high associations were found to exist for NOI with OI, OI with VI, PAS with PAW and SHT.

Table 4. 7: Contingency table for PAW against ROM and SHT

		PAW			
		Med	Thi	Wid	Total
ROM	Dme	32	60	11	103
	Dmr	36	59	9	104
	Sld	294	53	945	1292
	Slo	23	4	10	37
SHT	Med	9	1	216	226
	Nos	251	160	152	563
	Thi	123	14	601	738
	Wid	2	1	6	9
Total		385	176	975	1536

In order to better represent high associations between variables; Tables 4.8-4.10 show contingency tables for highly associated variables.

As seen from Table 4.8 higher number of accidents occurred when NOI were 1 and when the OI was 2.

Table 4. 8: Contingency table for NOI against OI

		NOI		Total
		[1-2)	[2-inf)	
OI	[1-2)	439	5	444
	[2-3)	535	133	668
	[3-inf)	105	319	424
Total		1079	457	1536

OI was found to be highly associated with VI, as shown in Table 4.9, accidents were found to be the most when involved 1 or 2 occupants and when involved 1 or 2 vehicles.

Table 4. 9: Contingency table for OI against VI

		OI			Total
		[1-2)	[2-3)	[3-inf)	
VI	[1-2)	443	111	53	607
	[2-3)	1	557	267	825
	[3-inf)	0	0	104	104
Total		444	668	424	1536

PAS were found to have a high association with PAW and with SHT. Table 4.10 shows that higher number of accidents occur when PAW= Wid in which the shoulder was paved, as well as, those that occur when the shoulder was thin.

Table 4. 10: Contingency table for PAS against PAW and SHT

		PAS		Total
		N	Y	
PAW	Med	195	190	385
	Thi	131	45	176
	Wid	117	858	975
SHT	Med	2	224	226
	Nos	414	149	563
	Thi	25	713	738
	Wid	2	7	9
Total		443	1093	1536

To that end, the association tests carried out herein is only used to find out the associations between pairs of variables in order to have an indication of existing relations for the variables considered. Thus, including more variables might change these associations.

4.2. Methodology

4.2.1. Phases of the research work

The research work carried out herein is composed of the following phases:

1. Firstly, three different search algorithms along with three different score metrics are used in order to build 9 different BNs which will be described in section 4.2.2.

2. Secondly, a comparison between these 9 different BNs is performed based on performance indicators as will be described in section 4.2.3.

3. Then, for the BN that shows improved results in terms of the performance indicators, the variables that significantly contribute to the occurrence of KSI are determined based on inference of the resulting BN.

4. Later, Based on the eighteen variables obtained from the accident reports (see Table 4.1), identification of the variables that affect injury severity in traffic accidents was performed using different methods of evaluator-search algorithms.

5. For each one of the selected subsets of variables, ten simplified BNs are built using the hillclimbing search algorithm and the MDL score.

6. The performance of the built BNs using the selected subsets of variables is compared with the performance of the original BN, which is built using the eighteen variables (BN-18).

7. From all the simplified built BNs, the selected ones are those whose results improve or maintain the results obtained by the performance indicators of BN-18 in 90% of the cases or more, and whose improvements are statistically significant.

8. For the selected BNs, the variables that repeat in more than 50% of the cases are identified and a new BN is built using these variables.

9. Finally, the results obtained by this new BN, based on a double process of variable selection procedure, are compared with those obtained by BN-18.

4.2.2. Methods of Learning of BNs

Researchers developed methods that could learn the DAG (structure) from data, in which the conditional probabilities could also be learned from data, where in a Bayesian network the values in the conditional probability distributions are called parameters. In the following; for more details, readers are referred to (Neapolitan, 2003).

4.2.3. Learning the Parameters (conditional probability estimation)

A Bayesian network with $\mathbb{G}$= (V, E), where V = {X_1, X_2, … , X_n) and the possible states of the variable X_i is C(X_i) = {X_{i1}, X_{i2}, …, X_{iri}}. In which r_i is the number of states of X_i, PA (X_i) is the set of parents of X_i in $\mathbb{G}$, and the number of parents' variables for X_i is s_i= |PA (X_i)|. q_i is the number of the possible configurations (states) of PA(X_i), q_i= $\prod X_j \epsilon$ PA (Xi)r_j. C (PA (Xi)) = {$\mathrm{PA}_{i1}, \mathrm{PA}_{i2}, \dots, \mathrm{PA}_{iqi}$} is the set of configurations of $\mathrm{PA}_{\mathbb{G}}$ (X_i).

It is needed to estimate P(X_i = x_{ik} | PA_{ij}), for all i= 1, 2, …, n, for all j = 1, 2, …, qi, for all k= 1, 2, …,ri. Where the set D= { $v^{(1)}$, $v^{(2)}$, …, $v^{(N)}$} of N observations of V, $v^{(h)}$= $(x_1^{(h)}, x_2^{(h)}, \dots, x_n^{(h)})$, with $x_i^{(h)} \epsilon\, C(X_i)$. Also, it is supposed that the data is complete and does not contain any missing data.

The likelihood is defined as follows:

$$L(D|\Theta) = P(D|\theta) = P\left(v^{(1)}, \dots, v^{(N)} \middle| \Theta\right) =$$

$$\prod_{i=1}^{n}\left(\prod_{h=1}^{N} P(x_i^{(h)}|PA(X_i)^{(h)}, \Theta_i)\right) = \prod_{i=1}^{n} L\,(D_i\colon \Theta_i) \tag{4.5}$$

Where, the global independence is assumed for the parameters.

Thus, the distributions could be estimated for each variable X_i independently:

$$L(D|\Theta) = \prod_{i=1}^{n}\left(L(D_i|\Theta_i)\right) = \prod_{i=1}^{n}\prod_{j=1}^{q_i} L\left(D_{ij}|\Theta_{ij}\right) \tag{4.6}$$

Where, it is assumed that there is a local independence between the parameters.

Where, N_{ij}= number of times that the configuration PA_{ij} has been observed in the data, N_{ij}=N if $PA(X_i)= \varphi$.

4.2.4. Learning the Structure

Three methods could be used to learn the BN structure:

1. Based on the detection of independencies:
   - A qualitative study is realized to determine the dependent and independent relations
   - Tries to find the network(s) that represent the majority of those relationships
2. Based on evaluation functions and searching techniques:
   - They try to find the network that represents adequately the data used.
   - Employ an evaluation function or a metric to measure the adequacy of each candidate structure.
   - The method of search (heuristics) is used to explore the space of possible solutions
3. Hybrid methods:
   - Jointly employ a search technique guided by a metric and detection of independencies.

The method that will be described herein is the one used in this research work, which is the evaluation functions and searching techniques. The method will be used because it is less sensitive to errors in individual tests as mentioned before in section 2.2.3.

This method generates models (candidate networks) by exploring the space of possible solutions (of a hyper-exponential size) through a searching technique. Each generated model is evaluated through the usage of a metric *g*, which measures the grade of adequacy between the graph DAG and the data DAG ($\mathbb{G}$ | D)

These method is characterized with the following:

1. The employed metric

2. The search space
3. The searching technique

1. The employed metric

In general, the metrics works to DAG $\mathbb{G}$ that approves:

$$\mathbb{G}^* = \arg max_{\mathbb{G} \epsilon Ç n} \; g\;(\mathbb{G}|D) \tag{4.7}$$

Where,

$\mathbb{G}^*$: is the DAG that we need to estimate

G (G|D) is the metric that measures the degree of fit between any candidate DAG $\mathbb{G}$ and the set of data D,

and $Ç_n$ is the family of DAGs that is defined for the set of variables $V_{n.}$

These metrics maximize the log-likelihood of the data (logarithms of the probability of the data given the structure). Maximizing the log-likelihood is equal to minimizing the entropy[1] of each variable given its parents, in which they try to search for the parents of each variable that provide the maximum information about it. In addition, they prefer configurations that favor the presence of the edges between variables, which express a high degree of dependency.

Minimum Description Length

The objective of these metrics is to reduce the number of elements that are necessary to represent or transmit a message. The frequent messages have short codes, and the largest codes are assigned to the least frequent messages. The principle of the minimum description length (MDL) is to select, in order to represent the messages, the encoding that the minimum length needs.

Thus, the complicated models need large MDL, but they reduce the MDL for the data given the model (are more precise). In addition, simple models require shorter MDL,

[1] entropy is a measure of the uncertainty associated with a random variable.

but the MDL of the data given the model increases. Actually, the principle of the MDL seeks to establish an appropriate balance between simplicity and precision.

For the Bayesian networks, the best network is that which minimizes the sum of the lengths of description of the network and of the data given the network.

The length of description for a network (probabilities) is proportional to the number of parameters of the joint factored distribution.

$$\sum_{i=1}^{n}(r_i - 1)q_i \tag{4.8}$$

Where,

r_i is the number of states of $X_{i,}$

q_i is the number of the possible configurations (states) of PA(X_i),

The metric MDL (log joint probability of structure and database) is given by:

$$gMDL\ (\mathbb{G}|D) = \sum_{i=1}^{n}\sum_{j=1}^{qi}\sum_{k=1}^{ri} N_{ijk}log\left(\frac{N_{ijk}}{N_{ij}}\right) - \frac{1}{2}log(N)\sum_{i=1}^{n}(r_i - 1)q_i \tag{4.9}$$

Where,

N_{ij}= number of times that the configuration PA_{ij} has been observed in the data,

N_{ijk}: number of occurrences of configurations of variables and their parents,

N: number of variables in the network.

Metrics based on Bayes

The best network is the one which maximizes the probability of obtaining a network conditioned on the database, P($\mathbb{G}$|D). These metrics employ the Bayes theorem. In which it is more convenient to work with the logarithmic space, where the following is used:

$$\log\big(P(\mathbb{G}, D)\big) = \log\big(P(\mathbb{G})\big) + \log\ (P(D|\mathbb{G})) \tag{4.10}$$

It is always assumed a uniform a priori distribution for P($\mathbb{G}$), where log(P($\mathbb{G}$)) is constant and can be eliminated.

$$P(D|\mathbb{G}) = \int_{\theta} P(D|\mathbb{G},\theta)\, P(\theta|\mathbb{G}) d\theta \qquad (4.11)$$

BD Metric (Bayesian Dirichlet)

This metric employs the a priori Dirichlet (instead of uniforms) for the parameters' distribution.

The BD metric is given by:

$$gBD(\mathbb{G}|D) = log(P(G)) + \sum_{i=1}^{n}\left[\sum_{j=1}^{qi}\left[log\left(\frac{\Gamma(\eta_{ij})}{\Gamma(N_{ij}+\eta_{ij})}\right) + \sum_{k=1}^{ri} log\left(\frac{\Gamma(N_{ijk}+\eta_{ijk})}{\Gamma(\eta_{ijk})}\right)\right]\right] \qquad (4.12)$$

Where,

η_{ijk}: are the hyper parameters of the Dirichlet distribution; $\eta_{ij} = \sum_{k}^{ri} \eta_{ijk}$,

N_{ijk}: number of occurrences of configurations of variables and their parents,

$N_{ij} = \sum_{k=1}^{ri} N_{ijk}$,

Γ: Gamma function, which satisfies $\Gamma(m) = \sum_{k=1}^{ri} N_{ijk}$

BDeu Metric

The Bayesian network a prior assigns a uniform probability to each configuration of $\{X_i\} \cup Pa_{\mathbb{G}}(X_i)$.

The BDeu metric is given by:

$$gBDeu(\mathbb{G}|D) = log(P(G)) + \sum_{i=1}^{n}\left[\sum_{j=1}^{qi}\left[log\left(\frac{\Gamma\left(\frac{1}{q_i}\right)}{\Gamma\left(N_{ij}+\frac{1}{q_i}\right)}\right) + \sum_{k=1}^{ri} log\left(\frac{\Gamma\left(N_{ijk}+\frac{1}{q_i}\right)}{\Gamma\left(\frac{1}{q_i}\right)}\right)\right]\right] \qquad (4.13)$$

Where,

r_i is the number of states of $X_{i,}$

q_i is the number of the possible configurations (states) of $PA(X_i)$,

N_{ijk}: number of occurrences of configurations of variables and their parents,

$N_{ij} = \sum_{k=1}^{ri} N_{ijk}$,

Akaike information criterion (AIC) Metric

The AIC metric gAIC of a Bayesian network structure for a database D is given by.

$$gAIC = -N \sum_{i=1}^{n} \sum_{j=1}^{qi} \sum_{k=1}^{ri} \frac{N_{ijk}}{N} log \frac{N_{ijk}}{N_{ij}} + K$$

Where,

N_{ijk}= number of cases in D where X_i= x_{ik} and PA (xi)=wij, N_{ij}=$\sum_{k=1}^{ri} N_{ijk}$,

Wij= jth instantiation of PA(xi) in D,

r_i is the number of states of $X_{i,}$

q_i is the number of the possible configurations (states) of PA(X_i)

2. The search space

A search algorithm requires a search space which contains all candidate solutions, and a set of operations that transforms one candidate solution to another. In Bayesian networks, the simplest search space consists of all DAGs containing the n variables.

3. Search algorithms

For simplified structures (like trees, in which each node has as a total one parent), finding the optimal structure is a simple task. However, in general, searching for the optimal structure is an NP-hard[2] problem.

The number of possible structures for a DAG with n nodes is (Table 4.11 illustrates a calculated example for Equation 4.14):

$$f(n) = \sum_{i=1}^{n} (-1)^{i+1} \frac{n!}{i!\,(n-i)!} 2^{(n-i)} f(n-i) \tag{4.14}$$

[2] NP-hard (non-deterministic polynomial-time hard), in computational complexity theory, is a class of problems that are, informally, "at least as hard as the hardest problems in NP"

Table 4. 11: An example illustrating the calculations of Equation 4.14

n	f(n)
0	1
1	1
2	3
3	25
4	543
5	29281
6	3781503
7	1138779265
8	783702329343
9	1213442454842880
10	41750989764305900000
11	316034593964189000000000
12	5219396513438290000000000000
13	186766007444320000000000000000000

Therefore, it is necessary to employ methods for heuristic[3] search. The most common way, is to apply a local search method, in which operators are defined on order to move from one configuration into another (neighborhood). The search is started with a determined structure that might be an empty network, a tree, or a random network.

At each iteration, all the neighbors are evaluated for the current network, in which the operator that mostly increases the metric is applied. The search ends when there is no possible improvement (local optimum). Figure 4.5 shows the typical operators that are normally used in local search.

One of the problems of the local search is that it might get stuck in either a maximum local (in which there is no any other neighbor that improves the metric). The second main problem of these algorithms is the possible presence of plateaus, which are regions of the space of assignments where no local move could improve the metric.

1. Hillclimbing search:

Is an iterative algorithm that starts with an arbitrary solution to a problem, then attempts to find a better solution by incrementally changing a single element of the solution. If the change produces a better solution, an incremental change is made to the new solution, repeating until no further improvements can be found.

[3] A heuristic is a method that might not always find the best solution, but is guaranteed to find a good solution in reasonable time.

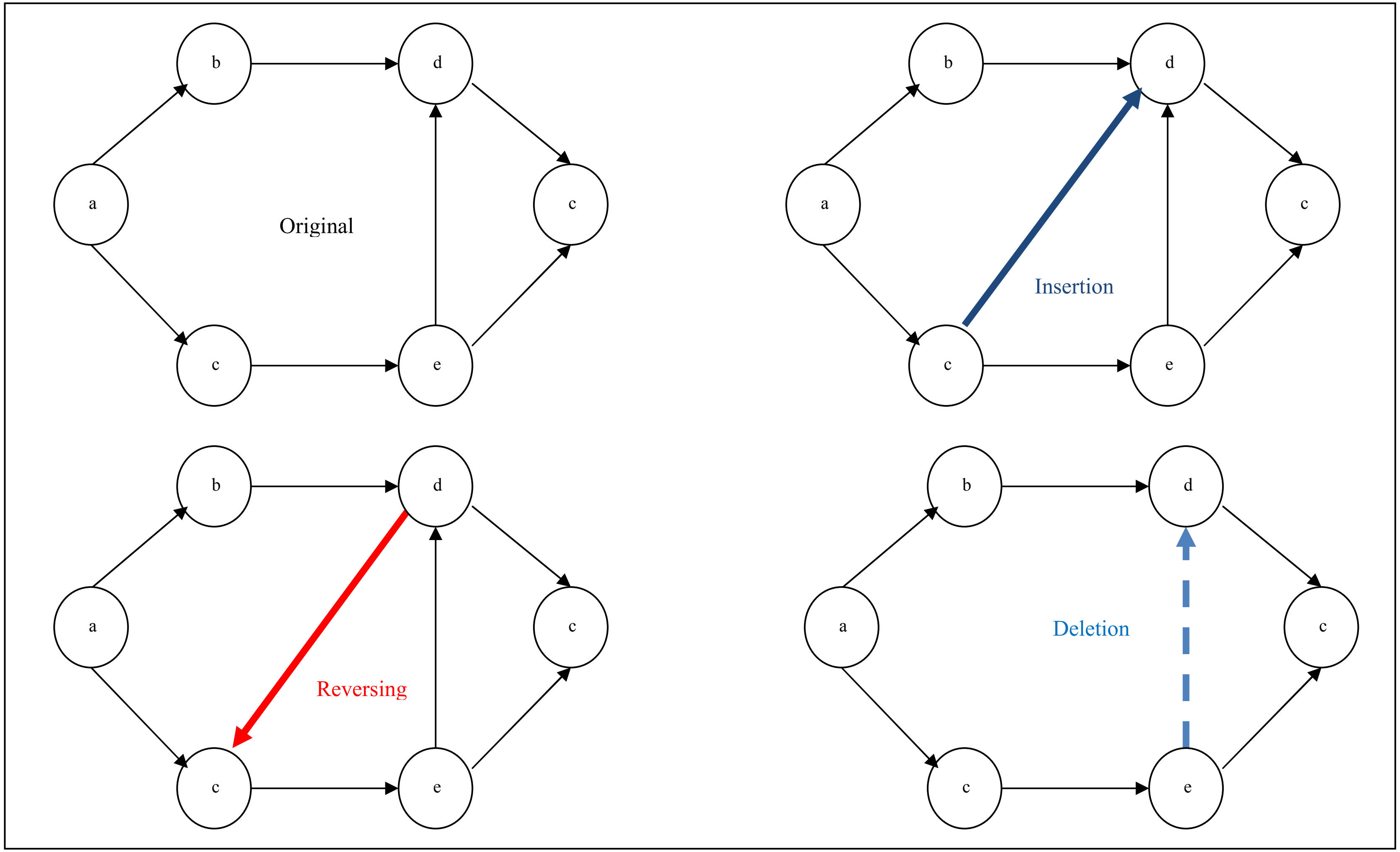


Figure 4. 5: Typical operator

Hill climbing is good for finding a local optimum (a good solution that lies relatively near the initial solution) but it is not guaranteed to find the best possible solution (the global optimum) out of all possible solutions.

2. Tabu search:

   Glover (1986) proposed a new approach, which he called Tabu Search, to allow local search methods to overcome local optima. The basic principle of Tabu search is to pursue local seacrh whenever it encounters a local optimum by allowing non-improving moves; cycling back to previously visited solutions is prevented by the use of memories, called tabu lists, that record the recent history of the search.

3. Simulated annealing

   This is a stochastic optimization approach based on the analogy to physical annealing. It tries to avoid being trapped in local optima by accepting both “good” and “bad” moves at the beginning of the iterations, and gradually lowering the probability of accepting “bad” moves. Even though in theory simulated annealing can find global optima if we lower the above probability slowly in exponential time, its performance in a practical time frame depends heavily on the parameters comprising its “cooling schedule”. In general simulated annealing is time-consuming, but has been successfully applied to many optimization problems (Tao et al., 1992).

The relative simplicity of the algorithm makes it a popular first choice amongst optimizing algorithms. Although more advanced algorithms such as simulated annealing or tabu search may give better results, in some situations hill climbing works just as well. Hill climbing can often produce a better result than other algorithms when the amount of time available to perform a search is limited, such as with real-time systems (Russell and Norvig, 2003).

In the research work carried out herein, three search algorithm (Hillclimbing, Simulated annealing and Tabu search) are used in combination with three score metrics (BDeu, MDL and AIC) to build the BNs, resulting in nine different BNs. Thus the usage of these combinations of search-score metrics was motivated by the advantages of each of which that have been described previously.

4.2.3. Bayesian networks evaluation indicators

Five indicators are used in this research work to compare the BNs built (see Equations. 4.15-4.18): accuracy, sensitivity, specificity, HMSS, and ROC area were calculated for each BN.

$$\text{Accuracy} = \frac{\text{tSI} + \text{tKSI}}{\text{tSI} + \text{tKSI} + \text{fSI} + \text{fKSI}} \times 100\% \tag{4.15}$$

$$\text{Sensitivity} = \frac{\text{tSI}}{\text{tSI} + \text{fKSI}} \times 100\% \tag{4.16}$$

$$\text{Specificity} = \frac{\text{tKSI}}{\text{tKSI} + \text{fSI}} \times 100\% \tag{4.17}$$

$$\text{HMSS} = \frac{2 \times \text{Sensitivity} \times \text{Specificity}}{\text{Sensitivity} + \text{Specificity}} \tag{4.18}$$

Where,

tSI: is true slight injured cases (cases observed to be SI and are classified (predicted) to be SI as well),

tKSI: is true killed or seriously injured cases (cases observed to be KSI and are classified (predicted) to be KSI as well),

fSI: false slight injured cases (cases observed to be SI but classified (predicted) to be KSI),

fKSI: false killed or seriously injured cases (cases observed to be KSI but classified (predicted) to be SI).

Accuracy in equation (4.15) is the proportion of instances that were correctly classified by the classifier, where it only gives information on the classifier's general performance.

Sensitivity represents the proportion of correctly predicted slight injured among the entire observed slight injured. Specificity represents the proportion of correctly predicted killed or seriously injured among all the observed killed or seriously injured (see equations 4.16 and 4.17). Another measure used to assess the performance of the Bayesian network built is the Harmonic Mean of Sensitivity and Specificity (HMSS), which gives an equal weight of both sensitivity and specificity (see equation 4.18).

Another indicator is the Receiver Operating Characteristic Curve (ROC) Area. What ROC curves represent is the true positive rate (sensitivity) vs. the false positive rate (1−specificity). ROC curves are more useful as descriptors of overall test performance, reflected by the area under the curve, with a maximum of 1.00 describing a perfect test and an ROC area of 0.50 describing a valueless test.

Other measures used in the literature to evaluate the performance of BNs specifically include both the Most Probable Explanation (MPE) (Simoncic, 2004) and the complexity or the total number of BN arcs (Cruz-Ramírez et al., 2007). MPE is a technique that is developed for generating explanation in BNs, in which the configuration with the maximum posterior probability is calculated (Pearl, 2004).

For the analysis of traffic accidents' injury severity, and in order to determine the optimal BN, the measures described above will be calculated first: accuracy, sensitivity, specificity, ROC area, the MPE and the complexity of the built BNs. Later, the best BN found in terms of these measures will be used for inference.

4.2.4. Variables Selection methods

Even after choosing the best Bayesian network to represent the data used, there would still be some redundant information that degrades the efficiency of the classifier. Thus, possibly not all the variables contribute to the determination of the class. In addition, the computational cost might be high to examine all the variables. Therefore, the incorporation of the variables' selection techniques (or selection of characteristics) eliminates irrelevant and/or redundant variables.

A relevant feature is neither irrelevant nor redundant to the target concept; an irrelevant feature does not affect the classification task in any way, and a redundant feature does not add anything new to the classification task (John et al., 1994).

The objective of using such techniques is to reduce the number of variables that characterize the data used. In return, the efficiency of the classification is enhanced, and the parameters' estimation becomes more robust.

In general, variables selection attempts to select the minimally sized subset of features according to the following criteria (Dash and Liu, 1997):

- The classification accuracy does not significantly decrease; and
- The resulting class distribution, given only the values for the selected features, is as close as possible to the original class distribution, given all features.

According to Dash and Liu (1997) there are four basic steps in a typical feature selection method:

1. A generation procedure: used to generate the next candidate subset, basically, it generates subsets of variables for evaluation. It can start with:
    a. no variables: variables are iteratively added or removed
    b. with all variables: variables are iteratively added or removed, or
    c. with a random set of variables: variables are either iteratively added or removed or produced randomly thereafter.
2. An evaluation function: to evaluate the subset under examination, in which it measures the goodness of a subset produced by some generation procedure, this value is then compared with the previous best. If the value was not found to be better, then it replaces the previous best subset.
3. A stopping criterion: to decide when to stop, where without stopping criterion the variable selection process may run exhaustively or forever through the space of subsets.

1. Generation (search) procedure

There are different approaches that are used to generate candidate subsets:

1. Complete

A complete search is done in order to find an optimal subset according to the evaluation function used. Different heuristic functions used to reduce the search without risking the chances of finding an optimal subset. However, it is a costly procedure if the variables space is too large.

2. Heuristic

All remaining variables that are to be selected (rejected) are considered for selection (rejection) in each iteration of this generation procedure. The generation of subsets is incremental (increasing or decreasing), and the search space is smaller and faster in producing results. However, it might miss out some features of high order relations.

3. Random

In this procedure, there is no predefined way to select candidate variables; it picks variables at random (i.e. probabilistic approach). Here, the optimal subset depends on the number of iterations used, which in their turn, depend on the available resource. However, it requires more user defined input parameters, in which results will depend on how these parameters are defined.

2. Evaluation Functions

The evaluation function tries to measure the discriminating ability of a variable or a subset to distinguish the different class labels.

Evaluation functions are divided into five categories according to Dash and Liu (1997):

1. Distance functions: This measure select those variables that support instances of the same class to stay within the same proximity. Thus, instances of the same class should be closer in terms of distance than those from different class. Methods using this evaluation function are called “filter methods”.
2. Information measures: this measure determines the information gain from a variable. For a variable X to be preferred over a variable Y, the information gain from variable X is to be greater than that from variable Y. Where, the information gain from variable X is defined as the difference between the prior uncertainty and expected posterior uncertainty using X. Methods using this evaluation function are called “filter methods”.
3. Dependence measures: they qualify the ability to predict the value of one variable from the value of another. If the correlation of variable X with class C is higher than

the correlation of variable Y with C, then feature X is preferred to Y. A slight variation of this is to determine the dependence of a variable on other variables; this value indicates the degree of redundancy of the variable. If the variable is heavily dependent on another, then it is redundant. Methods using this evaluation function are called “filter methods”.

4. Consistency measures: they are different from the other measures in which they rely heavily on the training dataset and the of Min-Variables bias in selecting a subset of variables. Where, Min-Variables bias prefers consistent hypotheses definable over as few variables as possible. These measures find out the minimally sized subset that satisfies the acceptable inconsistency rate that is usually set by the user. An example of inconsistency is that when we have two instances that have matching variable values but different class label. Methods using this evaluation function are called “filter methods”.
5. Classifier Error Rate Measures: variables are selected using the classifier that later on uses these variables in predicting the class labels of unseen instances, the resulting accuracy is very high, but the computations are quite slow. If the error rate resulting from selecting a subset is less than a predefined threshold, then select the variable subset. Methods using this evaluation function are called “wrapper methods”.

In general, the previously mentioned evaluators belong to either of two approaches, filters pr wrappers.

- Filters: a preliminary process, composed of:
  - A specific search in the space of variables
  - A specific function of evaluation

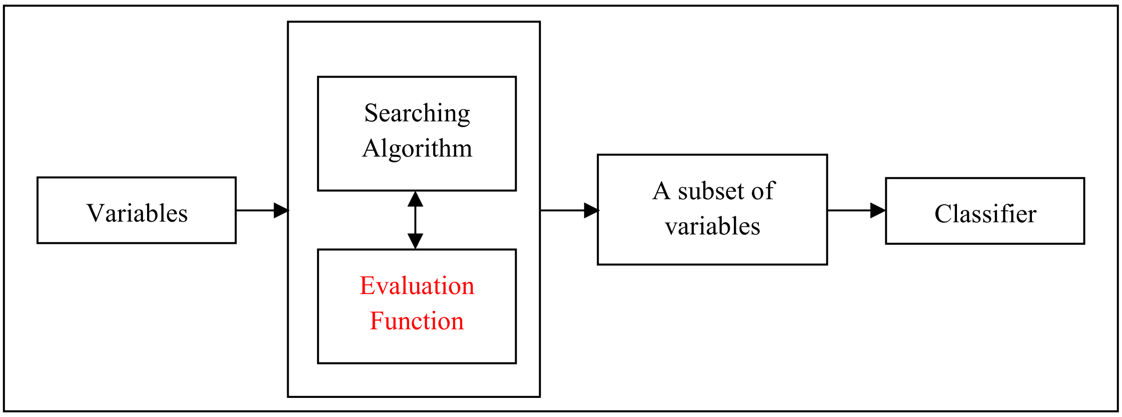


Figure 4. 6: An illustration of the Filter approach

- Wrappers: in this approach the selection of variables is a consequence of classification, in which:
    - A search in the space of variables is done in combination with a search in the space of configurations
    - The selection of variables is done based on classifier evaluation function.

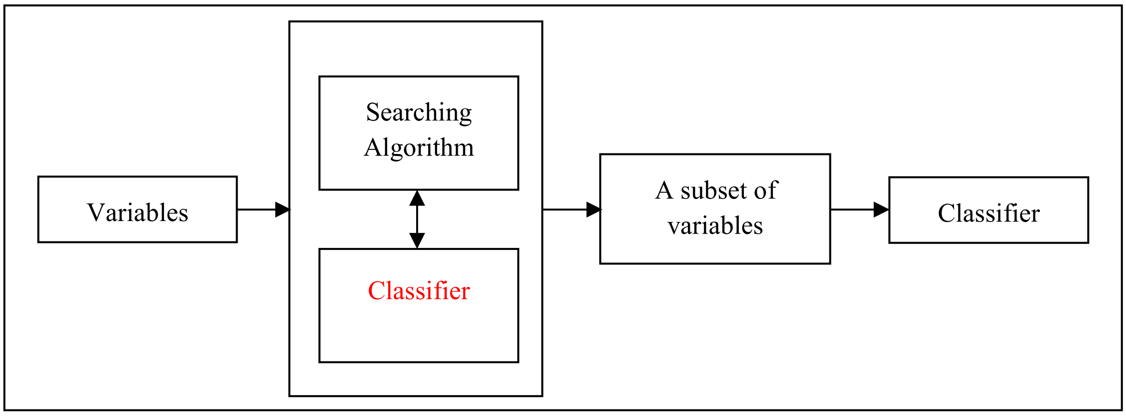


Figure 4. 7: An illustration of the Wrapper approach

3. Stopping criterion

Generation procedures and evaluation functions can influence the choice for a stopping criterion, the following describes the stopping criteria determination for each.
Stopping criteria based on a generation procedure include:

- whether a predefined number of features are selected,
- and whether a predefined number of iterations is reached.

Stopping criteria based on an evaluation function can be:

- whether addition (or deletion) of any feature does not produce a better subset;
- and whether an optimal subset according to some evaluation function is obtained. The loop continues until some stopping criterion is satisfied.

In this research work, six evaluators with eleven search methods were used. A brief description of each of the evaluators used is given below:

1. Correlation-based variable selection (CfsSubsetEval): this evaluator measures the predictive ability of each variable individually and the degree of redundancy among

them. It selects the sets of variables that are highly correlated with the class but have low inter-correlation with each other (Hall, 1998).

2. Consistency-based variable selection (ConsistencySubsetEval): this evaluator measures the degree of consistency of the variable sets in class values when the training values are projected onto the set. This evaluator is usually used in conjunction with a random or exhaustive search (Liu and Setiono, 1996).

3. Classifier Subset Evaluator (ClassifierSubsetEval): this evaluator uses the classifier specified in the object editor as a parameter, to evaluate sets of variables on the training data or on a separate holdout set (Witten and Frank, 2005).

4. Wrapper Subset Evaluator (WrapperSubsetEval): this evaluator uses a classifier to evaluate variable sets and it employs cross-validation to estimate the accuracy of the learning scheme for each set (Khhavi and John, 1997).

5. Filtered Subset Evaluator (FilteredSubsetEval): The filter model evaluates the subset of variables by examining the intrinsic characteristic of the data without involving any data-mining algorithm (Witten and Frank, 2005).

6. Cost Sensitive Subset Evaluator (CostSensitiveSubsetEval): This evaluator projects the training set into attribute set and measure consistency in class values, making the subset cost sensitive (Liu and Setiono, 1996).

A brief description of each of the search methods used is given below:

1. Best First: this Search method uses greedy hillclimbing augmented with a backtracking facility to search through the variables' subsets. Best first may start with an empty set of variables and search forward, or start with a full set of variables and search backward, or start at any point and search in both directions (Pearl, 1984).

2. Genetic Search: An initial population is formed by generating many individual solutions. During each successive generation, a proportion of the existing population is selected to breed a new generation. This process is repeated until increasing the average fitness, and reaching a termination condition (Goldberg, 1989).

3. Greedy Stepwise: Performs a greedy forward or backward search through the space of variables' subsets. The search could be initiated with none or with all the variables or from an arbitrary point in the space. Thus, the search is stopped when the addition or deletion of any variables that remains; results in a decrease in evaluation (Russell and Norvig, 2003).

4. Linear Forward Selection: this search method is an extension of Best First where a fixed number k of variables is selected, whereas k is increased in each step when fixed-width is selected. The search direction can be forward or floating forward selection (with optional backward search steps) (Gutlein et al., 2009).

5. Scatter Search V1: Starts with a population of many significant variables and stops when the result is higher than a given threshold or when no further improvement could be attained (García López et al., 2006).

6. Tabu Search: It explores the solution space beyond the local optimum, once a local optimum is reached; upward moves and those worsening the solutions are allowed (Hedar et. al., 2008).

7. Rank Search: Uses a variable/subset evaluator to rank all variables. If a subset evaluator is specified then a forward selection search is used to generate a ranked list. From the ranked list of variables, subsets of increasing size are evaluated (Witten and Frank, 2005).

8. Exhaustive Search: Performs an exhaustive search through the space of variables' subsets starting from the empty set of variables and reporting the best subset found (Witten and Frank, 2005).

9. Subset Size Forward Selection: Performs an interior cross-validation, where it is performed on each fold to determine the optimal subset-size. In the final step the search is performed on the whole data (Gutlein et al., 2009).

10. Random Search: Performs a random search in the space of variables' subsets. A random search is started from a random point, if no initial point is chosen, and reports the best subset found. If a start set is set, Random searches randomly for subsets that are as good as or better than the start point with the same number of variables or with a lower number of variables (Liu and Setiono, 1996).

11. Race Search: Races the cross validation error of competing subsets of variables, and is only used with a ClassifierSubsetEval (Moore and Lee, 1994).

The search and evaluation algorithms used herein are all of these that exist in the WEKA 3.7.0 software.

CHAPTER 5

Results and discussion

In this chapter, the results obtained by constructing Bayesian networks are presented, and their performance evaluation is presented as well. At all stages of building BNs and selecting the variables, WEKA 3.7.0 software was used (Witten and Frank, 2005).

5.1. Data preparation for Bayesian networks

The only preprocessing filter used on this dataset is the unsupervised variable filter for replacing missing values. This filter replaces the missing values with the modes and means from the training data.

The original dataset obtained from the DGT is divided into two subsets: a training set containing 2/3 of the data (1,024 records), and a testing set containing the rest of the data (512 records). The testing set is used to validate the results obtained using the training set. Multiple repetitions or trials (10 times) of cross validation are used to reduce variability, and the validation results are averaged over the trials.

5.2. Comparison of the BNs constructed using all the variables

After running the different learning search methods and score metrics, eighteen different average networks (average of 10 runs) were obtained, and they are displayed in (Figures 1-9) in the Appendix III. No causal interpretation of the arcs in the networks is assumed, such as other researchers consider (Acid et al., 2004). Thus, the arcs are interpreted as direct dependence relationships between the linked variables, and the absence of arcs means the existence of conditional independence relationships.

In order to measure the differences and resemblances between models, Table 5.1 shows the two numbers *l/a* for each pair of sore metrics that belong to the same search method, where *l* is the number of common edges (in either direction), and *a* number of common arcs between the networks learned by these algorithms.

Table 5.1: Number of common edges and arcs, *l/a*, between pairs of learned networks

	Hillclimbing		
	BDeu	MDL	AIC
BDeu	28/28	-	-
MDL	26/22	28/28	-
AIC	23/15	21/17	35/35
	Simulated Annealing		
	BDeu	MDL	AIC
BDeu	28/28	-	-
MDL	23/19	28/28	-
AIC	23/13	23/10	38/38
	Tabu		
	BDeu	MDL	AIC
BDeu	23/23	-	-
MDL	20/17	23/23	-
AIC	21/17	21/17	23/23

Figure 5.1 displays the edges in common to all the networks: three arcs and fourteen undirected edges. Five additional edges are also displayed in Figure 5.1 that are supported by all the networks except one. It should be noticed that the number of possible edges in this domain is 171, and only a total of 48 different edges appear in these models. Thus, the 18 models agree in the presence of 16 edges and the absence of 123 edges, in which this could be interpreted as the existence of 16 direct dependence and 123 conditional independence assertions between pairs of variables.

Moreover, Figure 5.1 shows that the existence of 16 direct dependence relationships that are common to all models indicates a strong relationships between the following pairs of variables: (ACT-SEV), (ACT-VI), (VI-OI), (NOI-SEV), (NOI-OI), (SEV-CAU), (SEV-DAY), (SEV-AGE), (SEV-LIG), (SEV-LAW), (SEV-ATF), (LIG-TIM), (LAW-PAW), (PAW-SHT), (SHT-PAS), (SID-ATF).

Thus, in order to compare the three different search methods (Hillclimbing, Simulated Annealing, and Tabu) that were used along with 3 different score metrics (BDeu, MDL, and AIC) to build Bayesian networks. The experiments were obtained for both the training and the test sets, in which the averages and the standard deviations of 10 trials for each are illustrated in Table 5.2.

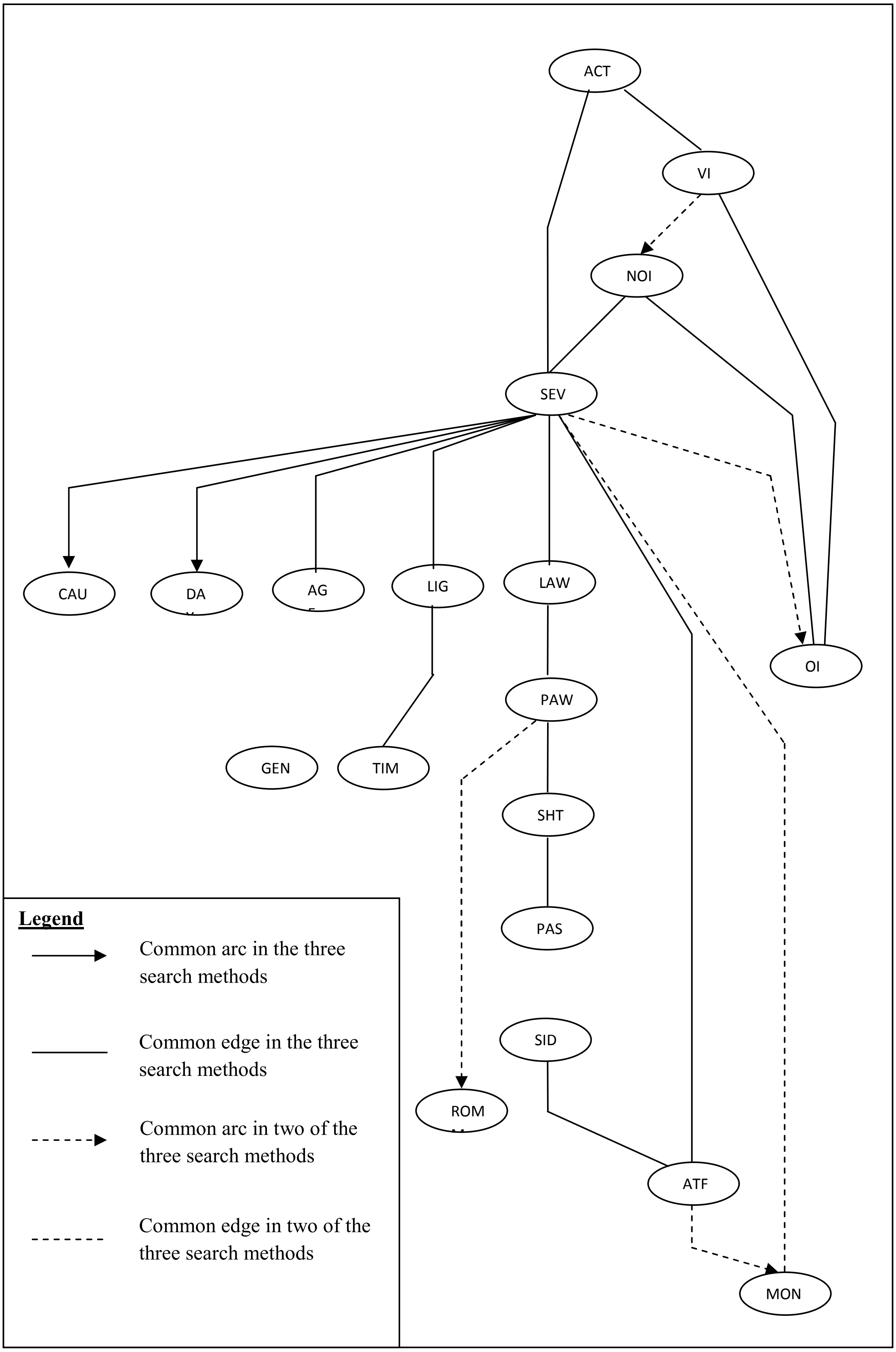

ACT
VI
NOI
SEV
CAU
DA
AG
LIG
LAW
OI
PAW
GEN
TIM
SHT
PAS
SID
ROM
ATF
MON
Legend
Common arc in the three search methods
Common edge in the three search methods
Common arc in two of the three search methods
Common edge in two of the three search methods

Figure 5.1: displays the edges in common between the networks built using all the search and score methods.

Table 5.2: Results of using search and score methods

Score	BDeu		MDL		AIC	
Dataset	training	test	training	test	training	test
Indicator	average±s.d.*	average±s.d.*	average±s.d.*	average±s.d.*	average±s.d.*	average±s.d.*
Search	Hillclimbing					
Accuracy	0.61±0.01	0.57±0.02	0.60±0.01	0.59±0.02	0.58±0.01	0.58±0.03
Sensitivity	0.74±0.02	0.65±0.04	0.73±0.02	0.65±0.03	0.66±0.02	0.63±0.04
Specificity	0.44±0.03	0.49±0.05	0.45±0.03	0.53±0.05	0.47±0.03	0.53±0.04
HMSS	0.55±0.02	0.56±0.03	0.56±0.02	0.58±0.03	0.55±0.02	0.58±0.02
ROC Area	0.62±0.04	0.58±0.02	0.61±0.02	0.62±0.02	0.58±0.02	0.61±0.03
no. of arcs	28±1	28±2	28±2	28±3	35±1	35±2
Search	Simulated Annealing					
Accuracy	0,60±0,02	0,58±0,02	0,60±0,02	0,59±0,02	0,58±0,01	0,57±0,02
Sensitivity	0,67±0,13	0,66±0,11	0,66±0,13	0,65±0,10	0,64±0,12	0,63±0,06
Specificity	0,49±0,14	0,48±0,09	0,50±0,14	0,51±0,09	0,49±0,11	0,49±0,08
HMSS	0,54±0,02	0,54±0,03	0,54±0,02	0,56±0,03	0,54±0,01	0,54±0,02
ROC Area	0,62±0,01	0,60±0,02	0,62±0,01	0,61±0,03	0,60±0,02	0,59±0,03
no. of arcs	28±1	28±1	28±2	28±2	38±2	38±2
Search	Tabu					
Accuracy	0,60±0,02	0,58±0,02	0,59±0,02	0,58±0,02	0,60±0,02	0,58±0,02
Sensitivity	0,69±0,16	0,67±0,14	0,69±0,15	0,68±0,14	0,70±0,17	0,67±0,15
Specificity	0,47±0,20	0,46±0,18	0,46±0,19	0,47±0,19	0,46±0,21	0,46±0,19
HMSS	0,50±0,17	0,50±0,18	0,51±0,13	0,51±0,18	0,49±0,17	0,50±0,18
ROC Area	0,62±0,04	0,59±0,05	0,62±0,05	0,60±0,05	0,61±0,05	0,60±0,04
no. of arcs	23±5	23±5	23±5	23±5	23±5	23±5

It can be seen that both the training and the test results are very similar. The accuracies performed in this study did not vary significantly, the maximum values obtained using all the search and score methods are summarized in Table 5.3. The highest accuracy obtained for the test set was for the Hillclimber search + MDL score metric (59%) while for the training set it was best obtained for the Hillclimbing search + Bdeu score metric. Abdel wahab and Abdel-Aty (2001) used artificial neural networks to model injury severity in traffic accidents. They obtained accuracies of 65.6% and 60.4% for training and testing sets respectively when using an MLP neural network, 56.2% when using fuzzy ARTMAP neural network and 58.1% when using O-ARTMAP. Thus, the results obtained in this paper were within the range of accuracies found by Abdel wahab and Abdel-Aty (2001).

Table 5.3: Maximum values of indicators as obtained using the different search and score methods

	Maximum value	Training		Maximum value	Test	
		Search	Score		Search	Score
Accuracy	0,61	Hillclimbing	Bdeu	0,59	Hillclimbing	MDL
Sensitivity	0,74	Hillclimbing	Bdeu	0,68	Tabu	MDL
Specificity	0,5	Simulated Annealing	MDL	0,53	Hillclimbing	MDL, AIC
HMSS	0,56	Hillclimbing	MDL	0,58	Hillclimbing	MDL, AIC
ROC Area	0,62	Hillclimbing	Bdeu	0,62	Hillclimbing	MDL

In addition, the highest sensitivity for the test set was obtained for the training set using the Hillclimbing search + BDeu score; where 74% of the cases observed to be slight were also predicted to be slight. The highest sensitivity for the test set was obtained using the Tabu + BDeu, where (68%) of the slightly injured cases were correctly classified. On the other hand, the specificity results indicated that the ability of all the search methods used to classify killed or seriously injured were relatively poor. None of the search methods achieved good results regarding the classification of killed or seriously injured (specificity); the best result obtained for the training set was using Simulated Annealing search + MDL, where the results for the test set indicated that 53% of the cases were correctly classified as killed or seriously injured using the Hillclimbing search + MDL and AIC score metrics.

The results of sensitivity for all the search methods were relatively better than those for specificity, thus indicating that the models were capable of classifying slight injured rather than killed or seriously injured. This, however, was expected, since the original dataset contained more slight injuries.

HMSS could be used as a single measure of performance of the BN instead of using sensitivity and specificity separately. The results indicated that the best HMSS was obtained using Hillclimbing search + MDL (56%) for the training set , while for the test set the highest HMSS was obtained using Hillclimbing search + MDL with (62%).

Figures (5.3-5.5) show the area under ROC curves for the BNs built using the three search methods based on the test set, where the X-axis represents false positive rate and the Y-axis represents the true positive rate. The maximum area under ROC for the training set was obtained using HIllclimbing search + Bdeu, while for the test set it was obtained using Hillclimbing search and MDL score with 0.62.

Finally, the least complicated structure was obtained using the Tabu search followed by Hillclimbing. DAGs for all the constructed BNs and for all the different search methods and score metrics are found in the Appendix III.

Thus, according to the results presented in Table 5.3 for the performance measures that are calculated for each BN using different search and scores methods, the best performance is obtained using the Hillclimbing search method according to the results of Accuracy, HMSS, and ROC area, and thus it will be used for further analysis.

Following Simoncic (2004), and in order to choose between the different score metrics, the most convenient way that could be used to analyze the graphical performance of the three metrics is to calculate the Most Probable Explanation (MPE) for the training dataset and compare it with the results obtained from the test dataset.

MPE is given by the most probable configuration of values for all variables in the BN. For the three estimated structures, the MPE is given by the following values for variables (see Table 4.1):

ACT=AS; AGE=(25-64]; ATF=GW; CAU=DC; DAY=WD; GEN=M; LAW=WID; LIG=DL; MON=SUM; NOI=1; OI=2; PAS=Y; PAW=WID; ROM=SLD; SEV=SI; SHT=THI; SID=WR; TIM=(12-18]; VI=2.

Given the estimated BN structures (BDeu, MDL and AIC) and the conditional probabilities for each node (see Figures 1-3 in Appendix III), the probability of the MPE can be computed as shown in Table 5.4.

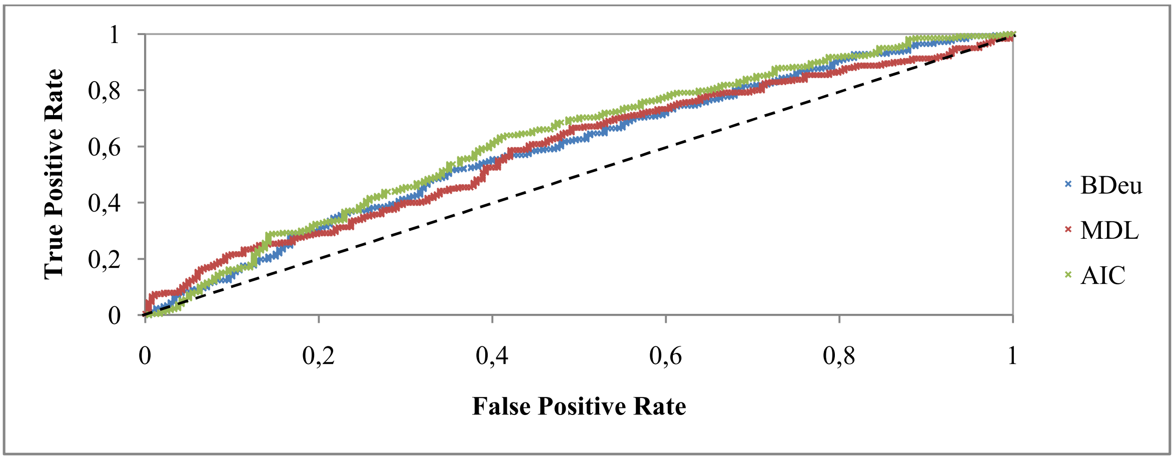


Figure 5.2: ROC Area for Hillclimber search method with BDeu, MDL, and AIC score metrics

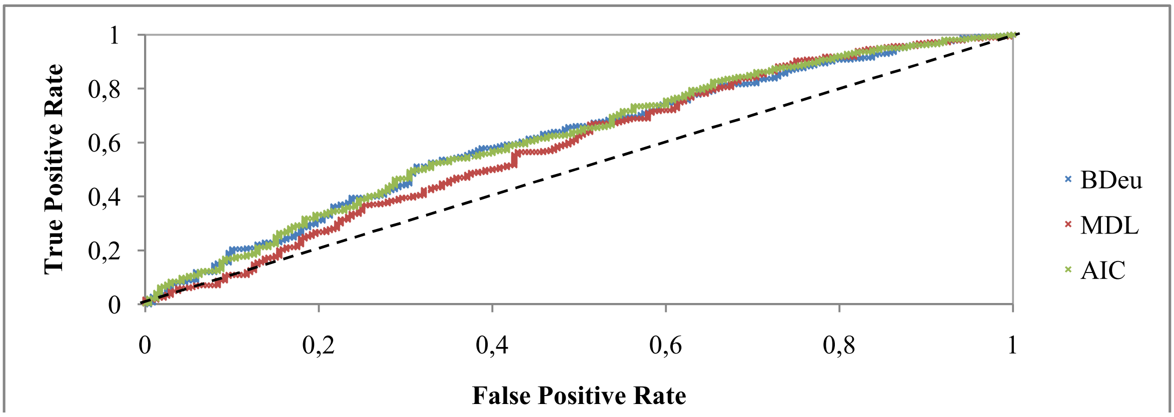


Figure 5.3: ROC Area for Tabu search method with BDeu, MDL, and AIC score metrics

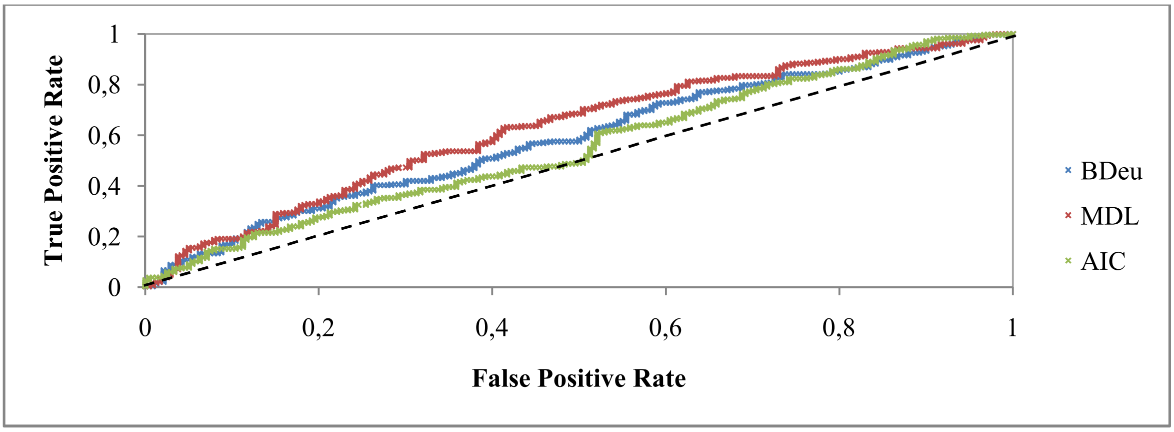


Figure 5.4: ROC Area for Simulated Annealing search method with BDeu, MDL, and AIC score metrics

For the network built by the BDe score metric, the MPE is given by the probability values shown in Table 5.4, column 2, row 2. Using these values, MPE for the BDe score equals 0.00088. The same calculations for the test dataset produced MPE_{test}=0.00081. This comparison of MPE and MPE_{test} can provide an indication of the quality of the estimated BN using BDe score metric; where it can be seen that there is a difference (8.2%) between the MPE produced by the training dataset and the test dataset.

The MPE for the MDL BN is given by the probability values shown in Table 5.4, column 2, row 3. Using these values, MPE equals 0.00076. The test dataset produced MPE_{test}=0.00073. Therefore, the MPE as explained by the MDL is closer to the test dataset estimation (4.4% of difference), thus representing a network that is more capable of explaining different data.

Table 5.4: MPE for Hillclimbing search method and the three score metrics.

Score metric	MPE Formulas	MPE	MPE$_{test}$
BDeu	P(ACT=AS)•P(AGE=(25-64]|SEV=SI)•P(ATF=GW|SEV=SI,SID=WR)• P(CAU=DC|SEV=SI)•P(DAY=WD|SEV=SI)•P(GEN=M|SEV=SI)• P(LAW=WID|SEV=SI)•P(LIG=DL|SEV=SI)•P(MON=SUM|SEV=SI,ATF=GW)• P(NOI=1|VI=2)•P(OI=2)|SEV=SI,NOI=1,VI=2)•P(PAS=Y|SEV=SI,SHT=THI)• P(PAW=WID|SEV=SI,LAW=WID)•P(ROM=SLD|SEV=SI,PAS=Y,PAW=WID)• P(SEV=SI|ACT=AS,NOI=1)•P(SHT=THI|PAW=WID)•P(SID=WR|PAS=Y)• P(TIM=(12-18]|SEV=SI,LIG=DL)•P(VI=2|ACT=AS)	0.00088	0.00081
MDL	P(ACT=AS|PAS=Y)•P(AGE=(25-64]|SEV=SI)•P(ATF=GW|SEV=SI,SID=WR)• P(CAU=DC|SEV=SI)•P(DAY=WD|SEV=SI)•P(GEN=M|SEV=SI)• P(LAW=WID|SEV=SI,PAW=WID)•P(LIG=DL|SEV=SI,TIM=(12-18])• P(MON=SUM|SEV=SI,ATF=GW)•P(NOI=1|VI=2)•P(OI=2|SEV=SI,NOI=1,VI=2)• P(PAS=Y|SHT=THI)•P(PAW=WID|SHT=THI)• P(ROM=SLD|SEV=SI,PAS=Y,PAW=WID)• P(SEV=SI|SHT=THI,PAS=Y,ACT=AS,NOI=1)•P(SHT=THI)•P(SID=WR|PAS=Y)• P(TIM=(12-18]|VI=2)•P(VI=2|ACT=AS)	0.00076	0.00073
AIC	P(ACT=AS|VI=2)•P(AGE=(25-64]|LIG=DL)•P(ATF=GW|SID=WR)• P(CAU=DC|SEV=SI,GEN=M)•P(DAY=WD|SEV=SI,VI=2)•P(GEN=M|DAY=WD)• P(LAW=WID|SEV=SI,ROM=SLD,PAW=WID)• P(LIG=DL|MON=SUM,TIM=(12-18])•P(MON=SUM|PAS=Y,ATF=GW)• P(NOI=1|AGE=(25-64],VI=2)•P(OI=2|SEV=SI,NOI=1,VI=2)• P(PAS=Y|PAW=WID,SHT=THI)•P(PAW=WID|SHT=THI)• P(ROM=SLD|PAS=Y,PAW=WID)• P(SEV=SI|MON=SUM,LIG=DL,ATF=GW,AGE=(25-64],NOI=1,ACT=AS)• P(SHT=THI|ACT=AS)•P(SID=WR|PAS=Y,ROM=SLD)•P(TIM=(12-18])• P(VI=2|TIM=(12-18])	0.00100	0.00092

The MPE for the AIC BN is given by the probability values shown in Table 5.4, column 2, row 4. Using these values, MPE is 0.00100. The test dataset produced MPE$_{test}$=0.00092. The most probable explanation has a higher probability than that produced by the test subset (8.7% of difference).

To conclude, the above calculations performed for the MPE and for the three score metrics when compared to the MPEs calculated for the test subset shows that, relatively speaking, the MDL score metric MPE gives the best explanation with regard to the MPE$_{test}$. Whereas the difference between MPE of the built network and that computed for the test subset is the least among all the other MPEs produced by BDeu, and AIC score metrics.

The last step in comparing the various score metrics and evaluating their performance was to compare the graphs' complexity, measured by the total number of arcs produced by the three score metrics studied.

Table 5.2 shows the number of arcs obtained using the three score metrics. The most complicated BN (having the highest number of arcs) is the BN built using the AIC score; this BN has 35 arcs, while the least complicated BN was the BN built by the BDeu score, with 28 arcs; followed by the BN built by the MDL score, with 29 arcs.

The results obtained by the different BNs built showed that the three different score metrics did not vary significantly in terms of their accuracy, specificity, sensitivity, HMSS and ROC area. This however, indicates that BNs are valid for analyzing traffic accident injury severities and builds on the results presented by Simoncic (2004), who indicated that BNs could effectively be used to analyze this specific problem.

On the other hand, the results for the complexity of the BN graphs, the number of arcs and the MPE show some differences between the three score metrics. MDL shows the best results in terms of MPE (smaller differences between training and test sets). BDeu and MDL show the best results in terms of complexity of BN graphs and number of arcs.

Thus, the results obtained by MDL score shows that it produced a network which was relatively successful in terms of classification and prediction, where it had the best total accuracy obtained for the test set (59%). Also, HMSS showed a relatively good result for both training and testing sets respectively (56-58%) and the ROC area results were good as well (61-62%).

In addition to the ability of BNs to represent available information and to make predictions when new data is received, they can also be useful tools for performing specific inference tasks. A network model can be used to compute the posterior probability of any variable in different contexts (Acid et al., 2004).

In order to illustrate this possibility, the posterior probability distribution were calculated using Hillclimbing search and the MDL score metric, so that some indications of the values of variables that contribute to the occurrence of a killed or seriously injured (KSI) individual in a traffic accident could be illustrated.

Table 5.5 assists in the identification of the variables and values that contribute the most to the occurrence of a KSI individual in a traffic accident. For each variable, the probability of a value was set to be 1.0 (setting evidence) and the other values of the same variable were set to be 0.0. Thus, the associated probability of severity was calculated. Underlined values in Table 5.5 show the values of variables in which the probability of a KSI was found to be higher than that of SI.

For example, this table shows that assigning a probability of 1.0 to the value AS (angle or side impact) of the variable ACT, the probability of SI becomes 0.6219 and the probability of KSI becomes 0.3780. These probabilities are calculated from the conditional probability table of the BN built using the MDL score. Since it is intended to determine which values of variables contribute the most to the occurrence of a KSI individual in a traffic accident, Table 5.5 does not include the variables in which the values of probabilities of SI are always higher than those of KSI. Setting evidences for the values of variables used to build the BN indicated that ACT, AGE, LIG and NOI were found to be significant.

Table 5.5: Inference results for variables that are associated with KSI in traffic accidents.

Variables	Values	Probabilities when setting evidences	
		SI	KSI
ACT	AS	0.6219	0.3780
	CF	0.5226	0.4773
	HO	0.3412	0.6587
	O	0.5808	0.4191
	PU	0.6683	0.3316
	R	0.4944	0.5055
	SP	0.6066	0.3933
AGE	[18-25]	0.4999	0.5000
	(25-64]	0.5567	0.4432
	≥64	0.5937	0.4062
LIG	D	0.5486	0.4513
	DL	0.5615	0.4384
	I	0.6239	0.3760
	S	0.6254	0.3745
	W	0.4527	0.5472
NOI	1	0.4957	0.5042
	>1	0.6545	0.3454

SI: slight injured; KSI: killed or seriously injured

A detailed discussion of the most significant variables that were found to contribute to the occurrence of a killed or seriously injured (KSI) individual in a traffic accident is given below.

1. *Accident type (ACT)*:

As shown in Table 5.5, when setting the probabilities of both HO (head on collisions) and R (rollover) values to be equal to 1.0, the probability of having KSI accidents increased, which means that these types of accidents are more significant in accidents with killed or seriously injured. Kockelman and Kweon (2002) found that head on crashes were more dangerous than angle crashes, left-side, and right-side crashes; they also found that they were significant in accidents that involved killed or seriously injured, but rollover crashes were more dangerous than all of the preceding crash types.

2. *Age (AGE)*:

The results shown in Table 5.5 indicate that drivers in the age group [18-25] years were found to be more involved in accidents that resulted in KSI. Tavris et al. (2001) found that male drivers in the age group (16–24) years were much more likely to be involved in killed or seriously injured accidents than those involving older drivers.

3. *Lighting (LIG)*:

Gray et al. (2008) found that among the factors that lead to a slight injury is driving in the daylight, and that more severe injuries are predicted during darkness. Helai et al. (2008) and Abdel-Aty (2003) found the same results. This coincides with the results found in this study, which indicate that roadways without lighting (W) are associated with accidents that had KSI individuals.

4. *Number of injuries (NOI)*:

The results obtained in this research work indicate that when an accident results in one injury, it is more likely to be a serious injury or even fatal. Scheetz et al. (2003) used classification and regression trees to model the injury severity of traffic accidents. They also found that the number of injured occupants was a significant factor in classifying injury severity.

5.3. BNs constructed using selected variables

A number of BNs are constructed using some selected variables in order to evaluate the performance of Bayesian networks when using only the most significant variables.

All the possible combinations of evaluator-search algorithms described in section 4.2.4. were applied (59 combinations). The total number of combinations was supposed to be 66; however, seven combinations were found to be incompatible. Table 5.6 shows the unused combinations.

Table 5. 6: Search-evaluator combinations that were found to be incompatible

Search Method	Evaluator
Race Search	CfsSubsetEval ConsistencySubsetEval WrapperSubsetEavl FilteredSubsetEval CostSensitiveSubsetEval
Tabu Search	ClassifierSubsetEval WrapperSubsetEavl

Table 5.7 shows the variables selected after running the 59 different combinations of the evaluator-search algorithms. The number of selected variables lies between four variables (ACT, ATF, LIG and NOI), as obtained using Correlation-based variable selection, Filtered Subset Evaluator and Cost Sensitive Subset Evaluator with several search methods, and a maximum number of sixteen selected variables.

Table 5.8 shows the number of times that each one of the eighteen variables has been selected. There are three variables that were selected approximately 95% of times. The variables ACT (accident type) and LIG (lighting) were selected 58 times over 59 combinations, which mean that they were selected almost by all the evaluator-search combinations.

Table 5. 7: Selected variables for different combinations of evaluator-search method.

Number of variables	Variable selection method		Variables selected	BN
	Evaluator	Search Method		
18	The original training set of variables		ACT, AGE, ATF, CAU, DAY, GEN, LAW, LIG, MON, NOI, OI, PAS, PAW, ROM, SHT, SID, TIM, VI	BN-18
4	CFS	Best first, Exhaustive, Greedy stepwise, Linear forward selection, Rank, Scatter V1, Tabu	ACT, ATF, LIG, NOI	BN-4
	Cost sensitive	Best first, Exhaustive, Greedy stepwise, Linear forward selection, Random, Rank, Scatter V1, Tabu		
	Filtered	Rank		
5	CFS	Genetic, Random	ACT, ATF, LIG, NOI, SID	BN-5a
	Classifier	Subset Size Forward Selection	ACT, ATF, CAU, LIG, NOI	BN-5b
	Wrapper	Scatter V1	ACT, DAY, LIG, NOI, OI	BN-5c
6	Classifier	Scatter V1	ACT, ATF, GEN, LIG, MON, NOI	BN-6a
	Cost sensitive	Genetic	ACT, ATF, LIG, NOI, SHT, SID	BN-6b
	Filtered	Random	ACT, ATF, LAW, LIG, NOI, SID	BN-6c
	Wrapper	Greedy stepwise, Subset Size Forward Selection	ACT, AGE, GEN, LIG, MON, SHT	BN-6d
7	CFS	Subset Size Forward Selection	ACT, ATF, GEN, LAW, LIG, NOI, SID	BN-7a
	Filtered	Best first, Exhaustive, Greedy stepwise, Linear forward selection, Scatter V1, Subset Size Forward Selection, Tabu		
	Cost sensitive	Subset Size Forward Selection	ACT, ATF, LAW, LIG, NOI, SHT, SID	BN-7b
	Classifier	Greedy stepwise	ACT, AGE, ATF, GEN, LIG, NOI, OI	BN-7c
	Wrapper	Random	ACT, GEN, LIG, NOI, OI, SHT, VI	BN-7d
8	Classifier	Race	ACT, AGE, GEN, LAW, LIG, NOI, OI, ROM	BN-8a
	Filtered	Genetic	ACT, ATF, GEN, LAW, LIG, MON, NOI, SID	BN-8b
9	Wrapper	Best first	ACT, ATF, GEN, LAW, LIG, NOI, OI, PAS, PAW	BN-9a
		Exhaustive	ACT, AGE, ATF, GEN, LIG, MON, NOI, OI, VI	BN-9b
		Genetic	ACT, ATF, DAY, GEN, LIG, NOI, OI, PAS, PAW	BN-9c
11	Classifier	Exhaustive	ACT, CAU, DAY, GEN, LIG, MON, NOI, OI, PAS, TIM, VI	BN-11a
	Wrapper	Linear forward selection	ACT, AGE, ATF, DAY, GEN, LIG, MON, NOI, OI, SHT, VI	BN-11b
12	Classifier	Random	ACT, AGE, CAU, GEN, LIG, MON, NOI, OI, PAS, SID, TIM, VI	BN-12
14	Consistency	Best first, Exhaustive, Greedy stepwise, Linear forward selection, Rank, Scatter V1, Subset Size Forward Selection, Tabu	ACT, AGE, DAY, GEN, LAW, LIG, MON, NOI, OI, PAS, PAW, SHT, SID, TIM	BN-14
15	Consistency	Genetic	ACT, ATF, CAU, DAY, GEN, LAW, LIG, NOI, OI, PAS, PAW, ROM, SHT, TIM, VI	BN-15a
	Classifier	Best first	ACT, AGE, CAU, DAY, GEN, LAW, LIG, MON, NOI, OI, PAS, PAW, SHT, SID, TIM	BN-15b
16	Classifier	Genetic	ACT, ATF, CAU, DAY, GEN, LAW, LIG, MON, NOI, OI, PAS, PAW, ROM, SHT, TIM, VI	BN-16a
		Liner forward selection	ACT, AGE, ATF, CAU, DAY, GEN, LIG, MON, NOI, OI, PAS, PAW, ROM, SHT, SID, VI	BN-16b
	Consistency	Random	ACT, AGE, ATF, CAU, DAY, GEN, LAW, LIG, MON, NOI, OI, PAS, PAW, SHT, SID, TIM	BN-16c

NOI (number of injuries) was the third most selected variable (56 times). The forth most selected variable was ATF (atmospheric factors) with 42 times. On the other hand, the least selected variable was ROM (pavement markings), with only five times.

For each subset of selected variables, 20 BNs were built representing 10 runs for the training set, and 10 for the testing set, for each of the selected variable groups (26 groups) and for the eighteen original variables. In total, 540 BNs have been built for this analysis. The averages of the performance indicators are calculated for each one of these BNs.

Table 5. 8: Number of times each variable is selected

Variable	Number of times variable has been selected
ACT	58
LIG	58
NOI	56
ATF	42
GEN	34
LAW	26
SID	26
OI	25
MON	21
SHT	20
AGE	19
DAY	18
PAS	18
PAW	16
TIM	15
CAU	9
VI	9
ROM	5

Table 5. 9: Average values for accuracy, sensitivity, specificity, HMSS and ROC area for the 27 built BNs (training and test data).

Number of selected variables	BN	BN results									
		training	test	training	test	training	test	training	test	training	test
		Accuracy		Sensitivity		Specifity		HMSS		ROC area	
All the variables (18)	**BN-18**	**0,59**	**0,58**	**0,71**	**0,65**	**0,45**	**0,49**	**0,55**	**0,56**	**0,62**	**0,61**
4	BN-4	**0,60**	**0,59**	0,71	**0,68**	**0,46**	_0,48_	**0,56**	0,56	**0,63**	0,61
5	BN-5a	0,59	0,57	_0,70_	**0,68**	0,45	_0,45_	0,55	_0,54_	0,62	_0,60_
	BN-5b	**0,60**	**0,59**	0,71	**0,68**	**0,47**	0,49	**0,56**	0,56	**0,63**	0,61
	BN-5c	**0,61***	**0,61***	**0,75***	**0,73***	_0,44_	_0,46_	0,55	0,56	**0,63***	**0,62**
6	BN-6a	**0,60**	**0,60***	_0,70_	**0,67**	**0,49***	0,49	**0,58**	**0,57**	**0,64***	**0,62**
	BN-6b	0,59	0,57	0,71	**0,67**	0,45	_0,45_	0,55	_0,54_	0,62	_0,60_
	BN-6c	0,59	0,57	_0,70_	**0,67**	0,45	_0,46_	0,55	_0,54_	0,62	_0,60_
	BN-6d	**0,60**	0,57	**0,74**	**0,69***	_0,42_	_0,42*_	_0,54_	_0,52*_	_0,60*_	_0,58*_
7	BN-7a	**0,60**	0,57	_0,70_	**0,66**	**0,47**	_0,48_	**0,56**	_0,55_	0,62	_0,60_
	BN-7b	0,59	0,57	_0,70_	**0,66**	0,45	_0,46_	0,55	_0,54_	0,62	_0,60_
	BN-7c	**0,62***	**0,60***	**0,74**	**0,70***	**0,47**	_0,48_	**0,57**	**0,57**	**0,64***	**0,63**
	BN-7d	**0,61***	**0,59**	**0,76**	**0,72**	_0,42_	_0,44_	_0,54_	_0,54_	**0,64***	**0,62**
8	BN-8a	**0,62***	**0,60***	**0,75***	**0,71***	**0,46**	_0,47_	**0,57**	0,56	**0,64***	**0,63***
	BN-8b	0,59	0,58	_0,69_	**0,67**	**0,47**	_0,48_	**0,56**	0,56*	**0,63**	_0,60_
9	BN-9a	**0,61***	**0,60***	**0,74***	**0,70***	**0,46**	_0,47_	**0,56**	0,56	**0,65***	**0,63***
	BN-9b	**0,62***	**0,59**	**0,75***	**0,71***	**0,46**	_0,47_	**0,57**	0,56	**0,64***	**0,62**
	BN-9c	**0,61***	**0,59**	**0,73**	**0,69***	**0,46**	_0,48_	**0,56**	0,56	**0,65***	**0,63**
11	BN-11a	**0,60**	0,58	**0,74***	**0,69**	_0,43_	_0,46_	_0,54_	_0,55_	0,62	0,61
	BN-11b	**0,61***	**0,59**	**0,74**	**0,69***	**0,46**	_0,46_	**0,56**	0,56	**0,64***	**0,62**
12	BN-12	**0,60**	**0,59**	**0,74**	**0,68**	_0,43_	_0,48_	_0,54_	0,56	0,62	0,61
14	BN-14	0,59	0,57	0,71	**0,66**	0,45	_0,48_	0,55	_0,55_	0,62	0,61
15	BN-15a	**0,60**	0,58	**0,73**	**0,67**	_0,44_	_0,47_	0,55	_0,55_	0,62	0,61
	BN-15b	**0,60**	0,58	0,71	**0,66**	**0,46**	_0,48_	**0,56**	0,56	0,62	_0,60_
16	BN-16a	**0,60**	0,57	**0,72**	0,65	0,45	_0,47_	0,55	_0,55_	0,62	_0,60_
	BN-16b	0,59	0,57	0,71	**0,66**	0,45	_0,47_	0,55	_0,55_	0,62	0,61
	BN-16c	**0,60**	0,58	0,71	0,65	**0,47**	0,49	**0,65**	0,56	**0,63**	_0,60_

Table 5.9 shows the average results of these indicators for the 27 BNs for the training and the testing sets of data. With respect to the results obtained by the training set, the following findings could be highlighted:

- The values obtained for the all the performance indicators lie in the range between 0.42 and 0.76. These values however, are in the range of values obtained by other researchers (Delen et al., 2006; Abdelwahab and Abdel-Aty, 2001).
- The highest values obtained for the indicators of Sensitivity and ROC area are 0.69-0.76 for the former, and in the range of 0.60-0.65 for the latter.
- The worst results were obtained for Specificity (between 0.42 and 0.49). This indicator is used to measure the ability of the model to classify the KSI cases. Since the number of SI cases is higher than the number of KSI cases, and thus the BNs are a data mining technique, better results are obtained for larger groups (SI in this case) from those obtained for smaller groups (Chang and Wang, 2006).
- As Accuracy and HMSS are indicators that take into account both Sensitivity and Specificity, their results are intermediate, ranging between 0.59 and 0.62 for Accuracy and 0.54 and 0.65 for HMSS.

In most of the cases (74%), the values of the performance indicators obtained for the simplified BNs maintain or improve the results as compared to those of BN-18. Average values for each of the 27 BNs (training and testing sets) were tested for statistical significance ($p<0.05$) using least significant difference (LSD) ANOVA test.

Table 5.9 shows seven BNs (BN-6a, BN-7c, BN-8a, BN-9a, BN-9b, BN-9c, and BN-11b) that present statistically significant improvements in their performance indicators with respect to BN-18, with only a worsened value in one of their indicators. Having a look at these seven BNs (see Table 5.7), it is observed that there are 7 variables (ACT, AGE, ATF, GEN, LIG, NOI, and OI) that repeat in more than 50% of these 7 BNs.

None of the previously built BNs was built using this set of variables; therefore, a new BN (BN-7) was built using these seven variables. Table 5.10 shows the average values of the indicators for this new BN. The results of all the performance indicators of this

BN improve with respect to BN-18 except for one indicator (specificity for the test set), and these improvements are statistically significant (p<0.05) in 60% of the cases (accuracy, sensitivity and ROC area).

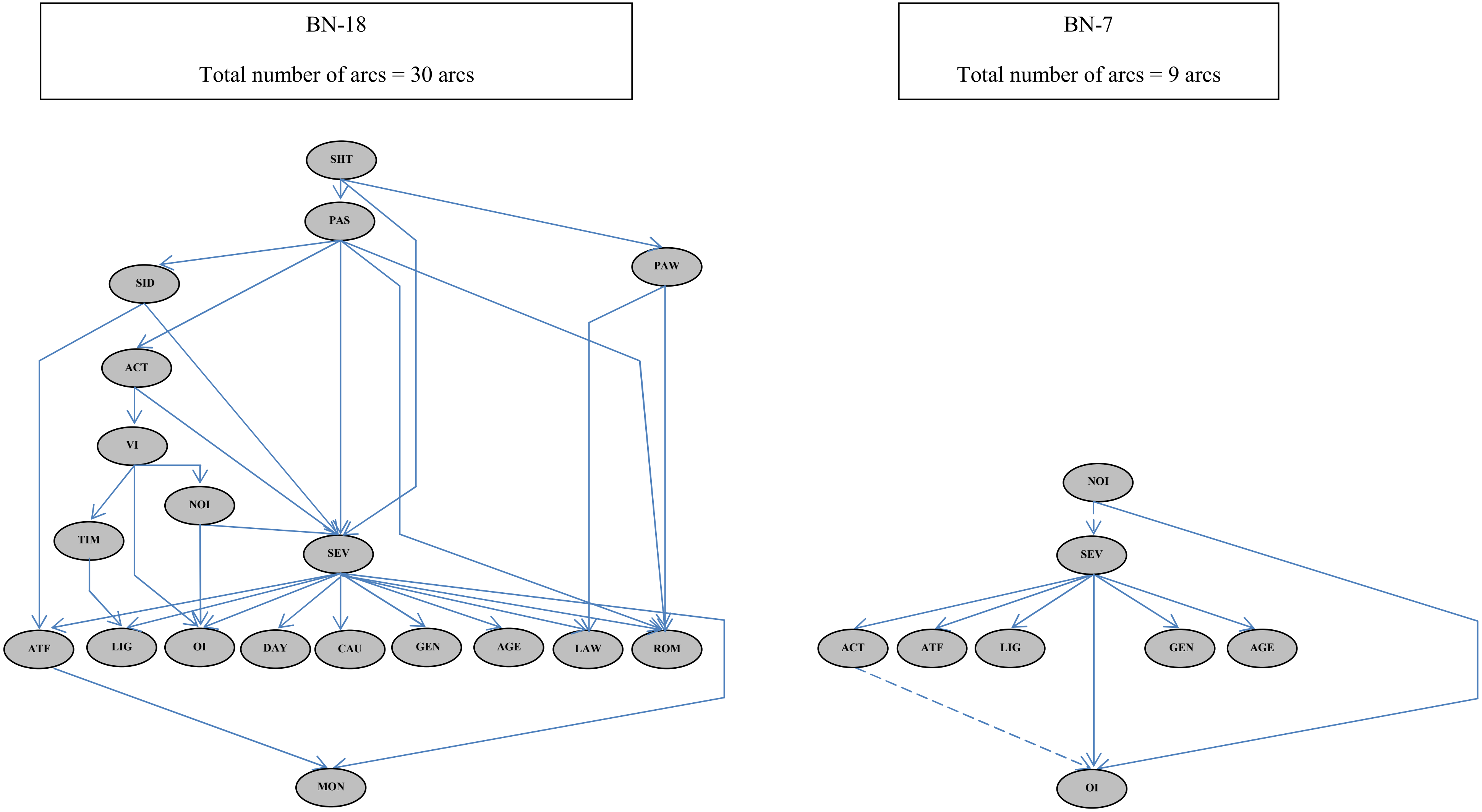
Table 5. 10: Average values for accuracy, sensitivity, specificity, HMSS and ROC area for BN-18 and BN-7 (training and test data).

Indicators	BN results			
	BN-18		BN-7	
	training	test	training	test
Accuracy	0,59	0,58	**0,61***	**0,60***
Sensitivity	0,71	0,65	**0,74***	**0,70***
Specificity	0,45	0,49	**0,46**	*0,48*
HMSS	0,55	0,56	**0,57**	**0,57**
ROC area	0,62	0,61	**0,65***	**0,63***

Thus, it could be said that a simplified BN has been identified (BN-7), with only seven variables, whose results are similar or even better than those obtained by the original BN (BN-18) which includes all variables obtained from the police accidents report.

Figure 5.5 shows the structure of both BN-18 and BN-7. It shows that the complexity of the built BN is relieved when built using the subset of seven variables. The indicator used to measure the complexity of the built network is the number of arcs. The number of arcs was found to be 30 in the original BN (BN-18), whereas it was decreased to 9 arcs for the network of BN-7. The arcs in the networks indicate the existence of a relationship between variables.

The structure of the BN-7 network is similar to the network structure built using the 18 variables (BN-18), keeping in mind that 11 variables have disappeared in BN-7. Thus, the relationships between the variables SEV-ACT, SEV-LIG, SEV-GEN, SEV-OI, SEV-AGE, SEV-ATF, NOI-OI are the same in both BN-18 and BN-7, except for the two new connections between SEV-NOI and ACT-OI that appeared in BN-7.



BN-18

Total number of arcs = 30 arcs

BN-7

Total number of arcs = 9 arcs

Figure 5. 5: The original BN (BN-18) and the simplified BN (BN-7)

The variables that appear in BN-7 (accident type, age, atmospheric factors, gender, lighting, number of injuries and occupants involved) could be considered the ones that significantly affect the injury severity in a traffic accident.

Kockelman and Kweon (2002) found accident type to be one of the significant variables that affect the injury severity of traffic accidents. They found that head on crashes were more dangerous than angle crashes, left-side, and right-side crashes; they also found that they were significant in accidents that involved killed or seriously injured.

Age was found to be a significant variable affecting the injury severity of traffic accidents by Tavris et al. (2001). They also found that male drivers in the age group (16–24) years were much more likely to be involved in killed or seriously injured accidents than those involving older drivers.

Xie et al. (2009) found that adverse weather could actually lead to lower probability of suffering the most severe category of injuries. They explained their results by the fact that under such conditions, drivers tend to drive at lower speeds and be more cautious. They also found gender to be a significant variable; their results indicated that the chance for male drivers to suffer the most severe category of injuries is less than female drivers under the same crash circumstances. Their results coincided with the results found by Kockelman and Kweon (2002).

Lighting has been found to be a significant variable defining injury severity in traffic accidents in several studies (Abdel-Aty, 2003; Helai et al., 2007 and Gray et al., 2008), where, they have found that more severe injuries are predicted during darkness.

Scheetz et al. (2003) found that the number of injured occupants was a significant factor in classifying injury severity.

Occupant involved in a traffic accident was found to be a significant variable by Dupont et al. (2010). They found that the higher the number of vehicles involved in the accident and the level of occupancy of these vehicles, the higher the probability for each car occupant to survive.

CHAPTER 6

Conclusions and future research

6.1. Conclusions

In this chapter the major conclusions for the injury severity Built Bayesian networks' models are given, as well as, those related to variables' selection and their associated Bayesian networks.

Conclusions for the Bayesian network model of injury severity, specified in section 2.2.1 and estimated in Section 5.2, are as follows. Traffic accident data was obtained from the DGT for a period of three years (2003-2005) for Granada (Spain). Nine BNs were built using three different search and score metrics: Hillclimbing, Simulated Annealing and Tabu with combination of BDeu, MDL and AIC.

Several indicators have been used in order to evaluate the performance of the built BNs: accuracy, sensitivity, specificity, HMSS, ROC Area, MPE and graph complexity (or number of arcs). The results obtained for these indicators do not vary significantly between the different search and score metrics used and they are within the range of previous studies (Abdel Wahab and Abdel-Aty, 2001; Simoncic, 2004). So, it is concluded that BNs might be a useful tool for classifying traffic accidents according to their injury severity.

Inference was used to identify the values of the variables that are associated with KSI in traffic accidents on Spanish rural highways. Based on the results, it would be possible to identify the factors that related with an accident being classified as KSI on Spanish rural highways. It would be a head-on or rollover traffic accident in a roadway without lighting with only one injury within the age of 18 and 25 years. These factors (head-on or rollover, unlit roadway, only one injury and within the age of 18 and 25 years) do not have to exist all at once in order to have a KSI accident. Any of these or a combination of them might increase the probability of a KSI accident. In general, these results are consistent with the literature (Tavris et al., 2001; Kockelman and Kweon, 2002; Abdel-Aty, 2003; Helai et al., 2007; Gray et al., 2008; Scheetz et al., 2009). However, this finding may vary for other countries and datasets.

Conclusions for variables' selection models, specified in Section 4.2.4 and estimated in Section 5.3, are as follows. The main objective of using variables' selection techniques was to determine if it is possible to maintain or improve the performance of a model that is used to predict the injury severity of a traffic accident based on BNs reducing the number of variables considered in the analysis. The performance of the model was measured using five indicators (accuracy, specificity, sensitivity, HMSS and ROC area).

To that end, 59 combinations of evaluator-search algorithms, which are commonly used in data mining, were used and 26 subsets of variables were identified. Within these subsets of variables the variable accident type (ACT), lighting (LIG) and number of injuries (NOI) were selected the most times (over 95%). Therefore, it could be said that these variables are the most significant ones in the classification of injury severity in traffic accidents, since they are included in almost all the selected subsets of variables.

Comparing the average values of the indicators for each one of the simplified BNs with respect to the average values obtained for the original BN (BN-18), it is observed that, in most of cases (74%), the performance indicators values for the simplified BNs maintained or improved in comparison with those of BN-18. Therefore, it could be said that, in most cases, simplified networks maintain the performance of the original BN.

Seven BNs were found to present statistically significant improvements in their performance indicators with respect to BN-18 and only one value of these indicators was worsened. In more than 50% of these BNs the following variables are repeated: ACT, AGE, ATF, GEN, LIG, NOI and OI.

These 7 variables were used to build a new BN (BN-7). The results of the performance indicators of this BN with respect to BN-18 improve practically in all the cases, and these improvements are statistically significant ($p<0.05$) in 60% of the cases (accuracy, sensitivity and ROC area).

To conclude, BNs, which have proved their effectiveness in different research areas, could be applied in the domain of injury severity of traffic accident modeling. Their effectiveness has been found to be similar to other data-mining techniques used to model severity in traffic accidents. Compared with other well-known statistical methods, the main advantage of the BNs seems to be their complex approach where system variables are interdependent and where no dependent and independent variables are needed (Simoncic, 2004).

In addition, this research work shows that, for the analysis of the severity of road accidents by Bayesian networks on rural roads, it is possible to reduce the number of variables considered in more than 60% (from 18 to 7 variables) maintaining the performance of the models and reducing their complexity. Thus the findings of this research work agrees with Chang and Wang (2006) where they stated that if a model is applied only on a few important variables, more useful results could be obtained. The procedure used to simplify BN models in order to analyze the severity of traffic accidents on rural highways could be also applied to other types of infrastructure (intersections, freeways, etc.) as well as to other models used to assess severity of traffic accidents (multinomial logit models, hierarchical logit models, probit models, etc.).

Finally, some limitations should be pointed out, such as the effect that the imbalanced dataset (slight injured versus killed or seriously injured) has on both sensitivity and specificity, and the need for large datasets.

6.2. Future research work

As illustrated within the elaboration of this research work that the analysis of traffic accidents' injury severity is not an easy task to be solved, and thus the research carried out and being carried on in this field has to be expanded in order to utilize techniques that has not been used yet, or is still in its early applications stages.

In the near future, two other studies are planned, in which the first one is related to analyzing traffic accidents injury severity on intersections using a clustering technique first in order to cluster accidents according to the type of the accident so that the characteristics of the database is homogenized, and then different Bayesian networks are applied to each resulting cluster. Different cluster are to be identified and to be used in order to build Bayesian networks, in which the built Bayesian networks are compared to an original Bayesian network that is built using the whole dataset (un-clustered) to find out if using a clustering technique first will affect the performance of the resulting model as measured using the performance evaluators indicators discussed in section 4.2.3.

Another study which is also planned to be performed in the near future is related to analyzing the same dataset used herein however growing decision trees instead. In which a new technique is to be applied which varies the root of the decision tree based on the variable used as the root. Three different splitting criteria are to be used in order to split the varied roots' trees into 3, 4, 5 and 6 levels in order to find out the effect of varying both the root and the split level of the tree on the extracted rules. Three different decision trees are to be used, ID3, IP-Tree and J-48 for this purpose, in which 216 different trees are to be grown to obtain different significant rules. These rules are supposed to represent at most 2% of the population used (1536) and an accuracy of at least of 80%, in order to represent the significant variables that are related to a killed or severely injured outcome in a traffic accident.

In addition, it is planned that the results presented by the extracted rules obtained from the decision trees will be compared to the significant variables obtained using the variables' selection algorithms, where different Bayesian networks will be built to measure the performance obtained using these variables.

One of the main findings obtained by building the Bayesian networks was that the imbalance classes affected the models' performance. This however, is one of the future research lines, in which emerging data mining techniques that deal with the problem of imbalance datasets will be used to deal with the problem encountered herein. These techniques will be used as preprocessing techniques on the datasets used prior to using any classifier or modeling techniques in order to decrease the imbalance ratio of the datasets used and hopefully to improve the performance obtained.

The research work carried out herein studied the injury severity when using accidents data for rural two lane highways only. In which future research work should focus on analyzing other types of roadways. In addition, future research line will be oriented towards using traffic accidents databases that belong to different countries, and compare the results obtained in Spain with these obtained elsewhere.

REFERENCES

LIST OF REFERENCES

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

[2] Abdel-Aty, M., and Abdelwahab, H. T. (2004). Predicting injury severity levels in traffic crashes: A modeling comparison. *Journal of Transportation Engineering*, 130 (2), 204-210.

[3] Abdel-Aty, M., and Keller, J. (2005). Exploring the overall and specific crash severity levels at signalized intersections. *Accident Analysis and Prevention*, 37(3), 417-425.

[4] Abdelwahab, H.T., and Abdel-Aty M.A. (2001). Development of Artificial Neural Network Models to Predict Driver Injury Severity in Traffic Accidents at Signalized Intersections. *Transportation Research Record*, 1746, 6-13.

[5] Acid, S., de Campos, L.M., Fernández-Luna, J.M., Rodríguez, S., J.M., R. and Salcedo, J.L. (2004). A comparison of Learning algorithms for Bayesian networks: a case study based on data from an emergency medical service, *Artificial Intelligence in Medicine*, 30, 215-232.

[6] Awadzi, K.D., Classen, S., Hall, A., Duncan, R.P., and Garvan, C.W. (2008). Predictors of injury among younger and older adults in fatal motor vehicle crashes. *Accident Analysis and Prevention*, 40 (6), 1804–1810.

[7] Bédard M., Guyatt, G.H., Stones, M.J., and Hirdes, J.P. (2002). The independent contribution of driver, crash, and vehicle characteristics to driver fatalities, *Accident Analysis and Prevention*, 34, 717–727.

[8] Brameier, M., and Banzhaf, W. (2007). *Linear Genetic Programming: Genetic and Evolutionary Computation Series*. Springer, New York.

[9] Bruin, J. (2006). newtest: command to compute new test. UCLA: Academic Technology Services, Statistical Consulting Group. Available at http://www.ats.ucla.edu/stat/stata/ado/analysis/ Accessed August March, 2011.

[10] Burham, K.P., Anderson, D.R. (2002). *Model selection and multimodel inference: A practical information –theoretic approach*. Springer, New York.

[11] Carpenter, G. A., Grossberg, S., Markuzon, N., Reynolds, J.H., and Rosen, D. B. (1992). Fuzzy ARTMAP: A neural-network architecture for incremental supervised learning of analog multidimensional maps. *IEEE Transactions on Neural Networks*, 3 (5), 698–713.

[12] Castillo, E. (2008). Traffic Estimation and Optimal Counting Location Without Path Enumeration Using Bayesian Networks. *Computer-Aided Civil and Infrastructure Engineering*, 23, 189–207.

[13] Castillo, E., Menédez, J.M., Sánchez-Cambronero, S. (2008). Predicting traffic flow using Bayesian networks. *Transportation Research Part B: Methodological*, 42 (5), 482-509.

[14] Chang, L.Y, and Wang, H.W. (2006). Analysis of traffic injury severity: An application of non-parametric classification tree techniques. *Accident Analysis and Prevention*, 38, 1019-1027.

[15] Chen, W.H, and Jovanis, P.P. (2000). Method for identifying factors contributing to driver-injury severity in traffic crashes. *Transportation Research Record*, 1717, 1-9.

[16] Cooper, G. F. and Herskovits, E. (1992). A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9, 309-347.

[17] Council, F., and Stewart, J. (1996). Severity Indexes for Roadside Objects. *Transportation Research Record*, 1528(1), 87-96.

[18] Crewson, P. (2006). *Applied statistics handbook*. [Online] (1.2) Available at http://www.acastat.com/Statbook/chisqassoc.htm [Accessed 4 April 2011].

[19] Cruz-Ramírez N, Acosta-Mesa H.G, Carrillo-Calvet H., Nava-Fernández L.A., and Barrientos-Martínez R.E. (2007) Diagnosis of breast cancer using Bayesian networks: A case study. *Computers in Biology and Medicine*, 37 (11), 1553-1564.

[20] Daniels, S., Brijs, T., Nuyts, E., and Wets, G. (2010). Externality of risk and crash severity at roundabouts. *Accident Analysis and Prevention*, 42 (6), 1966-1973.

[21] Das, A., Abdel-Aty, M. (2010). A genetic programming approach to explore the crash severity on multi-lane roads. *Accident Analysis and Prevention*, 42 (2), 548-557.

[22] Dash, M. and Liu, H. (1997). Feature selection for classification. *Intelligent Data Analysis*, 1, pp.131-56.

[23] Davis, G.A., Pei, J. (2003). Bayesian networks and traffic accident reconstruction. *Proceedings of the International Conference on Artificial Intelligence and Law*, 171-176.

[24] De Campos Ibáñez, L.M. (2011). *Modelos Probabilísticos para minería de datos* [lecture notes]. Retrieved from https://docto-si.ugr.es/master/intranet/src/Material Cursos.php?id_curso=M5 [Accessed March 2011].

[25] Delen, D., Sharda, R., and Bessonov, M. (2006). Identifying significant predictors of injury severity in traffic accidents using a series of artificial neural networks. *Accident Analysis and Prevention*, 38, 434–444.

[26] Dirección General de Tráfico – DGT. (2007). *Anuario Estadístico 2007*, NIPO: 128-08-161-7, Available at http://www.dgt.es/portal/es/seguridad vial/estadistica /accidentes 30dias/ [Accessed March 2011].

[27] Dissanayake, S. (2004). Comparison of severity affecting factors between young and older drivers in single vehicle crashes. *IATSS Research*, 28 (2), 48-54.

[28] Dissanayake, S., and Lu, J. (2002). Analysis of severity of young driver crashes: Sequential binary logistic regression modeling. *Transportation Research Record*, 1784 (1), 108-114.

[29] Donelson, A., Ramachandran, K., Zhao, K., and Kalinowski, A. (1999). Rates of occupant deaths in vehicle rollover: Importance of fatality-risk factors. *Transportation Research Record*, 1665(1), 109-117.

[30] Dupont, E., Martensen, H., Papadimitriou, E., and Yannis, G. (2010). Risk and Protection factors in fatal accidents. *Accident Analysis and Prevention*, 42, 645-653.

[31] Fisher, R.A. (1934). *Statistical methods for research workers*. (5th ed). Oliver and Boyd, Edinburgh.

[32] Freund, R.J., Wilson, W.J., and Sa, P. (2006). *Regression Analysis: Statistical modeling of a response variable*. (2^{nd} ed.), Elsevier, Burlington.

[33] García López, F., García Torres, M., Melián Batista, B., Moreno Pérez, J., and Moreno-Vega, J. (2006). Solving feature subset selection problem by a Parallel Scatter Search. *European Journal of Operational Research* , 477-489.

[34] Gårder, P. (2006). Segment characteristics and severity of head-on crashes on two-lane rural highways in Maine. *Accident Analysis and Prevention*, 38(4), 652-661.

[35] Glover, F., (1986). Future paths for integer programming and links to artificial intelligence. *Computer and Operations Research*, 13, pp.533-49.

[36] Goldberg, D. (1989). *Genetic algorithms in search, optimization and machine learning*. Addison-Wesley.

[37] Gray, R.C., Quddus, M.A. and Evans, A. (2008). Injury severity analysis of accidents involv- ing young male drivers in Great Britain. *Journal of Safety Research*, 39, 483–495.

[38] Gray, R.C., Quddus, M.A., and Evans, A. (2008). Injury severity analysis of accidents involving young male drivers in Great Britain. *Journal of Safety Research*, 39(5), 483-495.

[39] Gregoriades A. (2007). Road safety assessment using Bayesian belief networks and agent-based simulation. *Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics* , art. no. 4413954, 615-620.

[40] Gregoriades, A., Sutcliffe, A., Papageorgiou, G., Louvieris, P. (2010). Human-centered safety analysis of prospective road designs. *IEEE Transactions on Systems, Man, and Cybernetics Part A: Systems and Humans*, 40 (2), 236-250.

[41] Guo, Z., Chen, F., Wang, X. (2010). A method coupling fuzzy algorithm and Bayesian network for calculation of operation risk of major highway infrastructure. *Proceedings - 2010 7th International Conference on Fuzzy Systems and Knowledge Discovery, FSKD*, art. no. 5569176, pp. 1095-1100

[42] Gurney, K. (1997). *An introduction to neural networks*. UCL Press Limited, London.

[43] Gutlein, M., Frank, E., Hall, M., and Karwath, A. (2009). Large scale attribute selection using wrappers. In: *Proceedings of 2009 IEEE Symposium on Computational Intelligence and Data Mining, CIDM 2009*, Washington.

[44] Haleem, K., Abdel-Aty, M. (2010). Examining traffic crash injury severity at unsignalized intersections. *Journal of Safety Research*, 41(4), 347-357.

[45] Hall, M. (1998). *Correlaion-based Feature Subset Selection for Machine Learning*. Doctoral Dissertation, The University of Waikato, Hamilton, New Zeland.

[46] Hedar, A.-R., Wang, J., Fukushima, M. (2008). Tabu search for attribute reduction in rough set theory. *Soft Computing*, 12 (9), 909-918.

[47] Helai, H., Chor, C.H., and Haque, M.M. (2008). Severity of driver injury and vehicle dam- age in traffic crashes at intersections: a Bayesian hierarchical analysis. *Accident Analysis and Prevention*, 40, 45–54.

[48] Howard, R.A. (1988). Decision analysis. Practice and promise. *Management Science* , 34(6), 679-95.

[49] Janssens, D., Wets, G., Brijs, T., Vanhoof, K., Arentze, T., Timmermans, H. (2006). Integrating Bayesian networks and decision trees in a sequential rule-based transportation model. *European Journal of Operational Research*, 175 (1), 16-34.

[50] Jin, Y., Wang, X., and Chen, X. (2010). Right-angle crash injury severity analysis using ordered probability models. In: *2010 International Conference on Intelligent Computation Technology and Automation ICICTA 2010.*

[51] Jin, S., Wang, D.-H., Qi, H.-S. (2010). Bayesian network method of speed estimation from single-loop outputs. *Journal of Transportation Systems Engineering and Information Technology*, 10 (1), 54-58

[52] John, G.H., Kohavi, R. and Pfleger, K. (1994). Irrelevant features and subset selection problem. In: *Proceedings of the Eleventh International Conference on Machine Learning.*

[53] Jung, S., Qin, X., and Noyce, D.A. (2010). Rainfall effect on single-vehicle crash severities using polychotomous response models. *Accident Analysis and Prevention*, 42(1), 213-224.

[54] Keele, L., and Park, D.K. (2006). Difficult choices: An evaluation of heterogenous choice models. In: *2004 Meeting of the American Political Science Association*, Chicago, IL. Available at http://www.allacademic.com/meta/p59491_index.html [Accessed December 2010].

[55] Khattak, A., and Rocha, M. (2003). Are SUVs "Supremely unsafe vehicles"? Analysis of rollovers and injuries with sport utility vehicles. *Transportation Research Record*, 1840(1), 167-177.

[56] Khattak, A.J. (2001). Injury severity in multivehicle rear-end crashes. *Transportation Research Record*, 1746 (1), 59-68.

[57] Khattak, A.J., Pawlovich, M.D., Souleyrette, R.R., and Hallmark, S.L. (2002). Factors related to more severe older driver traffic crash injuries. *Journal of Transportation Engineering*, 128(3), 243-249.

[58] Khhavi, R., and John, G. (1997). Wrappers for feature subset selection. *Artificial Intelligence*, 97, 273-324.

[59] Kleinbaum, D.G., and Klein, M. (2002). *Logistic regression: A self-learning text*. Springer, New York.

[60] Kockelman, K.M., and Kweon, Y.J. (2002). Driver injury severity: an application of ordered probit models. *Accident Analysis and Prevention*, 34, 313–321.

[61] Kolmogorov, A.N., Foundations of the Theory of Probabilty, Chelsea, New York, 1950 (originally published in 1933 as Grundbegriffe der Wahrscheinlichkeitsrechnung, Spring, Berlin).

[62] Kononen, D.W., Flannagan, C.A.C., and Wang, S.C. (2011). Identification and validation of a logistic regression model for predicting serious injuries associated with motor vehicle crashes. *Accident Analysis and Prevention*, 43(1), 112-122.

[63] Kopelias, P., Papadimitriou, F, Papandreou, K, and Prevedouros, P. (2007). Urban Freeway Crash Analysis. *Transportation Research Record*, 2015, 123-131.

[64] Krull, K., Khattak, A.J., and Council, F.M. (2000). Injury effects of rollovers and events sequence in single-vehicle crashes. *Transportation Research Record*, 1717 (1), 46-54.

[65] Kruskal, W.H. (1958). Ordinal measures of association. *Journal of the American Statistical Association*, 53 (284), 814–861.

[66] Lemp, J.D., Kockelman, K.M., and Unnikrishnan, A. (2011). Analysis of large truck crash severity using heteroskedastic ordered probit models. *Accident Analysis and Prevention*, 43(1), 370-380.

[67] Lenguerrand, E., Martin, J. L., and Laumon, B. (2006). Modelling the hierarchical structure of road crash data--application to severity analysis. *Accident Analysis and Prevention*, 38(1), 43-53.

[68] Liu, H., and Setiono, R. (1996). A probabilistic approach to feature selection-A filter solution. *Proceedings of 13th International Conference on Machine Learning*.

[69] Lucas, P. (1999). Bayesian Networks in Medicine: a Model-based Approach to Medical Decision Making, Available at http://citeseer.ist.psu.edu/467626.html.

[70] Malyshkina, N.V. (2008). Markov switching models: An application to roadway safety (Doctoral dissertation). Available at http://proquest.umi.com/pqdlink?did=1888978671&Fmt=7&clientId=54638&RQT=309&VName=PQD.

[71] Malyshkina, N.V., and Mannering, F.L. (2009). Markov switching multinomial logit model: An application to accident-injury severities. *Accident Analysis and Prevention*, 41(4), 829-838.

[72] McCrea, J. and Moutari, S. (2010). A hybrid macroscopic-based model for traffic flow in road networks. *European Journal of Operational Research*, 207, 676–684.

[73] McFadden, D. (1979). *Quantitaive methods for analyzing travel behaviour of individuals: some recent developments*. In: Hensher, D.A., Stopher, P.R.: (eds): Behavioural travel modelling. Croom Helm, London.

[74] Mehta, C.R. and Patel, N.R. (1996). *SPSS exact tests 7.0 for windows*. Chicago, IL: SPSS Inc.

[75] Milton, J.C., Shankar, V.N., and Mannering, F.L. (2008). Highway accident severities and the mixed logit model: An exploratory empirical analysis. *Accident Analysis and Prevention*, 40, 260–266.

[76] Mittal, A., Kassim, A., and Tan, T. (2007). *Bayesian network technologies: Applications and graphical models*. IGI Publishing.

[77] Montgomery, D.C., and Runger, G.C. (2003). *Applied statistics and probability for engineers*. John Whiley and Sons, Inc, New York.

[78] Moore, A., and Lee, M. (1994). Efficient algorithms for minimizing cross validation error. *Proceedings of 11th International Conference on Machine Lerning*.

[79] Moore, D.N., Schneider IV, W.H., Savolainenb, P.T., Farzaneh, M. (2010). Mixed logit analysis of bicyclist injury severity resulting from motor vehicle crashes at intersection and non-intersection locations. *Accident Analysis and Prevention*, Article in press: doi:10.1016/j.aap.2010.09.01.

[80] Morgan, A. (2009). What factors affect crash injury severity under specific weather conditions? Available at https://engineering.purdue.edu/ITE/research/seminarfiles09-10/studentpresentation_abbymorgan.pdf Accessed January, 2011.

[81] Neapolitan, R.E. (2003). *Learning Bayesian Networks*, Prentice Hall.

[82] Ni, D., Leonard II, J.D. (2005). Markov chain Monte Carlo multiple imputation using Bayesian networks for incomplete intelligent transportation systems data. *Transportation Research Record*,1935, 57-67.

[83] NIST/SEMATECH. (2003). e-Handbook of statistical methods, Available at http://www.itl.nist.gov/div898/handbook/ [Accessed March 2011].

[84] Oh, J.T. (2006). Development of severity models for vehicle accident injuries for signalized intersection in rural areas. *KSCE Journal of Civil Engineering*, 10 (3), 219-225.

[85] Ortúzar, J.D.D., and Willumsen, L.G. (2001). *Modelling transport*. John Wiley and Sons Ltd, West Sussex.

[86] Ouyang, Y., Shankar, V., and Yamamoto, T. (2002). Modeling the Simultaneity in Injury Causation in Multivehicle Collisions. *Transportation Research Record*, 1784 (1), 143-152.

[87] Ozbay K, and Noyan N., (2006). Estimation of incident clearance times using Bayesian Networks approach. *Accident Analysis and Prevention*, 38, 542–555.

[88] Paleti, R., Eluru, N., and Bhat, C.R. (2010). Examining the influence of aggressive driving behavior on driver injury severity in traffic crashes. *Accident Analysis and Prevention*, 42 (6), 1839–1854.

[89] Pearl J. (2000). *Causality: Models, Reasoning and Inference*. Cambridge: Cambridge University Press.

[90] Pearl, J. (1984). *Heuristics: Intelligent search strategies for computer problem solving*. Addison-Wesley.

[91] Peek-Asa, C., Britton, C., Young, T., Pawlovich, M., and Falb, S. (2010). Teenage driver crash incidence and factors influencing crash injury by rurality. *Journal of Safety Researc*h, 41 (6), 487–492.

[92] Quddus, M.A., Wang, C., and Ison, S.G. (2010). Road traffic congestion and crash severity: Econometric analysis using ordered response models. *Journal of Transportation Engineering*, 136 (5), 424-435.

[93] Renski, H., Khattak, A.J., and Council, F.M. (1999). Effect of speed limit increase on crash injury severity: analysis of single-vehicle crashes on north Carolina Interstate highways. *Transportation Research Record*, 1665 (1), 100-108.

[94] Rockach, L., and Maimon, O. (2008). *Data mining with decision trees theory and applications*. World Scientific Publishing Co. Pte. Ltd., Singapore.

[95] Ruggeri, F., Kenett, R., Faltan, F.W. (2007). Encyclopedia of Statistics in Quality and Reliability. John Wiley and Sons Ltd, West Sussex.

[96] Rumelhart, D. E.,Hinton, G. E., and Williams, R. J. (1986). *Learning internal representation by error propagation, parallel distributed processing: explorations in the microstructure of cognition*. MIT Press, Cambridge.

[97] Russell, S., and Norvig, P. (2003). *Artificial Intelligence: A modern approach* . Upper Saddle River, New Jersey: Prentice Hall.

[98] Saccomanno, F.F., Nassar, S.A., and Shortreed, J.H. (1996). Reliability of statistical road accident Injury severity models. *Transportation Research Record*, 1542(1), 14-23.

[99] Sánchez-Cambronero, S., Rivas, A., Gallego, I., Menéndez, J.M. (2010). Predicting traffic flow in road networks using Bayesian networks with data from an optimal plate scanning device location. *Proceedings 1: ICAART 2010 - 2nd International Conference on Agents and Artificial Intelligence,* 1, 552-559.

[100] Savolainen, P.T., Mannering, F.L., Lord, Dominique, L., Quddus, M. (2011). The statistical analysis of highway crash-injury severities. A review and assessment of methodological alternatives. *Accident Analysis and Prevention*, 43 (5), 1666-1676.

[101] Scheetz, L.J., Zhang, J., and Kolassa, J. (2003). Classification tree to identify severe and moderate injuries in young and middle aged adults. *Artificial Intelligence in Medicine*, 45, 1–10.

[102] Schneider, W.H., Savolainen, P.T., and Zimmerman, K. (2009). Driver injury severity resulting from single-vehicle crashes along horizontal curves on rural two-lane highways. *Transportation Research Record*, 2012 (1), 85-92.

[103] Shanker, V., Mannering, F., and Barfield, W. (1996). Statistical analysis of accidents severity on rural freeways. *Accident Analysis and Prevention*, 28 (3), 391-401.

[104] Simoncic, M. (2004). A Bayesian network model of two-car accidents. *Journal of transportation and Statistics*, 7, 13-25.

[105] Smaili, C., El Najjar, M.E., Charpillet, F. (2007). Multi-sensor fusion method using dynamic bayesian network for precise vehicle localization and road matching. *Proceedings: International Conference on Tools with Artificial Intelligence, ICTAI 1*, art. no. 4410276, 146-151.

[106] Smaili, C., El Najjar, M.E., Charpillet, F. (2008). A road matching method for precise vehicle localization using hybrid Bayesian network. *Journal of Intelligent Transportation Systems: Technology, Planning, and Operations*, 12 (4), 176-188.

[107] Srinivasan, K.K. (2002). Injury severity analysis with variable and correlated thresholds: ordered mixed logit formulation. *Transportation Research Record*, 1784 (1), 132-142.

[108] Sun, S., Zhang, C. and Zhang, Y. (2005). Traffic Flow Forecasting Using a Spatio-temporal Bayesian Network Predictor. *Artificial Neural Networks: Formal models and their applications – ICANN 2005 Lecture Notes in Computer Science*, 3697/2005,752, DOI: 10.1007/11550907_43.

[109] Sun, S, Zhang, C., Yu, G. (2006). A bayesian network approach to traffic flow forecasting. *IEEE Transactions on Intelligent Transportation Systems*, 7 (1), 124 – 132.

[110] Tavris, D.R., Kuhn, E.M., and Layde, P.M. (2001). Age and gender patterns in motor vehicle crash injuries: importance of type of crash and occupant role. *Accident Analysis and Prevention*, 33, 167–172.

[111] Tao, L., Zhao, C., Thulasiraman, K. and Swamy, M.N.S. (1992). Simulated annealing and Tabu search algorithms for multiway graph partition. *Journal of Circuits, Systems, and Computers*, 2(2), 159-85.

[112] The World Bank. (2003). Traffic fatalities and economic growth. Washington, DC. World Bank. (Policy Research Working Paper No. 3035). Retrieved from http://econweb.umd.edu/~cropper/publications/wp3.pdf

[113] Train, K. (2009). *Discrete choice methods with simulation*. Cambridge University Press, Cambridge.

[114] Wang, X., and Kockelman, K.M. (2005). Use of Heteroscedastic Ordered Logit model to study severity of occupant Injury: Distinguishing effect of vehicle weight and type. *Transportation Research Record*, 1908 (1), 195-204.

[115] Wang, Z., Chen, H., and Lu, J.J. (2009). Exploring impacts of factors contributing to injury severity at freeway diverge areas. *Transportation Research Record*, 2102 (1), 43-52.

[116] Washington, S., Karlaftis, M. & Mannering, F. (2003). *Statistical and econometric methods for transportation data analysis*. (2nd ed.). Boca Raton: Florida, Champan and Hall/CRC.

[117] Witten, I. H., and Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques* (2nd Edition). San Francisco, California: Morgan Kaufmann Publishers.

[118] World Health Organization (WHO). (2004). World health organization, world report on road traffic injury prevention Geneva. Available at http://whqlibdoc.who.int/publications/2004/9241562609.pdf [Accessed February 2011].

[119] Wuensch, K.L. (2011). *East Carolina University: Common univariate and bivariate applications of the chi-square distribution*. [Online] Available at: http://core.ecu.edu/psyc/wuenschk/docs□□/Chi- square.doc [Accessed 4 April 2011].

[120] Xie, Y., Zhang, Y., and Liang, F. (2009). Crash Injury Severity Analysis Using Bayesian Ordered Probit Models. *Journal of Transportation Engineering ASCE*, 135 (1), 18-25.

[121] Xu, H., Zong, F., Zhang, H. (2010). Bayesian network-based road traffic accident causality analysis. *Proceedings - 2010 WASE International Conference on Information Engineering, ICIE 2010 3*, art. no. 5571612, 413-417.

[122] Yates, D., Moore, D. and McCabe, G. (1998). *The practice of statistics: TI-83 graphing calculator enhanced*. New York: W.H. Freedman and Company.

[123] Zadeh, L. (1994). Preface, in (Marks II R.J.), *Fuzzy logic technology and applications*, IEEE Publications.

[124] Zhang, K., Taylor, M.A.P. (2006). Effective arterial road incident detection: A Bayesian network based algorithm. *Transportation Research Part C: Emerging Technologies*, 14 (6), 403-417.

[125] Zhu, X., and Srinivasan, S. (2011). A comprehensive analysis of factors influencing the injury severity of large-truck crashes. *Accident Analysis and Prevention*, 43 (1), 49–57.

APPENDICES

APPENDIX I

i. Studies using Logit models

Twenty five studies of accident injury severity were found to use one or more of the previously mentioned LM described in 2.3.1.1. As mentioned previously the most used types of LMs were BLM, OLM, and OPM.

The first study found to analyze injury severity used a HL; this study was done by Shanker et al. (1996). In their study they used HL to model the severity of traffic accidents on rural highways, their results indicated that when the HL accounted for the shared un-observables between property damage and possible injury crashes the best structural fit for the observed distribution of crash severities was provided. Also, they found that the estimation results provided evidence on the effect of environmental conditions, highway design, accident type, driver characteristics, and vehicle attributes on accident severity.

In 2010 Haleem and Abdel-Aty (2010) and Daniels et al. (2010) used HL; Haleem and Abdel-Aty (2010) analyzed injury severity for accidents of single vehicles occurring on three and four-legged unsignalized intersections. However, in their analysis they only presented the results obtained for the three legged intersection. They found that the natural logarithm of the upstream distance, speed limit, at fault driver's age, and size of intersection to be negatively related to injury severity, where, the width of the left shoulder near the median on the major road was found to be positively related to injury severity.

Daniels et al. (2010) used HL to model injury severity of accidents occurring on roundabouts. Their results indicated that injury severity increases with higher age, accidents at night and accidents outside built-up areas are more severe and that single-vehicle accidents seem to have more severe outcomes than multiple-vehicle accidents.

BLM were the second models found to be used to model injury severity, Donelson et al. (1999) used BLM to predict fatality for occupants of light duty trucks in single vehicle rollover accidents. Their results indicated that weekend (10 p.m. Friday to 4 a.m. Monday), divided highway, rural area, male driver, curve in road, and nighttime hours (10 p.m. – 4 a.m.) were associated with fatality.

Krull et al. (2000) used BLM to explore the effect of rollover for single vehicles accidents. Their results indicated that driver injury severity increases with rollovers, failure to use a seat belt, passenger cars (as opposed to pickup trucks), alcohol use, daylight, rural roads (as opposed to urban), posted speed limit, and dry pavement (as opposed to slick pavement). Restricting the injury models to rollover crashes only, it was determined that hitting point objects or longitudinal objects before rolling over resulted in a more severe injury than rolling over first.

Dissanayake and Lu (2002) used BLM to analyze injury severity of single vehicles' accidents with fixed objects for young drivers. Their results indicated that alcohol or drugs, ejection in the crash, point of impact, rural crash locations, existence of curve or grade at the crash location, and speed of the vehicle significantly increased the probability of having a more severe crash. Restraint device usage and being a male clearly reduced the tendency of high severity, and some other variables, such as weather condition, residence location, and physical condition, were not important at all.

Fatalities in single vehicles' accidents were modeled by Bédard et al. (2002), they used BLM as well to do the analysis. They found that older drivers, female and older female drivers, Alcohol, not using the seat belt and speed larger than 112 km/ hour are related to fatal accidents where larger trucks were found to be protective against fatal accidents.

Ouyang et al. (2002) studied the outcomes of injury for accidents that occur between cars and trucks. Using two forms of the BLM, a simultaneous and a joint form, they found that parameter estimates for key roadway design variables may vary greatly between a simultaneous BLM and its single-equation counterparts. They also found those accidents' types of head on, rear end, high speed opposite direction for cars; increase the injury severity for occupants, exceeding 72 km/hr on intersections, and that injury severity of occupants in cars increases with older trucks, slippery pavement, and heavy trucks. Also, increased injury severity is associated with older drivers, drivers under influence of alcohol, aggressive actions of drivers, not using the seat belt, ejection of car occupants, speed limits exceeding 88 km/hr, occupants in cars and trucks in straight segments with a speed limit of 72 km/hr. Factors that were found to decrease the injury severity were rainy weather, snowing and fog.

Injury severity of young and older drivers involved in single vehicles' traffic accidents were studied by Dissanayake (2004). According to his findings, injury severity increases for younger drivers when they are at fault, driving on curves or grades, being involved in a side impact accident, being under the influence of alcohol, being ejected, and at daylight. Where older drivers were found to be more severely injured at higher speeds, in rural areas, and if they were male drivers, thus, the injury severity decreases if they were seat belted.

Lenuguerrand et al. (2006) compared BLM, HL, and GEE models results which were used to analyze the fatalities if car occupants' involved in traffic accidents. Their results indicated that HL models provide the lowest biases followed by GEE and finally BLM when using correlated data. Their findings indicated that main and secondary highways, single vehicles accidents at or outside intersections, and nighttime or twilight increase the risk of fatality in traffic accidents.

Most recently Jung et al. (2010), Dupont et al. (2010), Peek-Asa et al. (2010), and Kononen et al. (2011) used BLM to model the resulting injury severity in traffic accidents.

Evaluating the effect of rainfall on the injury severity of single vehicles accidents was done by Jung et al. (2010). They used a sequential BLM in which property and damage only (PDO) accidents were removed at the second stage in the forward format while fatal and injury crashes were removed in the backward format. Moreover the severity of accidents occurring on dry pavement in cloudy or clear weather was estimated to compare with its counterpart in rainy condition. The results indicated that the backward format of sequential BLM model is recommended as the final model for predicting all levels of single vehicle accidents that occurred on high-speed highways in rainy weather. The findings of this research work indicated that 15-min rainfall intensity, wind speed, horizontal/vertical curve, female driver, and safety belt usage are significantly associated with injury severity of traffic accidents.

Dupont et al. (2010) used BLM to model the survivability of occupants in fatal traffic accidents and the factors reacted to the survivability. The findings indicated that the car occupants' survival chances are negatively associated with their own age and that of their vehicle. The survival chances are also lower when seatbelt is not used. Front damage, as compared to other damaged car areas, appears to be associated with increased survival probability, but mostly in the case in which the accident opponent was another car.

Peek-Asa et al. (2010) studied the factors that are associated with the increased odds of fatal or severe injury for traffic accidents occurring in urban and rural. For younger teen drivers (age 10 through 15), overall crash rates were higher for more rural areas, although for older teen drivers (age 16 through 18) the overall crash rates were lower for rural areas. Rural teen crashes were nearly five times more likely to lead to a fatal or severe injury crash than urban teen crashes. Rural crashes were more likely to involve single vehicles, be late at night, and involve a failure to yield the right-of-way and crossing the center divider.

The latest study found to use BLM to model injury severity was done by Kononen et al. (2011). They used BLM to predict the severity of occupants involved in traffic accidents. Results indicated that Delta-V (mph), unbelted, direction of impact is left, and occupants ≥55 years are more associated with higher injury severity.

OLM was used by Abdelwahab and Abdel-Aty (2001), Srinivasan (2002), Khattak and Rocha (2003), Jin et al. (2010), Jung et al., 2010, and Quddus et al. (2010) to analyze the injury severity of traffic accidents.

Where the significant variables found by the OLM models were not reported by both Abdelwahab and Abdel-Aty (2001) and Srinivasan (2002). Khattak and Rocha (2003) studied the influence of various vehicles' platforms on rollover of single vehicles' accidents and analyzed the injury severity of drivers. Their results indicated that sport utility vehicles, rollover, driving off the road, losing control of the vehicle, speed, young drivers, female drivers, and ejection are all related with increased injury severity, where being belted decreases the injury severity.

Right angle accidents on signalized intersection were modeled using OLM models by Jung et al. (2010). Their findings indicated that person ejected or not, alcohol and/or drug use, the driver's age, point of impact, and standardized yellow time for entering movement were significant variables affecting the average severity of accidents.

Quddus et al. (2010) used OLM to explore the relationship between the severity of road crashes and the level of traffic congestion using disaggregated crash records and a measure of traffic congestion while controlling for other contributory factors. Results indicated that increased traffic flow, radius of road curvature, darkness, wet pavement, and time trend are all correlated with decreased injury severity; where three-lane stretches of the highway, single-vehicle crash, and weekdays were correlated with increased injury severity.

Srinivasan (2002) compared the performance obtained by OLM and that obtained by OMXL. The Chi-square tests results indicated that the more general OMXL formulation provides a statistically superior representation of observed injury severity data than corresponding OLM. Results indicated that occupants of a moped are at a particularly high risk of injury severity given an accident. Where, The results indicate that heavy-duty trucks are safer for all injury severity levels than passenger cars. Also, the model implies that elderly passengers (over 65) have an increased chance of sustaining moderately severe injuries than other drivers. Moreover, no significant differences in injury severity between younger and middle-aged drivers are observed. The influence of gender is also only seen for mild injury levels, drivers with fatigue were roughly five times as likely to experience severe injuries, and face a roughly 30% lower chance of experiencing a property damage only event than the baseline victims. The results also indicate that drunk drivers face a 3.5 times larger risk of fatal injuries than other drivers, while the chance of moderate injuries increases by 16%. Drivers involved in accidents on weekends, particularly during late night on Saturday, are more prone to higher injury severity levels compared to drivers involved in accidents on weekdays.

HKL was used by Wang and Kockelman (2005) to study the effects of various vehicle, environmental, roadway, and occupant characteristics on the severity of injuries sustained by vehicle occupants in one and two vehicle accidents. their results indicated that older occupants, female occupants, older occupants navigating curved roadway sections with higher speed limits, female occupants navigating curved roadway sections with higher speed limits are all associated with higher injury severity.

Paleti et al. (2010) also used HKL to capture the moderating effect of aggressive driving behavior while assessing the influence of a comprehensive set of variables on injury severity. They found that aggressive driving, young drivers pursuing aggressive driving, young drivers, female drivers, accident types (rollover, head-on, fixed object), not being seat belted, pick-ups' drivers, and being under influence of alcohol are all associated with higher injury severity levels. While, having two or more occupants where at least one of them is above the age of 20 years, traffic congestion, adverse weather, sport utility vehicles' drivers, low speed limits <50km/hr, rear-end accidents, sideswipe/angle accidents are all associated with decreased injury severity levels.

Awadzi et al. (2008), Malyshkina and Mannering (2009), and Schneider et al. (2009) used MNL to model injury severity. Awadzi et al. (2008) modeled the injury severity of younger and older drivers in traffic accidents. They found that angle crashes of 1-3o'clock, and 7-9 o'clock for older drivers, occupants with older drivers, daylight for older drivers, female older drivers, through intersections, under influence, unbelted, adverse weather, passenger cars, and fixed objects impacts are all associated with higher injury severity in traffic accidents.

Malyshkina and Mannering (2009) analyzed injury severity for accidents of two and single vehicles. Their findings indicated that summer, construction at the accident location, and precipitation all are associated with higher injury severity in traffic accidents.

Schneider et al. (2009) assessed driver injury severity resulting from single-vehicle crashes on rural two-lane highways. They indicated that Run-off-road, older drivers, older drivers on curves of smaller radius, unbelted drivers, uninsured drivers, under influence of drugs or alcohol, and fatigued drivers, were more likely to be seriously injured.

Milton et. al. (2008) used MXL to study the variation that the influence of variables has on injury severity of accidents on roadway segments. Their results indicated that pavement friction, number of interchange/mile are all associated with higher injury severity, where % trucks, number of horizontal curves/mile, number of grade/mile, and snowfall are associated with decreased injury severity.

Finally, Quddus et al. (2010) compared the performance obtained by OLM with that of a HCM. The results obtained indicated that both model types were suitable for both ordinal dependent variables and disaggregate accidents' data are used. The results obtained regarding significant variables found were mentioned previously.

ii. Studies using probit models

Fourteen studies of accident injury severity were found to use one or more of the previously mentioned PM described in 2.3.1.2. The most used type of PMs was OPM, where it was used in all the studies that applied the PM.

In 1999 Renksi et al. (1999) used OPM to analyze effect of speed limit on occupant injury in single vehicles accidents; excluding pedestrian, bicyclists, or motorcycles' accidents. Their results indicated that speed limit increased by 16.1 km/hr, more vehicle occupants, Run-off-road; drivers under the influence of alcohol, and hitting fixed objects all are associated with increased injury severity of traffic accidents.

Khattak (2001) used OPM to analyze the effect of information and vehicle technology on injury severity in rear end crashes in two and three vehicles crashes. The results found in his research work indicated that newer vehicles, vehicle types (vans, pickups, and station wagons), center high mounted stop light (CHMSL) all decrease the injury severity of traffic accidents.

Kockelman and Kweon (2002) analyzed injury severity based on the number of vehicles involved in the traffic accident, where the models were build for all vehicles, two vehicles, and for single vehicles. The results indicated that manner of collision, number of vehicles involved, driver gender, vehicle type, and driver alcohol use play major roles. Rollover and head-on collisions are particularly serious, contributing to more severe injury levels than speed increases of 50 mph and more. And males tend to fare significantly better than females. In contrast, the effects of late-night driving on weekends and daylight conditions had rather negligible effects on injuries sustained by drivers, after controlling for the other variables.

Khattak et al. (2002) used OPM to isolate factors that contribute to injuries to older drivers involved in accidents. The findings indicated that being under influence of alcohol, farm vehicles, rural area, darkness, curves in level terrain, increased age, and driving with no occupants, are all associated with higher injury, while animal related accidents are associated with less injury.

Abdel-Aty and Abdelwahab (2004) used OPM to compare its performance with that obtained by using ANN. Where the results indicated that the peak period and weather variables were found to be insignificant. Driving under the influence factor was not significant, but the interaction between it and the seat belt factor was found to be significant. The marginal effect of seat belt use has the highest effect on the probability of not being injured. Female drivers tend to be more likely to suffer severe injuries than male drivers. Speed ratio, a measure of speeding, increases the probability of severe injuries. Drivers in passenger cars are more likely to experience higher injury severity levels than those in passenger vans or pickup trucks.

Abdel-Aty and Keller (2005) analyzed accidents' injury severity on signalized intersections, where OPM was used. Findings indicated that having a divided minor roadway or a higher speed limit on the minor roadway is associated with a decreased injury severity while accidents involving a pedestrian/bicyclists and left turn accidents had the highest probability of a more severity accident.

Oh (2006) established a statistical relationship correlating crash severity with weather, traffic maneuvers, and specific roadway geometrics at four-legged signalized intersections in rural areas. Four models were built: single vehicle, two vehicles, three or more vehicles, and multiple vehicles. The results indicated that for all accidents' model: sharper horizontal curves, angle and head on, rear end crashes, and higher speed limits on major roads, are all associated with higher injury severity. While AADT on major roads, presence of protected left turn, more vehicles' occupants, wider medians on major roads, and more commercial driveways on minor road are all associated to less injury severity. For three or more vehicles' accidents: number of vehicles' involved is associated with higher injury severity, while right turn lane on minor roads, longer sight distance for minor roads, higher AADT on minor roads are associated with less injury severity. For two vehicles' accidents: more vehicles' occupants, sharper horizontal curves, higher speed limit, and more vertical curves on minor roads are all associated with higher injury severity while lower number of commercial driveways, higher traffic flows on major roads, and crash type are associated with lower injury severity. For single vehicles' crash: more driveways on major roads, protected left turn lane, and more AADT on major roads are associated with less injury severity, where highest crest curves on minor roads are associated with higher injury severity.

Gårder (2006) analyzed the statistical association between head-on accident severity and potential causal factors. Using OPM the results indicated that wet pavement, narrow road segments, high density of access points, night time, and braking performance on wet pavement are all associated with higher injury severity while wider lanes are associated with less injury severity.

Gray et al. (2008) studied accidents for young male drivers and modeled these accidents using OPM. Findings indicated that time period (0:00-7:00), Days of the week (Thursday, Friday, Saturday, and Sunday), nighttime, Months of (January, May, April, and August), good weather with strong wind, slippery and wet pavement , volatile movement, object impacts on the roadway, hazards on the roadway where a previous accident has occurred, two-lane highways, and vehicle to be on the main road than entering or leaving the main road are all associated with higher injury severity.

Xie et al. (2009) analyzed the relationship between accident injury severity and factors such as driver's characteristics, vehicle type, and roadway conditions. Two models (BOP and OPM) were built and compared based on datasets with different sample sizes. comparison results show that these two types of models produce similar results for large sample data. When the sample data size is small, with proper prior setting, the BOP model can produce more reasonable parameter estimations and better prediction performance than the OP model. This research also found that the BOP model provides a flexible framework that can combine information contained in the data with the prior knowledge of the parameters to improve model performance. Modeling results indicated that weekends, off-peak periods, two-way streets, involves only one vehicle, accidents involving pedestrian, and accidents involving other vulnerable road users are all associated with higher injury severity. While, locations with police control, locations with road median, crashes involving hit object, parked vehicles, and sideswipes are associated with lower injury severity.

Wang et al. (2009) used OPM to identify factors contributing to injury severity at freeway diverge areas and to evaluate impacts of the factors. The results indicated that factors that significantly influence injury severity at freeway diverge areas include length of deceleration and ramp lanes, curve and grade at diverge areas, light and weather conditions, alcohol or drug involvement, heavy-vehicle involvement, number of lanes on main lines, average daily traffic on main lines, surface condition, land type, and crash type.

Haleem and Abdel-Aty (2010) compared the previously mentioned results of using HL with others using OPM and PM to analyze accident injury severity at three- and four-legged un-signalized intersections. Several important factors affecting crash severity at unsignalized intersections were identified. These include the traffic volume on the major approach, and the number of through lanes on the minor approach (surrogate measure for traffic volume), and among the geometric factors, the upstream and downstream distance to the nearest signalized intersection, left and right shoulder width, number of left turn movements on the minor approach, and number of right and left turn lanes on the major approach. As for driver factors, young and very young at-fault drivers were associated with the least fatal probability compared to other age groups.

Lemp et al. (2011) studied the impact of vehicle, occupant, driver, and environmental characteristics on injury outcomes for those involved in crashes with heavy-duty trucks. They used an OPM and a HOP to model the injury severity, the results indicated that the HOP's likelihood dominates the OPM's 100% of the time. Findings indicated were generally consistent by both models, where the following factors were found to be significant, mean number of trailers, and two-trailer long-combination vehicles are associated with higher injury severity, where truck length, and gross vehicle weight rating were found to be associated with less injury severity.

Zhu and Srinivasan (2011) analyzed the empirical factors affecting injury severity of large-truck. Two measures of severity were used: PAR (determined from police accident reports), and RES (determined by researchers). Several similarities and some differences were observed across the two models which underscore the need for improved accuracy in the assessment of injury severity of accidents. The models indicated the impacts of several driver behavior variables on the severity of the accidents, after controlling for a variety of other factors. For example, driver distraction (truck drivers), alcohol use (car drivers), and emotional factors (car drivers) are found to be associated with higher severity crashes. A further interesting finding is the strong statistical significance of several dummy variables that indicate missing data – these reflect how the nature of the crash itself could affect the completeness of the data.

iii. Studies using other models

Other studies found in literature that used different modelling techniques than the ones described previously include, decision trees, Bayesian networks, neural networks, linear genetic algorithms and log-linear models.

Council and Stewart (1996), Chen and Jovanis (2000), and Chang and Wang (2006) all used decision trees t analyze injury severity of traffic accidents.

To analyze severity of accident of single vehicles with fixed objects and to develop an injury indexes for various fixed objects that are struck when vehicles leave the roadway, CART were used by Council and Stewart (1996). The results found indicated that area type, speed limit, vehicle type were significantly associated with injury severity.

Chen and Jovanis (2000) used CHIAD to identify significant variables that contribute to the occurrence of a specific injury severity for bus drivers. They used CHAID to select the variables prior to applying a LLM to the data used. Their modeling results indicated that late-night or early morning driving increases the risk for bus drivers of being severely injured, particularly when the drivers caused the accident or when the drivers were involved in rear-end accidents. Bus accidents involving large trucks or tractor-trailers also increase the risk.

CART was also used by Chang and Wang (2006) to establish the relationship between injury severity and driver/vehicle characteristics, highway/environmental variables and accident variables. The results indicated that pedestrians, motorcycle and bicycle riders are identified to have higher risks of being injured than other types of vehicle drivers in traffic accidents.

In 2001 the first application of ANNs to the injury severity analysis was performed by Abdelwahab and Abdel-Aty (2001) on two-vehicle accidents that occurred at signalized intersections. They built a MLP, a Fuzzy ARTMAP ANNs, and they compared the results obtained by both to a model built using OLM as mentioned previously. The results indicated that the MLP ANN outperforms that of Fuzzy ARTMAP and OLM. Where the findings indicated that at fault drivers, and using the seat belts decreases the injury severity while passenger car drivers, impact at the driver side, and intersections located in rural areas increase the injury severity of the traffic accident.

Abdel-Aty and Abdelwahab (2004) used MLP ANN and Fuzzy ARTMAP to investigate the viability and potential benefits of using the ANN in predicting driver injury severity conditioned on the premise that a crash has occurred. An OPM was also compared to the results obtained by ANN model. The results indicated the superiority of the MLP ANN model once more over both the Fuzzy ARTMAP and the OPM. Findings indicated that speed ratio increase, female drivers, rural areas, older drivers, being under the influence of alcohol, passenger cars, nighttime, side impact of drivers, and drivers on curved segments are all associated with increased injury severity while using the seat belt was found to decrease the injury severity.

Delen et. al. (2006) used a series of MLP ANNs to model the potentially non-linear relationships between the injury severity levels and accident related factors. Results indicated that the use of a restraint system like a seat belt, use of alcohol or drugs, persons' age and gender, and vehicle role in the accident were found to have an important influence on the outcome of the accident. At the same time, weather conditions or the time of the accident did not seem to affect the severity risk of injury.

Bayesian networks were used in the analysis of injury severity of traffic accidents by Simoncic (2004). He built a BN to model the severity resulting from two-car accidents. the findings indicated that older and younger drivers, night time, accidents outside built up areas, rainy weather, weekends, inexperienced drivers, and male under 25 years with less than one year of experience are all associated with an increased risk of higher injury severity while driving experience >11 years), and working days were found to be associated with less injury severity.

Most recently, LGP was applied to the analysis and modeling of injury severity of traffic accidents. Das and Abdel-Aty (2010) aimed to understand the relationship of geometric and environmental factors with injury related accidents as well as with severe accidents. Their results indicated that dry pavement, good pavement conditions, wider shoulder, and wider sidewalk width are associated with less injury severity. They also found that vision obstruction was found to be a leading factor for severe accidents, percentage of trucks, even if small, was found to be more likely to make the accident injury prone. On the other hand, interaction terms among variables like on-street parking with higher posted speed limit have been found to make injuries more probable.

APPENDIX II

Table 1: Associations between variables

SEV	Chi-square	1,000			
	p-value				
	Fisher's test				
	p-value				
	Cramer´s V				
	p-value				
	Phi correlation				
	p-value				
ACT	Chi-square	45,880	1,000		
	p-value	0,000			
	Fisher's test	46,103			
	p-value	0,000			
	Cramer´s V	**0,173**			
	p-value	0,000			
	Phi correlation	0,170			
	p-value	0,000			
AGE	Chi-square	6,877	27,196	1,000	
	p-value	0,034	0,010		
	Fisher's test	6,866	27,311		
	p-value	0,034	0,006		
	Cramer´s V	**0,067**	**0,094**		
	p-value	0,034	0,007		
	Phi correlation	0,067	0,133		
	p-value	0,034	0,010		
ATF	Chi-square	7,939	20,608a	7,593	1,000
	p-value	0,046	0,295	0,257	
	Fisher's test	7,851	19,799b	7,567	
	p-value	0,048	0,254	0,231	
	Cramer´s V	**0,072**	0,067	0,050	
	p-value	0,046	0,295	0,257	
	Phi correlation	0,072	0,116	0,070	
	p-value	0,047	0,295	0,257	
		SEV	ACT	AGE	ATF

CAU	Chi-square	5,081a	30,720a	3,589a	29,857a	1,000			
	p-value	0,165	0,060	0,715	0,017				
	Fisher's test	5,057b	27,098b	2,142b	18,947b				
	p-value	0,159	0,036	0,892	0,023				
	Cramer´s V	0,058	**0,082**	0,034	**0,080**				
	p-value	0,165	0,060	0,715	0,017				
	Phi correlation	0,058	0,141	0,048	0,139				
	p-value	0,165	0,060	0,715	0,017				
DAY	Chi-square	5,632	58,333a	11,022	38,715a	10,139a	1,000		
	p-value	0,232	0,000	0,198	0,000	0,581			
	Fisher's test	5,590	56,673b	10,504	27,795b	7,685b			
	p-value	0,234	0,000	0,219	0,003	0,780			
	Cramer´s V	0,061	**0,970**	0,060	**0,092**	**0,047**			
	p-value	0,232	0,000	0,198	0,000	0,581			
	Phi correlation	0,061	0,195	0,085	0,159	0,081			
	p-value	0,232	0,000	0,198	0,000	0,581			
GEN	Chi-square	7,666	8,925	8,028	6,656a	12,893a	14,010	1,000	
	p-value	0,006c	0,182	0,017	0,085	0,010	0,007		
	Fisher's test	7,275*	8,889	9,391	6,390b	11,663b	14,466		
	p-value	0,007*	0,180	0,008	0,091	0,006	0,006		
	Cramer´s V	**0,071**	0,076	**0,072**	0,066	**0,092**	**0,096**		
	p-value	0,006	0,182	0,017	0,085	0,010	0,007		
	Phi correlation	0,071d	0,076	0,072	0,066	0,092	0,096		
	p-value	0,006	0,182	0,017	0,085	0,010	0,007		
LAW	Chi-square	4,769	47,763a	12,593	4,896a	37,412a	2,879a	0,131	1,000
	p-value	0,089	0,000	0,014	0,529	0,002	0,945	0,959	
	Fisher's test	4,715	43,304b	11,681	4,565b	23,352b	2,015b	0,153	
	p-value	0,091	0,000	0,016	0,530	0,000	0,979	0,944	
	Cramer´s V	**0,056**	**0,125**	**0,064**	0,040	**0,110**	0,031	0,009	
	p-value	0,089	0,000	0,014	0,529	0,002	0,945	0,959	
	Phi correlation	0,056	0,176	0,091	0,056	0,156	0,043	0,009	
	p-value	0,089	0,000	0,014	0,529	0,002	0,945	0,959	
		SEV	ACT	AGE	ATF	CAU	DAY	GEN	LAW

LIG	Chi-square	21,035	89,645a	62,970	13,997a	25,745a	17,426	10,352	9,318a
	p-value	0,000	0,000	0,000	0,297	0,034	0,351	0,036	0,303
	Fisher's test	20,903	83,645b	61,563	14,695a	16,497b	17,231	9,951	10,470b
	p-value	0,000	0,000	0,000	0,186	0,108	0,335	0,039	0,186
	Cramer´s V	**0,117**	**0,121**	**0,143**	0,055	**0,075**	0,053	**0,082**	0,055
	p-value	0,000	0,000	0,000	0,297	0,034	0,351	0,036	0,303
	Phi correlation	0,117	0,242	0,202	0,095	0,129	0,107	0,082	0,078
	p-value	0,000	0,000	0,000	0,297	0,034	0,351	0,036	0,303
MON	Chi-square	3,949	17,483	7,034	93,958	11,544a	36,722	0,811	6,287
	p-value	0,262	0,480	0,316	0,000	0,240	0,001	0,846	0,393
	Fisher's test	3,947	17,544	7,012	113,456	10,760b	33,576	0,810	6,209
	p-value	0,262	0,467	0,316	0,000	0,225	0,001	0,847	0,396
	Cramer´s V	0,051	0,062	0,048	**0,143**	0,050	**0,089**	0,023	0,045
	p-value	0,262	0,490	0,316	0,000	0,231	0,001	0,846	0,393
	Phi correlation	0,051	0,107	0,068	0,247	0,087	0,155	0,023	0,064
	p-value	0,262	0,490	0,316	0,000	0,231	0,001	0,846	0,393
NOI	Chi-square	44,657	51,608	0,841	8,062	2,079a	11,917	6,999	1,432
	p-value	0,000c	0,000	0,657	0,047	0,563	0,017	0,010c	0,479
	Fisher's test	43,910*	48,430	0,906	7,963	1,357b	11,744	6,593*	1,529
	p-value	0,000*	0,000	0,641	0,049	0,694	0,018	0,010*	0,468
	Cramer´s V	0,171	**0,183**	0,023	**0,072**	0,037	**0,088**	0,068	0,031
	p-value	0,000	0,000	0,657	0,047	0,563	0,017	0,010	0,479
	Phi correlation	**-0,171**d	0,183	0,023	0,072	0,037	0,088	**-0,068**d	0,031
	p-value	0,000	0,000	0,657	0,047	0,563	0,017	0,010	0,479
OI	Chi-square	4,618	929,998	18,426	19,204	4,562a	21,529	7,460	2,976
	p-value	0,099	0,000	0,001	0,005	0,613	0,006	0,025	0,567
	Fisher's test	4,608	1073,835	21,532	18,509	4,641b	21,575	7,199	3,075
	p-value	0,099	0,000	0,000	0,006	0,582	0,006	0,028	0,546
	Cramer´s V	**0,055**	**0,550**	**0,077**	**0,079**	0,039	**0,084**	**0,070**	0,031
	p-value	0,099	0,000	0,001	0,005	0,613	0,006	0,025	0,567
	Phi correlation	0,055	0,778	0,110	0,112	0,055	0,118	0,070	0,044
	p-value	0,099	0,000	0,001	0,005	0,613	0,006	0,025	0,567
		SEV	ACT	AGE	ATF	CAU	DAY	GEN	LAW

MON	Chi-square	48,804	1,000		
	p-value	0,000			
	Fisher's test	47,761			
	p-value	0,000			
	Cramer´s V	**0,103**			
	p-value	0,000			
	Phi correlation	0,178			
	p-value	0,000			
NOI	Chi-square	3,734	3,591	1,000	
	p-value	0,443	0,310		
	Fisher's test	3,811	3,558		
	p-value	0,430	0,315		
	Cramer´s V	0,049	0,048		
	p-value	0,443	0,310		
	Phi correlation	0,049	0,048		
	p-value	0,443	0,310		
OI	Chi-square	47,781	5,949	624,723	1,000
	p-value	0,000	0,437	0,000	
	Fisher's test	46,540	5,948	670,481	
	p-value	0,000	0,435	0,000	
	Cramer´s V	**0,125**	0,044	**0,638**	
	p-value	0,000	0,437	0,000	
	Phi correlation	0,176	0,062	0,638	
	p-value	0,000	0,437	0,000	
		LIG	MON	NOI	OI

PAS	Chi-square	0,679	48,596	5,045	6,704	1,335a	0,146	0,641	236,978
	p-value	0,428c	0,000	0,081	0,085	0,745	0,997	0,432c	0,000
	Fisher's test	0,589*	48,519	5,104	6,344	1,526b	0,132	0,521*	221,082
	p-value	0,443*	0,000	0,077	0,101	0,713	0,998	0,470*	0,000
	Cramer´s V	0,021	**0,178**	0,057	0,066	0,029	0,010	0,020	**0,393**
	p-value	0,435	0,000	0,081	0,085	0,745	0,997	0,433	0,000
	Phi correlation	0,021d	0,178	0,057	0,066	0,029	0,010	0,020d	0,393
	p-value	0,435	0,000	0,081	0,085	0,745	0,997	0,433	0,000
PAW	Chi-square	0,591	63,651	4,175	5,270	5,298a	2,574	0,645	687,670
	p-value	0,747	0,000	0,382	0,512	0,490	0,958	0,728	0,000
	Fisher's test	0,601	61,196	4,091	5,188	5,088b	2,434	0,728	669,621
	p-value	0,743	0,000	0,389	0,511	0,483	0,964	0,690	0,000
	Cramer´s V	0,020	**0,144**	0,037	0,041	0,042	0,029	0,020	**0,473**
	p-value	0,747	0,000	0,382	0,512	0,490	0,958	0,728	0,000
	Phi correlation	0,020	0,204	0,052	0,059	0,059	0,014	0,020	0,669
	p-value	0,744	0,000	0,382	0,512	0,490	0,958	0,728	0,000
ROM	Chi-square	1,250	75,326a	5,941	10,114a	21,620a	13,372	1,137	390,174a
	p-value	0,741	0,000	0,423	0,324	0,031	0,332	0,766	0,000
	Fisher's test	1,256	71,563b	6,533	8,928b	18,058b	12,635	1,114	332,387b
	p-value	0,743	0,000	0,344	0,349	0,019	0,358	0,769	0,000
	Cramer´s V	0,029	**0,128**	0,044	0,047	**0,068**	0,054	0,027	**0,356**
	p-value	0,741	0,000	0,423	0,324	0,031	0,332	0,766	0,000
	Phi correlation	0,029	0,221	0,062	0,081	0,119	0,093	0,027	0,504
	p-value	0,741	0,000	0,423	0,324	0,031	0,332	0,766	0,000
SHT	Chi-square	1,812a	71,912a	14,815a	9,148a	5,246a	12,809a	4,768	295,420a
	p-value	0,635	0,000	0,025	0,424	0,682	0,367	0,180	0,000
	Fisher's test	1,777b	66,872b	13,888b	8,404b	6,919b	11,272b	5,122	300,364b
	p-value	0,638	0,000	0,024	0,464	0,746	0,460	0,150	0,000
	Cramer´s V	0,034	**0,125**	**0,069**	0,045	0,034	0,053	0,056	**0,310**
	p-value	0,635	0,000	0,025	0,400	0,682	0,367	0,180	0,000
	Phi correlation	0,034	0,216	0,098	0,077	0,058	0,091	0,056	0,439
	p-value	0,635	0,000	0,025	0,400	0,682	0,367	0,180	0,000
		SEV	ACT	AGE	ATF	CAU	DAY	GEN	LAW

PAS	Chi-square	5,683	30,721	0,410	6,377	1,000			
	p-value	0,228	0,000	0,538c	0,037				
	Fisher's test	5,683	30,305	0,335*	6,283				
	p-value	0,225	0,000	0,563*	0,038				
	Cramer´s V	0,061	**0,141**	0,016	**0,064**				
	p-value	0,228	0,000	0,541	0,037				
	Phi correlation	0,061	0,141	-0,016d	0,064				
	p-value	0,228	0,000	0,522	0,037				
PAW	Chi-square	11,145	6,458	1,323	15,925	402,210	1,000		
	p-value	0,190	0,378	0,524	0,003	0,000			
	Fisher's test	11,354	6,260	1,310	15,111	395,614			
	p-value	0,120	0,401	0,519	0,004	0,000			
	Cramer´s V	0,060	0,046	0,029	**0,072**	**0,512**			
	p-value	0,194	0,378	0,524	0,003	0,000			
	Phi correlation	0,085	0,065	0,029	0,102	0,512			
	p-value	0,190	0,378	0,524	0,003	0,000			
ROM	Chi-square	13,354a	10,638	2,525	16,537	342,715	593,456	1,000	
	p-value	0,331	0,303	0,477	0,011	0,000	0,000		
	Fisher's test	11,659b	11,366	2,649	15,737	312,436	471,292		
	p-value	0,417	0,250	0,450	0,014	0,000	0,000		
	Cramer´s V	0,054	0,048	0,041	**0,073**	**0,472**	**0,440**		
	p-value	0,331	0,303	0,477	0,011	0,000	0,000		
	Phi correlation	0,093	0,083	0,041	0,104	0,472	0,622		
	p-value	0,331	0,303	0,447	0,011	0,000	0,000		
SHT	Chi-square	8,481a	12,665a	4,730	9,294a	867,202	562,819	389,807a	1,000
	p-value	0,730	0,176	0,187	0,158	0,000	0,000	0,000	
	Fisher's test	6,953	12,141b	4,686	9,284b	938,883	600,157	401,119b	
	p-value	0,819	0,183	0,191	0,148	0,000	0,000	0,000	
	Cramer´s V	0,043	0,052	0,055	0,055	**0,751**	**0,428**	**0,291**	
	p-value	0,730	0,176	0,187	0,156	0,000	0,000	0,000	
	Phi correlation	0,074	0,091	0,055	0,780	0,751	0,605	0,504	
	p-value	0,730	0,176	0,187	0,156	0,000	0,000	0,000	
		LIG	MON	NOI	OI	PAS	PAW	ROM	SHT

SID	Chi-square	9,378a	96,894a	7,866a	404,141a	32,149a	33,389a	3,038a	16,001
	p-value	0,089	0,000	0,615	0,000	0,089	0,037	0,707	0,110
	Fisher's test	9,835b	97,396b	6,922	198,606b	24,888b	29,932b	3,640b	14,853
	p-value	0,073	0,000	0,654	0,000	0,072	0,032	0,581	0,103
	Cramer´s V	**0,078**	**0,112**	0,051	**0,296**	**0,084**	**0,074**	0,044	0,072
	p-value	0,089	0,000	0,615	0,000	0,089	0,037	0,707	0,110
	Phi correlation	0,780	0,251	0,072	0,513	0,145	0,147	0,044	0,102
	p-value	0,089	0,000	0,615	0,000	0,089	0,037	0,707	0,110
TIM	Chi-square	9,589	111,999	47,299	9,466	10,387a	50,074	11,696	4,096
	p-value	0,023	0,000	0,000	0,406	0,311	0,000	0,009	0,665
	Fisher's test	9,530	108,192	48,441	9,482	8,641b	47,596	12,688	4,007
	p-value	0,023	0,000	0,000	0,393	0,411	0,000	0,006	0,672
	Cramer´s V	**0,079**	**0,156**	**0,124**	0,045	0,047	**0,104**	**0,087**	0,037
	p-value	0,023	0,000	0,000	0,406	0,311	0,000	0,009	0,665
	Phi correlation	0,079	0,270	0,175	0,079	0,082	0,181	0,087	0,052
	p-value	0,023	0,000	0,000	0,406	0,311	0,000	0,009	0,665
VI	Chi-square	7,525	1847,335	20,678	5,174	5,075a	38,593	2,525	15,744
	p-value	0,023	0,000	0,000	0,528	0,520	0,000	0,280	0,005
	Fisher's test	7,529	1852,398	20,767	5,331	5,290b	38,431	2,621	15,536
	p-value	0,024	0,000	0,000	0,486	0,461	0,000	0,265	0,003
	Cramer´s V	**0,070**	**0,775**	**0,082**	0,041	0,041	**0,112**	0,041	**0,072**
	p-value	0,023	0,000	0,000	0,528	0,520	0,000	0,280	0,005
	Phi correlation	0,070	1,097	0,116	0,058	0,057	0,159	0,041	0,101
	p-value	0,023	0,000	0,000	0,528	0,520	0,000	0,280	0,005
		SEV	ACT	AGE	ATF	CAU	DAY	GEN	LAW

SID	Chi-square	38,823a	33,196	5,283	37,433a	54,334	41,251	69,431a
	p-value	0,020	0,003	0,388	0,000	0,000	0,000	0,000
	Fisher's test	37,517b	37,442	5,265	37,085b	52,240	38,434	62,570b
	p-value	0,003	0,000	0,384	0,000	0,000	0,000	0,000
	Cramer´s V	**0,079**	**0,085**	0,059	**0,110**	**0,188**	**0,116**	**0,123**
	p-value	0,020	0,003	0,388	0,000	0,000	0,000	0,000
	Phi correlation	0,159	0,147	0,059	0,156	0,188	0,164	0,213
	p-value	0,020	0,003	0,388	0,000	0,000	0,000	0,000
TIM	Chi-square	931,969	6,440	2,181	54,100	2,537	4,800	6,307
	p-value	0,000	0,697	0,538	0,000	0,463	0,575	0,712
	Fisher's test	1045,828	6,433	2,192	50,705	2,555	4,859	60,455
	p-value	0,000	0,698	0,537	0,000	0,459	0,567	0,692
	Cramer´s V	**0,450**	0,037	0,038	**0,133**	0,041	0,040	0,037
	p-value	0,000	0,697	0,538	0,000	0,463	0,575	0,712
	Phi correlation	0,779	0,065	0,038	0,189	0,041	0,056	0,064
	p-value	0,000	0,697	0,538	0,000	0,463	0,575	0,712
VI	Chi-square	64,186	2,120	51,115	1200,421	23,136	40,847	56,524
	p-value	0,000	0,913	0,000	0,000	0,000	0,000	0,000
	Fisher's test	64,184	2,178	47,700	1322,533	23,355	70,801	57,557
	p-value	0,000	0,908	0,000	0,000	0,000	0,000	0,000
	Cramer´s V	**0,145**	0,026	**0,182**	**0,625**	**0,123**	**0,115**	**0,136**
	p-value	0,000	0,913	0,000	0,000	0,000	0,000	0,000
	Phi correlation	0,204	0,037	0,182	0,884	0,123	0,163	0,192
	p-value	0,000	0,913	0,000	0,000	0,000	0,000	0,000
		LIG	MON	NOI	OI	PAS	PAW	ROM

SID	Chi-square	62,213a	1,000		
	p-value	0,003			
	Fisher's test	62,268b			
	p-value	0,000			
	Cramer´s V	**0,116**			
	p-value	0,003			
	Phi correlation	0,201			
	p-value	0,003			
TIM	Chi-square	7,340a	10,723a	1,000	
	p-value	0,604	0,781		
	Fisher's test	6,643b	10,616b		
	p-value	0,664	0,772		
	Cramer´s V	0,040	0,048		
	p-value	0,604	0,781		
	Phi correlation	0,069	0,084		
	p-value	0,604	0,781		
VI	Chi-square	41,153a	39,964a	98,608	1,000
	p-value	0,000	0,001	0,000	
	Fisher's test	40,985b	39,800	99,498	
	p-value	0,000	0,000	0,000	
	Cramer´s V	**0,116**	**0,114**	**0,179**	
	p-value	0,000	0,001	0,000	
	Phi correlation	0,164	0,161	0,253	
	p-value	0,000	0,001	0,000	
		SHT	SID	TIM	VI

a: chi-square does not apply

b: Fisher´s test applies

c: exact results are provided for the p-value

d: phi applies

*: continuity correction (Yate’s correction) is used

— Numbers in violate are significantly little

— Numbers in red are significantly low

— Numbers in green are significantly moderate

— Numbers in blue are significantly high

APPENDIX III

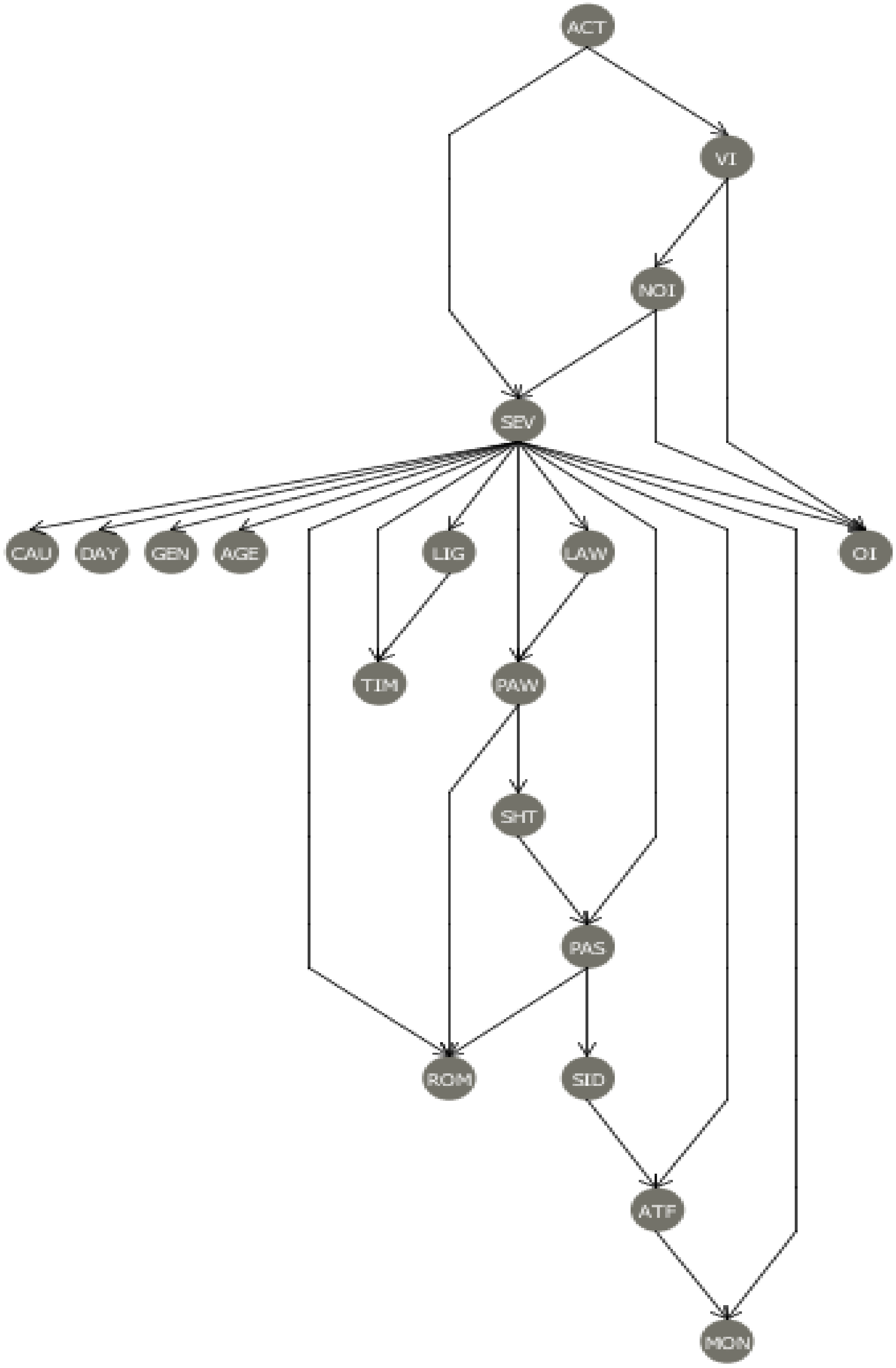


Figure 1: Bayesian network for Hillclimibing + BDeu

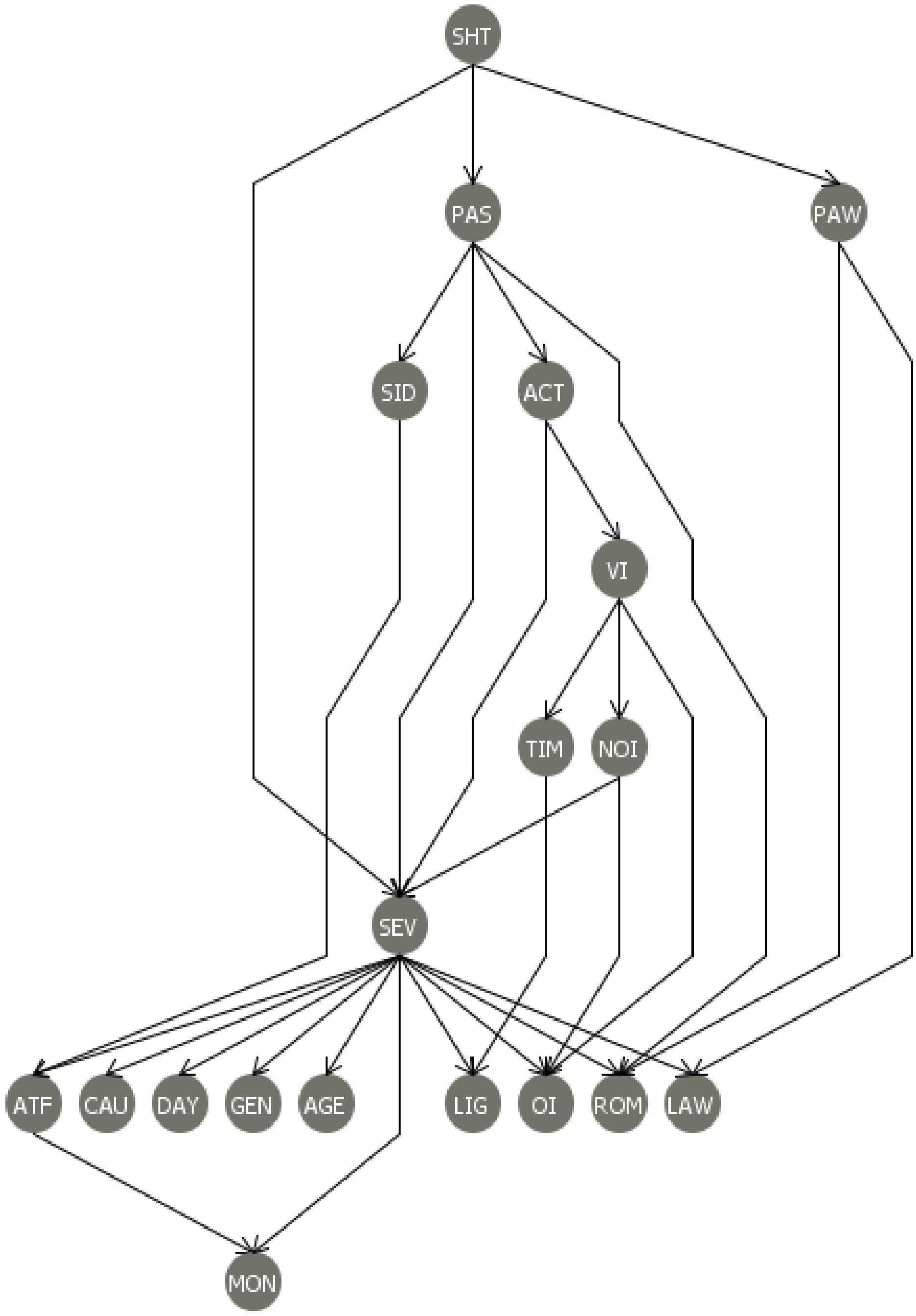


Figure 2: Bayesian network for Hillclimibing + MDL

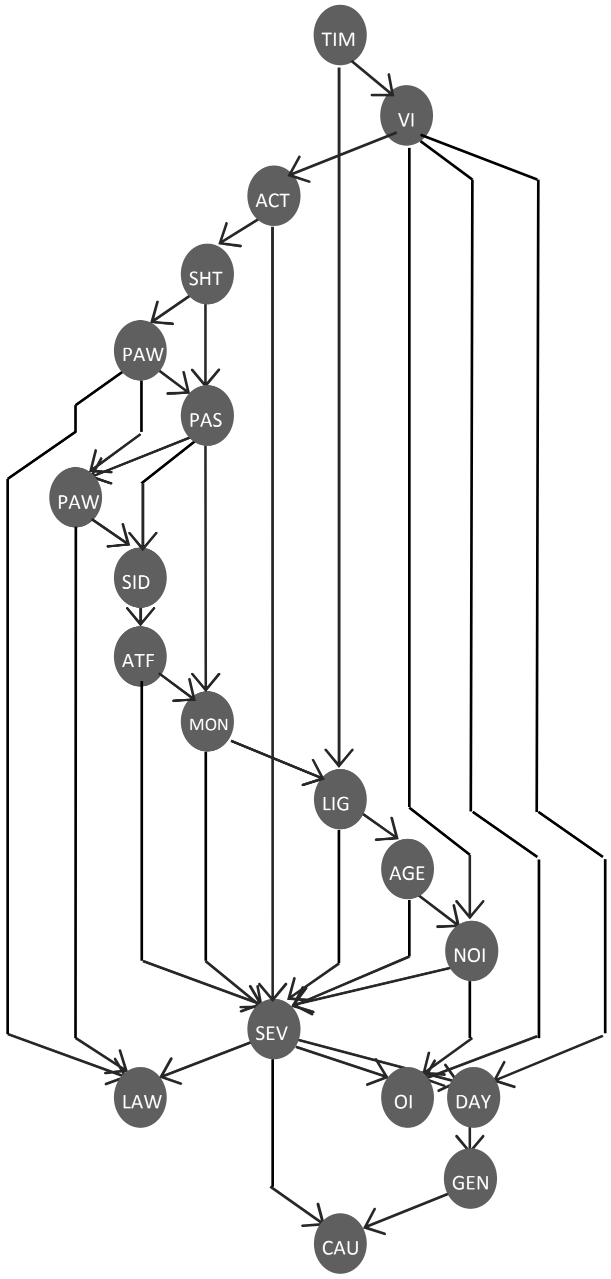


Figure 3: Bayesian network for Hillclimibing + AIC

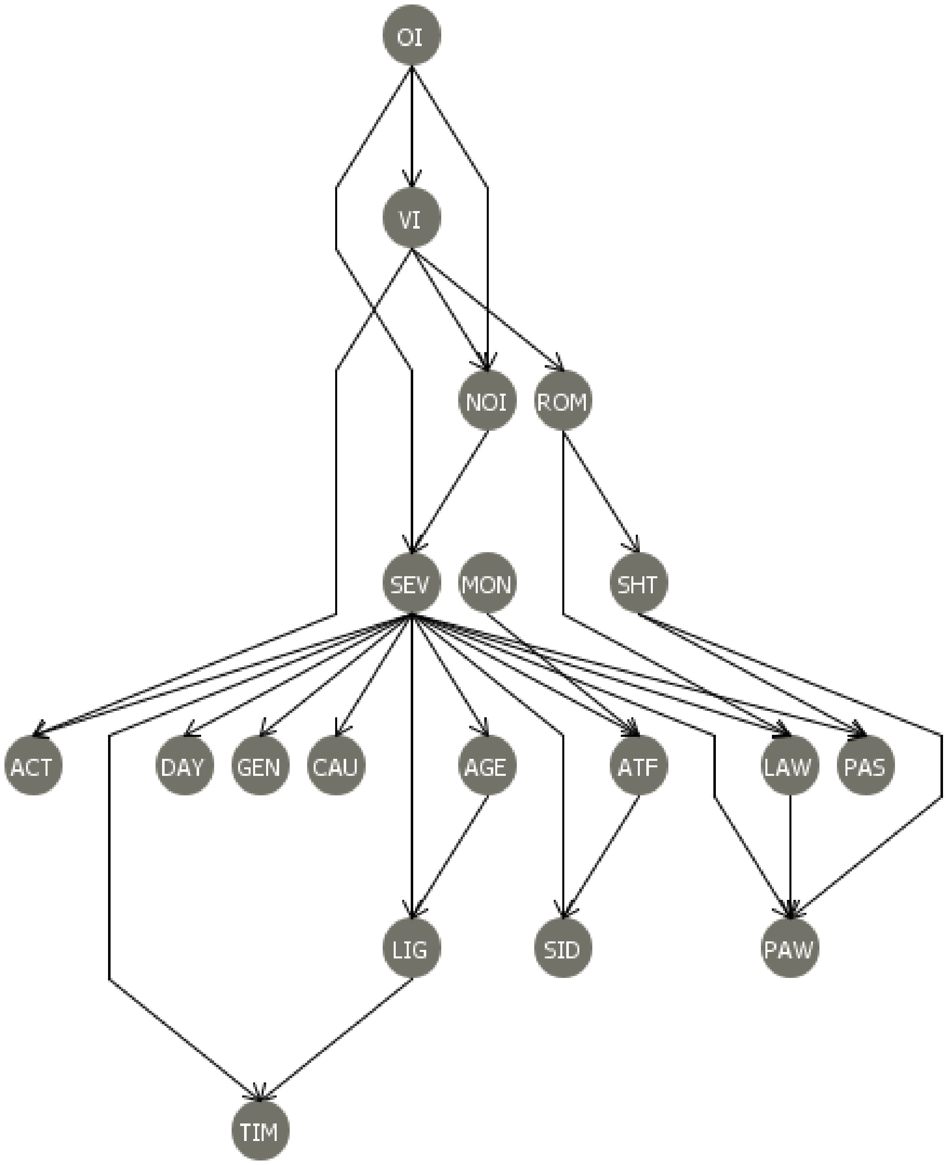


Figure 4: Bayesian network for Simulated Annealing + BDeu

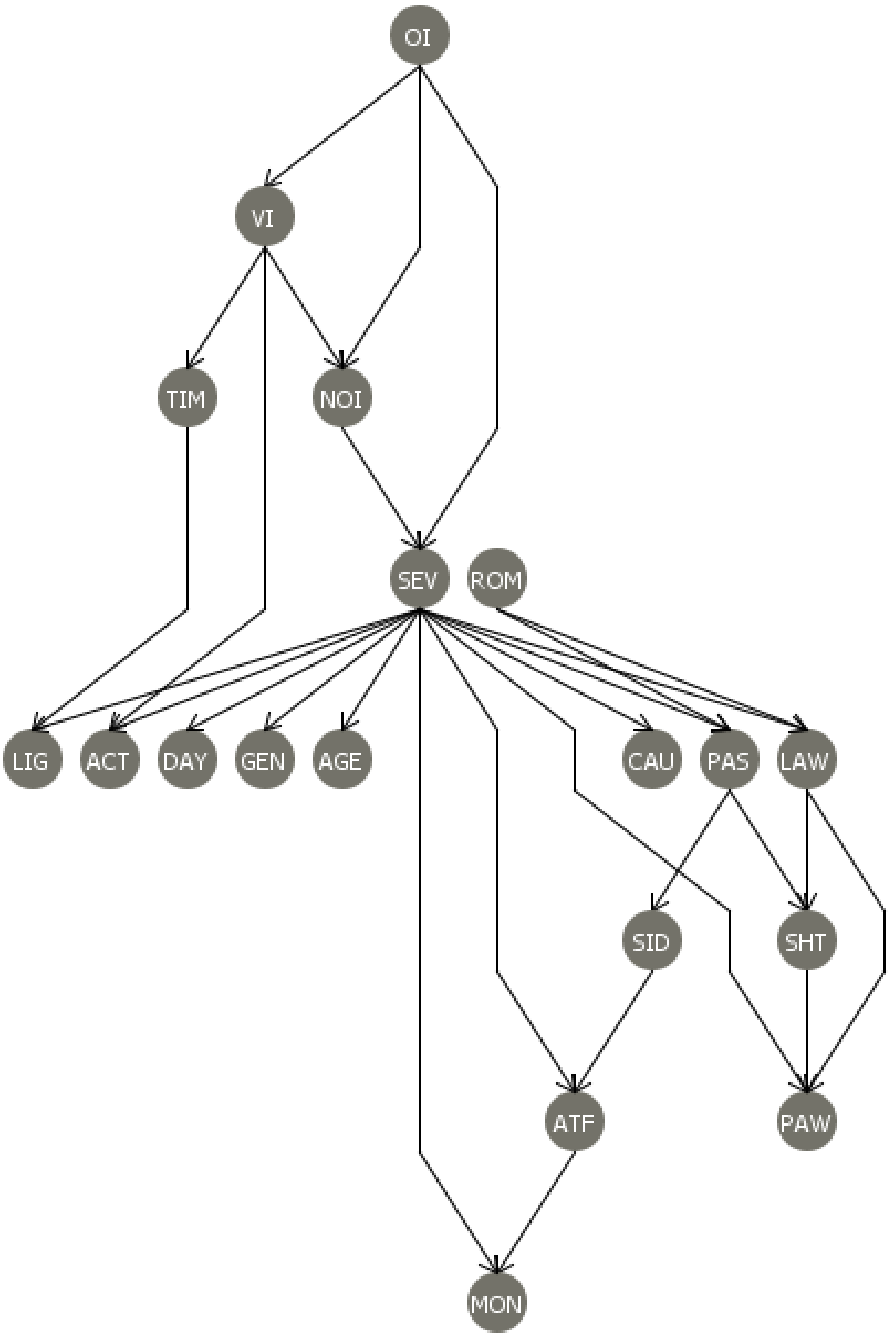


Figure 5: Bayesian network for Simulated Annealing + MDL

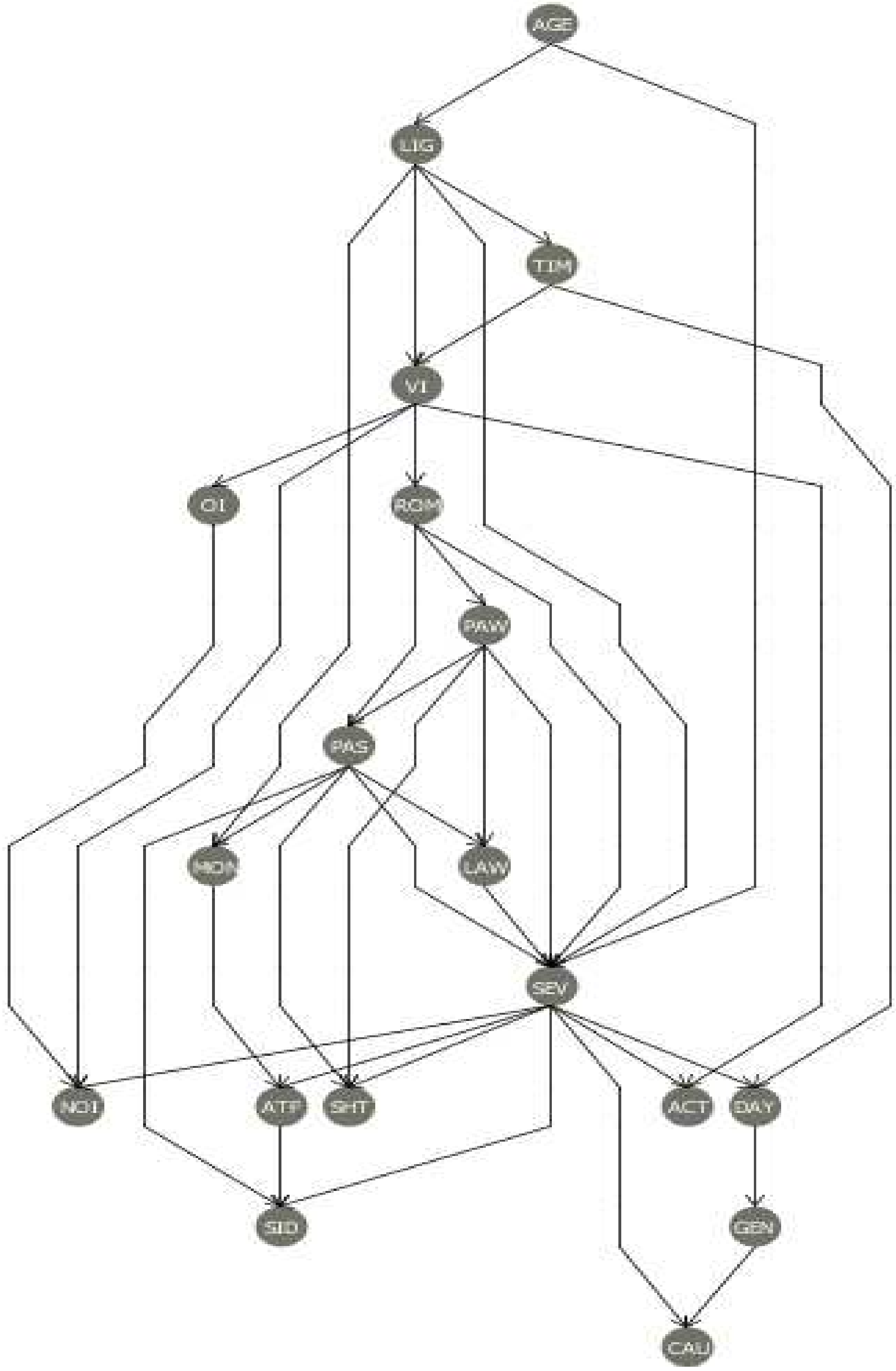


Figure 6 Bayesian network for Simulated Annealing + AIC

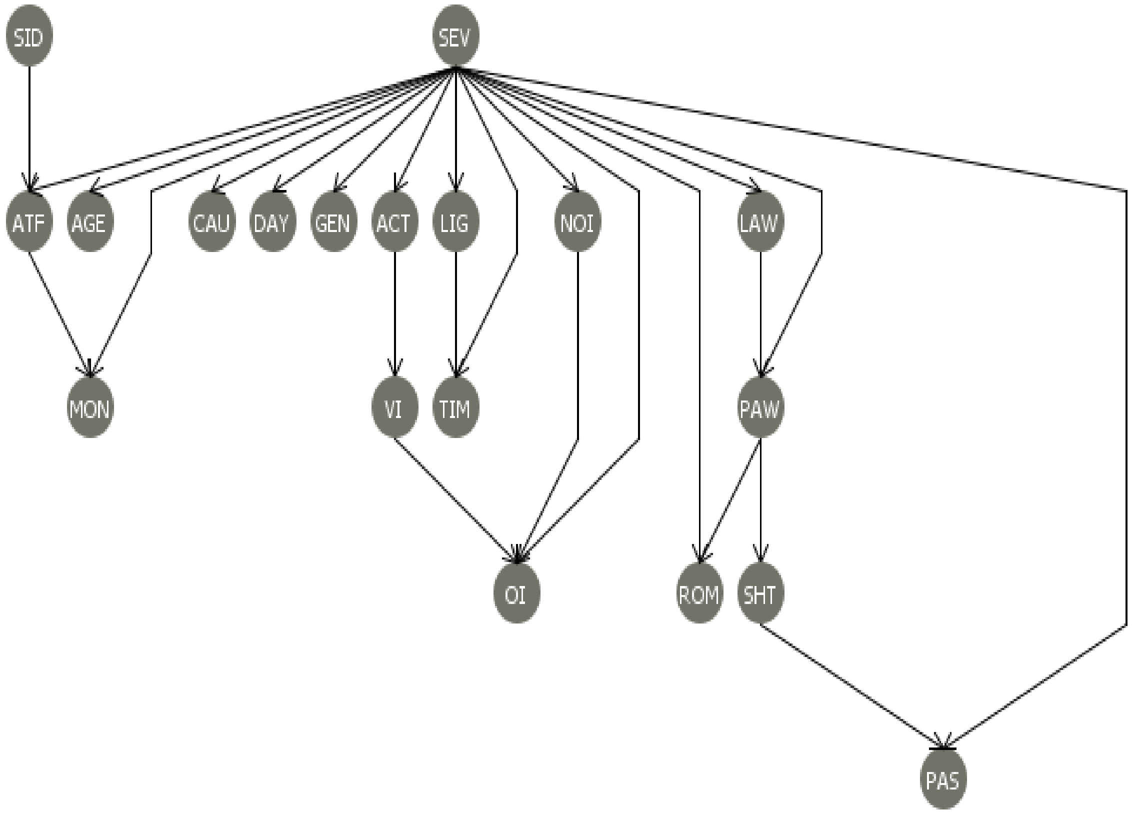


Figure 7: Bayesian network for Tabu + BDeu

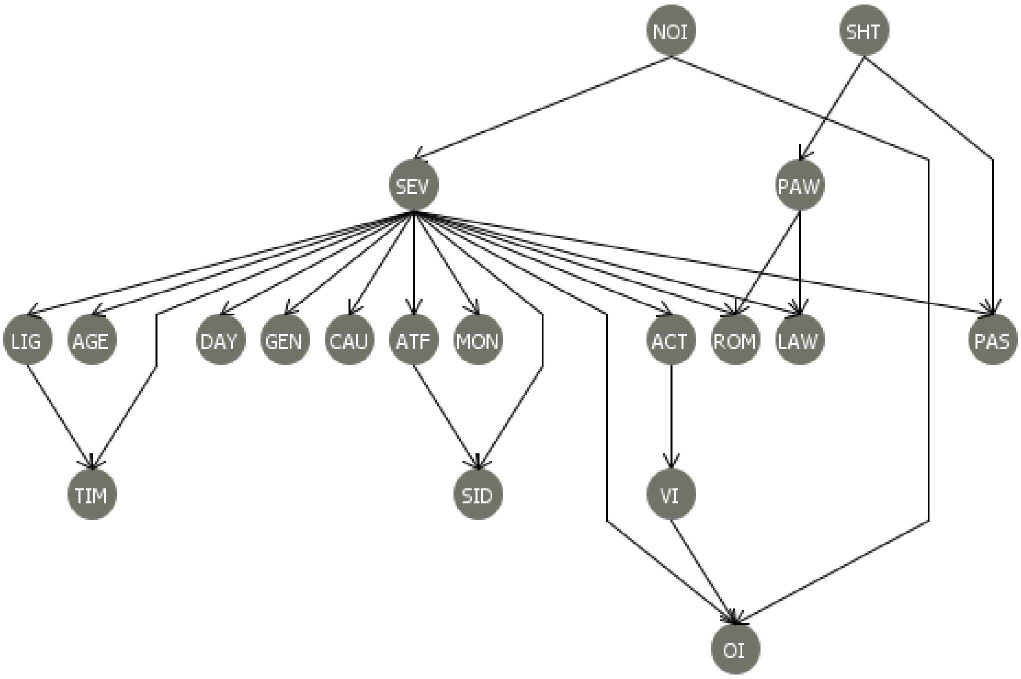


Figure 8: Bayesian network for Tabu + MDL

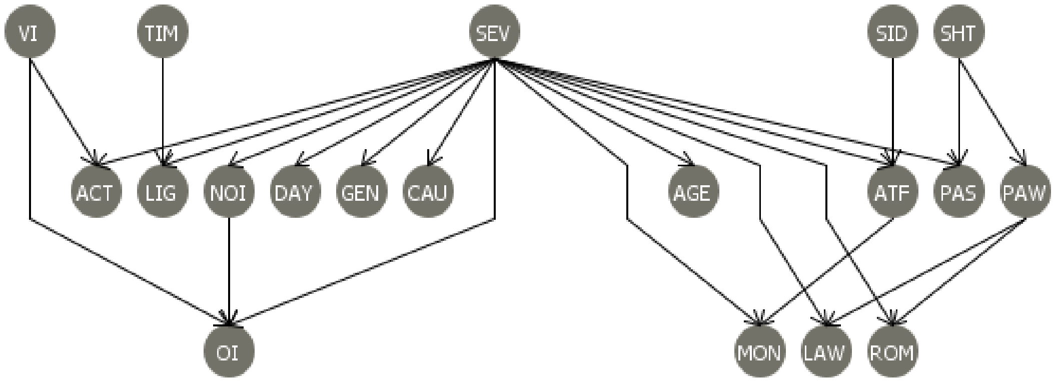


Figure 9: Bayesian network for Tabu + AIC

APPENDIX IV

Published and Accepted Papers for Publishing

This appendix includes the three papers that were prepared during the research work period:

Paper 1 Randa Oqab Mujalli and Juan de Oña. Injury Severity Models for Motorized Vehicle Accidents: A review. Paper accepted in *Transport*

Paper 2 Juan de Oña, Randa Oqab Mujalli, Francisco J. Calvo. Analysis of traffic accident injury severity on Spanish rural highways using Bayesian networks. Published in *Accident Analysis and Prevention*, Volume 43, Issue 1, January 2011, Pages 402-411.

Paper 3 Randa Oqab Mujalli and Juan de Oña. A method for simplifying the analysis of traffic accidents injury severity on two-lane highways using Bayesian networks. Paper accepted in *Journal of Safety Research*

1. Injury Severity Models for Motorized Vehicle Accidents: A review

Randa Oqab Mujalli and Juan de Oña López: Injury Severity Models for Motorized Vehicle Accidents: A review. ***Accepted in Transport with minor revisions***

This paper presents a comprehensive review of the existing literature of accidents' injury severity. Several modeling techniques were found to be used by researchers analyzing injury severity of traffic accidents. In which the modeling techniques used to analyze injury severity of traffic accidents were classified into 4 groups: discrete outcome models, data mining techniques, soft computing techniques and other techniques. The analysis and the comparison between models was performed based on seven criteria (modeling technique, number of records, number of variables, area type, features, injury level and model fit).

A statistical analysis was performed on the number of variables used, number of records and the number of injury severity levels. The statistical analysis is aimed to find out if there exists a significant statistical difference in these magnitudes using different modeling techniques.

The results indicated that neither the number of variables used in a study nor the number of records showed a significant statistical difference with respect to the modeling technique used. The only magnitude that showed a significant statistical difference between the probit models and the other models was the number of categories used for the injury severity.

Thus, this paper indicated that each modeling technique has its own characteristics and limitations, in which the researcher should be aware of which when using a specific model.

De: em.tran.0.23e273.e0f8a734@editorialmanager.com
[mailto:em.tran.0.23e273.e0f8a734@editorialmanager.com] En nombre de Transport
Enviado el: miércoles, 22 de junio de 2011 19:02
Para: Juan de Ona
Asunto: Editor's comments on your article submitted to Transport

Title: Severity models for 4-wheels vehicles accidents: a review
Article number: TRAN-D-11-00026

Dear Dr de Ona,

I am pleased to inform you that your article has been accepted for publication in Transport, subject to minor revisions.

I enclose the comments received from external peer review at the bottom of this message. Please revise your article, taking these into account, and resubmit a revised version along with a letter detailing, point by point,the changes you have made (or not made and why).

I would be grateful if you would resubmit your article by 13 Jul 2011.

CHECKLIST OF ITEMS FOR RESUBMISSION
a) Your post-nominal letters and affiliation (eg: BEng, CEng, MICE, senior engineer at Arup, London).

b) Your co-authors' post-nominal letters and affiliations:
(it is the lead author's responsibility to provide these)

c) Keywords, available to download here
<http://www.icevirtuallibrary.com/upload/proceedingskeywords.pdf> .

d) Author photographs (it is the responsibility of lead author to provide these; the publisher will not include author photographs unless a complete
set is provided)

e) Figures/tables/illustrations.
Line drawings, graphs and figures should be no larger than A4 size and should be capable of bearing reduction (ie: sharp black lines).

To be suitable for reproduction, the width of the image should be at least
1500 pixels or 'dots', for which you will need a digital camera setting of

1500 x 1000 pixels or scan a 150 x 100 mm (6" x 4") photographic print or drawing at a scanner resolution of at least 250 dots per inch (dpi). Please remember that the journal is printed in black and white. Detailed guidelines can be downloaded here
<http://www.icevirtuallibrary.com/upload/figure.pdf>
.

f) It is your responsibility to obtain copyright permission to reproduce figures that you submit with your paper. Download a request form here
<http://www.icevirtuallibrary.com/upload/Figureclearanceform.pdf> .

g) You (and any co-authors) have signed a copyright transfer form. We cannot publish your article without this. Download the form here
<http://www.icevirtuallibrary.com/upload/jnlcopyright.pdf> .

h) You have supplied references in Harvard style. For details, click here
<http://www.icevirtuallibrary.com/upload/reference.pdf> .

SUPPLEMENTARY DATA
You may place data, supplementary data to your peer reviewed article, on the journal homepage. For example, extra material such as tables of raw data.

This supplementary data will only be available online only.

If you wish to do this, please upload your file(s) as a separate document called 'Supplementary Data' when you upload your revision.

To submit a revision, go to http://tran.edmgr.com/ and log in as 'author'.
You will see a menu item call 'Submission Needing Revision'. You will find
your submission record there.

Best wishes

Agnes Alvite
Journal Coordinator

Text written: 2011

Injury Severity Models for Motorized Vehicle Accidents: A review

Randa Oqab Mujalli (M.Sc Engg) and Juan de Oña[#] (Ph.D. Associate Professor)

TRYSE Research Group. Department of Civil Engineering, University of Granada

[#] Corresponding author, ETSI Caminos, Canales y Puertos, c/ Severo Ochoa, s/n, 18071 Granada (Spain), Phone: +34 958 24 99 79, email: jdona@ugr.es

Main text: 7,070 words. Number of tables: 5. Number of figures: 2.

Injury Severity Models for Motorized Vehicle Accidents: A review

Abstract

Modeling of traffic accidents injury severity is a complex task. In the last few years the number and variety of studies that analyze injury severity of traffic accidents have increased considerably. In this paper 19 modeling techniques used to model injury severity of traffic accidents where at least a 4-wheeled vehicle is involved have been analyzed. The analysis and the comparison between models was performed based on seven criteria (modeling technique, number of records, number of variables, area type, features, injury level and model fit). In general, it is not possible to recommend a method that could be identified as the best one. Each modeling technique has its own limitations and characteristics, being aware of which will help analysts to decide the best method to be used in each particular modeling problem. However, some general conclusions could be established: in most of cases the results of models' fits are found to be satisfactory, though not excellent; in the case of data mining models accuracy improves with balanced datasets; and no correlation was found to exist between the number of accidents' records and the number of analyzed variables.

Keywords: health & safety, reviews, safety & hazards, statistical analysis, traffic engineering

1. Introduction

Road accidents constitute a major public health problem worldwide, causing around 1.2 million deaths and over 50 million injuries each year (WHO, 2004).

Identifying the main factors that are related to injury severity of traffic accidents, especially to fatalities, has been of the main interests to injury severity analysts. Figure 1 shows that the number of studies about injury severity of traffic accidents has been increasing with time, with the maximum number in the last five years.

(insert figure 1)

Many techniques have been used to analyze the injury severity of traffic accidents. The methods that have been mostly used include ordered probit models, binary logit models, ordered logit models, and hierarchical logit models. However, in recent years other types of models have appeared: Artificial Neural Networks, Bayesian Networks, Trees, and Genetic Programming.

The knowledge of the advantages and disadvantages of each method would help safety analysts on deciding the most appropriate method for each particular analysis. The scope of this paper is to provide an insight on each one of the methods already used to analyze injury severity of traffic accidents where at least a motorized 4-wheeled vehicle is involved, excluding traffic accidents' studies that analyze injury severity from a medical point of view, or those that discuss the vehicle design and equipments and their relation with the injury outcome and those studies that analyze accidents in urban areas only.

The analysis of the different models is performed based on the following aspects: type of modeling technique, number of crashes considered in the analysis, number of variables for analyzing the severity, area type of the road (urban, suburban or rural), features considered in the analysis (basic segment and/or intersection), type and number of categories for injury levels, and model fit.

This paper is organized as follows: section 2 describes briefly the techniques used in the literature to analyze and/or model injury severity; a discussion of the studies found in literature is presented in section 3, where summary and conclusions are given in section 4.

2. Modeling techniques

In this section the modeling techniques used to analyze injury severity of traffic accidents were classified into 4 groups: discrete outcome models, data mining techniques, soft computing techniques and other techniques.

2.1. Discrete Outcome Models (DOM)

DOM are used to represent probabilities of having an outcome based on certain factors or characteristics. In general, these models cannot be calibrated using standard curve-fitting techniques, such as least squares, because their dependent variable is an un-observed probability (between 0 and 1) and the observations are the individual outcomes (either 0 or 1) (Ortúzar and Willumsen, 2001).

2.1.1. Logit Models (LM)

A special case of general linear regression is logistic regression or logit model, which assumes that the response variable follows the logit-function. Logistic model is an approach that is used in order to describe the relationship of single or several independent variables to a binary outcome variable. This modeling approach is usually preferred by researchers, since the logistic function must lie in the range between zero and one, and this is not usually the case with other possible functions (Kleinbaum and Klein, 2002).

The simplest form of the LM is the binary form, where the outcome variable is one of two outcomes. Binary logit model (BLM) and other extensions of it which were found to be mostly used in the literature of accidents' injury severity analysis.

A brief description of these extensions is the following (Ortúzar and Willumsen, 2001; Kleinbaum and Klein, 2002; Train, 2009; and Keele and Park, 2006; Washington et al., 2011):

1. Multinomial logit model (MNL) is used when the outcome variable has more than two unordered categories.
2. Hierarchical logit, nested logit or multi-level logit model (HL) is used when certain assumptions valid for the MNL are violated, e.g., when either the outcome are not independent, or when there are variations among individuals.
3. Mixed logit model (MXL) is a generalized extreme value (GEV) model where, this distribution allows for correlations over outcomes and it is a generalization of the univariate extreme value distribution that is used for standard LM. This model alleviates the three limitations for the standard LM by allowing for random variation, unrestricted substitution patterns and correlation in unobserved factors over time. MXL are actually the integrals of the standard logit probabilities over density parameters.
4. Ordered logit model (OLM), also known as proportional odds model, this model has an observed ordinal variable (Y), where Y is a function of another latent continuous unmeasured variable (Y*). Values of Y* determine the values of the resulting Y. Y* has various threshold or cut points, where the value of Y depends on these thresholds. The random disturbance or the error term here follows a logistic distribution (Washington et al., 2011)
5. Heteroskedastic logit model (HKL) is also a GEV model; however, instead of capturing correlations among outcomes, it allows the variance of unobserved factors to differ over outcomes.
6. Heterogeneous model (HM) is used when dealing with categorical dependent variables. If the variants of the error term are non-constant, the standard error will be incorrect and the parameters will be biased and inconsistent. In order to deal with unequal error variances the HM is used.
7. Generalized estimating equations (GEE) are an extension of the logistic model to handle outcome variables that have binary correlated outcomes. GEE takes into account the correlated nature of the outcome.

25 studies of accident injury severity used one or more of the previously mentioned LM. These studies are listed in detail in Table 1.

(insert table 1)

2.1.2. Probit models (PM)

PM deal with the three limitations of LM: they can handle random variation, they allow any pattern of substitution, and they are applicable to panel data with temporally correlated errors (Train, 2009).

The most used type of PM in the analysis of accidents' severity is the ordered probit models (OPM). OPM is a generalization of the PM to the case of more than two outcomes of an ordinal dependent variable. In the case the model cannot be estimated using the ordinary least square, it is usually estimated using the maximum likelihood (Train, 2009).

Other PM used are the following:

- Heteroskedastic probit model (HOP): it is used when the error terms are not homoskedastic and their variance may be parametrized as a function of covariates. HOP offers more flexibility than OPM, since they capture the effect of the independent variables on the variance or uncertainty in the outcome (Lemp, et al., 2011).
- Bayesian ordered probit model (BOP): it is an extension of the Bayesian inference into the OPM, in which the parameters to be estimated are assumed to follow certain prior distributions. Based on the data, the likelihood function is used to update the prior distribution and obtain the posterior distribution (Xie et al., 2009).

14 studies of accident injury severity used one or more of the previously mentioned PM. These studies are listed in detail in Table 2.

(insert table 2)

(insert table 3)

2.2. Data Mining Techniques

Data mining is defined as the process of discovering patterns in data. The patterns discovered must be meaningful in that they lead to some advantages (Witten and Frank, 2005). Many data mining techniques are being in use in different fields of science, economy, engineering, etc. Decision trees and Bayesian networks have been also used to analyze the injury severity of traffic accidents.

2.2.1. Decision trees

Decision trees are nonlinear predictive models that use the tree to represent the recursive partition. Within the literature of the accident injury severity studies, two types have been used (see Table 3):

1. Classification and regression trees (CART): it constructs binary trees, in which each internal node has exactly two outgoing edges. CART can consider misclassification costs in the tree induction. It also enables users to provide prior probability distribution. An important feature of CART is its ability to generate regression trees, where the leaf predicts a real number and not a class (Rokach and Maimon, 2008).
2. Chi squared automatic interaction detection (CHAID): it is a procedure used to generate decision trees. For each input variable, CHAID finds the pair of values that is least significantly different with respect to the target variable (Rokach and Maimon, 2008).

2.2.2. Bayesian networks (BNs)

BNs are graphical models of interactions among a set of variables, where the variables are represented as nodes of a graph and the interactions as directed links between the nodes. Any pair of unconnected/nonadjacent nodes of such a graph indicates (conditional) independence between the variables represented by these nodes under particular circumstances (Mittal and Kassim, 2007).

Two studies were found to use BNs to analyze injury severity of accidents (see Table 3).

2.3. Soft computing techniques

Soft computing is a mix of distinct methods which in a way or another cooperate in their fundamentals. The principal objective of soft computing is to exploit the tolerance for imprecision and uncertainty in order to achieve manageability, robustness and solutions at low cost (Zadeh, 1994).

2.3.1. Artificial Neural Networks (ANN)

A neural network (NN) is an interconnected assembly of simple processing elements, units or nodes. The processing ability of the network is stored in the inter-unit connection strengths, or weights, obtained by a process of adaptation to, or learning from, a set of training patterns (Gurney, 1997). NN are composed of neurons which in turn are composed of a number of inputs, and each input comes with a connection that has a weight and a threshold value.

A number of ANN types have been used by the researchers of accident severity analysis (see Table 3):

1. Multi-layer perceptron ANN (MLP): they usually consist of three layers: input layer, hidden layer, and output layer. The connections in MLP are feed-forward type in which they are allowed from a certain index to layers of a higher index. In order to train MLP, the back propagation algorithm is used (Rumelhart, et al., 1986).
2. Fuzzy ART MAP ANN (ARTMAP): they are based on adaptive resonance theory. It is a clustering algorithm that maps a set of input vectors to a set of clusters. Models built by fuzzy ARTMAP have a fast, stable learning in response to binary input patterns (Carpenter et al., 1992).

2.3.2. Evolutionary algorithms (EA)

EA mimic the natural evolution in order to optimize a solution to a problem (Brameier and Banzhaf, 2007). These algorithms exploit differential fitness advantages in a population of solutions to gradually improve the state of that population.

Genetic programming (GP) is defined as any direct evolution or breeding of computer programs for the purpose of inductive learning. Unlike other EAs, GPs can complete missing parts of existing model.

Linear genetic programming (LGP) is a GP variant that evolves sequences of instructions from an imperative programming language or from a machine language. Linear refers to the structure of the imperative program representation, where the nodes do not have to be linearly listed nor the method itself needs to be linear.

Das and Abdel-Aty (2010) used LGP to classify injury severity according to the accident type in order to find the geometric and environmental factors that are related to this classification (see Table 3).

2.4. Generalized linear models (GLM)

The log-linear model (LLM) is one of the specialized cases of GLM for Poisson-distributed data. In log-linear models, there is no distinction between independent and dependent variables, all variables are treated as response variables. Chen and Jovanis (2000) used LLM to identify significant variables that contribute to the occurrence of a specific injury severity for bus drivers (see Table 3).

3. Discussion

Figure 1 shows the different models which have been used in the field of traffic accidents' injury severity analysis. The most used are the logit and probit, followed by other models. However, Tables 1, 2 and 3 show that there is a large dispersion in the principal magnitudes used for the models.

(insert Table 4)

It should be taken into account that the following sections are not intended to make a comparison amongst different modeling techniques or different studies, however, the sections are presented as a guidance for analysts, since a comparison is not possible due to the differences that exist in data sources, study objectives, and certain conditions that might apply to a study while not for others.

3.1. Number of records considered in the analysis

The number of records considered in the analysis ranges between 255 and 622,432. However, four extreme outliers were identified (Montgomery and Runger, 2003). Table 4 shows the statistical analysis results without extreme outliers.

Table 4 shows that the number of records (without extreme outliers) for all the studies ranges between 255 and 81,172, with a median value of 3,955. PM present the lowest median, with 3,136 crashes, followed by LM with 4,552 records. The other models present the highest median (4,713) for the number of records considered in the analysis.

All the values are very similar and no significant statistical difference was observed between the three groups (logit, probit and others) based on the Mann-Whitney U test.

3.2. Number of variables for analyzing severity

The number of variables used in the modeling of injury severity range between 5 and 58. However, one extreme outlier was identified, which was not considered for the statistical analysis.

Table 4 shows that the number of variables (without extreme outliers) for all the studies ranges between 5 and 36, with a median value of 15. PM present the highest median, with 16 variables, followed by LM with 15 variables. The other models present the lowest median (14) for the number of variables used for analyzing severity.

All the values are very similar and no significant statistical difference was observed between the three groups (logit, probit and others) based on the Mann-Whitney U test.

No correlation was found to exist between the number of variables for analyzing severity and the number of records considered in the analysis.

3.3. Focus of the study

There is a relatively high dispersion in the type of analyzed roadway segment. Figure 2 shows that 14 studies analyzed only basic segments, 9 studies analyzed only intersections, and 13 studies analyzed both intersections and roadways segments in the same study. However, Moore et al. (2010) recommend that intersections and road segments should not be analyzed together, since the factors related to accidents occurring on intersections are different from those occurring on roadway segments.

(insert figure 2)

Regarding the area type in which the roadway or the intersection exists. Only six studies analyze rural areas, while 22 studies mixed the data for rural, urban and/or suburban highways, keeping in mind that the characteristics of the roadways and intersections differ significantly between urban and rural areas.

3.4. Type and number of categories for injury levels

The definition of the injury severity might refer to the emphasis of the study (Krull et al., 2000), for convenience (Ouyang et al., 2002), or because of the small counts of certain category with respect to other categories (Peek-Asa et al., 2010).

Most of the studies used the KABCO scale, which is the scale used in police observed accident records (Morgan, 2009). Others combined one or more categories into one.

Table 4 shows that the number of injury levels for all the studies ranges between 1 and 7, with a median value of 3. In this case no extreme outliers were identified. PM present the highest median, with 5 levels, followed by LM with 3 levels. The other models present the lowest median (2) for the number of injury levels.

In this case, significant statistical differences were observed ($p<0.05$), based on the Mann-Whitney U test, between the number of injury levels considered in the probit models with respect to the other two types of models (logit and others).

3.5. Model fit

In general, the statistical tests used to validate the performance of the model vary with the study. However, the number of different tests and their characteristics does not permit comparing the results of one study with another.

In the following, descriptions of the fit parameters used for the analyzed models are presented:

3.5.1. R-squared

R-squared (R^2) is a statistic that is generated in ordinary least squares (OLS) regression that is often used as a goodness-of-fit measure. Its value ranges between 0 and 1, where 1 indicates a high level of explanation of the variance as explained by the regression model and zero as a low level of explanation (Bruin, 2006).

Chen and Jovanis (2000) used R^2 to test LLM fit (see Table 3). The results indicated that the log-linear model fitted the data very well (R^2=0.95).

3.5.2. Pseudo R-square

When analyzing data with a logistic regression, an equivalent statistic to R-squared does not exist. The model estimates from a logistic regression are maximum likelihood estimates arrived at through an iterative process. However, to evaluate the goodness-of-fit of logistic models, several pseudo R-squareds (ρ^2) have been developed (McFadden, adjusted McFadden, Efron's, Cox & Snell, Negelkerke/ Cragg & Uhler's, McKelvey & Zavoina, Count, adjusted count, etc.). They look like R^2 in the sense that they are on a similar scale, ranging from 0 to 1 (though some ρ^2 never achieve 0 or 1) with higher values indicating better model fit (Bruin, 2006).

In this paper (see Table 1, 2 and 3) several studies used a ρ^2 to test the fit of their models. Others studies supplied the log-likelihood (LL) of the model. Thus, when information about LL was available, the McFadden ρ^2 was calculated in order to homogenize the model fit results.

However, Bruin (2006) indicated that ρ^2 cannot be interpreted independently or compared across datasets; a ρ^2 only has meaning when compared to another ρ^2 of the same type, on the same data, predicting the same outcome. Thus, it is only possible to indicate if the results of a study are among the satisfactory range for such parameter for model fit. As indicated by McFadden (1979), the satisfactory range for the McFadden's ρ^2 lies between 0.2 and 0.4.

Most of the models in this paper that use McFadden's ρ^2 present values below 0.2. However, there are five studies that present values over 0.2: Shanker et al. (1996) (ρ^2=0.52), Wang and Kockelman (2005) (ρ^2=0.235-0.257), Abdel-Aty and Keller (2005) (ρ^2=0.24), Oh (2006) (ρ^2=0.378-0.480) and Schneider et al. (2009) (ρ^2= 0.23-0.258).

3.5.3. Accuracy

Accuracy measures the percentage of cases in the accident data correctly predicted by the model. Therefore, accuracy is obtained at the case-specific level, that is, cases that are correctly classified as fatal or nonfatal according to their observed injury experience (Saccomanno et al., 1996).

Most of the studies used this parameter to test the capability of their models to correctly classify the injury severity into a specific level (see Table 1, 2 and 3). Global accuracy range lies between 0.41 and 0.89. The highest global accuracy achieved was for a BLM model built by Dissanayake and Lu (2002) and the lowest global accuracy was obtained by Abdel-Aty and Abdelwahab (2004) when they constructed a Fuzzy ARTMAP ANN model.

The results presented by Dissanayake and Lu (2002) indicates that the number of accidents classified under each severity level was homogeneous along all the levels. On the other hand, the lowest accuracy obtained for a specific level (fatal accidents) with a CART model was practically zero (Chang and Wang, 2006). The authors referred this result to the fact that their dataset was imbalanced, where the fatal accidents accounted only for about 0.4% of the whole sample used to build the model.

Delen et al. (2006) also obtained relatively low accuracy results (40.7%) for their model (MLP ANN). They explained their results by a multi-class classification problem. A possible solution would be reducing the multi-class problem into series of binary classification problems. Applying such a solution, the complete dataset was separated into eight subsets with binary output variables. In which a top-down (more serious injury versus the less serious injuries) and a bottom-up (less serious injury versus more serious injuries) were built. The method applied by Delen et al. (2006) could be compared to that applied by Dissanayake and Lu (2002) where, after developing two formats of binary logistic models (top-down and a bottom-up format with four models in each), they found that the method of selecting the models' format did not drastically affect model reliability, however, they chose to use the top-down format in their analysis since it achieved better model accuracies (73.4-98.0%).

3.5.4. Other Measures

Other measures used to test the model fit are: Akaike Information Criterion (AIC), log-likelihood (LL), chi-squared (χ^2), and Kendall rank correlation coefficient (Kendall's tau (τ) coefficient).

The likelihood is the probability of the data given the parameter estimates. The goal of a model is to find values for the parameters (coefficients) that maximize value of the likelihood function. Many procedures use the log-likelihood, because it is easier to work with (Bruin, 2006). Two studies used the LL as the fit test. However, only Lemp et al. (2011) used it to compare two models (OPM against HOP).

The Akaike Information Criterion (AIC) is used only once (Haleem and Abdel-Aty, 2010) to select a model from a set of models. The chosen model is the one that minimizes the Kullback-Leibler distance between the model and the truth (low AIC and high LL indicates good fit). AIC is a criterion that seeks a model that has a good fit to the truth with few parameters (Burham and Anderson, 2002).

χ^2 test is used to verify if a sample of data came from a population with a specific distribution. χ^2 is applied to data put into classes. However, the value of χ^2 statistic is dependent on how the data is classified. Another disadvantage of χ^2 is that it requires a sufficient sample size in order for χ^2 approximation to be valid (NIST/SEMATECH, 2003). Three studies (Srinivasan, 2002; Malyshkina and Mannering, 2009; and Daniels et al., 2010) were found to use χ^2 as the model goodness of fit test.

Only Donelson, et al. (1999) used Kendall's tau (τ) coefficient, which is a statistic used to measure the association between two quantities. A τ test is a non-parametric hypothesis test which uses the coefficient to test for statistical dependence Kruskal (1958). The results indicated that the BLM fitted the data used.

3.6. Modeling techniques

The usage of the modeling techniques varied with time, thus DOM continued to be dominant. Since 2001 new methods started to be applied in the analysis of injury severity, such as Neural Networks, Bayesian Networks, and most recently Genetic Algorithms.

Table 5 shows the frequency of usage of each model. In this paper 19 modeling techniques used to model injury severity of traffic accidents, applied in 58 case-studies, have been analyzed. The most used techniques are the DOM (46 cases), highlighting the models BLM, OLM and OPM over all the others. These three models were used in more than 54% of the cases.

(insert table 5)

3.6.1. Logit Models

Table 5 shows that the most used logit models are BLM followed by OLM, while the least used were MXL, GEE and OMXL.

The frequent usage of BLM to analyze accident severity might refer to the fact that most of the studies used the outcome variable as binary (Dissanayake and Lu, 2002; Delen et al., 2006; Jung et al., 2010). This refers to the fact that BLM is easily interpretable.

An OLM restriction is that regression parameters have to be the same for different accident severity levels, called proportional odds. However, it is not always clear if the distance between accident severity levels is equal, and hence it is arbitrary to assume that all coefficients of ordered probability models are the same (Jung et al., 2010).

Moreover, Srinivasan (2002) stated that the primary restriction of the ordered models comes from the assumption of deterministic thresholds that are often identical across all observations for each ordinal outcome level. Also, it is assumed that the outcome is homogeneous and independent from exogenous variables. In addition, these models disregard possible correlations across the thresholds of different outcomes.

Consequently, these assumptions could lead to significant bias and inconsistency in ordered outcome models. Therefore, Srinivasan (2002) used an OMXL, where she compared OMXL to OLM using a χ^2 test. The results indicated that the χ^2 test rejected the restrictive OLM.

Lenuguerrand et al. (2006) compared different models (BLM, HL and GEE). HL was found to be more adequate for problems with correlated data than BLM and GEE, clusters and sub-clusters, since BLM and GEE both underestimate parameters and confidence intervals. Thus, they recommended the use of HL when the number of vehicles per accident or the number of occupants per accident is high.

3.6.2. Probit Models

The most frequently used probit model is the OPM (see Table 5). OPM have been used to model injury severity of accident on roadways and intersections. Some researchers have used models that combined the accident occurring at intersections with accident off intersections (Gray et al., 2008; Xie et al., 2009; Zhu and Srinivasan, 2011).

OPM proved to be a good choice for modeling injury severity of accidents even when compared with other models such as the BOP, the OPM still performed as well (Xie et al., 2009). BOP and OPM produced similar results for large data size, where they recommended using BOP for smaller data sizes, as it can produce more reasonable parameter estimation and better prediction performance. On the contrary, when comparing OPM with HOP, the HOP was preferred over the OPM in terms of log-likelihoods (Lemp et al., 2011).

Haleem and Abdel-Aty (2010) used OPM, PM and HL to analyze accident injury severity at intersections. The results indicated that the PM fits the data better than the OPM.

3.6.3. Other modeling techniques

CART were used by Council and Stewart (1996) and Chang and Wang (2006) to model injury severity of accident. The results presented by Chang and Wang (2006) indicated that CART can effectively handle multi-collinearity problems, and they could handle the outliers that exist in the data by isolating them into a node.

However Chang and Wang (2006) indicated that one of the problems with applying the CART is that they do not provide confidence intervals for the risk factors (splitters) and predictions. Also, they have a difficulty in applying the sensitivity analysis which does not permit examining the marginal effects of the predictors on the response variable. In addition the CART models are unstable, the structure and the accuracy alter if different strategies are followed to create learning and test sets.

BNs were used by Simoncic (2004) and de Oña et al. (2011) to model injury severity of accident. The work presented by Simoncic (2004) based the conclusion upon one single network, which was not validated using a test set. de Oña et al. (2011) built several BNs to model injury severity of accidents. These networks were compared to each others in terms of MPE, complexity, accuracy, sensitivity, specificity, ROC area, and HMSS. Each of the networks was validated using a test set.

MLP and Fuzzy ARTMAP ANNs have been compared twice to analyze injury severity on road segment and on intersections (Abdelwahab and Abdel-Aty, 2001; Abdel-Aty and Abdelwahab, 2004). Both studies indicated that MLP ANN performance is superior to Fuzzy ARTMAP ANN. Delen et al. (2006) also used MLP ANN to model injury severity on roadways. They used more injury levels and their results (in terms of accuracy) were worse than those of previous studies (Abdelwahab and Abdel-Aty, 2001; and Abdel-Aty and Abdelwahab, 2004).

Abdelwahab and Abdel-Aty (2001) compared the performance of MLP ANN and fuzzy ARTMAP ANN with the performance of OLM. Their results indicated that the best in terms of accuracy was the MLP ANN followed by OLM, and finally by fuzzy ARTMAP. Thus, OLM was superior in performance with respect to certain types of ANNs. Abdel-Aty and Abdelwahab (2004) used MLP ANN, fuzzy ARTMAP and OPM. The results showed once again the superiority of MLP ANN over all the other techniques, however, this time OPM did not perform better than the fuzzy ARTMAP.

4. Summary and conclusions

The review of several studies on models used in the modeling of traffic accidents injury severity indicates that each method has its advantages and disadvantages.

Many modeling techniques have been in use to analyze the injury severity of accidents. The most used models are the logit and probit. However, in recent years, methods based on data mining techniques, as well as other models based on soft computing techniques, have appeared.

Within the discrete outcome models, the most used are OPM, BLM and OLM. BLM are commonly used when the study uses a binary variable for severity. When the severity is ordered (killed, severe injury, slight injury, possible injury, or property damage only) OPM and OLM are commonly used.

There is a large diversity in the number of accidents' records and the number of variables used. However, no significant statistical difference was found between logit, probit and other models. The number of records and the number of variables are found to be mostly dependent upon the availability of data.

Most of the studies use the KABCO scale or a modification. Based on the studies analyzed, the probit models use a higher number of injury levels (5) than the logit models (3 levels) or the rest of the models (2 levels). In this case, significant statistical differences were observed ($p<0.05$) between the probit models and the other types of models.

The model fit results are satisfactory in most of the cases (e.g. global accuracy in the range of 0.41 and 0.89; McFadden's pseudo R-square values between 0.2 and 0.4), although some exceptional results can be observed (e.g. Chen and Jovanis (2000) obtained $R^2=0.95$), while others were not so satisfactory (e.g. many studies with McFadden's pseudo R-square below 0.2).

Different factors affect the accuracy obtained by data mining and soft computing models, such as the balance of cases among the different categories that lie under the injury severity levels. If the number of observed cases classified among the different levels is not relatively different, this identifies a balanced dataset; and accuracy would improve since the classification will not be biased towards a specific injury severity level.

In general, it is not possible to identify which is the best method to be used. Using a certain model might be suitable under certain circumstances, while not under others. Many examples are available in the literature (Lenuguerrand et. al., 2006; Xie et al., 2009). Probably this is one of the main reasons why, in recent years, the number of studies that analyze injury severity of traffic accidents has greatly increased. Documentation of characteristics and limitations of each modeling technique will help analysts to decide the best method to be used in each particular modeling problem.

References

Abdel-Aty, M. & Abdelwahab, H. T. (2004). Predicting injury severity levels in traffic crashes: A modeling comparison. *Journal of Transportation Engineering* 130, No. 2, 204-210.

Abdel-Aty, M. & Keller, J. (2005). Exploring the overall and specific crash severity levels at signalized intersections. *Accident Analysis and Prevention* 37, No. 3, 417-425.

Abdelwahab, H. & Abdel-Aty, M. (2001). Development of Artificial Neural Network Models to Predict Driver Injury Severity in Traffic Accidents at Signalized Intersections. *Transportation Research Record* 1746, No. 1, 6-13.

Awadzi, K.D., Classen, S., Hall, A., Duncan, R.P. & Garvan, C.W. (2008). Predictors of injury among younger and older adults in fatal motor vehicle crashes. *Accident Analysis and Prevention* 40, No. 6, 1804–1810.

Bédard, M., Guyatt, G.H., Stones, M.J. & Hirdes, J.P. (2002). The independent contribution of driver, crash, and vehicle characteristics to driver fatalities. *Accident Analysis and Prevention* 34, No. 6, 717-727.

Brameier, M. & Banzhaf, W. *Linear Genetic Programming: Genetic and Evolutionary Computation Series*. Springer, New York, 2007.

Bruin, J. newtest: command to compute new test, 2006. See www.ats.ucla.edu/stat/stata/ado/analysis/ for further details. Accessed 01/03/ 2011.

Burham, K.P. & Anderson, D.R. *Model selection and multimodel inference: A practical information –theoretic approach*. Springer, New York, 2002.

Carpenter, G. A., Grossberg, S., Markuzon, N., Reynolds, J.H. & Rosen, D. B. (1992). Fuzzy ARTMAP: A neural-network architecture for incremental supervised learning of analog multidimensional maps. *IEEE Transactions on Neural Networks* 3, No. 5, 698–713.

Chang, L.-Y. & Wang, H.-W. (2006). Analysis of traffic injury severity: an application of non-parametric classification tree techniques. *Accident Analysis and Prevention* 38, No. 5, 1019-1027.

Chen, W.-H. & Jovanis, P.P. (2000). Method for identifying factors contributing to driver-injury severity in traffic crashes. *Transportation Research Record* 1717, No. 1, 1-9.

Council, F. & Stewart, J. (1996). Severity Indexes for Roadside Objects. *Transportation Research Record* 1528, No. 1, 87-96.

Daniels, S., Brijs, T., Nuyts, E. & Wets, G. (2010). Externality of risk and crash severity at roundabouts. *Accident Analysis and Prevention* 42, No. 6, 1966-1973.

Das, A. & Abdel-Aty, M. (2010). A genetic programming approach to explore the crash severity on multi-lane roads. *Accident Analysis and Prevention* 42, No. 2, 548-557.

de Oña, J., Mujalli, R.O., & Calvo, F.J. (2011). Analysis of traffic accident injury severity on Spanish rural highways using Bayesian networks. *Accident Analysis and Prevention* 43, No. 1, 402-411.

Delen, D., Sharda, R. & Bessonov, M. (2006). Identifying significant predictors of injury severity in traffic accidents using a series of artificial neural networks. *Accident Analysis and Prevention* 38, No. 3, 434-444.

Dissanayake, S. (2004). Comparison of severity affecting factors between young and older drivers in single vehicle crashes. *IATSS Research* 28, No. 2, 48-54.

Dissanayake, S. & Lu, J. (2002). Analysis of severity of young driver crashes: Sequential binary logistic regression modeling. *Transportation Research Record* 1784, No. 1, 108-114.

Donelson, A., Ramachandran, K., Zhao, K. & Kalinowski, A. (1999). Rates of occupant deaths in vehicle rollover: Importance of fatality-risk factors. *Transportation Research Record* 1665, No. 1, 109-117.

Dupont, E., Martensen, H., Papadimitriou, E. & Yannis, G. (2010). Risk and protection factors in fatal accidents. *Accident Analysis and Prevention* 42, No. 2, 645-653.

Gårder, P. (2006). Segment characteristics and severity of head-on crashes on two-lane rural highways in Maine. *Accident Analysis and Prevention* 38, No. 4, 652-661.

Gray, R.C., Quddus, M.A. & Evans, A. (2008). Injury severity analysis of accidents involving young male drivers in Great Britain. *Journal of Safety Research* 39, No. 5, 483-495.

Gurney, K. *An introduction to neural networks*. UCL Press, London, 1997.

Haleem, K. & Abdel-Aty, M. (2010). Examining traffic crash injury severity at unsignalized intersections. *Journal of Safety Research* 41, No. 4, 347-357.

Jin, Y., Wang, X. & Chen, X. Right-angle crash injury severity analysis using ordered probability models. *Proceedings 2010 International Conference on Intelligent Computation Technology and Automation ICICTA 2010*, 2010.

Jung, S., Qin, X. & Noyce, D.A. (2010). Rainfall effect on single-vehicle crash severities using polychotomous response models. *Accident Analysis and Prevention* 42, No. 1, 213-224.

Keele, L. & Park, D.K. Difficult choices: An evaluation of heterogeneous choice models. *Proceedings 2004 Meeting of the American Political Science Association*, 2006.

Khattak, A.J. (2001). Injury severity in multivehicle rear-end crashes. *Transportation Research Record* 1746, No. 1, 59-68.

Khattak, A.J., Pawlovich, M.D., Souleyrette, R.R. & Hallmark, S.L. (2002). Factors related to more severe older driver traffic crash injuries. *Journal of Transportation Engineering* 128, No. 3, 243-249.

Khattak, A. & Rocha, M. (2003). Are SUVs "Supremely unsafe vehicles"? Analysis of rollovers and injuries with sport utility vehicles. *Transportation Research Record* 1840, No. 1, 167-177.

Kleinbaum, D.G. & Klein, M. *Logistic regression: A self-learning text.* Springer, New York, 2002.

Kockelman, K.M. & Kweon, Y.-J. (2002). Driver injury severity: an application of ordered probit models. *Accident Analysis and Prevention* 34, No. 3, 313-321.

Kononen, D.W., Flannagan, C.A.C. & Wang, S.C. (2011). Identification and validation of a logistic regression model for predicting serious injuries associated with motor vehicle crashes. *Accident Analysis and Prevention* 43, No. 1, 112-122.

Krull, K., Khattak, A.J. & Council, F.M. (2000). Injury effects of rollovers and events sequence in single-vehicle crashes. *Transportation Research Record* 1717, No. 1, 46-54.

Kruskal, W.H. (1958). Ordinal measures of association. *Journal of the American Statistical Association* 53, No. 284, 814–861.

Lemp, J.D., Kockelman, K.M. & Unnikrishnan, A. (2011). Analysis of large truck crash severity using heteroskedastic ordered probit models. *Accident Analysis and Prevention* 43, No. 1, 370-380.

Lenguerrand, E., Martin, J. L. & Laumon, B. (2006). Modelling the hierarchical structure of road crash data--application to severity analysis. *Accident Analysis and Prevention* 38, No. 1, 43-53.

Malyshkina, N.V. & Mannering, F.L. (2009). Markov switching multinomial logit model: An application to accident-injury severities. *Accident Analysis and Prevention* 41, No. 4, 829-838.

McFadden, D. Quantitaive methods for analyzing travel behaviour of individuals: some recent developments. In *Behavioral travel modeling* (Hensher, D.A., Stopher, P.R.: (eds.)). Croom Helm, London, 1979.

Milton, J.C., Shankar, V.N. & Mannering, F.L. (2008). Highway accident severities and the mixed logit model: An exploratory empirical analysis. *Accident Analysis and Prevention* 40, No. 1, 260-266.

Mittal, A. & Kassim, A. *Bayesian network technologies: applications and graphical models*. IGI publishing, Hershey, 2007.

Montgomery, D.C. & Runger, G.C. *Applied statistics and probability for engineers*. Whiley, New York, 2003.

Moore, D.N., Schneider IV, W.H., Savolainenb, P.T. & Farzaneh, M. (2010). Mixed logit analysis of bicyclist injury severity resulting from motor vehicle crashes at intersection and non-intersection locations. *Accident Analysis and Prevention*, *Accident Analysis and Prevention* 43, No. 3, 621-630.

Morgan, A. What factors affect crash injury severity under specific weather conditions?. 2009. See www.engineering.purdue.edu/ITE/research/seminarfiles09-10/studentpresentation abbymorgan.pdf. Accessed 08/01/2011.

NIST/SEMATECH. e-Handbook of statistical methods. 2003. See www.itl.nist.gov/div898/ handbook/. Accessed 25/03/ 2011.

Oh, J.T. (2006). Development of severity models for vehicle accident injuries for signalized intersection in rural areas. *KSCE Journal of Civil Engineering* 10, No. 3, 219-225.

Ortúzar, J.D.D. & Willumsen, L.G. *Modelling transport*. Wiley, West Sussex, 2001.

Ouyang, Y., Shankar, V. & Yamamoto, T. (2002). Modeling the Simultaneity in Injury Causation in Multivehicle Collisions. *Transportation Research Record* 1784, No. 1, 143-152.

Paleti, R., Eluru, N. & Bhat, C.R. (2010). Examining the influence of aggressive driving behavior on driver injury severity in traffic crashes. *Accident Analysis and Prevention* 42, No. 6, 1839–1854.

Peek-Asa, C., Britton, C., Young, T., Pawlovich, M. & Falb, S. (2010). Teenage driver crash incidence and factors influencing crash injury by rurality. *Journal of Safety Research* 41, No. 6, 487–492.

Quddus, M.A., Wang, C. & Ison, S.G. (2010). Road traffic congestion and crash severity: Econometric analysis using ordered response models. *Journal of Transportation Engineering* 136, No. 5, 424-435.

Renski, H., Khattak, A.J. & Council, F.M. (1999). Effect of speed limit increase on crash injury severity: analysis of single-vehicle crashes on north Carolina Interstate highways. *Transportation Research Record* 1665, No. 1, 100-108.

Rockach, L. & Maimon, O. *Data mining with decision trees theory and applications*. World Scientific Publishing, Singapore, 2008.

Rumelhart, D. E.,Hinton, G. E. & Williams, R. J. *Learning internal representation by error propagation, parallel distributed processing: explorations in the microstructure of cognition*. MIT Press, Cambridge, 1986.

Saccomanno, F.F., Nassar, S.A. & Shortreed, J.H. (1996). Reliability of statistical road accident Injury severity models. *Transportation Research Record* 1542, No. 1, 14-23.

Schneider, W.H., Savolainen, P.T. & Zimmerman, K. (2009). Driver injury severity resulting from single-vehicle crashes along horizontal curves on rural two-lane highways. *Transportation Research Record* 2012, No. 1, 85-92.

Shanker, V., Mannering, F. & Barfield, W. (1996). Statistical analysis of accidents severity on rural freeways. *Accident Analysis and Prevention* 28, No. 3, 391-401.

Simoncic M. (2004). A Bayesian network model of two-car accidents. *Journal of Transportation and Statistics* 7, No. 2-3, 13-25.

Srinivasan, K.K. (2002). Injury severity analysis with variable and correlated thresholds: ordered mixed logit formulation. *Transportation Research Record* 1784, No. 1, 132-142.

Train, K. *Discrete choice methods with simulation*. Cambridge University Press, Cambridge, 2009.

Wang, X. & Kockelman, K.M. (2005). Use of Heteroscedastic Ordered Logit model to study severity of occupant Injury: Distinguishing effect of vehicle weight and type. *Transportation Research Record* 1908, No. 1, 195-204.

Wang, Z., Chen, H. & Lu, J.J. (2009). Exploring impacts of factors contributing to injury severity at freeway diverge areas. *Transportation Research Record* 2102, No. 1, 43-52.

Washington, S., Karlaftis, M. & Mannering, F. *Statistical and econometric methods for transportation data analysis*. (2nd ed.), Champan and Hall/CRC, Florida, 2011.

WHO. World health organization, world report on road traffic injury prevention Geneva. 2004. See www.whqlibdoc.who.int/publications/2004/9241562609.pdf. Accessed 10/01/2011.

Witten, I.H. & Frank, E. *Data mining: Practical machine learning tools and techniques*. Morgan Kaufmann Publishers, San Francisco, 2005.

Xie, Y., Zhang, Y. & Liang, F. (2009). Crash injury severity analysis using Bayesian ordered probit models. *Journal of Transportation Engineering* 135, No. 1, 18-25.

Zadeh, L. *Fuzzy logic technology and applications*. In (Marks II R.J.), IEEE Publications, 1994.

Zhu, X. & Srinivasan, S. (2011). A comprehensive analysis of factors influencing the injury severity of large-truck crashes. *Accident Analysis and Prevention* 43, No. 1, 49–57.

List of Figures:

List of Tables:

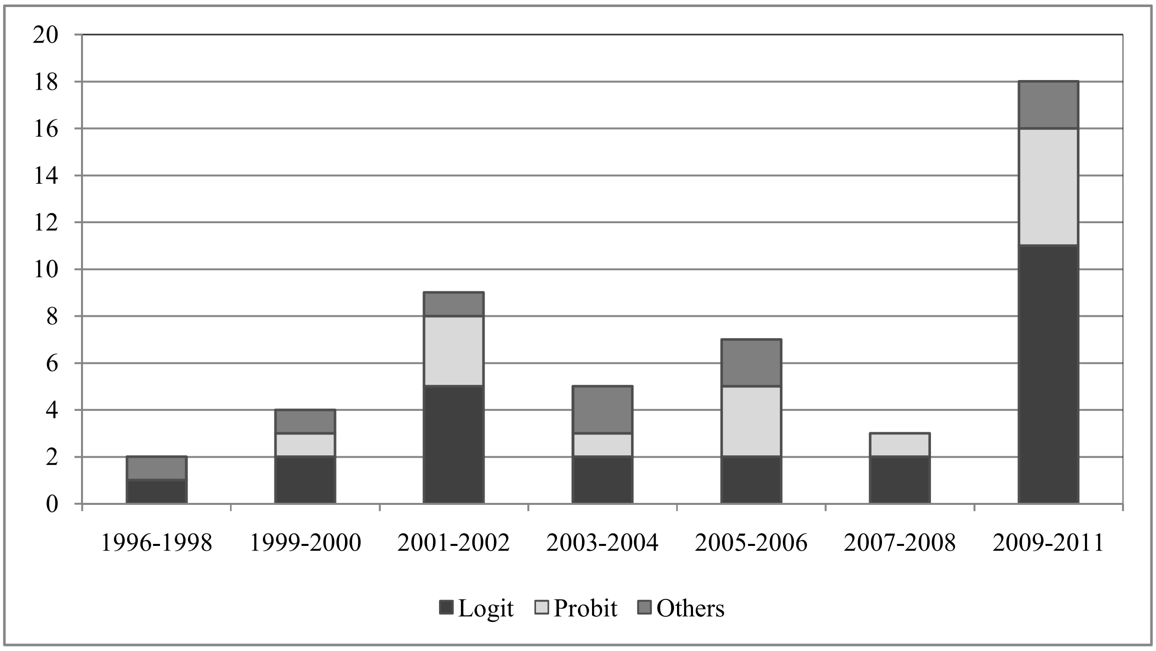


Figure 1: Case-studies by type of model analyzed from 1996 to 2011

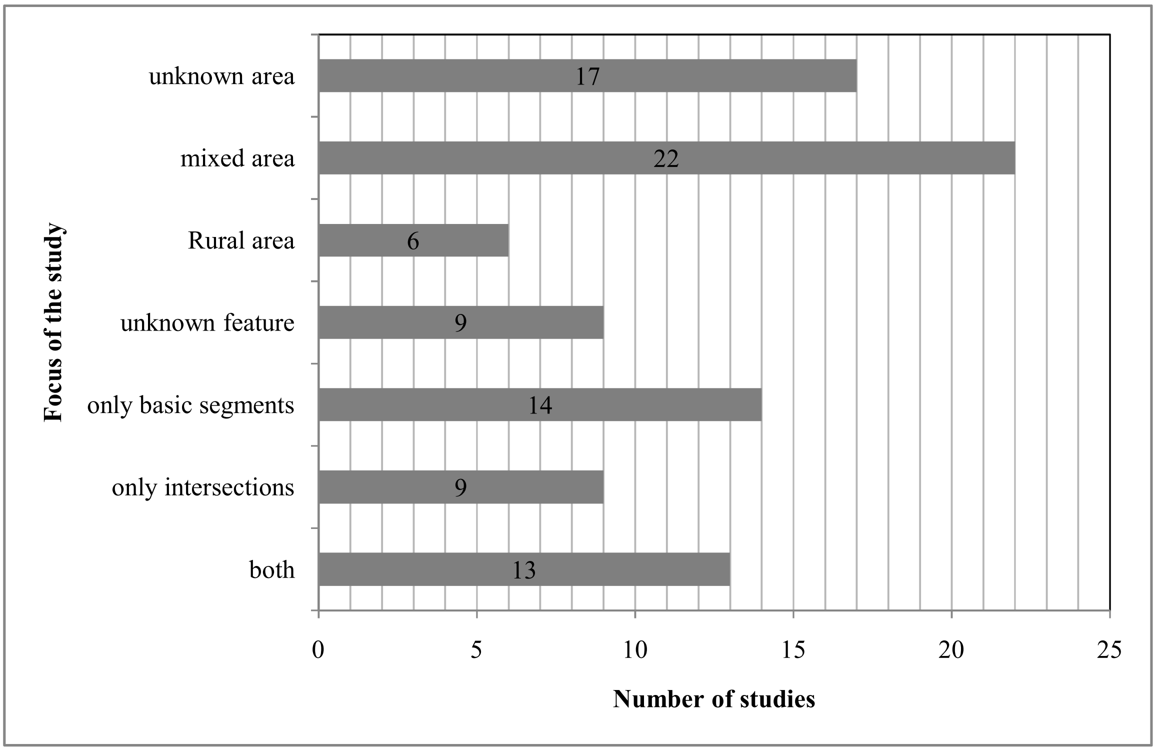


Figure 2: Case-studies according to the focus of the study

Table 1: Studies that analyze injury severity of traffic accidents using Logit Models

Study Authors (publication year)	Objectives of the study	Model type	No. records	No. variables	Area type	Features	Injury level	No. injury levels	Model fit test
Shanker et al., 1996	analyze severity on rural highways	HL	1505	21	Rural	SEG	KABCO	5	$\rho^2 = 0.52$ (Probably McFadden's)
Donelson et al., 1999	to predict fatality for occupants of light-duty trucks in single-vehicle rollover crashes	BLM	55000	11	mixed	SEG	K	1	Kendall's tau (τ) coefficient c statistic=[0.882-0.916] for all models , concordance) = [88.4%-85.5%] for all models, discordance =[6.4%-9.8%]for all models, tied pair of a fatal crash=[4%-5.6%] for all models
Krull et al., 2000	to explore the effect of rollover crashes for single vehicles	BLM	59743	16	mixed	SEG	K+A, B+C+O	2	$\rho^2 = 0.147$ (McFadden's)
Abdelwahab and Abdel-Aty, 2001	to analyze the injury severity of crashes of two-vehicles that occurred at signalized intersections	OLM	1168	14	n.a.	INT	A, C+B, O	3	accuracy =58.9%
Dissanayake and Lu, 2002	to analyze injury severity of young drivers for single vehicles fixed objects crashes	BLM	8382	16	mixed	SEG	KABCO	5	accuracy =89.2%
Bédard et al., 2002	to determine the independent contributions of driver, vehicle, and accidents characteristics on fatalities in single vehicles' fixed objects crashes	BLM	109837	12	n.a.	n.a.	K	1	n.a.
Srinivasan, 2002	to model injury severity	OMXL, OLM	3492	6	mixed	BOTH	KACO	4	OMXL against OLM: $\chi^2 > \chi^2$ critical (60.86>28.86 at 0.05) for the observed data, $\chi^2 > \chi^2$ critical (31.92>28.86 at 0.05) for the predictive data
Ouyang et al., 2002	to study the simultaneity of injury severity outcomes in two vehicles' crashes of car-truck combination	BLM	2986	24	mixed	BOTH	A+K, O+C	2	ρ^2= 0.172 (Probably McFadden's)
Khattak and Rocha, 2003	to study the influence of various vehicles platforms on rollover single vehicle crashes and driver injuries	OLM	4552	5	n.a.	n.a.	AIS	7	$\rho^2 = 0.1040$ (McFadden's)
Dissanayake, 2004	to identify roadway, environmental, vehicle, and driver related characteristics affecting the injury severity for single vehicles' crashes by young and older drivers	BLM	n.a.	15	mixed	SEG	KABCO	5	n.a.

Wang and Kockelman, 2005	to study the effects of various vehicle, environmental, roadway, and occupant characteristics on the severity of injuries sustained by vehicle occupants in one and two vehicle crashes	HKL	n.a.	25	n.a.	SEG	KABCO	5	ρ^2 (McFadden's)
									for one vehicle: HOP= 0.237, OP= 0.235
									for two vehicle: HOP= 0.257, OP= 0.251

Lenuguerrand et al., 2006	Classify severity of occupant into dead or not dead for cars' accidents	GEE, HL, BLM	12030	9	n.a.	INT	K, not K	2	n.a.
Awadzi et al., 2008	to model injury severity of younger and older drivers	MNL	n.a.	18	mixed	BOTH	KBO	3	n.a.
Milton et. al.,	To study the variation that the influence of variables has on injury	MXL	22568	26	mixed	BOTH	K+A, BO	2	$\rho^2 = 0.1145$ (Calculated

2008	severity the roadway segments								McFadden's)
Malyshkina and Mannering, 2009	Analysis of injury severity of accidents for two- vehicles or less	MNL	81172	16	mixed	SEG	KBO	3	*p*-value for χ^2= 0.20-0.50
Schneider et al., 2009	to assess driver injury severity resulting from single-vehicle crashes on rural two-lane highways in Texas	MNL	10029	24	Rural	SEG	ABCO	4	ρ^2 (Probably McFadden's) for small radius model: ρ^2=0.258 for medium radius model: ρ^2=0.230 for large radius model: ρ^2=0.253
Jung et al., 2010	to assess the effects of rainfall on the severity of single-vehicle crashes on Wisconsin interstate highways	OLM, BLM	255	30	n.a.	n.a.	K+A, B+C, O	3	accuracy=88% for the KA, accuracy=68% for the B+C
Haleem and Abdel-Aty, 2010	To analyze crash injury severity at three- and four-legged un-signalized intersections	HL	2043	21	mixed	INT	K+A, B+ C+ O	2	AIC= 34040
Jin et al., 2010	To analyze the factors affecting right-angle crash injury severity on four-legged signalized intersections	OLM	13218	7	n.a.	INT	KABCO	5	ρ^2 = 0.0542
Daniels et al., 2010	to investigate which factors might explain the severity of crashes or injuries on roundabouts	HL	1491	7	n.a.	INT	K+ A, K	2	χ^2 for K+A= 10.88 (DF=8, *p*-value= 0.21), for K = 4.86 (DF= 6, *p*-value= 0.56)
Paleti et al., 2010	to capture the moderating effect of aggressive driving behavior while assessing the influence of a comprehensive set of variables on injury severity	HKL	6950	15	n.a.	n.a.	KABCO	5	ρ^2 =0.1188 (Calculated McFadden's)
Quddus et al., 2010	to explore the relationship between the severity of road crashes and the level of traffic congestion using disaggregated crash records and a measure of traffic congestion while controlling for other contributory factors	OLM, HM	3998	17	n.a.	SEG	KAC	3	ρ^2 (McFadden's) ρ^2 = 0.096 for the OLM ρ^2 = 0.099 for the HM
Dupont et al., 2010	to predict the chances for occupants involved in traffic accidents to end among the survivors given that the accident was fatal and to examine the features of the road users or of the vehicles that are positively or negatively associated with survival chances risk factor	BLM	1296	14	n.a.	BOTH	K	1	n.a.
Peek-Asa et al., 2010	to identify driver and crash characteristics associated with increased odds of fatal or severe injury among urban and rural crashes	BLM	87185	12	mixed	BOTH	KA	2	n.a.
Kononen et al., 2011	to predict the probability that a crash-involved vehicle will contain one or more occupants with serious or incapacitating injuries	BLM	n.a.	7	n.a.	n.a.	A	1	sensitivity= 40% specificity= 98% ROC area= 0.84

n.a.: not available data; KABCO (K=killed, A=incapacitating, B=non-incapacitating, C=possible injury, O=no injury); AIS (0=no injury, 1=minor, 2=moderate, 3=serious, 4=severe; 5=critical; 6=unsurvivable); SEG=basic segment only, INT=intersection only, BOTH=intersection + segments

Probably McFadden's: ρ^2 value was given in the study, and was found to apply to McFadden's formula; Calculated McFadden's: ρ^2 value was not given, but was calculated using log-likelihoods given in the study

Table 2: Studies that analyze injury severity of traffic accidents using Probit Models

Study Authors (publication year)	Objectives of the study	Model type	No. records	No. variables	Area type	Features	Injury level	No. injury levels	Model fit test
Renksi et al. 1999	to analyze effect of speed limit on occupant injury in single vehicles crashes. Excluding pedestrian, bicyclists, or motorcycles'' crashes	OPM	2729	7	mixed	SEG	KABCO	5	$\rho^2 = 0.116$ (Probably McFadden's)
Khattak, 2001	to analyze the effect of information and vehicle technology on injury severity in rear end crashes in two and three vehicles crashes	OPM	3912	36	mixed	SEG	KABCO	5	ρ^2 (McFadden's) ρ^2= 0.0319, 0.0671, 0.0660 for drivers 1, 2, 3 respectively
Kockelman and Kweon, 2002	to analyze the injury severity of all crashes, two vehicles crashes, single vehicle' crashes	OPM	n.a.	13	n.a.	n.a.	KACO	4	ρ^2= (0.0451-0.0868) (McFadden's)
Khattak et al., 2002	to isolate factors that contribute to injuries to older drivers involved in crashes	OPM	17045	16	mixed	BOTH	KABC	4	ρ^2= 0.057 (Probably McFadden's)
Abdel-Aty and Abdelwahab, 2004	to investigate the viability and potential benefits of using the ANN in predicting driver injury severity conditioned on the premise that a crash has occurred	OPM	7891	12	mixed	SEG	K+A, BCO	2	accuracy=61.7%
Abdel-Aty and Keller, 2005	to analyze crashes' injury severity on signalized intersections, where the ordial probit model was used to find the expected injury severity level	OPM	21371	34	n.a.	INT	KABCO	5	ρ^2= 0.24 (Calculated McFadden's)
Oh, 2006	to establish a statistical relationship correlating crash severity with weather, traffic maneuvers, and specific roadway geometrics at four-legged signalized intersections in rural areas. Four models were built: single vehicle, two vehicles, three or more vehicles, multiple vehicles	OPM	449	16	Rural	INT	KACO	4	ρ^2= 0.176 for all crashes model ρ^2= 0.480 for 3 or more vehicles ρ^2= 0.197 for 2 vehicles ρ^2=0.378 for single vehicle
Gårder, 2006	to analyze the statistical association between head-on crash severity and potential causal factors	OPM	3136	7	Rural	SEG	KABCO	5	n.a.
Gray et al., 2008	to study accidents for young male drivers	OPM	622431	13	mixed	BOTH	KACO	4	LL =-33665.05 for London model LL =-267706.85 for UK
Xie et al., 2009	analyze the relationship between accident injury severity and factors such as driver´s characteristics, vehicle type, and roadway conditions	OPM, BOP	76994	14	n.a.	INT	KABCO	5	accuracy for BOP for small data size= [55%-68%], for OP for small data size= [58%-68%], for BOP for predicted rest of the data= [61.8%-65.4%], for OP for predicted rest of the data =[59.9%-62.9%]
Wang et al., 2009	to identify factors contributing to injury severity at freeway diverge areas and to evaluate impacts of the factors	OPM	10946	17	n.a.	INT	KABCO	5	$\rho^2 = 0.0273$

Haleem and Abdel-Aty, 2010	To analyze crash injury severity at three- and four-legged un-signalized intersections	OPM, PM	2043	21	mixed	INT	for the OPM: KABCO for the PM: K+A, BCO	5, 2	AIC= 17091 (OPM + 3-legged) AIC= 9423 (OPM + 4-legged) AIC= 3804 (PM + 3-legged) AIC= 2100 (PM + 4-legged)

Lemp et al., 2011	to study the impact of vehicle, occupant, driver, and environmental characteristics on injury outcomes for those involved in crashes with heavy-duty trucks	OPM, HOP	1849	27	mixed	n.a.	KABCO	5	LL
									for OPM= -1993
									for HOP= -1896

Zhu and Srinivasan, 2011	to analyze the empirical factors affecting injury severity of large-truck. Two measures of severity were used: PAR= determined from police accident reports, RES= determined by researchers	OPM	953	28	mixed	BOTH	KA, B+C	2	ρ²
									for PAR model= 0.1780
									for RES model= 0.1827

n.a.: not available data; KABCO (K=killed, A=incapacitating, B=non-incapacitating, C=possible injury, O=no injury); SEG=basic segment only, INT=intersection only, BOTH=intersection + segments
Probably McFadden's: ρ² value was given in the study, and was found to apply to McFadden's formula; Calculated McFadden's: ρ² value was not given, but was calculated using log-likelihoods given in the study

Table 3: Studies that analyze injury severity of traffic accidents using other techniques

Study Authors (publication year)	Objectives of the study	Model type	No. records	No. variables	Area type	Features	Injury level	No. injury levels	Model fit test
TREES									
Council and Stewart, 1996	analyze severity of accident of single vehicles with fixed objects (for occupants)	CART	n.a.	7	mixed	BOTH	KBO	3	n.a.
Chen and Jovanis, 2000	to identify significant variables that contribute to the occurrence of a specific injury severity for bus drivers	CHAID	408	24	Rural	BOTH	KB	2	n.a.

Chang and Wang, 2006	To model the injury severity of an individual involved in a traffic accident	CART	26831	14	mixed	BOTH	KBO	3	accuracy
									for fatal (0%),
									for injury (94%),
									for no injury (68%)

BAYESIAN NETWORKS									
Simoncic, 2004	to model two car accident injury severity	BN	17558	12	mixed	n.a.	K+A, other	2	n.a.
de Oña et al., 2011	to classify crashes according to their injury severity	BN	1536	18	Rural	SEG	K+A, C	2	accuracy= 60% sensitivity= 73% specificity= 45% ROC area= 61%
NEURAL NETWORKS									
Abdelwahab and Abdel-Aty, 2001	to analyze the injury severity of crashes of two-vehicles that occurred at signalized intersections	MLP, Fuzzy ARTMAP	1168	14	n.a.	INT	A, B+C, O	3	accuracy =65.6%

Abdel-Aty and Abdelwahab, 2004	to investigate the viability and potential benefits of using the ANN in predicting driver injury severity conditioned on the premise that a crash has occurred	MLP, Fuzzy ARTMAP	7891	12	mixed	SEG	K+A, BCO	2	accuracy
									for MLP= 73.5%
									for fuzzy ARTMAP 40.6%

Delen et. al., 2006	To model the potentially non-linear relationships between the injury severity levels and accident-related factors	MLP	30358	13	n.a.	n.a.	KABCO	5	accuracy= 40.73%
LINEAR GENETIC PROGRAMMING									
Das and Abdel-Aty, 2010	To understand the relationship of geometric and environmental factors with injury related crashes as well as with severe crashes	LGP	104952	58	mixed	BOTH	B+O, A+C	2	accuracy= 60.4%
OTHERS									
Chen and Jovanis, 2000	to identify significant variables that contribute to the occurrence of a specific injury severity for bus drivers	LLM	408	24	Rural	BOTH	KB	2	R^2= 0.95

n.a.: not available data; KABCO (K=killed, A=incapacitating, B=non-incapacitating, C=possible injury, O=no injury); SEG=basic segment only, INT=intersection only, BOTH=intersection + segments

Table 4. Maximum, minimum and median for number of accident records, number of variables and number of injury levels

	No. of accident records	No. of Variables	No. of Injury Levels
All studies			
Max	81,172	36	7
Min	255	5	1
Median	3,955	15	3
Logit studies			
Max	81,172	30	7
Min	255	5	1
Median	4,552	15	3 a
Probit studies			
Max	76,994	36	5
Min	449	7	2
Median	3,136	16	5 b
Other studies			
Max	30,358	24	5
Min	408	7	2
Median	4,713	14	2 a

a, b: denotes differences statistically significant ($p < 0.05$), based on Mann-Whitney U test

Table 5: Frequency of usage of each model

	Family of models	Model	Frequency
Discrete Model	Logit Models	BLM OLM HL MNL HKL MXL GEE OMXL	11 6 4 3 2 1 1 1
	Probit Models	OPM BOP HOP PM	14 1 1 1
Other Models	Decision Trees	CART CHAID	2 1
	Bayesian Networks	BN	2
	Artificial Neural Networks	MLP Fuzzy ARTMAP	3 2
	Evolutionary algorithms	LGP	1
	Log-linear models	LLM	1

2. Analysis of traffic accident injury severity on Spanish rural highways using Bayesian networks

Juan de Oña, Randa Oqab Mujalli, Francisco J. Calvo. Analysis of traffic accident injury severity on Spanish rural highways using Bayesian networks. Published in ***Accident Analysis and Prevention****, Volume 43, Issue 1, January 2011, Pages 402-411.*

This paper illustrates the potential of using Bayesian Networks (BNs) to classify traffic accidents according to their injury severity. In which an analysis of 1536 accidents on rural highways in Spain, where 18 variables representing the aforementioned contributing factors were used to build 3 different BNs that classified the severity of accidents into two classes slightly injured (SI) and killed or severely injured (KSI).

First, a combination of the Hillclimbing search algorithm with each of the following metrics: BDeu, MDL and AIC was made in order to build the Bayesian networks.

Three different Bayesian networks were built and the corresponding performance evaluation indicators (accuracy, sensitivity, specificity, HMSS and ROC area) were calculated for each of the Bayesian networks built. Also, the complexity was calculated based on the number of arcs that exists in the graph of the Bayesian network structure.

Thus, the results obtained by these Bayesian networks in terms of the aforementioned performance measures were close. In order to better distinguish the Bayesian network with the best performance, the most probable explanation was calculated for each of them.

The results indicated that the Hillclimbing with the MDL score metric represent relatively the best performance, and hence the Bayesian network built using this combination was used for further analysis.

In order to find out which are the significant variables that are related to the occurrence of a killed or severe injury in a traffic accident, an inference for the Bayesian network was done. The results of this inference indicated that the following variables are the most significant: accident type, driver age, lighting and number of injuries.

Accident Analysis and Prevention 43 (2011) 402–411

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

Analysis of traffic accident injury severity on Spanish rural highways using Bayesian networks

Juan de Oña*,[1], Randa Oqab Mujalli[1], Francisco J. Calvo

TRYSE Research Group, Department of Civil Engineering, University of Granada, ETSI Caminos, Canales y Puertos, c/Severo Ochoa, s/n, 18071 Granada, Spain

ARTICLE INFO

Article history:
Received 18 June 2010
Received in revised form
16 September 2010
Accepted 23 September 2010

Keywords:
Bayesian networks
Injury severity
Traffic accidents
Classification

ABSTRACT

Several different factors contribute to injury severity in traffic accidents, such as driver characteristics, highway characteristics, vehicle characteristics, accidents characteristics, and atmospheric factors. This paper shows the possibility of using Bayesian Networks (BNs) to classify traffic accidents according to their injury severity. BNs are capable of making predictions without the need for pre assumptions and are used to make graphic representations of complex systems with interrelated components. This paper presents an analysis of 1536 accidents on rural highways in Spain, where 18 variables representing the aforementioned contributing factors were used to build 3 different BNs that classified the severity of accidents into slightly injured and killed or severely injured. The variables that best identify the factors that are associated with a killed or seriously injured accident (accident type, driver age, lighting and number of injuries) were identified by inference.

1. Introduction

The number of traffic accidents and their effects, mainly human fatalities and injuries, justify the importance of analyzing the factors that contribute to their occurrence. Identifying the factors that significantly influence the injury severity of traffic accidents was the main objective of many previous studies. Factors affecting injury severity of a traffic accident are usually caused by one or more of the following factors: driver characteristics, highway characteristics, vehicle characteristics, accidents characteristics and atmospheric factors (Kopelias et al., 2007; Chang and Wang, 2006).

Regression analysis has been widely used to determine the contributing factors that cause a specific injury severity. The most commonly used regression models in traffic injury analysis are the logistic regression model and the ordered Probit model (Al-Ghamdi, 2002; Milton et al., 2008; Bédard et al., 2002; Yau et al., 2006; Yamamoto and Shankar, 2004; Kockelman and Kweon, 2002). However, most of the regression models that are used to model traffic injury severity have their own model assumptions and pre-defined underlying relationships between dependent and independent variables (i.e. linear relations between the variables) (Chang and Wang, 2006). If these assumptions are violated, the model could lead to erroneous estimations of the likelihood of severe injury.

Gregoriades (2007) highlighted the interest of using Bayesian Networks (BNs) to model traffic accidents and discussed the need to not consider traffic accidents as a deterministic assessment problem. Instead, researchers should model the uncertainties involved in the factors that can lead to road accidents. He listed a number of candidate approaches for modeling uncertainty, such as, Bayesian probability.

BNs make it easy to describe accidents that involve many interdependent variables. The relationship and structure of the variables can be studied and trained from accident data. They do not need to know any pre-defined relationships between dependent and independent variables.

The three main advantages of BNs are bi-directional induction, incorporation of missing variables and probabilistic inference. By using BNs, it is relatively easy to discover the underlying patterns of data, to investigate the relationships between variables and to make predictions using these relationships. Incident data used in a study by Ozbay and Noyan (2006) were collected from incident clearance survey forms to understand incident clearance characteristics and then used to develop incident duration prediction models. The researchers modeled the incidents' clearance durations using BNs and were able to represent the stochastic nature of incidents.

Using BNs to analyze traffic accident injury severity is scarce. A two car accident injury severity model was constructed using BNs (Simoncic, 2004). A BN was built using several variables, and the Most Probable Explanation (MPE) was calculated for the most probable configuration of values for all the variables in the BN, in order to serve as an indication of the quality of the estimated BN. The results pointed out that BNs could be applied in road accident

* Corresponding author. Tel.: +34 958 24 99 79.
E-mail address: jdona@ugr.es (J. de Oña).

[1] These authors have contributed equally to this work.

0001-4575/$ – see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.aap.2010.09.010

modeling, and some improvements, such as using more variables and larger datasets, were recommended. Although this study highlighted the possibility of using BNs to model traffic accidents, the results were based on building only one possible network, without measuring the performance of the Bayesian classifier.

The scope of this paper is to validate the possibility of using BNs to classify traffic accidents according to their injury severity, and to find out the best BN classification performance along with the best graphical representation, in order to be capable of identifying the relevant variables that affect the injury severity of traffic accidents.

This paper is organized as follows. Section 2 presents the data used and briefly reviews the concept of BNs and Bayesian learning. The methods used for preprocessing and evaluating the data are also presented; finally a brief description of inference is presented. In Section 3, the results and their discussion are presented. In Section 4, summary and conclusions are given.

2. Materials and methods

2.1. Accident data

Accident data were obtained from the Spanish General Traffic Directorate (DGT) for rural highways for the province of Granada (South of Spain) for three years (2003–2005). The total number of accidents obtained for this period was 3302. The data were first checked out for questionable data, and those which were found to be unrealistic were screened out. Only rural highways were considered in this study; data related to intersections were not included, since intersections have their own specific characteristics and need to be analyzed separately. Finally, the database used to conduct the study contained 1536 records. Table 1 provides information on the data used for this study.

Eighteen variables were used with the class variable of injury severity (SEV) in an attempt to identify the important patterns of an accident that usually require an explanation.

The data included variables describing the conditions that contributed to the accident and injury severity.

- Injury severity variables: number of injuries (e.g., passengers, drivers and pedestrians), severity level of injuries (e.g., fatal, severe, slight). Following previous studies (Chang and Wang, 2006; Milton et al., 2008) the injury severity of an accident is determined according to the level of injury to the worst injured occupant.
- Roadway information: characteristics of the roadway on which the accidents occurred (e.g., grade, pavement width, lane width, shoulder type, pavement markings, sight distance, if the shoulder was paved or not, etc.).
- Weather information: weather conditions when the accident occurred (e.g., good weather, rain, fog, snow and windy).
- Accident information: contributing circumstances (e.g., type of accident, time of accident (hour, day, month and year), and vehicles involved in the accident).
- Driver data: characteristics of the driver, such as age or gender.

2.2. BN definition

Over the last decade, BNs have become a popular representation for encoding uncertain expert knowledge in expert systems. The field of BNs has grown enormously, with theoretical and computational developments in many areas (Mittal et al., 2007) such as: modeling knowledge in bioinformatics, medicine, document classification, information retrieval, image processing, data fusion, decision support systems, engineering, gaming, and law.

Let $U = \{x_1, \ldots, x_n\}$, $n \geq 1$ be a set of variables. A BN over a set of variables U is a network structure, which is a Directed Acyclic Graph (DAG) over U and a set of probability tables $B_p = \{p(x_i|pa(x_i), x_i \in U)\}$ where $pa(x_i)$ is the set of parents or antecedents of x_i in BN and $i = 1, 2, 3, \ldots, n$. A BN represents joint probability distributions $P(U) = \prod_{x_i \in U} p(x_i|pa(x_i))$.

The classification task consists in classifying a variable $y = x_0$ called the class variable, given a set of variables $U = x_1, \ldots, x_n$, called attribute variables. A classifier h: $U \rightarrow y$ is a function that maps an instance of U to a value of y. The classifier is learned from a dataset D consisting of samples over (U, y). The learning task consists of finding an appropriate BN given a data set D over U.

BNs are graphical models of interactions among a set of variables, where the variables are represented as nodes (also known as vertices) of a graph and the interactions (direct dependences) as directed links (also known as arcs and edges) between the nodes. Any pair of unconnected/nonadjacent nodes of such a graph indicates (conditional) independence between the variables represented by these nodes under particular circumstances that can easily be read from the graph. Each node contains the states of the random variable and it represents a conditional probability table. The conditional probability table of a node contains the probabilities of the node being in a specific state, given the states of its parents.

Fig. 1 shows that the dependencies and independencies among the factors that affect the time of journey (the class variable) are represented in the form of direct edges (arrows) between factors that are represented as nodes. For example, the variable (vehicle type) is a parent (antecedent) of the two variables (cost and velocity) called children or descendents. Any knowledge (evidence) about the parent variable affects the probabilities of occurrence of the children or descendent variables.

It should be noticed that the edges in a BN are not necessarily causal. That is, a BN can satisfy the probability distribution of the variables in the BN without the edges being causal (Neapolitan, 2009). Thus, the edges between variables in a non-causal BN could imply a sort of interrelationship(s) among these variables.

2.3. BN learning and the scoring metrics used

When there are masses of data available and it is necessary to interpret them and to provide a model for predicting the behavior of unobserved cases, the learning of both structure and parameters is used (Cooper and Herskovits, 1992). There are two main approaches to structure learning in BNs:

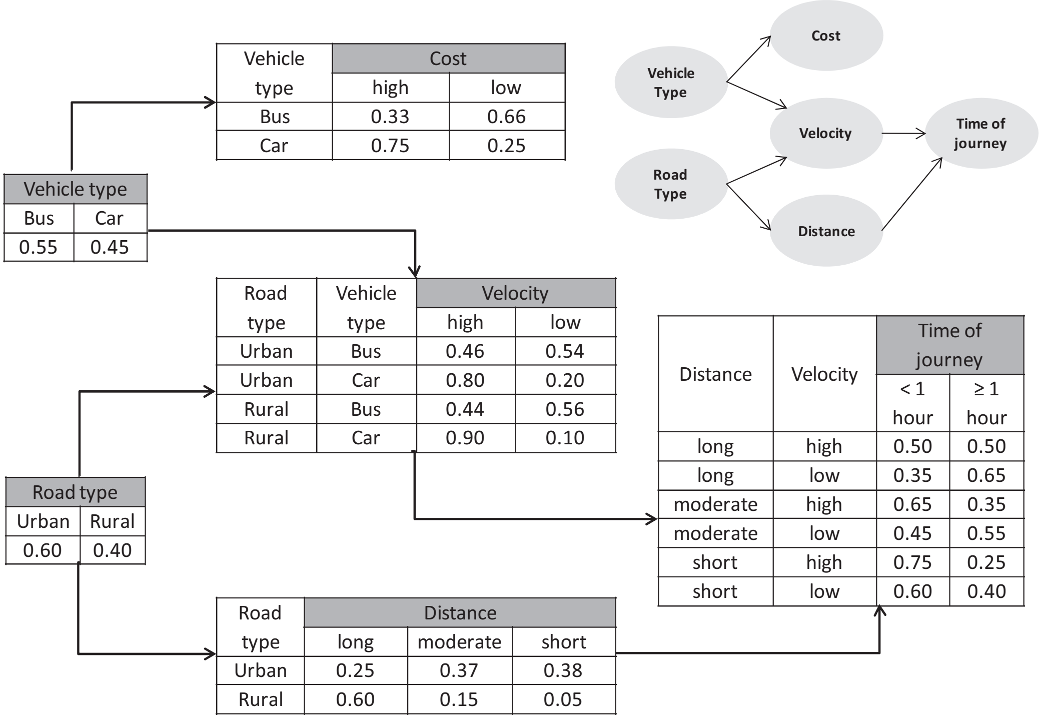
- **Constraint based**: Perform tests of conditional independence on the data, and search for a network that is consistent with the observed dependencies and independencies.
- **Score based**: Define a score that evaluates how well the dependencies or independencies in a structure match the data and search for a structure that maximizes the score.

The advantage of score-based methods over the constraint-based methods is that they are less sensitive to errors in individual tests; compromises can be made between the extent to which variables are dependent in the data and the cost of adding the edge. Because of the aforementioned advantages, the score based method is followed in this study.

Weka software (Witten and Frank, 2005) was used in this study to build the BN. This software is freely available, it is implemented in Java language, it contains a collection of data processing and modeling techniques and it contains a graphical user interface. The BNs built here used all the nineteen variables of the 1536 records.

Table 1
Variables, values and actual classification by severity.

Variables	Values	SEV*				Total
		SI		KSI		
ACT: accident type	AS: angle or side collision	381	61.45%	239	38.55%	620
	CF: fixed objects	99	52.94%	88	47.06%	187
	HO: head on	84	40.58%	123	59.42%	207
	O: other	75	59.06%	52	40.94%	127
	PU: pile up	33	78.57%	9	21.43%	42
	R: rollover	163	49.39%	167	50.61%	330
	SP: straight path	17	73.91%	6	26.09%	23
AGE: age	[18–25]	225	50.34%	222	49.66%	447
	(25–64]	586	57.73%	429	42.27%	1015
	>64	41	55.41%	33	44.59%	74
ATF: atmospheric factors	GW: good weather	730	54.23%	616	45.77%	1346
	HR: heavy rain	23	71.88%	9	28.13%	32
	LR: light rain	84	61.76%	52	38.24%	136
	O: other	15	68.18%	7	31.81%	22
CAU: cause	DC: driver characteristics	791	54.93%	649	45.07%	1440
	OF: other factors	50	66.67%	25	33.33%	75
	RC: road characteristics	3	75.00%	1	25.00%	4
	VC: vehicle characteristics	8	47.06%	9	52.94%	17
DAY: day	BW: beginning of week	123	60.29%	81	39.71%	204
	EW: end of week	132	57.14%	99	42.86%	231
	F: festive	29	61.70%	18	38.30%	47
	WD: week day	325	55.65%	259	44.35%	584
	WE: week end	243	51.70%	227	48.30%	470
GEN: gender	F: female	148	63.79%	84	36.21%	232
	M: male	704	53.99%	600	46.01%	1304
LAW: lane width	THI: thin: <3.25 m	19	67.86%	9	32.14%	28
	MED: medium: 3.25 m $\leq L \leq$ 3.75 m	176	51.16%	168	48.84%	344
	WID: wide: >3.75 m	657	56.44%	507	43.56%	1164
LIG: lighting	D: dusk	52	61.18%	33	38.82%	85
	DL: daylight	573	58.65%	404	41.35%	977
	I: insufficient	27	54.00%	23	46.00%	50
	S: sufficient	36	59.02%	25	40.98%	61
	W: without lighting	164	45.18%	199	54.82%	363
MON: month	AUT: autumn	218	54.23%	184	45.77%	402
	SPR: spring	206	59.03%	143	40.97%	349
	SUM: summer	246	56.55%	189	43.45%	435
	WIN: winter	182	52.00%	168	48.00%	350
NOI: number of injuries	1	539	49.95%	540	50.05%	1079
	>1	313	68.49%	144	31.51%	457
OI: occupants involved	1	229	51.58%	215	48.42%	444
	2	374	55.99%	294	44.01%	668
	>2	249	58.73%	175	41.27%	424
PAS: paved shoulder	Missing values	66	51.56%	62	48.44%	128
	N: no	253	57.11%	190	42.89%	443
	Y: yes	533	55.23%	432	44.77%	965
PAW: pavement width	THI: thin: <6 m	95	53.98%	81	46.02%	176
	MED: medium: 6 m $\leq$ law $\leq$ 7 m	209	54.29%	176	45.71%	385
	WID: wide: >7 m	548	56.21%	427	43.79%	975
ROM: pavement markings	DME: does not exist or was deleted	60	58.25%	43	41.75%	103
	DMR: define margins of roadway	60	57.69%	44	42.31%	104
	SLD: separate lanes and defined road margins	714	55.26%	578	44.74%	1292
	SLO: separate lanes only	18	48.65%	19	51.35%	37
SHT: Shoulder type	NOS: does not exist	311	55.24%	252	44.76%	563
	THI: thin: <1.5 m	402	54.47%	336	45.53%	738
	MED: medium: 1.5 m $\leq$ sht <2.50 m	133	58.85%	93	41.15%	226
	WID: wide $\geq$2.50 m	6	66.67%	3	33.33%	9
SID: sight distance	A: atmospheric	26	81.25%	6	18.75%	32
	B: building	10	55.56%	8	44.44%	18
	O: other	6	66.67%	3	33.34%	9
	T: topological	187	55.49%	150	44.51%	337
	V: vegetation	6	54.55%	5	45.45%	11
	WR: without restriction	617	54.65%	512	45.35%	1129
TIM: time	[0–6]	99	46.26%	115	53.74%	214
	(6–12]	236	57.99%	171	42.01%	407
	(12–18]	314	57.72%	230	42.28%	544
	(18–24)	203	54.72%	168	45.28%	371
VI: vehicles involved	1	316	52.06%	291	47.94%	607
	2	468	56.73%	357	43.27%	825
	>2	68	65.38%	36	34.62%	104
Total		852	55.47%	684	44.53%	1536



Vehicle type	
Bus	Car
0.55	0.45

Vehicle type	Cost	
	high	low
Bus	0.33	0.66
Car	0.75	0.25

Road type	
Urban	Rural
0.60	0.40

Road type	Vehicle type	Velocity	
		high	low
Urban	Bus	0.46	0.54
Urban	Car	0.80	0.20
Rural	Bus	0.44	0.56
Rural	Car	0.90	0.10

Road type	Distance		
	long	moderate	short
Urban	0.25	0.37	0.38
Rural	0.60	0.15	0.05

Distance	Velocity	Time of journey	
		< 1 hour	≥ 1 hour
long	high	0.50	0.50
long	low	0.35	0.65
moderate	high	0.65	0.35
moderate	low	0.45	0.55
short	high	0.75	0.25
short	low	0.60	0.40

Fig. 1. An example of a BN with the corresponding CPTs for each node.

In order to build BN structures; BDe Score metric, Minimum Description Length (MDL) and the Akaike Information Criterion (AIC) score functions were run, based on the hill climbing algorithm.

Let r_i $(1 \le i \le n)$ be the cardinality of x_i, q_i is used to denote the cardinality of the parent set of x_i in BN, that is, the number of different values to which the parents of x_i can be instantiated. So, q_i can be calculated as the product of cardinalities of nodes in $pa(x_i)$, $q_i = \prod_{x_j \in pa(x_i)} r_j$. Note $pa(x_i) = \phi$ implies $q_i = 1$.

$N_{ij}(1 \le i \le n,\ 1 \le j \le q_i)$ denotes the number of records in D for which $pa(x_i)$ takes its jth value. $N_{ijk}(1 \le i \le n,\ 1 \le j \le q_i,\ 1 \le k \le r_i)$ denotes the number of records in D for which $pa(x_i)$ takes its jth value and for which x_i takes its kth value. So, $N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$. N denotes the number of records in D.

Let the *entropy metric* H (BN,D) of a network structure and database be defined as:

$$H(BN, D) = -N \sum_{i=1}^{n} \sum_{j=1}^{q_i} \sum_{k=1}^{r_i} \frac{N_{ijk}}{N} \log \frac{N_{ijk}}{N_{ij}} \tag{1}$$

and the number of parameters K as:

$$K = \sum_{i=1}^{n} (r_i - 1) \cdot q_i \tag{2}$$

The AIC metric $Q_{AIC}(BN, D)$ of a Bayesian network structure for a database D is:

$$Q_{AIC}(BN, D) = H(BN, D) + K \tag{3}$$

A term $P(BN)$ can be added representing prior information over network structures, but will be ignored for simplicity in the Weka implementation (Bouckaert, 1995).

The MDL metric $Q_{MDL}(BN, D)$ of a Bayesian network structure BN for a database D is defined as:

$$Q_{MDL}(BN, D) = H(BN, D) + \frac{K}{2} \log N \tag{4}$$

The BDe metric $Q_{BDe}(BN, D)$ of a BN structure for a database D is:

$$Q_{BDe}(BN, D) = P(BN) \prod_{i=0}^{n} \prod_{j=1}^{q_i} \frac{\Gamma(1/q_i)}{\Gamma((1/q_i) + N_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma((1/r_i) \cdot q_i + N_{ijk})}{\Gamma((1/r_i) \cdot q_i)} \tag{5}$$

where $P(BN)$ is the prior on the network structure (taken to be constant hence ignored in the Weka implementation) (Bouckaert, 1995) and $\Gamma(-)$ the gamma-function.

Using hill climbing algorithm, the states of search space are possible models. Operations are the insertion, deletion and reverse of an edge in the network to modify a model. The hill climbing search algorithm was applied in this study mainly because it is fast and widely used, and also produces good results in terms of network complexity and accuracy (Madden, 2009).

2.4. BN data preprocessing

The variables obtained from the DGT were further refined and categorized into distinct values in order to be able to work with them. Other variables were merged or abstracted on the basis of procedures followed in previous studies (Simoncic, 2004; Helai et al., 2008), where the class variable was injury severity (slight injured –SI– and killed or seriously injured –KSI), and the severity was considered for the most severe case in the accident (Chang and Wang, 2006; Simoncic, 2004).

The only preprocessing filter used on this dataset was the unsupervised variable filter for replacing missing values. This filter replaces the missing values with the modes and means from the training data. The cross validation method was used to split the data into ten equal folds (or subsets), the BN was built on the fold (called training set) and the analysis was validated on the other subset (called the validation set or testing set). Multiple repetitions or trials (10 times) of cross validation are used to reduce variability, and the validation results are averaged over the trials.

2.5. BN evaluation indicators

Five indicators are used in this study to compare the BNs built (see Eqs. (6–9)): accuracy, sensitivity, specificity, HMSS, and ROC area were calculated for each BN.

$$\text{Accuracy} = \frac{tSI + tKSI}{tSI + tKSI + fSI + fKSI} 100\% \qquad (6)$$

$$\text{Sensitivity} = \frac{tSI}{tSI + fKSI} 100\% \qquad (7)$$

$$\text{Specificity} = \frac{tKSI}{tKSI + fSI} 100\% \qquad (8)$$

$$\text{HMSS} = \frac{2 \times \text{Sensitivity} \times \text{Specificity}}{\text{Sensitivity} + \text{Specificity}} \qquad (9)$$

where *tSI* is true slight injured cases, *tKSI* true killed or seriously injured cases, *fSI* false slight injured cases, and *fKSI* false killed or seriously injured cases.

Accuracy (see Eq. (6)) is proportion of instances that were correctly classified by the classifier. Accuracy only gives information on the classifier's general performance.

Sensitivity represents the proportion of correctly predicted slight injured among all the observed slight injured. Specificity represents the proportion of correctly predicted killed or seriously injured among all the observed killed or seriously injured (see Eqs. (7–8)). Another measure used to assess the performance of the BN built was the Harmonic Mean of Sensitivity and Specificity (HMSS), which gives an equal weight of both sensitivity and specificity (see Eq. (9)).

Another indicator is the Receiver Operating Characteristic Curve (ROC) Area. What ROC curves represent is the true positive rate (sensitivity) vs. the false positive rate (1 − specificity). ROC curves are more useful as descriptors of overall test performance, reflected by the area under the curve, with a maximum of 1.00 describing a perfect test and an ROC area of 0.50 describing a valueless test.

Other measures used in the literature to evaluate the performance of BNs specifically include both the Most Probable Explanation (MPE) (Simoncic, 2004) and the complexity or the total number of BN arcs (Cruz-Ramírez et al., 2007). MPE is a technique that is developed for generating explanation in BNs, in which the configuration with the maximum posterior probability is calculated (Pearl, 2004).

For the analysis of traffic accident injury severity and to determine the optimal BN, the measures described above will be calculated first: accuracy, sensitivity, specificity, ROC area, the MPE and the complexity of the built BNs. Later, the best BN found in terms of these measures will be used for inference.

2.6. BN inference

Inference in BNs consists of computing the conditional probability of some variables, given that other variables are set to evidence. Inference may be done for a specific state or value of a variable, given evidence on the state of other variable(s). Thus, using the conditional probability table for the BN built, their values can be easily inferred. Fig. 1 shows an example of a conditional probability table, where it could be seen that given evidence for the distance to be "short" and the velocity to be "high", the probability that the time of journey will be less than 1 h is 0.75. Thus, other inferences could be extracted using this figure, where the example presented here is used to explain how inference in BNs works.

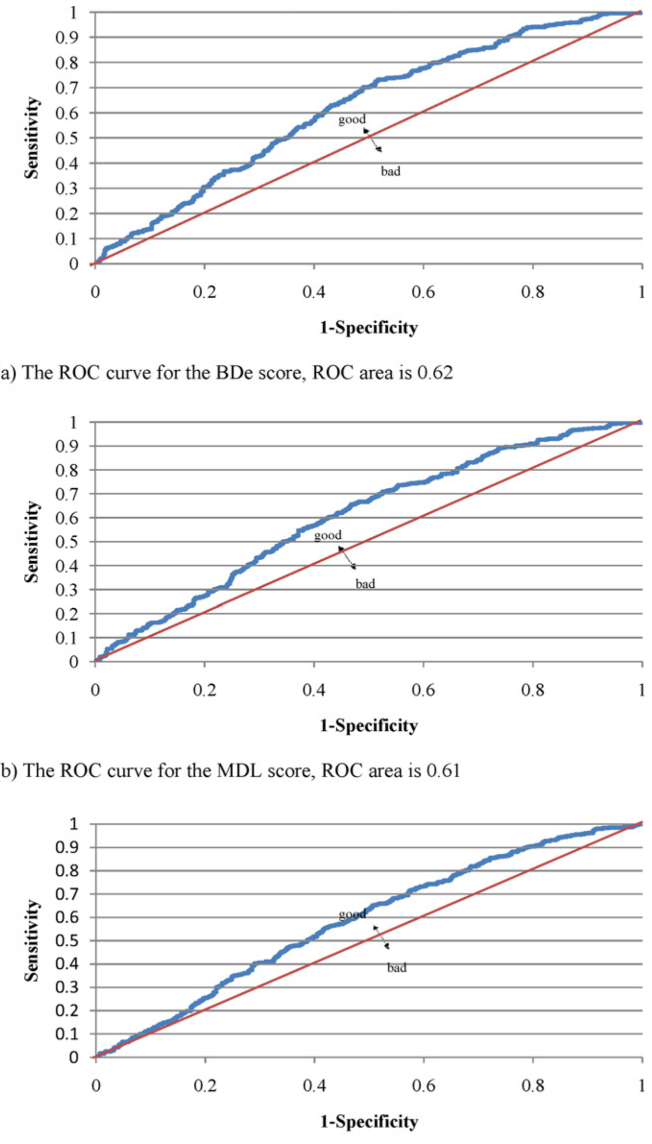
In this paper, inference is used to determine the most significant variables that are associated with KSI in traffic accidents.

3. Results and discussion

Table 2 shows the results obtained from building BNs using the hill climbing search method and three different score metrics (BDe, MDL and AIC) using both the training and the test set to validate the results. From the original dataset, 2/3 of the data was held for training the BNs and the other 1/3 was used for testing them.

Ten different schemes of training/testing datasets were used to analyze the effect of swapping training and test datasets. Table 2 shows the average and the standard deviation of each one of the indicators for the score metrics used.

It can be seen that both the training and the test results are very similar. The accuracies performed in this study did not vary significantly; the highest accuracy was for the BDe score (61%). Abdel Wahab and Abdel-Aty (2001) used some data mining techniques to model injury severity in traffic accidents. They obtained accuracies of 60.4% and 65.6% for training and testing sets respectively when using an MLP neural network, 56.2% when using fuzzy ARTMAP neural network and 58.1% when using O-ARTMAP. Thus, the results



(a) The ROC curve for the BDe score, ROC area is 0.62

(b) The ROC curve for the MDL score, ROC area is 0.61

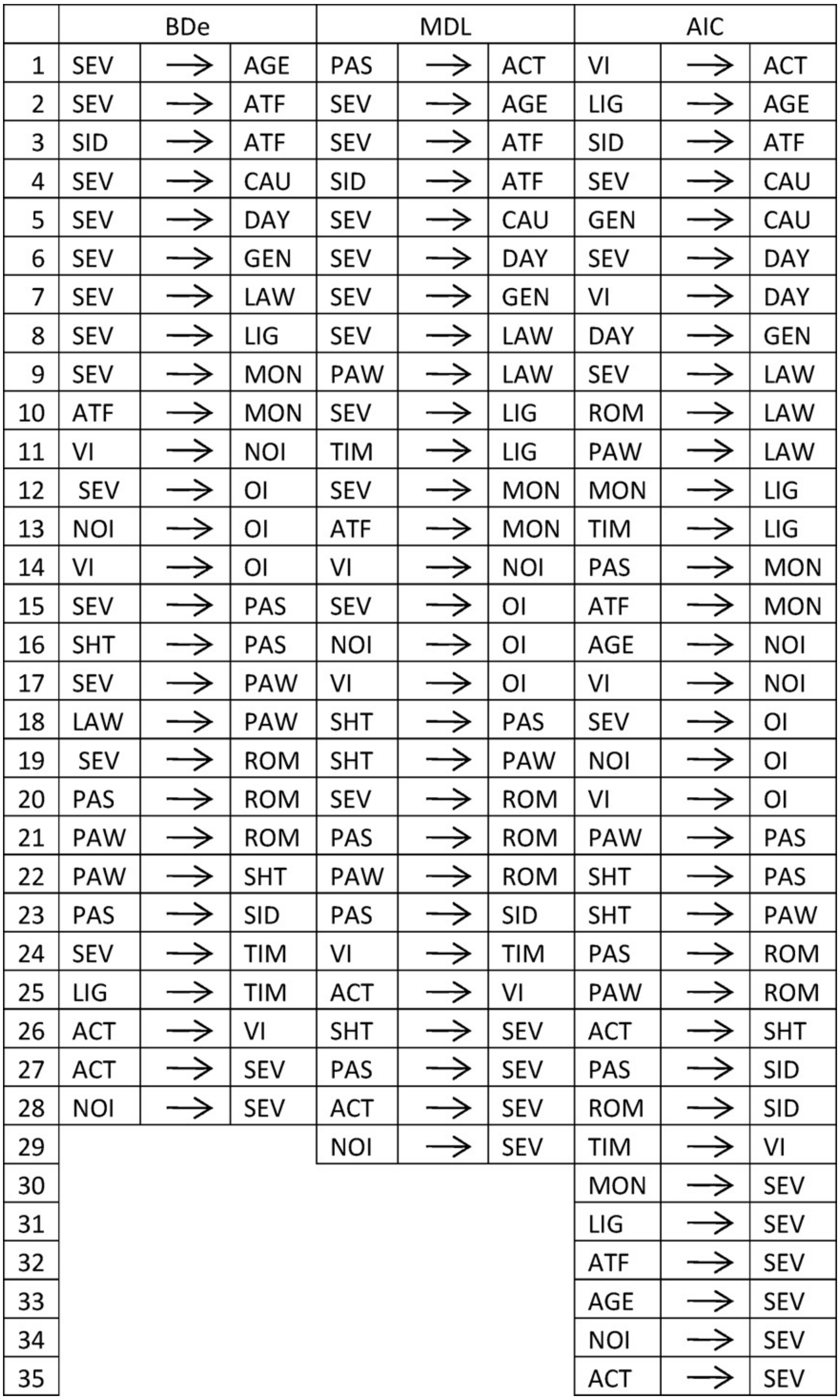
(c) The ROC curve for the AIC score, the ROC area is 0.59

Fig. 2. The ROC curves for the three score methods and one dataset. (a) The ROC curve for the BDe score, ROC area is 0.62. (b) The ROC curve for the MDL score, ROC area is 0.61. (c) The ROC curve for the AIC score, the ROC area is 0.59.

Table 2
Accuracy, sensitivity, specificity, HMSS and ROC Area for BDe, MDL and AIC score metrics (training and test sets).

Score metric	BDe		MDL		AIC	
Dataset	Training	Test	Training	Test	Training	Test
Indicator	Average ± s.d.[a]	Average ± s.d.[a]	Average ± s.d.[a]	Average ± s.d.[a]	Average ± s.d.[a]	Average ± s.d.[a]
Accuracy	0.61 ± 0.01	0.57 ± 0.02	0.60 ± 0.01	0.59 ± 0.02	0.58 ± 0.01	0.58 ± 0.03
Sensitivity	0.74 ± 0.02	0.65 ± 0.04	0.73 ± 0.02	0.65 ± 0.03	0.66 ± 0.02	0.63 ± 0.04
Specificity	0.44 ± 0.03	0.49 ± 0.05	0.45 ± 0.03	0.53 ± 0.05	0.47 ± 0.03	0.53 ± 0.04
HMSS	0.55 ± 0.02	0.56 ± 0.03	0.56 ± 0.02	0.58 ± 0.03	0.55 ± 0.02	0.58 ± 0.02
ROC area	0.62 ± 0.04	0.58 ± 0.02	0.61 ± 0.02	0.62 ± 0.02	0.58 ± 0.02	0.61 ± 0.03

[a] s.d.: standard deviation.



	BDe			MDL			AIC		
1	SEV	→	AGE	PAS	→	ACT	VI	→	ACT
2	SEV	→	ATF	SEV	→	AGE	LIG	→	AGE
3	SID	→	ATF	SEV	→	ATF	SID	→	ATF
4	SEV	→	CAU	SID	→	ATF	SEV	→	CAU
5	SEV	→	DAY	SEV	→	CAU	GEN	→	CAU
6	SEV	→	GEN	SEV	→	DAY	SEV	→	DAY
7	SEV	→	LAW	SEV	→	GEN	VI	→	DAY
8	SEV	→	LIG	SEV	→	LAW	DAY	→	GEN
9	SEV	→	MON	PAW	→	LAW	SEV	→	LAW
10	ATF	→	MON	SEV	→	LIG	ROM	→	LAW
11	VI	→	NOI	TIM	→	LIG	PAW	→	LAW
12	SEV	→	OI	SEV	→	MON	MON	→	LIG
13	NOI	→	OI	ATF	→	MON	TIM	→	LIG
14	VI	→	OI	VI	→	NOI	PAS	→	MON
15	SEV	→	PAS	SEV	→	OI	ATF	→	MON
16	SHT	→	PAS	NOI	→	OI	AGE	→	NOI
17	SEV	→	PAW	VI	→	OI	VI	→	NOI
18	LAW	→	PAW	SHT	→	PAS	SEV	→	OI
19	SEV	→	ROM	SHT	→	PAW	NOI	→	OI
20	PAS	→	ROM	SEV	→	ROM	VI	→	OI
21	PAW	→	ROM	PAS	→	ROM	PAW	→	PAS
22	PAW	→	SHT	PAW	→	ROM	SHT	→	PAS
23	PAS	→	SID	PAS	→	SID	SHT	→	PAW
24	SEV	→	TIM	VI	→	TIM	PAS	→	ROM
25	LIG	→	TIM	ACT	→	VI	PAW	→	ROM
26	ACT	→	VI	SHT	→	SEV	ACT	→	SHT
27	ACT	→	SEV	PAS	→	SEV	PAS	→	SID
28	NOI	→	SEV	ACT	→	SEV	ROM	→	SID
29				NOI	→	SEV	TIM	→	VI
30							MON	→	SEV
31							LIG	→	SEV
32							ATF	→	SEV
33							AGE	→	SEV
34							NOI	→	SEV
35							ACT	→	SEV

Fig. 3. The arcs as obtained by applying the three score metrics.

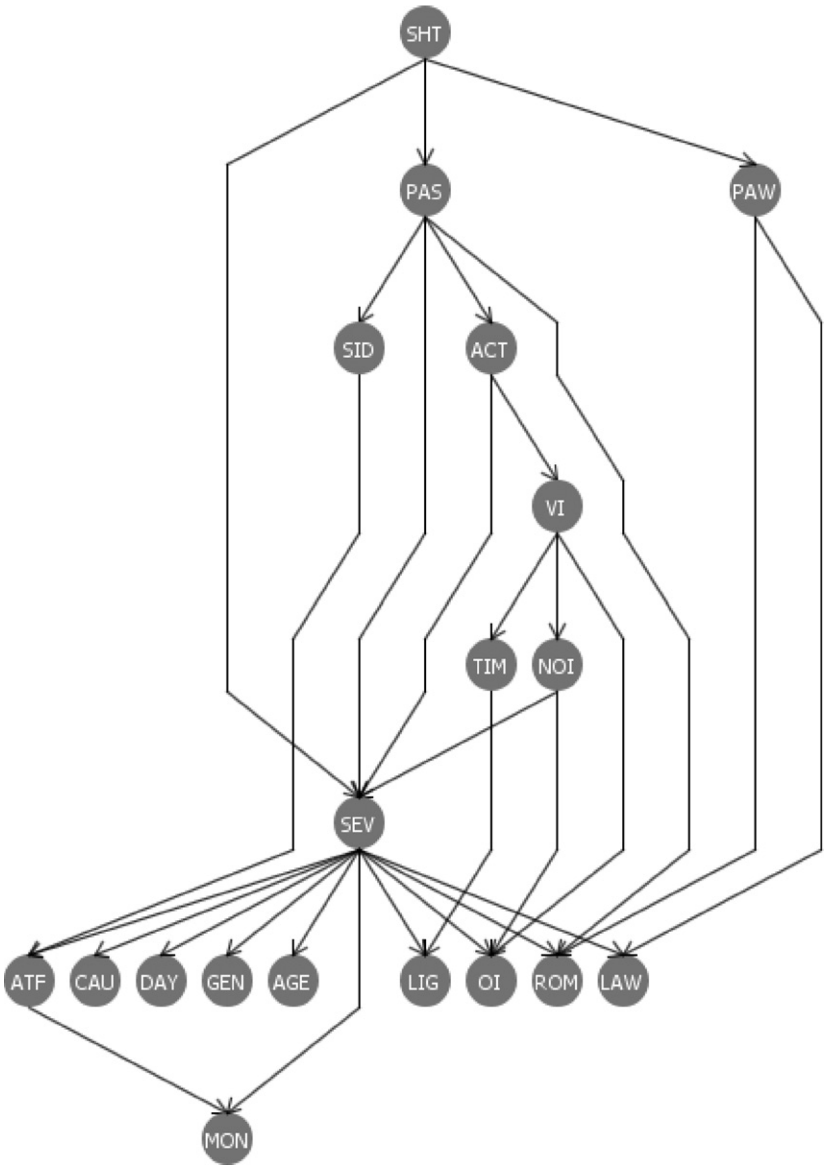


Fig. 4. BN structure for the MDL score.

obtained in this paper were within the range of accuracies found by Abdel Wahab and Abdel-Aty (2001).

Also, the highest sensitivity was for BDe score; where 74% of the cases observed to be slight were also predicted to be slight. Although the BDe was capable of classifying 74% of the slight injured correctly, its specificity results indicated that its ability to classify killed or seriously injured were relatively poor. None of the score metrics achieved good results regarding the classification of killed or seriously injured (specificity); the best was for MDL and AIC scores, and test dataset with 53% of correctly classified killed or seriously injured.

The results of sensitivity for all the score metrics were relatively better than those of specificity, thus indicating that the models were able to classify slight injured rather than killed or seriously injured. This, however, was expected, since the original dataset contained more slight injuries.

HMSS could be used as a single measure of performance of the BN instead of using sensitivity and specificity separately. The results indicated that the best HMSS was achieved by using MDL and AIC scores (58%).

Fig. 2 shows the ROC curves for the BNs built using the three score methods, where the *X*-axis represents (1 − specificity) and the *Y*-axis represents the sensitivity.

The best ROC area obtained by BDe and MDL scores was 62%.

Table 2 suggests that the three score metrics were valid and equally effective on average.

Following Simoncic (2004), the most convenient way to analyze the graphical performance of the three metrics is to calculate the Most Probable Explanation (MPE) for the training dataset and compare it with the results obtained from the test dataset. The training/testing dataset that showed the best results for the previous indicators was used for this purpose.

MPE is given by the most probable configuration of values for all variables in the BN. For the three estimated structures, the MPE is given by the following values for variables (see Table 1):

ACT = AS; AGE = (25–64]; ATF = GW; CAU = DC; DAY = WD; GEN = M; LAW = WID; LIG = DL; MON = SUM; NOI = 1; OI = 2; PAS = Y; PAW = WID; ROM = SLD; SEV = SI; SHT = THI; SID = WR; TIM =(12–18]; VI = 2

Table 3
MPE for the three score metrics.

Score metric	MPE formulas	MPE	MPE_{test}
BDe	$P(ACT=AS)\cdot P(AGE=(25-64]\mid SEV=SI)\cdot P(ATF=GW\mid SEV=SI,SID=WR)\cdot$ $P(CAU=DC\mid SEV=SI)\cdot P(DAY=WD\mid SEV=SI)\cdot P(GEN=M\mid SEV=SI)\cdot$ $P(LAW=WID\mid SEV=SI)\cdot P(LIG=DL\mid SEV=SI)\cdot P(MON=SUM\mid SEV=SI,ATF=GW)\cdot$ $P(NOI=1\mid VI=2)\cdot P(OI=2)\mid SEV=SI,NOI=1,VI=2)\cdot P(PAS=Y\mid SEV=SI,SHT=THI)\cdot$ $P(PAW=WID\mid SEV=SI,LAW=WID)\cdot P(ROM=SLD\mid SEV=SI,PAS=Y,PAW=WID)\cdot$ $P(SEV=SI\mid ACT=AS,NOI=1)\cdot P(SHT=THI\mid PAW=WID)\cdot P(SID=WR\mid PAS=Y)\cdot$ $P(TIM=(12-18]\mid SEV=SI,LIG=DL)\cdot P(VI=2\mid ACT=AS)$	0.00088	0.00081
MDL	$P(ACT=AS\mid PAS=Y)\cdot P(AGE=(25-64]\mid SEV=SI)\cdot P(ATF=GW\mid SEV=SI,SID=WR)\cdot$ $P(CAU=DC\mid SEV=SI)\cdot P(DAY=WD\mid SEV=SI)\cdot P(GEN=M\mid SEV=SI)\cdot$ $P(LAW=WID\mid SEV=SI,PAW=WID)\cdot P(LIG=DL\mid SEV=SI,TIM=(12-18])\cdot$ $P(MON=SUM\mid SEV=SI,ATF=GW)\cdot P(NOI=1\mid VI=2)\cdot P(OI=2\mid SEV=SI,NOI=1,VI=2)\cdot$ $P(PAS=Y\mid SHT=THI)\cdot P(PAW=WID\mid SHT=THI)\cdot$ $P(ROM=SLD\mid SEV=SI,PAS=Y,PAW=WID)\cdot$ $P(SEV=SI\mid SHT=THI,PAS=Y,ACT=AS,NOI=1)\cdot P(SHT=THI)\cdot P(SID=WR\mid PAS=Y)\cdot$ $P(TIM=(12-18]\mid VI=2)\cdot P(VI=2\mid ACT=AS)$	0.00076	0.00073
AIC	$P(ACT=AS\mid VI=2)\cdot P(AGE=(25-64]\mid LIG=DL)\cdot P(ATF=GW\mid SID=WR)\cdot$ $P(CAU=DC\mid SEV=SI,GEN=M)\cdot P(DAY=WD\mid SEV=SI,VI=2)\cdot P(GEN=M\mid DAY=WD)\cdot$ $P(LAW=WID\mid SEV=SI,ROM=SLD,PAW=WID)\cdot$ $P(LIG=DL\mid MON=SUM,TIM=(12-18])\cdot P(MON=SUM\mid PAS=Y,ATF=GW)\cdot$ $P(NOI=1\mid AGE=(25-64],VI=2)\cdot P(OI=2\mid SEV=SI,NOI=1,VI=2)\cdot$ $P(PAS=Y\mid PAW=WID,SHT=THI)\cdot P(PAW=WID\mid SHT=THI)\cdot$ $P(ROM=SLD\mid PAS=Y,PAW=WID)\cdot$ $P(SEV=SI\mid MON=SUM,LIG=DL,ATF=GW,AGE=(25-64],NOI=1,ACT=AS)\cdot$ $P(SHT=THI\mid ACT=AS)\cdot P(SID=WR\mid PAS=Y,ROM=SLD)\cdot P(TIM=(12-18])\cdot$ $P(VI=2\mid TIM=(12-18])$	0.00100	0.00092

Given the estimated BN structures (BDe, MDL and AIC) and the conditional probabilities for each node (see Fig. 3), the probability of the MPE can be computed as shown in Table 3.

For the network built by the BDe score metric, the MPE is given by the probability values shown in Table 3, column 2, row 2. Using these values, MPE for the BDe score equals 0.00088. The same calculations for the test dataset produced $MPE_{test} = 0.00081$. This comparison of MPE and MPE_{test} can provide an indication of the quality of the estimated BN using BDe score metric; where it can be seen that there is a difference (8.2%) between the MPE produced by the training dataset and the test dataset.

The MPE for the MDL BN is given by the probability values shown in Table 3, column 2, row 3. Using these values, MPE equals 0.00076. The test dataset produced $MPE_{test} = 0.00073$. So, the MPE as explained by the MDL is closer to the test dataset estimation (4.4% of difference), thus representing a network that is more capable of explaining different data.

The MPE for the AIC BN is given by the probability values shown in Table 3, column 2, row 4. Using these values, MPE is 0.00100. The test dataset produced $MPE_{test} = 0.00092$. The most probable explanation has a higher probability than that produced by the test subset (8.7% of difference).

The conclusion from the above calculations of the MPE for the three score metrics as compared to the MPEs calculated for the test subset is that, in relative terms, the MDL score metric MPE gives the best explanation with regard to the MPE_{test}, whereas the difference between MPE of the built network and that computed for the test subset is the least among all the other MPEs produced by BDe, and AIC score metrics.

The last step in comparing the various score metrics and evaluating their performance was to compare the graphs' complexity, measured by the total number of arcs produced by the three score metrics studied.

Fig. 3 shows the number of arcs obtained by using the three score metrics. The most complicated BN (having the highest number of arcs) is the BN built using the AIC score; this BN has 35 arcs, while the least complicated BN was the BN built by the BDe score, with 28 arcs; followed by the BN built by the MDL score, with 29 arcs.

The results of building the BNs showed that the three different score metrics did not vary significantly in terms of their accuracy, specificity, sensitivity, HMSS and ROC area. This however, indicates that BNs are valid for analyzing traffic accident injury severities and builds on the results presented by Simoncic (2004), who indicated that BNs could effectively be used to analyze this specific problem.

On the other hand, the results for the complexity of the BN graphs, the number of arcs and the MPE show some differences between the three score metrics. MDL shows the best results in terms of MPE (smaller differences between training and test sets). BDe and MDL show the best results in terms of complexity of BN graphs and number of arcs.

A closer look at the results obtained by MDL score shows that it produced a network that was relatively successful in terms of classification and prediction, where it had the second best total accuracy (59–60%). Also, HMSS showed a relatively good result for both training and testing sets respectively (56–58%,) and the ROC area results were good as well (61–62%). The BN built by the MDL score is shown in Fig. 4.

Setting evidences for the variables used to build the BN using the MDL score could give indications of the values of variables that contribute to the occurrence of a killed or seriously injured (KSI) individual in a traffic accident.

Table 4 assists in the identification of the variables and values that contribute the most to the occurrence of a KSI individual in a traffic accident. For each variable, the probability of a value was set to be 1.0 (setting evidence) and the other values of the same variable were set to be 0.0. Thus, the associated probability of severity was calculated. Underlined values in Table 4 show the values of variables in which the probability of a KSI was found to be higher than that of SI.

For example, this table shows that assigning a probability of 1.0 to the value AS (angle or side impact) of the variable ACT, the probability of SI becomes 0.6219 and the probability of KSI becomes 0.3780. These probabilities are calculated from the conditional probability table of the BN built using the MDL score. Since it is intended to determine which values of variables contribute the most to the occurrence of a KSI individual in a traffic accident, Table 4 does not include the variables in which the values of probabilities of SI are always higher than those of KSI.

Table 4
Inference results for variables that are associated with KSI in traffic accidents.

Variables	Values	Probabilities when setting evidences	
		SI	KSI
ACT	AS	0.6219	0.3780
	CF	0.5226	0.4773
	HO	0.3412	0.6587
	O	0.5808	0.4191
	PU	0.6683	0.3316
	R	0.4944	0.5055
	SP	0.6066	0.3933
AGE	[18–25]	0.4999	0.5000
	(25–64]	0.5567	0.4432
	≥64	0.5937	0.4062
LIG	D	0.5486	0.4513
	DL	0.5615	0.4384
	I	0.6239	0.3760
	S	0.6254	0.3745
	W	0.4527	0.5472
NOI	1	0.4957	0.5042
	>1	0.6545	0.3454

SI: slight injured; KSI: killed or seriously injured.

Setting evidences for the values of variables used to build the BN indicated that ACT, AGE, LIG and NOI were found to be significant.

A detailed discussion of the most significant variables that were found to contribute to the occurrence of a killed or seriously injured (KSI) individual in a traffic accident is given below.

3.1. *Accident type (ACT)*

As shown in Table 4, when setting the probabilities of both HO (head on collisions) and R (rollover) values to be equal to 1.0, the probability of having KSI accidents increased, which means that these types of accidents are more significant in accidents with killed or seriously injured. Kockelman and Kweon (2002) found that head on crashes were more dangerous than angle crashes, left-side, and right-side crashes; they also found that they were significant in accidents that involved killed or seriously injured, but rollover crashes were more dangerous than all of the preceding crash types.

3.2. *Age (AGE)*

The results shown in Table 4 indicate that drivers in the age group [18–25] years were found to be more involved in accidents that resulted in KSI. Tavris et al. (2001) found that male drivers in the age group (16–24) years were much more likely to be involved in killed or seriously injured accidents than those involving older drivers.

3.3. *Lighting (LIG)*

Gray et al. (2008) found that among the factors that lead to a slight injury is driving in the daylight, and that more severe injuries are predicted during darkness. Helai et al. (2008) and Abdel-Aty (2003) found the same results. This coincides with the results found in this study, which indicate that roadways Without lighting (W) are associated with accidents that had KSI individuals.

3.4. *Number of injuries (NOI)*

The results obtained in this study indicate that when an accident results in one injury, it is more likely to be a serious injury or even fatal. Scheetz et al. (2009) used classification and regression trees to model the injury severity of traffic accidents. They also found that the number of injured occupants was a significant factor in classifying injury severity.

4. Limitations of the study

Before conclusions, some limitations should be pointed out:

- The need for large datasets when working with Bayesian networks, and the effect that imbalanced dataset (slight injured versus killed or seriously injured) has on both sensitivity and specificity.
- The data collection is based on the standard traffic police report used in Spain. So, the variable cause of the accident (CAU) was determined and judged based on the experience of the traffic police. However, a different person might have determined the same cause differently, since different time and person might lead to a different judgment.

5. Summary and conclusions

This paper uses BNs to analyze traffic accident data in order to validate the ability of this data-mining technique to classify traffic accidents according to their injury severity, and to identify the significant factors that are associated with KSI in traffic accidents.

Traffic accident data was obtained from the DGT for a period of three years (2003–2005) for Granada (Spain). Three BNs were built using three different score metrics: BDe, MDL and AIC.

Several indicators have been used in order to evaluate the performance of the built BNs: accuracy, sensitivity, specificity, HMSS, ROC Area, MPE and graph complexity (or number of arcs). The results obtained for these indicators do not vary significantly between the different score metrics used and they are within the range of previous studies (Abdel Wahab and Abdel-Aty, 2001; Simoncic, 2004). So, it could be concluded that BNs might be a useful tool for classifying traffic accidents according to their injury severity.

Inference was used to identify the values of the variables that are associated with KSI in traffic accidents on Spanish rural highways. Based on the results, it would be possible to identify the type of accident that would most probably be classified as KSI on Spanish rural highways. It would be a head-on or rollover traffic accident in a roadway without lighting with only one injury within the age of 18 and 25 years. These factors (head-on or rollover, unlit roadway, only one injury and within the age of 18 and 25 years) do not have to exist all at once in order to have a KSI accident. Any of these or a combination of them might increase the probability of a KSI accident. In general, these results are consistent with the literature (Tavris et al., 2001; Kockelman and Kweon, 2002; Abdel-Aty, 2003; Helai et al., 2008; Gray et al., 2008; Scheetz et al., 2009). However, this finding may vary for other countries and datasets.

BNs, which have proved their effectiveness in different research areas, could be usefully applied in the domain of traffic accident modeling. Their effectiveness has been found to be similar to other data-mining techniques used to model severity in traffic accidents. Compared with other well-known statistical methods, the main advantage of the BNs seems to be their complex approach where system variables are interdependent and where no dependent and independent variables are needed (Simoncic, 2004).

Acknowledgements

The authors are grateful to the Spanish General Directorate of Traffic (DGT) for supporting this research and offering all the resources that are available to them. The authors appreciate the reviewers' comments and effort in order to improve the paper.

References

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

Abdel wahab, H.T., Abdel-Aty, M.A., 2001. Development of artificial neural network models to predict driver injury severity in traffic accidents at signalized intersections. Transportation Research Record 1746, 6–13.

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

Bédard, M., Guyatt, G.H., Stones, M.J., Hirdes, J.P., 2002. The independent contribution of driver, crash, and vehicle characteristics to driver fatalities. Accident Analysis and Prevention 34, 717–727.

Bouckaert, R.R., 1995. Bayesian Belief Networks: From Construction to Inference. Ph.D. Thesis, University of Utrecht.

Chang, L.Y., Wang, H.W., 2006. Analysis of traffic injury severity: an application of non-parametric classification tree techniques. Accident Analysis and Prevention 38, 1019–1027.

Cooper, G.F., Herskovits, E., 1992. A Bayesian method for the induction of probabilistic networks from data. Machine Learning 9, 309–347.

Cruz-Ramírez, N., Acosta-Mesa, H.G., Carrillo-Calvet, H., Nava-Fernández, L.A., Barrientos-Martínez, R.E., 2007. Diagnosis of breast cancer using Bayesian networks: a case study. Computers in Biology and Medicine 37, 1553–1564.

Dirección General de Tráfico – DGT [online]. Available from World Wide Web: http://www.dgt.es/portal/es/seguridad_vial/estadistica/accidentes_30dias/anuario_estadistico/.

Gray, R.C., Quddus, M.A., Evans, A., 2008. Injury severity analysis of accidents involving young male drivers in Great Britain. Journal of Safety Research 39, 483–495.

Gregoriades, A., 2007. Towards a user-centred road safety management method based on road traffic simulation. In: Proceedings of the 39th Conference on Winter Simulation: 40 years! The Best is Yet to come, Washington, DC, pp. 1905–1914.

Helai, H., Chor, C.H., Haque, M.M., 2008. Severity of driver injury and vehicle damage in traffic crashes at intersections: a Bayesian hierarchical analysis. Accident Analysis and Prevention 40, 45–54.

Kockelman, K.M., Kweon, Y.J., 2002. Driver injury severity: an application of ordered probit models. Accident Analysis and Prevention 34, 313–321.

Kopelias, P., Papadimitriou, F., Papandreou, K., Prevedouros, P., 2007. Urban freeway crash analysis. Transportation Research Record 2015, 123–131.

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

Milton, J.C., Shankar, V.N., Mannering, F.L., 2008. Highway accident severities and the mixed logit model: an exploratory empirical analysis. Accident Analysis and Prevention 40, 260–266.

Mittal, A., Kassim, A., Tan, T., 2007. Bayesian Network Technologies: Applications and Graphical Models. IGI Publishing, New York.

Neapolitan, R.E., 2009. Probabilistic Methods for Bioinformatics. Morgan Kaufmann Publishers, San Francisco, CA.

Ozbay, K., Noyan, N., 2006. Estimation of incident clearance times using Bayesian Networks approach. Accident Analysis and Prevention 38, 542–555.

Pearl, J., 2004. Probabilistic Reasoning in Intelligent Systems. Morgan Kaufmann, San Francisco, CA.

Scheetz, L.J., Zhang, J., Kolassa, J., 2009. Classification tree to identify severe and moderate injuries in young and middle aged adults. Artificial Intelligence in Medicine 45, 1–10.

Simoncic, M., 2004. A Bayesian network model of two-car accidents. Journal of transportation and Statistics 7 (2/3), 13–25.

Tavris, D.R., Kuhn, E.M., Layde, P.M., 2001. Age and gender patterns in motor vehicle crash injuries: importance of type of crash and occupant role. Accident Analysis and Prevention 33, 167–172.

Witten, I.H., Frank, E., 2005. Data Mining: Practical Machine Learning Tools and Techniques, 2nd ed. Morgan Kaufmann, San Francisco, CA.

Yamamoto, T., Shankar, V.N., 2004. Bivariate ordered-response probit model of driver's and passenger's injury severities in collisions with fixed objects. Accident Analysis and Prevention 36, 869–876.

Yau, K.K.W., Lo, H.P., Fung, S.H.H., 2006. Multiple-vehicle traffic accidents in Hong Kong. Accident Analysis and Prevention 38, 1157–1161.

3. A method for simplifying the analysis of traffic accidents injury severity on two-lane highways using Bayesian networks.

Randa Oqab Mujalli and Juan de Oña. A method for simplifying the analysis of traffic accidents injury severity on two-lane highways using Bayesian networks. Paper accepted in ***Journal of Safety Research***

This paper presents a method for reducing the number of variables used in order to model injury severity of traffic accidents using a set of algorithms. All the possible combinations of evaluator-search algorithms were applied (59 combinations).

The work developed in this paper is divided into two stages, the first stage is performed based on selecting subsets of variables using all the existing variables' selection algorithms.

11 subsets of selected variables representing (4, 5, 6, 7, 8, 9, 11, 12, 14, 15 and 16) variables in each subset were obtained. In which some subsets constituted of different combinations of variables for the same number of variables used. Twenty six different Bayesian networks were built using these variables, where these Bayesian networks were compared against the original network built using all the variables.

The comparison criterion is based on comparing the performance indicators (accuracy, sensitivity, specificity, HMSS and ROC area) between the original network and the networks built using the selected subsets if variables. The Bayesian networks that improve the indicators' values are further analyzed in order to for identify the most significant variables (accident type, age, atmospheric factors, gender, lighting, number of injured and occupant involved).

The second stage is composed of building a new Bayesian networks using the selected variables that showed a statistical significant improvement with respect to the original Bayesian network in the first stage. The results of the performance indicators indicated in most of the cases, a statistically significant improvement with respect to the original Bayesian network. In which using only seven variables (accident type, age, atmospheric factors, gender, lighting, number of injured and occupant involved) a Bayesian network could be built to classify the severity of a traffic accident.

-----Mensaje original-----
De: ees.jsr.0.11f9ff.d7bfb65b@eesmail.elsevier.com
[mailto:ees.jsr.0.11f9ff.d7bfb65b@eesmail.elsevier.com] En nombre de Journal
of Safety Research
Enviado el: martes, 21 de junio de 2011 17:57
Para: jdona@ugr.es
Asunto: Your Submission

Ms. Ref. No.: JSR-D-11-00053R1

Dear Juan,

Your revised manuscript, A method for simplifying the analysis of traffic accidents injury severity on two-lane highways using Bayesian networks, has been accepted and is being incorporated into our schedule for publication in a future issue of the Journal of Safety Research.

Below are comments from the editor and reviewers.

Thank you for giving us the opportunity to publish your work.

Sincerely,

Thomas W. Planek, Ph.D., Editor-in-Chief
Mei-Li Lin, Ph.D., Editor
Kathleen Porretta, Managing Editor

Comments from the editors and reviewers:

Reviewer #1: PLEASE ANSWER THE FOLLOWING QUESTIONS:
(Please place an "X" next to the appropriate answer below)

1. Is the subject relevant? Significant?
X Yes ___ No

2. Are the problem and proposed solution(s) clearly stated?
X Yes ___ No

3. Is the research design appropriate and adequately defined?
X Yes ___ No

4. Is the data analysis correct? Are statistical tests of significance appropriate and used correctly?
X Yes ___ No

5. Are the conclusions adequately supported by the findings?
X Yes ___ No

6. Is the paper well organized and logically presented?
X Yes ___ No

7. Is there adequate referencing?
X Yes ___ No

8. Are the tables and figures accurate and appropriate?

X Yes ___ No

9. Is the paper:
X Acceptable in its present form
___ Acceptable with revisions
___ Not acceptable in present form-resubmit
___ Not acceptable

I believe my comments have all been addressed and I have no further comments.

****
For further assistance, please visit our customer support site at http://support.elsevier.com. Here you can search for solutions on a range of topics, find answers to frequently asked questions and learn more about EES via interactive tutorials. You will also find our 24/7 support contact details should you need any further assistance from one of our customer support representatives.

A method for simplifying the analysis of traffic accidents injury severity on two-lane highways using Bayesian networks

Randa Oqab Mujalli and Juan de Oña[#]

TRYSE Research Group. Department of Civil Engineering, University of Granada

[#] Corresponding author, ETSI Caminos, Canales y Puertos, c/ Severo Ochoa, s/n, 18071 Granada (Spain), Phone: +34 958 24 99 79, email: jdona@ugr.es

Abstract

Introduction: This study describes a method for reducing the number of variables frequently considered in modeling the severity of traffic accidents. The method's efficiency is assessed by constructing Bayesian networks (BN). *Method:* It is based on a two stage selection process. Several variable selection algorithms, commonly used in data mining, are applied in order to select subsets of variables. BNs are built using the selected subsets and their performance is compared with the original BN (with all the variables) using five indicators. The BNs that improve the indicators' values are further analyzed for identifying the most significant variables (accident type, age, atmospheric factors, gender, lighting, number of injured and occupant involved). A new BN is built using these variables, where the results of the indicators indicate in most of the cases, a statistically significant improvement with respect to the original BN. *Conclusions:* It is possible to reduce the number of variables used to model traffic accidents injury severity through BNs without reducing the performance of the model. *Impact on Industry:* The study provides the safety analysts a methodology that could be used to minimize the number of variables used in order to determine efficiently the injury severity of traffic accidents without reducing the performance of the model.

Keywords: injury severity, variable selection, Bayesian networks, data mining, classification

1. Introduction

A lot of information on traffic accidents exists, extracted from different sources, in which many variables that are expected to affect injury severity in traffic accidents are considered. The number of variables used in research work could be enormous, and in some cases this number could be even higher than 100 variables (Delen et. al., 2006). This might complicate the manner of dealing with a certain problem, where some of the variables considered might hide the effect of other more significant ones. A lot of different types of studies tried to identify the most significant variables in order to only consider them in the analysis of traffic accidents (Xie et al., 2009; Kopelias et al., 2007; Chang and Wang, 2006; Chen and Jovanis, 2000). Therefore, researchers in the field of traffic accidents and specifically in the domain of traffic accident injury severity focused their research on trying to identify the most significant variables that contribute to the occurrence of a specific injury severity in a traffic accident.

Most previous research utilized regression analysis techniques, such as logistic and ordered probit models (Al-Ghamdi, 2002; Milton et al. 2008; Bédard et al. 2002; Yau et al., 2006; Yamamoto and Shankar, 2004; Kockelman and Kweon, 2002). These techniques have their own drawbacks. Chang and Wang (2006) indicated that these regression models use certain assumptions, and if any of these assumptions were violated, the ability of the model to predict the factors that contribute to the occurrence of a specific injury severity would be affected.

Recently researchers used data mining techniques, such as artificial neural networks, regression trees and Bayesian networks.

For instance, Abdelwahab and Abdel-Aty (2001) used artificial neural networks to model the relationship between driver injury severity and crash factors related to driver, vehicle, roadway, and environment characteristics. Thirteen variables were tested first for significance using the χ^2 test, and the results indicated that only six variables were found to be significant: driver gender, fault, vehicle type, seat belt, point of impact, and area type. They compared the classification performance of Multi-Layer Perceptron (MLP) neural networks and that of the Ordered Probit Model (OPM). Their findings indicated that classification accuracy of MLP neural networks outperformed that of the OPM, where 65.6% and 60.4% of cases were correctly classified for the training and testing phases, respectively, compared to 58.9% and 57.1% correctly classified cases for the training and testing phases, respectively, by the OPM.

Another study that used the neural networks to model injury severity in traffic accidents (Delen et. al., 2006) classified the injury severity of a traffic accident into five categories (no injury, possible injury, minor non-incapacitating injury, incapacitating and fatality) and they used certain techniques, such as χ^2 test, stepwise logistic regression and decision tree induction to select the most significant variables. Out of 150 variables, they selected 17 variables as important in influencing the level of injury severity of drivers involved in accidents. They used the MLP neural networks to classify the injury severity level, where their data included "no injury" cases 10 times more than "fatal cases"; they faced an unbalanced dataset situation which affected their total accuracy (40.71%).

Other researchers used classification tree techniques to model injury severity in traffic accidents (Chang and Wang, 2006). In their study they developed a Classification and Regression Tree (CART) model to establish the relationship between injury severity and twenty explanatory variables that represented: driver/vehicle characteristics, highway/environmental variables and accident variables, where they aimed to model the injury severity of an individual involved in a traffic accident.

Use of Bayesian Networks (BN) as the modeling approach in analysis of crash-related injury severity has been relatively scarce. De Oña et al. (2011) employed BN to model the relationship between injury severity and 18 variables related to driver, vehicle, roadway, and environment characteristics.

Some of these studies tend to apply the models on the datasets without selecting the most significant variables (Delen et. al., 2006; Chang and Wang 2006; Simoncic, 2004). However, Chang and Wang (2006) stated that if the model was applied on a few important variables, much more useful results could be obtained. Others like Abdelwahab and Abdel-Aty (2001) used some statistical techniques to choose the most significant variables before applying their model.

The scope of this research is to build BNs using some selected variables in order to evaluate the performance of BNs when using only the most significant variables, and to compare the results with a base model that is built using all the variables in the original dataset, in order to find out whether using only the most significant variables would affect values of the measures used to assess the built model.

This paper is organized as follows. Section 2 presents the data used. In section 3, the method followed is presented and described, and a brief review of variable selection methods and the basic concept of BNs are presented, along with a description of the performance indicators used to assess the performance of the built BNs. In section 4, the results and their discussion are provided. In section 5, some conclusions are given.

2. Accident data

Accident data were obtained from the Spanish General Traffic Directorate (DGT) for rural highways in the province of Granada (southern Spain) for three years (2003-2005). The total number of accidents obtained for this period was 3,302. The data was first checked out for questionable data, and those which were found to be unrealistic were screened out. Only rural highways were considered in this study; data related to intersections were not included, since intersections have their own specific characteristics and need to be analyzed separately. Finally, the database used to conduct the study contained 1,536 records. Table 1 provides information on the data used for this study.

(insert Table 1)

Eighteen (18) variables were used with the class variable of injury severity (SEV) in an attempt to identify the important variables that affect injury severity in traffic accidents.

The data contained information related to the accidents and other information related to the drivers.

The data included variables describing the conditions that contributed to the accident and injury severity.

- Injury severity variables: number of injuries (e.g., passengers, drivers and pedestrians), severity level of injuries (e.g., slight injured –SI– and killed or seriously injured –KSI). Following previous studies (Chang and Wang, 2006; Milton et al., 2008) the injury severity of an accident is determined according to the level of injury to the worst injured occupant.
- Roadway information: characteristics of the roadway on which the accidents occurred (e.g., pavement width, lane width, shoulder type, pavement markings, sight distance, if the shoulder was paved or not, etc.)
- Weather information: weather conditions when the accident occurred (e.g., good weather, rain, fog, snow and windy)
- Accident information: contributing circumstances (e.g., type of accident, time of accident (hour, day, month and year), and vehicles involved in the accident).
- Driver data: characteristics of the driver, such as age or gender

3. Method

The procedure used in this study has been the following:

1. The original dataset obtained from the DGT was divided into two subsets: a training set containing 2/3 of the data (1,024 records), and a testing set containing the rest of the data (512 records). The testing set was used to validate the results obtained using the training set.
2. Based on the eighteen variables taken from the accident reports (see Table 1), identification of the variables that affect injury severity in traffic accidents was performed using different methods of evaluator-search algorithms.
3. For each one of the selected subsets of variables, ten simplified BNs were built using the hill climbing search algorithm and the MDL score (De Oña et al., 2011).
4. The performance of the built BNs using the selected subsets of variables was compared with the performance of the original BN which was built using the eighteen variables (BN-18). Five performance evaluation indicators were used.
5. Of all the simplified built BNs, the selected ones are those whose results improve or maintain the results obtained by the performance indicators of BN-18 in 90% of the cases or more, and whose improvements are statistically significant.

6. For the selected BNs, the variables that repeat in more than 50% of the cases are identified and a new BN is built using these variables.
7. Finally, the results obtained by this new BN, based on a double process of variable selection procedure, are compared with those obtained by BN-18.

3.1. Variable selection methods

In machine learning, variable selection is a process that is used to select a subset of variables and to remove variables that do not contribute to the performance of the machine learning technique used.

In this study, we used six evaluators with eleven search methods. Weka's Select Variable Panel (Witten and Frank, 2005) was used to perform the variable selection.

A brief description of each of the evaluators used is given below:

1. Correlation-based variable selection (CfsSubsetEval): this evaluator measures the predictive ability of each variable individually and the degree of redundancy among them. It selects the sets of variables that are highly correlated with the class but have low inter-correlation with each other (Hall, 1998).
2. Consistency-based variable selection (ConsistencySubsetEval): this evaluator measures the degree of consistency of the variable sets in class values when the training values are projected onto the set. This evaluator is usually used in conjunction with a random or exhaustive search (Liu and Setiono, 1996).
3. Classifier Subset Evaluator (ClassifierSubsetEval): this evaluator uses the classifier specified in the object editor as a parameter, to evaluate sets of variables on the training data or on a separate holdout set (Witten and Frank, 2005).
4. Wrapper Subset Evaluator (WrapperSubsetEval): this evaluator uses a classifier to evaluate variable sets and it employs cross-validation to estimate the accuracy of the learning scheme for each set (Khhavi and John, 1997).
5. Filtered Subset Evaluator (FilteredSubsetEval): The filter model evaluates the subset of variables by examining the intrinsic characteristic of the data without involving any data mining algorithm (Witten and Frank, 2005).
6. Cost Sensitive Subset Evaluator (CostSensitiveSubsetEval): This evaluator projects the training set into attribute set and measure consistency in class values, making the subset cost sensitive (Liu and Setiono, 1996).

A brief description of each of the search methods used is given below:

1. Best First: this Search method uses greedy hillclimbing augmented with a backtracking facility to search through the variables' subsets. Best first may start with an empty set of variables and search forward, or start with a full set of variables and search backward, or start at any point and search in both directions (Pearl, 1984).
2. Genetic Search: An initial population is formed by generating many individual solutions. During each successive generation, a proportion of the existing population is selected to breed a new generation. This process is repeated until increasing the average fitness, and reaching a termination condition (Goldberg, 1989).
3. Greedy Stepwise: Performs a greedy forward or backward search through the space of variables' subsets. The search could be initiated with none or with all the variables or from an arbitrary point in the space. Thus, the search is stopped when the addition or deletion of any variables that remains; results in a decrease in evaluation (Russell and Norvig, 2003).

4. Linear Forward Selection: this search method is an extension of Best First where a fixed number k of variables is selected, whereas k is increased in each step when fixed-width is selected. The search direction can be forward or floating forward selection (with optional backward search steps) (Gutlein et al., 2009).
5. Scatter Search V1: Starts with a population of many significant variables and stops when the result is higher than a given threshold or when no further improvement could be attained (García López et al.,2006).
6. Tabu Search: It explores the solution space beyond the local optimum, once a local optimum is reached; upward moves and those worsening the solutions are allowed (Hedar, 2008).
7. Rank Search: Uses a variable/subset evaluator to rank all variables. If a subset evaluator is specified then a forward selection search is used to generate a ranked list. From the ranked list of variables, subsets of increasing size are evaluated (Witten and Frank, 2005).
8. Exhaustive Search: Performs an exhaustive search through the space of variables' subsets starting from the empty set of variables and reporting the best subset found (Witten and Frank, 2005).
9. Subset Size Forward Selection: Performs an interior cross-validation, where it is performeded on each fold to determine the optimal subset-size. In the final step the search is performed on the whole data (Gutlein et al., 2009).
10. Random Search: Performs a random search in the space of variables' subsets. A random search is started from a random point, if no initial point is chosen, and reports the best subset found. If a start set is set, Random searches randomly for subsets that are as good as or better than the start point with the same number of variables or with a lower number of variables (Liu and Setiono, 1996).
11. Race Search: Races the cross validation error of competing subsets of variables, and is only used with a ClassifierSubsetEval (Moore and Lee,1994).

3.2. Bayesian Networks definition

Over the last decade, BNs have become a popular representation for encoding uncertain expert knowledge in expert systems. The field of BNs has grown enormously, with theoretical and computational developments in many areas (Mittal et al., 2007) such as: modeling knowledge in bioinformatics, medicine, document classification, information retrieval, image processing, data fusion, decision support systems, engineering, gaming, and law.

Let $U=\{x_1, \ldots, x_n\}$, $n\geq 1$ be a set of variables. A BN over a set of variables U is a network structure, which is a Directed Acyclic Graph (DAG) over U and a set of probability tables $B_p = \{p(x_i|pa(x_i), x_i \in U)\}$ where $pa(x_i)$ is the set of parents or antecedents of x_i in BN and $i=(1,2,3,\ldots,n)$. A BN represents joint probability distributions $P(U) = \prod_{x_i \in U} p\big(x_i|pa(x_i)\big)$.

Based on the theory of Bayesian networks (Neapolitan, 2004) the relations between variables, represented by arcs in the graph, could represent causality, relevance or relations of direct dependence between variables. In accordance with other authors (Acid et al., 2004), we do not assume a causal intrpretation of the arcs in the networks, although in some cases this might be reasonable (other approaches that explicitly try to detect causal influence are discussed in Glymour et al. (1999) and Pearl (2000)). Instead, the arcs are interpreted as direct dependence relationships between the linked variables, and the absence of arcs means the absence of direct dependence between variables, however indirect dependence relationships between variables could exist.

The classification task consists in classifying a variable $y = x_0$ called the class variable, given a set of variables $U = x_1 \ldots x_n$, called attribute variables. A classifier $h : U \rightarrow y$ is a function that maps an instance of *U* to a value of *y*. The classifier is learned from a dataset D consisting of samples over *(U, y)*. The learning task consists of finding an appropriate BN given a data set D over *U*.

For each one of the variable subsets selected in the previous step, BNs were built using the training dataset, the hill climbing search algorithm and the MDL score. The search algorithm and the score were applied in this study mainly because they are fast and widely used, and also produce good results in terms of network complexity and accuracy (Madden, 2009).

3.3. Performance Evaluation Indicators

In order to measure the performance of the BNs built for each one of the variable subsets with the training data several indicators were used. The performance evaluation indicators used in this study were accuracy, specificity, sensitivity, the harmonic mean of sensitivity and specificity (HMSS) and the ROC area.

$$\text{Accuracy} = \frac{\text{tSI+tKSI}}{\text{tSI+tKSI+fSI+fKSI}} 100\% \quad (1)$$

$$\text{Sensitivity} = \frac{\text{tSI}}{\text{tSI+fKSI}} 100\% \quad (2)$$

$$\text{Specificity} = \frac{\text{tKSI}}{\text{tKSI+fSI}} 100\% \quad (3)$$

$$\text{HMSS} = \frac{2 \times \text{sensitivity} \times \text{specificity}}{\text{sensitivity+specificity}} \quad (4)$$

Where *tSI* is true slight injured cases, *tKSI* true killed or seriously injured cases, *fSI* false slight injured cases, and *fKSI* false killed or seriously injured cases.

Accuracy (see Eq. 1) is proportion of instances that were correctly classified by the classifier. Accuracy only gives information on the classifier’s general performance. In some cases the accuracy might be high because the classifier is able to classify values that belonged to only one class correctly. Because the dataset used herein had an imbalanced distribution of KSI and SI, the overall accuracy alone is somewhat misleading. In order to assess the performance of the classifier, other measures should be used along with the overall accuracy, such as the quantities presented in Eqs. (2-4).

Sensitivity represents the proportion of correctly predicted SI among all the observed SI. Specificity represents the proportion of correctly predicted KSI among all the observed KSI (see Eqs. 2 and 3). Another measure used to assess the performance of the BN built was the Harmonic Mean of Sensitivity and Specificity (HMSS), which gives an equal weight of both sensitivity and specificity (see Eq. 4).

However, there is a trade-off between sensitivity and specificity, meaning it is necessary to calculate another evaluation method. Therefore, we used the area under a Receiver Operating Characteristic (ROC) curve as a target performance method. What ROC curves represent is the true positive rate (sensitivity) vs. the false positive rate (1-specificity). ROC curves are more useful as descriptors of overall performance, reflected by the area under the curve, with a maximum of 1.00 describing a perfect test and an ROC area of 0.50 describing a valueless test.

Finally, to validate the results, the measures described above are also calculated for each BN and the testing dataset.

4. Results and Discussion

All the possible combinations of evaluator-search algorithms described in section 3.1 were applied (59 combinations). The total number of combinations was supposed to be 66; however, seven combinations were found to be incompatible. Table 2 shows the unused combinations.

(insert table 2)

Table 3 shows the variables selected after running the 59 different combinations of the evaluator-search algorithms. The number of selected variables lies between four variables (ACT, ATF, LIG and NOI), as obtained using Correlation-based variable selection, Filtered Subset Evaluator and Cost Sensitive Subset Evaluator with several search methods, and a maximum number of sixteen selected variables.

(insert table 3)

The same subset of variables could be selected by several evaluator-search combinations. As an example, one of the subset composed of four variables (ACT, ATF, LIG and NOI) is selected by seventeen different combinations of evaluator-search algorithms.

(insert table 4)

Table 4 shows the number of times that each one of the eighteen variables has been selected. There are three variables that were selected approximately 95% of times. The variables ACT (accident type) and LIG (lighting) were selected 58 times over 59 combinations, which mean that they were selected almost by all the evaluator-search combinations. NOI (number of injuries) was the third most selected variable (56 times). The forth most selected variable was ATF (atmospheric factors) with 42 times. On the other hand, the least selected variable was ROM (pavement markings), with only 5 times.

For each subset of selected variables 20 BNs were built representing 10 runs for the training set, and 10 for the testing set, for each one of the selected variable groups (26 groups) and for the eighteen original variables. In total, 540 BNs have been built for this analysis. The averages of the performance evaluation indicators described in section 3.3 are calculated for each one of these BNs.

(insert table 5)

Table 5 shows the average results of these indicators for the 27 BNs for the training and the testing sets of data. With respect to the results obtained by the training set, the following findings could be highlighted:

- The values obtained for the all the performance indicators lie in the range between 0.42 and 0.76. These values however, are in the range of values obtained by other researchers (Delen et al., 2006; Abdelwahab and Abdel-Aty, 2001).
- The highest values obtained for the indicators of Sensitivity and ROC area are 0.69-0.76 for the former, and in the range of 0.60-0.65 for the latter.
- The worst results were obtained for Specificity (between 0.42 and 0.49). This indicator is used to measure the ability of the model to classify the KSI cases. Since the number of SI cases is higher than the number of KSI cases, and thus the BNs are a data mining technique, better results are obtained for larger groups (SI in this case) from those obtained for smaller groups (Chang and Wang, 2006).
- As Accuracy and HMSS are indicators that take into account both Sensitivity and Specificity, their results are intermediate, ranging between 0.59 and 0.62 for Accuracy and 0.54 and 0.65 for HMSS.

In most of the cases (74%), the values of the performance indicators obtained for the simplified BNs maintain or improve the results as compared to those of BN-18. Average values for each of the 27 BNs (training and testing sets) were tested for statistical significance ($p<0.05$) using least significant difference (LSD) ANOVA test.

Table 5 shows 7 BNs (BN-6a, BN-7c, BN-8a, BN-9a, BN-9b, BN-9c and BN-11b) that present statistically significant improvements in their performance indicators with respect to BN-18, with only a worsened value in one of their indicators. Having a closer look at these 7 BNs (see Table 3) it is observed that there are 7 variables (ACT, AGE, ATF, GEN, LIG, NOI, and OI) that repeat in more than 50% of these 7 BNs.

None of the previously built BNs was formed using this set of variables; therefore, a new BN (BN-7) was built using these 7 variables. Table 6 shows the average values of the indicators for this new BN. The results of all the performance indicators of this BN improve with respect to BN-18 except for one indicator (specificity for the test set), and these improvements are statistically significant ($p<0.05$) in 60% of the cases (accuracy, sensitivity and ROC area).

(insert Table 6)

Thus, it could be said that a simplified BN has been identified (BN-7), with only seven variables, whose results are similar or even better than those obtained by the original BN (BN-18) which includes all variables obtained from the police accidents report.

Figure 1 shows the structure of both BN-18 and BN-7. It shows that the complexity of the built BN is relieved when built using the subset of seven variables. The indicator used to measure the complexity of the built network is the number of arcs. The number of arcs was found to be 30 in the original BN (BN-18), whereas it was decreased to 9 arcs for the network of BN-7. The arcs in the networks indicate the existence of a relationship between variables. Where the arcs in a BN are not necessarily causal, that is, a BN can satisfy the probability distribution of the variables in the BN without the arcs being causal (Neapolitan, 2009). Thus, the arcs between variables in a non causal BN could imply a sort of interrelationship(s) among these variables.

The structure of the BN-7 network is similar to the network structure built using the 18 variables (BN-18), keeping in mind that 11 variables have disappeared in BN-7. Thus, the structure of relationships between the variables SEV-ACT, SEV-LIG, SEV-GEN, SEV-OI, SEV-AGE, SEV-ATF, NOI-OI are the same in both BN-18 and BN-7, except for the two new connections between SEV-NOI and ACT-OI that appeared in BN-7.

The variables that appear in BN-7 (accident type, age, atmospheric factors, gender, lighting, number of injuries and occupants involved) could be considered the ones that significantly affect the injury severity in a traffic accident.

(insert Table 7)

Setting evidences for the variables in BN-7 could give indications of the values of variables that contribute to the occurrence of a killed or seriously injured (KSI) individual in a traffic accident. Table 7 assists in the identification of the variables and values that contribute the most to the occurrence of a KSI individual in a traffic accident. For each variable, the probability of a value was set to be 1 (setting simple evidence). Thus, the associated probability of severity was calculated.

Bold values in Table 7 highlight the highest probability values of a KSI for each variable. For example, Table 7 shows that the highest probability of KSI occurs when the lighting is insufficient

(LIG=I). BN-7 allows predicting that the probability of having a KSI accident is 58.91% if LIG=I. For the other six variables (ACT, AGE, ATF, GEN, NOI and OI), when information is only available about the variable itself, the highest probability of SEV=KSI occurs in the case of: ACT=head on; AGE=[18-25]; ATF=good weather; GEN=male; NOI=1; and OI=1.

Although some of these results could seem to be strange (for example, the probability of having KSI is higher when NOI=1 or when OI=1, instead of increasing the severity as NOI or OI increases), they are consistent with other results found in the literature:

- Kockelman et al. (2002) found accident type to be one of the significant variables that affect the injury severity of traffic accidents. They found that head on crashes were more dangerous than angle crashes, left-side, and right-side crashes; they also found that they were significant in accidents that involved killed or seriously injured.

- Age was found to be a significant variable affecting the injury severity of traffic accidents by Tavris et al. (2001). They also found that male drivers in the age group (16–24) years were much more likely to be involved in killed or seriously injured accidents than those involving older drivers.

- Xie et al. (2009) found that adverse weather can actually lead to lower probability of suffering the most severe category of injuries. They explained their results by the fact that under such conditions, drivers tend to drive at lower speeds and be more cautious. They also found gender to be a significant variable; their results indicated that the chance for male drivers to suffer the most severe category of injuries is less than female drivers under the same crash circumstances. Their results coincided with the results found by Kockelman et al. (2002).

- Lighting has been found to be a significant variable defining injury severity in traffic accidents in several studies (Abdel-Aty, 2003; Helay et al., 2007 and Gray et al., 2008). They have found that more severe injuries are predicted during darkness.

- Scheetz et al. (2009) found that the number of injured occupants was a significant factor in classifying injury severity.

- Occupant involved in a traffic accident was found to be a significant variable by Dupont et al. (2010). They found that the higher the number of vehicles involved in the accident and the level of occupancy of these vehicles, the higher the probability for each car occupant to survive.

However, relationships between the variables and injury severity in traffic accidents are more subtle. The effect of variable’s value does not always lead to the same outcome (e.g. not always that OI decreases the probability of KSI decreases). Simplified BN allow and facilitate analyzing these subtleties. For example, Table 7 shows that, in general, the probability of a KSI accident decreases with the number of OI:

P(SEV=KSI │ OI=1)=51.00%

P(SEV=KSI │ OI=2)=42.50%

P(SEV=KSI │ OI=>2)=42.13%

However, Table 8 shows that the probability of SEV=KSI in the case of HO accidents that occurred in conditions of insufficient lighting (LIG=I) and with only one injury (NOI=1) increases with the number of OI:

P(SEV=KSI │ LIG=I, ACT=HO, NOI=1, OI=1)=57.13%

P(SEV=KSI │ LIG=I, ACT=HO, NOI=1, OI=2)=67.14%

P(SEV=KSI │ LIG=I, ACT=HO, NOI=1, OI=>2)=95.02%

(insert Table 8)

Table 7 shows the probability of having an accident with SEV=KSI when knowing a priori the value of only one variable (simple evidences). Table 8 shows the probability of KSI when knowing a priori the value of more than one variable (multiple evidences). Based on the case that presents the highest probability in Table 7 (LIG=I), the probabilities with multiple evidences were calculated in a descending order (ACT, NOI, OI, AGE, GEN and ATF) (see Table 8). In each step, the highest probability value was selected and used for the next step.

Simplified BN allow this kind of analysis with multiple evidences. This analysis provides added value information with regard to the analysis with simple evidences. For example, Table 7 shows that the probability of an accident with SEV=KSI is 51% in the case of NOI=1 and 30% in the case of NOI>1. However, in HO accidents under insufficient lighting (ACT=HO and LIG=I) the probability of SEV=KSI increases to 75.45% for NOI=1 and becomes 55.82% for NOI>1. Table 8 also shows that an accident with LIG=I, ACT=HO, NOI=1, OI>2, AGE=[18-25], GEN=M and ATF=GW has a probability of 96.51% of having SEV=KSI.

5. Summary and conclusions

The main objective of this research work was to determine if it is possible to maintain or improve the performance of a model that is used to predict the injury severity of a traffic accident based on BNs reducing the number of variables considered in the analysis. The performance of the model was measured using five indicators (accuracy, specificity, sensitivity, HMSS and ROC area).

In order to perform this analysis 1,536 records of traffic accidents on rural highways with information about 18 variables that are related with the severity of the accidents based on the standard police reports used in Spain were used. 59 combinations of evaluator-search algorithms, which are commonly used in data mining, were used and 26 subsets of variables were identified.

Within these subsets of variables the variable accident type (ACT), lighting (LIG) and number of injuries (NOI) were selected the most times (over 95%). Therefore, it could be said that these variables are the most significant ones in the classification of injury severity in traffic accidents, since they are included in almost all the selected subsets of variables.

For each one of these subsets of variables, 10 simplified BNs were built for the training stage and another 10 for the testing stage. In total, 540 BNs were built using the hill climbing search algorithm and the MDL score (de Oña et al., 2011).

Comparing the average values of the indicators for each one of the simplified BNs with respect to the average values obtained for the original BN (BN-18), it is observed that, in most of cases (74%), the performance indicators values for the simplified BNs maintained or improved in comparison with those of BN-18. Therefore, it could be said that, in most cases, simplified networks maintain the performance of the original BN.

Seven BNs were found to present statistically significant improvements in their performance indicators with respect to BN-18 and only one value of these indicators get worsened. In more than 50% of these BNs the following variables are repeated: ACT, AGE, ATF, GEN, LIG, NOI and OI.

These 7 variables were used to built a new BN (BN-7). The results of the performance indicators of this BN with respect to BN-18 improve practically in all the cases, and these improvements are statistically significant ($p<0.05$) in 60% of the cases (accuracy, sensitivity and ROC area).

Therefore, this research work shows that, for the analysis of the severity of road accidents by Bayesian networks on rural roads, it is possible to reduce the number of variables considered in more than 60% (from 18 to 7 variables) maintaining the performance of the models and reducing their complexity. Thus the findings of this research work agrees with Chang and Wang (2006) where they stated that if a model is applied only on a few important variables, more useful results could be obtained.

The procedure used to simplify BN models to analyze the severity of traffic accidents on rural highways could be also applied to other types of infrastructure (intersections, freeways, etc.) as well as to other models used to assess severity of traffic accidents (multinomial logit models, hierarchical logit models, probit models, etc.).

Acknowledgements

Support from Spanish Ministry of Science and Innovation (Research Project TRA2007-63564) as well as the data provided by the Spanish General Directorate of Traffic (DGT) are gratefully acknowledged.

References

1. Abdel-Aty, M. (2003). Analysis of driver injury severity levels at multiple locations using ordered probit models. *Journal of Safety Research*, 34, 597–603.
2. Abdelwahab, H.T., & Abdel-Aty M.A. (2001). Development of Artificial Neural Network Models to Predict Driver Injury Severity in Traffic Accidents at Signalized Intersections. *Transportation Research Record*, 1746, 6-13.
3. Acid, S., de Campos, L.M., Fernández-Luna, J.M., Rodríguez, S., J.M., R. & Salcedo, J.L. (2004) 'A comparison of Learning algorithms for Bayesian networks: a case study based on data from an emergency medical service, *Artificial Intelligence in Medicine*, 30, 215-232
4. Al-Ghamdi, A.S. (2002). Using logistic regression to estimate the influence of accident factors on accident severity, *Accident Analysis and Prevention*, 34, 729–741.
5. Bédard M., Guyatt, G.H., Stones, M.J., & Hirdes, J.P. (2002). The independent contribution of driver, crash, and vehicle characteristics to driver fatalities, *Accident Analysis and Prevention*, 34, 717–727.
6. Chang, L.Y, & Wang, H.W. (2006). Analysis of traffic injury severity: An application of non-parametric classification tree techniques. *Accident Analysis and Prevention*, 38, 1019-1027.
7. Chen, W.H, & Jovanis, P.P. (2000). Method for identifying factors contributing to driver-injury severity in traffic crashes. *Transportation Research Record*, 1717, 1-9.
8. De Oña, J., Mujalli, R.O., & Calvo, F. (2011). Analysis of traffic accident injury severity on Spanish rural highways using Bayesian networks. *Accident Analysis and Prevention*, 43 (1), 402-411.
9. Delen, D., Sharda, R., & Bessonov, M. (2006). Identifying significant predictors of injury severity in traffic accidents using a series of artificial neural networks. *Accident Analysis and Prevention*, 38, 434–444.

10. Dirección General de Tráfico – DGT. (2007). *Anuario Estadístico 2007, NIPO: 128-08-161-7*, Retrieved from http://www.dgt.es/.
11. Dupont, E., Martensen, H., Papadimitriou, E., & Yannis, G. (2010). Risk and Protection factors in fatal accidents. *Accident Analysis and Prevention*, 42, 645-653.
12. García López, F., García Torres, M., Melián Batista, B., Moreno Pérez, J., & Moreno-Vega, J. (2006). Solving feature subset selection problem by a Parallel Scatter Search . *European Journal of Operational Research* , 477-489.
13. Gray, R.C., Quddus, M.A. & Evans, A. (2008). Injury severity analysis of accidents involv- ing young male drivers in Great Britain. *Journal of Safety Research*, 39, 483–495.
14. Glymour, C., Cooper, G. & Chickering, D.M. (ed.) (1999). *Computation Causation and Discovery*, The AAAI Press.
15. Goldberg, D. (1989). *Genetic algorithms in search, optimization and machine learning*. Addison-Wesley.
16. Gutlein, M., Frank, E., Hall, M., & Karwath, A. (2009). Large scale attribute selection using wrappers. - *Proceedings of 2009 IEEE Symposium on Computational Intelligence and Data Mining, CIDM 2009*, Washington.
17. Hall, M. (1998). *Correlaion-based Feature Subset Selection for Machine Learning*. Hamilton, New Zeland (Doctoral Dissertation).
18. Hedar, A.-R. W. (2008). Tabu search for attribute reduction in rough set theory. *Soft Computing* , 12 (9) 909-918.
19. Helai, H., Chor, C.H., & Haque, M.M. (2008). Severity of driver injury and vehicle dam- age in traffic crashes at intersections: a Bayesian hierarchical analysis. *Accident Analysis and Prevention*, 40, 45–54.
20. Khhavi, R., & John, G. (1997). Wrappers for feature subset selection. *Artificial Intelligence*, 97, 273-324.
21. Kockelman, K.M., & Kweon, Y.J. (2002). Driver injury severity: an application of ordered probit models. *Accident Analysis and Prevention*, 34, 313–321.
22. Kopelias, P., Papadimitriou, F, Papandreou, K, & Prevedouros, P. (2007). Urban Freeway Crash Analysis. *Transportation Research Record: Journal Transportation Research Board*, 2015, 123-131.
23. Liu, H., & Setiono, R. (1996). A probabilistic approach to feature selection- A filter solution. *Proceedings of 13th International Conference on Machine Learning*.
24. Madden, M. G. (2009). On the classification performance of TAN and general Bayesian networks. *Journal of Knowledge-Based Systems*, 22, 489-495.
25. Milton, J.C., Shankar, V.N., & Mannering, F.L. (2008). Highway accident severities and the mixed logit model: An exploratory empirical analysis. *Accident Analysis and Prevention*, 40, 260–266.
26. Mittal, A., Kassim, A., & Tan, T. (2007). *Bayesian network technologies: Applications and graphical models*. IGI Publishing.
27. Moore, A., & Lee, M. (1994). Efficient algorithms for minimizing cross validation error. *Proceedings of 11th International Conference on Machine Lerning*.
28. Neapolitan, R.E. (2004). *Learning Bayesian Networks*, Prentice Hall.
29. Neapolitan, R.E. (2009). *Probabilistic Methods for Bioinformatics*. San Francisco, California: Morgan Kaufmann Publishers.
30. Pearl, J. (1984). *Heuristics: Intelligent search strategies for computer problem solving*. Addison-Wesley.
31. Pearl J. (2000). *Causality: Models, Reasoning and Inference*. Cambridge: Cambridge University Press.
32. Russell, S., & Norvig, P. (2003). *Artificial Intelligence: A modern approach* . Upper Saddle River, New Jersey: Prentice Hall.
33. Scheetz, L.J., Zhang, J., & Kolassa, J. (2003). Classification tree to identify severe and moderate injuries in young and middle aged adults. *Artificial Intelligence in Medicine*, 45, 1–10.
34. Simoncic, M. (2004). A Bayesian network model of two-car accidents. *Journal of transportation and Statistics*, 7, 13-25.

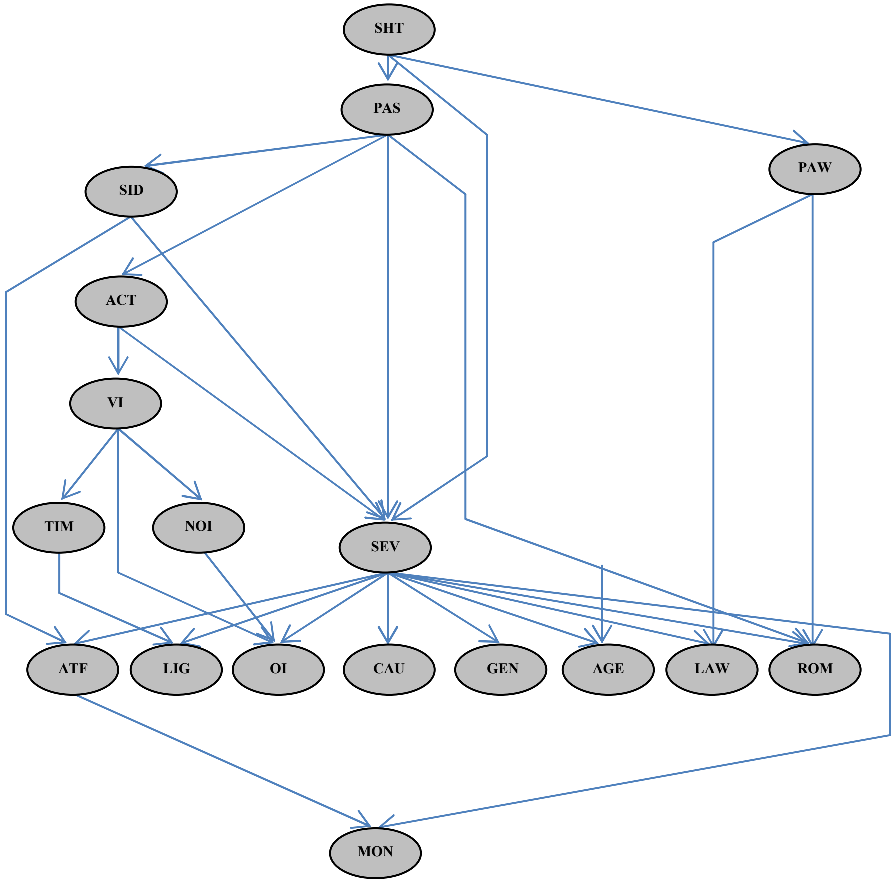
35. Tavris, D.R., Kuhn, E.M., & Layde, P.M. (2001). Age and gender patterns in motor vehicle crash injuries: importance of type of crash and occupant role. *Accident Analysis and Prevention*, 33, 167–172.
36. Witten, I. H., & Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques* (2nd Edition). San Francisco, California: Morgan Kaufmann Publishers.
37. Xie, Y., Zhang, Y., & Liang, F. (2009). Crash Injury Severity Analysis Using Bayesian Ordered Probit Models. *Journal of Transportation Engineering ASCE*, 135 (1),18-25.
38. Yamamoto, T., & Shankar, V.N. (2004). Bivariate ordered-response probit model of driver's and passenger's injury severities in collisions with fixed objects. *Accident Analysis and Prevention*, 36, 869–876.
39. Yau, K.K.W., Lo, H.P., & Fung, S.H.H. (2006). Multiple-vehicle traffic accidents in Hong Kong, *Accident Analysis and Prevention*, 38,1157–1161.

List of Figures:

List of Tables:

BN-18

Total number of arcs = 30 arcs



BN-7

Total number of arcs = 9 arcs

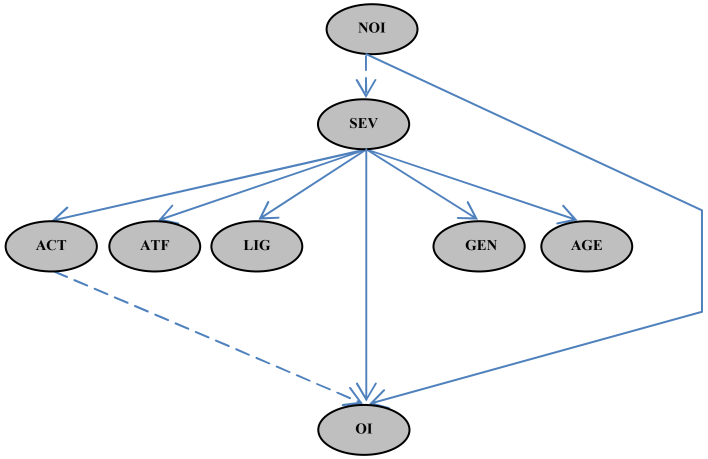


Table 1: Variables, values and actual classification by severity

Variables	Values	SEV*				Total
		SI		KSI		
ACT: accident Type	AS: angle or side collision	381	61.45%	239	38.55%	620
	CF: fixed objects	99	52.94%	88	47.06%	187
	HO: head on	84	40.58%	123	59.42%	207
	O: other	75	59.06%	52	40.94%	127
	PU: pile up	33	78.57%	9	21.43%	42
	R: rollover	163	49.39%	167	50.61%	330
	SP: straight path	17	73.91%	6	26.09%	23
AGE: age	[18-25]	225	50.34%	222	49.66%	447
	(25-64]	586	57.73%	429	42.27%	1015
	>64	41	55.41%	33	44.59%	74
ATF: atmospheric Factors	GW: good weather	730	54.23%	616	45.77%	1346
	HR: heavy rain	23	71.88%	9	28.13%	32
	LR: light rain	84	61.76%	52	38.24%	136
	O: other	15	68.18%	7	31.81%	22
CAU: cause	DC: driver characteristics	791	54.93%	649	45.07%	1440
	OF: other factors	50	66.67%	25	33.33%	75
	RC: road characteristics	3	75.00%	1	25.00%	4
	VC: vehicle charactersitics	8	47.06%	9	52.94%	17
DAY: day	BW: beginning of week	123	60.29%	81	39.71%	204
	EW: end of week	132	57.14%	99	42.86%	231
	F: festive	29	61.70%	18	38.30%	47
	WD: week day	325	55.65%	259	44.35%	584
	WE: week end	243	51.70%	227	48.30%	470
GEN : gender	F: female	148	63.79%	84	36.21%	232
	M: male	704	53.99%	600	46.01%	1304
LAW: lane width	THI: thin: <3.25m	19	67.86%	9	32.14%	28
	MED: medium: 3.25m<=L<=3.75m	176	51.16%	168	48.84%	344
	WID: wide: >3.75m	657	56.44%	507	43.56%	1164
LIG: lighting	D: dusk	52	61.18%	33	38.82%	85
	DL: daylight	573	58.65%	404	41.35%	977
	I: insufficient	27	54.00%	23	46.00%	50
	S: sufficient	36	59.02%	25	40.98%	61
	W: without lighting	164	45.18%	199	54.82%	363
MON: month	AUT: autumn	218	54.23%	184	45.77%	402
	SPR: spring	206	59.03%	143	40.97%	349
	SUM: summer	246	56.55%	189	43.45%	435
	WIN: winter	182	52.00%	168	48.00%	350
NOI: number of Injuries	1	539	49.95%	540	50.05%	1079
	>1	313	68.49%	144	31.51%	457

OI: occupants involved	1	229	51.58%	215	48.42%	444
	2	374	55.99%	294	44.01%	668
	>2	249	58.73%	175	41.27%	424
PAS: paved shoulder	missing values	66	51.56%	62	48.44%	128
	N: no	253	57.11%	190	42.89%	443
	Y: yes	533	55.23%	432	44.77%	965
PAW: pavement width	THI: thin: <6m	95	53.98%	81	46.02%	176
	MED: medium: 6 m<=law<=7m	209	54.29%	176	45.71%	385
	WID: wide: >7m	548	56.21%	427	43.79%	975
ROM: pavement markings	DME: does not exist or was deleted	60	58.25%	43	41.75%	103
	DMR: define margins of roadway	60	57.69%	44	42.31%	104
	SLD: separate lanes and defined road margins	714	55.26%	578	44.74%	1292
	SLO: separate lanes only	18	48.65%	19	51.35%	37
SHT: Shoulder type	NOS: does not exist	311	55.24%	252	44.76%	563
	THI: thin:<1.5m	402	54.47%	336	45.53%	738
	MED: medium: 1.5m<=sht<2.50m	133	58.85%	93	41.15%	226
	WID: wide >= 2.50 m	6	66.67%	3	33.33%	9
SID: sight distance	A: atmospheric	26	81.25%	6	18.75%	32
	B: building	10	55.56%	8	44.44%	18
	O: other	6	66.67%	3	33.34%	9
	T: topological	187	55.49%	150	44.51%	337
	V: vegetation	6	54.55%	5	45.45%	11
	WR: without restriction	617	54.65%	512	45.35%	1129
TIM: time	[0-6]	99	46.26%	115	53.74%	214
	(6-12]	236	57.99%	171	42.01%	407
	(12-18]	314	57.72%	230	42.28%	544
	(18-24)	203	54.72%	168	45.28%	371
VI: vehicles involved	1	316	52.06%	291	47.94%	607
	2	468	56.73%	357	43.27%	825
	>2	68	65.38%	36	34.62%	104
Total		852	55.47%	684	44.53%	1536

* SEV: injury severity; SI: slight injured; KSI: killed or seriously injured

Table 2: Search-evaluator combinations which were found to be incompatible

Search Method	Evaluator
Race Search	CfsSubsetEval ConsistencySubsetEval WrapperSubsetEavl FilteredSubsetEval CostSensitiveSubsetEval
Tabu Search	ClassifierSubsetEval WrapperSubsetEavl

Table 3: Selected variables for different combinations of evaluator-search method.

Number of variables	Variable selection method		Variables selected	BN
	Evaluator	Search Method		
18	The original training set of variables		ACT, AGE, ATF, CAU, DAY, GEN, LAW, LIG, MON, NOI, OI, PAS, PAW, ROM, SHT, SID, TIM, VI	BN-18
4	CFS	Best first, Exhaustive, Greedy stepwise, Linear forward selection, Rank, Scatter V1, Tabu	ACT, ATF, LIG, NOI	BN-4
	Cost sensitive	Best first, Exhaustive, Greedy stepwise, Linear forward selection, Random, Rank, Scatter V1, Tabu		
	Filtered	Rank		
5	CFS	Genetic, Random	ACT, ATF, LIG, NOI, SID	BN-5a
	Classifier	Subset Size Forward Selection	ACT, ATF, CAU, LIG, NOI	BN-5b
	Wrapper	Scatter V1	ACT, DAY, LIG, NOI, OI	BN-5c
6	Classifier	Scatter V1	ACT, ATF, GEN, LIG, MON, NOI	BN-6a
	Cost sensitive	Genetic	ACT, ATF, LIG, NOI, SHT, SID	BN-6b
	Filtered	Random	ACT, ATF, LAW, LIG, NOI, SID	BN-6c
	Wrapper	Greedy stepwise, Subset Size Forward Selection	ACT, AGE, GEN, LIG, MON, SHT	BN-6d
7	CFS	Subset Size Forward Selection	ACT, ATF, GEN, LAW, LIG, NOI, SID	BN-7a
	Filtered	Best first, Exhaustive, Greedy stepwise, Linear forward selection, Scatter V1, Subset Size Forward Selection, Tabu		
	Cost sensitive	Subset Size Forward Selection	ACT, ATF, LAW, LIG, NOI, SHT, SID	BN-7b
	Classifier	Greedy stepwise	ACT, AGE, ATF, GEN, LIG, NOI, OI	BN-7c
	Wrapper	Random	ACT, GEN, LIG, NOI, OI, SHT, VI	BN-7d
8	Classifier	Race	ACT, AGE, GEN, LAW, LIG, NOI, OI, ROM	BN-8a
	Filtered	Genetic	ACT, ATF, GEN, LAW, LIG, MON, NOI, SID	BN-8b
9	Wrapper	Best first	ACT, ATF, GEN, LAW, LIG, NOI, OI, PAS, PAW	BN-9a
		Exhaustive	ACT, AGE, ATF, GEN, LIG, MON, NOI, OI, VI	BN-9b
		Genetic	ACT, ATF, DAY, GEN, LIG, NOI, OI, PAS, PAW	BN-9c
11	Classifier	Exhaustive	ACT, CAU, DAY, GEN, LIG, MON, NOI, OI, PAS, TIM, VI	BN-11a
	Wrapper	Linear forward selection	ACT, AGE, ATF, DAY, GEN, LIG, MON, NOI, OI, SHT, VI	BN-11b
12	Classifier	Random	ACT, AGE, CAU, GEN, LIG, MON, NOI, OI, PAS, SID, TIM, VI	BN-12
14	Consistency	Best first, Exhaustive, Greedy stepwise, Linear forward selection, Rank, Scatter V1, Subset Size Forward Selection, Tabu	ACT, AGE, DAY, GEN, LAW, LIG, MON, NOI, OI, PAS, PAW, SHT, SID, TIM	BN-14
15	Consistency	Genetic	ACT, ATF, CAU, DAY, GEN, LAW, LIG, NOI, OI, PAS, PAW, ROM, SHT, TIM, VI	BN-15a
	Classifier	Best first	ACT, AGE, CAU, DAY, GEN, LAW, LIG, MON, NOI, OI, PAS, PAW, SHT, SID, TIM	BN-15b
16	Classifier	Genetic	ACT, ATF, CAU, DAY, GEN, LAW, LIG, MON, NOI, OI, PAS, PAW, ROM, SHT, TIM, VI	BN-16a
		Liner forward selection	ACT, AGE, ATF, CAU, DAY, GEN, LIG, MON, NOI, OI, PAS, PAW, ROM, SHT, SID, VI	BN-16b
	Consistency	Random	ACT, AGE, ATF, CAU, DAY, GEN, LAW, LIG, MON, NOI, OI, PAS, PAW, SHT, SID, TIM	BN-16c

Table 4: Number of times each variable is selected

Variable	Number of times variable has been selected
ACT	58
LIG	58
NOI	56
ATF	42
GEN	34
LAW	26
SID	26
OI	25
MON	21
SHT	20
AGE	19
DAY	18
PAS	18
PAW	16
TIM	15
CAU	9
VI	9
ROM	5

Table 5: Average values for accuracy, sensitivity, specificity, HMSS and ROC area for the 27 built BNs (training and test data).

Number of selected variables	BN	BN results									
		training	test	training	test	training	test	training	test	training	test
		Accuracy		Sensitivity		Specifity		HMSS		ROC area	
All the variables (18)	**BN-18**	**0,59**	**0,58**	**0,71**	**0,65**	**0,45**	**0,49**	**0,55**	**0,56**	**0,62**	**0,61**
4	BN-4	**0,60**	**0,59**	0,71	**0,68**	**0,46**	*0,48*	**0,56**	0,56	**0,63**	0,61
5	BN-5a	0,59	0,57	*0,70*	**0,68**	0,45	*0,45*	0,55	*0,54*	0,62	*0,60*
	BN-5b	**0,60**	**0,59**	0,71	**0,68**	**0,47**	0,49	**0,56**	0,56	**0,63**	0,61
	BN-5c	**0,61***	**0,61***	**0,75***	**0,73***	*0,44*	*0,46*	0,55	0,56	**0,63***	**0,62**
6	BN-6a	**0,60**	**0,60***	*0,70*	**0,67**	**0,49***	0,49	**0,58**	**0,57**	**0,64***	**0,62**
	BN-6b	0,59	0,57	0,71	**0,67**	0,45	*0,45*	0,55	*0,54*	0,62	*0,60*
	BN-6c	0,59	0,57	*0,70*	**0,67**	0,45	*0,46*	0,55	*0,54*	0,62	*0,60*
	BN-6d	**0,60**	0,57	**0,74**	**0,69***	*0,42*	*0,42**	*0,54*	*0,52**	*0,60**	*0,58**
7	BN-7a	**0,60**	0,57	*0,70*	**0,66**	**0,47**	*0,48*	**0,56**	*0,55*	0,62	*0,60*
	BN-7b	0,59	0,57	*0,70*	**0,66**	0,45	*0,46*	0,55	*0,54*	0,62	*0,60*
	BN-7c	**0,62***	**0,60***	**0,74**	**0,70***	**0,47**	*0,48*	**0,57**	**0,57**	**0,64***	**0,63**
	BN-7d	**0,61***	**0,59**	**0,76**	**0,72**	*0,42*	*0,44*	*0,54*	*0,54*	**0,64***	**0,62**
8	BN-8a	**0,62***	**0,60***	**0,75***	**0,71***	**0,46**	*0,47*	**0,57**	0,56	**0,64***	**0,63***
	BN-8b	0,59	0,58	*0,69*	**0,67**	**0,47**	*0,48*	**0,56**	0,56*	**0,63**	*0,60*
9	BN-9a	**0,61***	**0,60***	**0,74***	**0,70***	**0,46**	*0,47*	**0,56**	0,56	**0,65***	**0,63***
	BN-9b	**0,62***	**0,59**	**0,75***	**0,71***	**0,46**	*0,47*	**0,57**	0,56	**0,64***	**0,62**
	BN-9c	**0,61***	**0,59**	**0,73**	**0,69***	**0,46**	*0,48*	**0,56**	0,56	**0,65***	**0,63**
11	BN-11a	**0,60**	0,58	**0,74***	**0,69**	*0,43*	*0,46*	*0,54*	*0,55*	0,62	0,61
	BN-11b	**0,61***	**0,59**	**0,74**	**0,69***	**0,46**	*0,46*	**0,56**	0,56	**0,64***	**0,62**
12	BN-12	**0,60**	**0,59**	**0,74**	**0,68**	*0,43*	*0,48*	*0,54*	0,56	0,62	0,61
14	BN-14	0,59	0,57	0,71	**0,66**	0,45	*0,48*	0,55	*0,55*	0,62	0,61
15	BN-15a	**0,60**	0,58	**0,73**	**0,67**	*0,44*	*0,47*	0,55	*0,55*	0,62	0,61
	BN-15b	**0,60**	0,58	0,71	**0,66**	**0,46**	*0,48*	**0,56**	0,56	0,62	*0,60*
16	BN-16a	**0,60**	0,57	**0,72**	0,65	0,45	*0,47*	0,55	*0,55*	0,62	*0,60*
	BN-16b	0,59	0,57	0,71	**0,66**	0,45	*0,47*	0,55	*0,55*	0,62	0,61
	BN-16c	**0,60**	0,58	0,71	0,65	**0,47**	0,49	**0,65**	0,56	**0,63**	*0,60*

* Statistically significant when tested against BN-18 using 95% LSD ANOVA test (p<0.05).

Table 6: Average values for accuracy, sensitivity, specificity, HMSS and ROC area for BN-18 and BN-7 (training and test data).

Indicators	BN results			
	BN-18		BN-7	
	training	test	training	test
Accuracy	0,59	0,58	**0,61***	**0,60***
Sensitivity	0,71	0,65	**0,74***	**0,70***
Specificity	0,45	0,49	**0,46**	*0,48*
HMSS	0,55	0,56	**0,57**	**0,57**
ROC area	0,62	0,61	**0,65***	**0,63***

* Statistically significant when tested against BN-18 using 95% LSD ANOVA test (p<0.05).

Table 7: Inference results for simple evidences in BN-7

Variable	Evidence	Prob. KSI	Condition satisfied
ACT	AS	0.3979	P(SEV=KSI | ACT=AS)=0.3979
	CF	0.4840	P(SEV=KSI | ACT=CF)= 0.4840
	HO	**0.5778**	P(SEV=KSI | ACT=HO)= 0.5778
	O	0.3965	P(SEV=KSI | ACT=O)= 0.3965
	PU	0.2676	P(SEV=KSI | ACT=PU)= 0.2676
	R	0.4910	P(SEV=KSI | ACT=R)= 0.4910
	SP	0.2998	P(SEV=KSI | ACT=SP)= 0.2998
AGE	**[18-25]**	**0.5034**	P(SEV=KSI | AGE=[18-25])=0.5034
	(25-64]	0.4233	P(SEV=KSI | AGE=(25-64])=0.4233
	>64	0.4803	P(SEV=KSI | AGE=>64)=0.4803
ATF	**GW**	**0.4653**	P(SEV=KSI | ATF=GW)=0.4653
	HR	0.3124	P(SEV=KSI | ATF=HR)=0.3124
	LR	0.3640	P(SEV=KSI | ATF=LR)=0.3640
	O	0.2894	P(SEV=KSI | ATF=O)=0.2894
GEN	F	0.3258	P(SEV=KSI | GEN=F)= 0.3258
	M	**0.4715**	P(SEV=KSI | GEN=M)= 0.4715
LIG	D	0.3415	P(SEV=KSI | LIG=D)= 0.3415
	DL	0.4113	P(SEV=KSI | LIG=DL)= 0.4113
	I	**0.5891**	P(SEV=KSI | LIG=I)=0.5891
	S	0.4728	P(SEV=KSI | LIG=S)= 0.4728
	W	0.5508	P(SEV=KSI | LIG=W)= 0.5508
NOI	**1**	**0.5104**	P(SEV=KSI | NOI=1)= 0.5104
	>1	0.3000	P(SEV=KSI | NOI=>1)= 0.3000
OI	**1**	**0.5100**	P(SEV=KSI | OI=1)=0.5100
	2	0.4250	P(SEV=KSI | OI=2)= 0.4250
	>2	0.4213	P(SEV=KSI | OI=>2)=0.4213

Table 8: Inference results for multiple evidences in BN-7

No.Var.	Condition satisfied	Prob. KSI
1	**P(SEV=KSI | LIG=I)**	**0.5891**
2	P(SEV=KSI | LIG=I, ACT=AS)	*0.5374*
	P(SEV=KSI | LIG=I, ACT=CF)	0.6225
	P(SEV=KSI | LIG=I, ACT=HO)	**0.7063**
	P(SEV=KSI | LIG=I, ACT=O)	*0.5359*
	P(SEV=KSI | LIG=I, ACT=PU)	*0.3911*
	P(SEV=KSI | LIG=I, ACT=R)	0.6290
	P(SEV=KSI | LIG=I, ACT=SP)	*0.4294*
3	**P(SEV=KSI | LIG=I, ACT=HO, NOI=1)**	**0.7545**
	P(SEV=KSI | LIG=I, ACT=HO, NOI>1)	*0.5582*
4	P(SEV=KSI | LIG=I, ACT=HO, NOI=1, OI=1)	*0.5713*
	P(SEV=KSI | LIG=I, ACT=HO, NOI=1, OI=2)	*0.6714*
	P(SEV=KSI | LIG=I, ACT=HO, NOI=1, OI=>2)	**0.9502**
5	**P(SEV=KSI | LIG=I, ACT=HO, NOI=1, OI=>2, AGE=[18-25])**	**0.9595**
	P(SEV=KSI | LIG=I, ACT=HO, NOI=1, OI=>2, AGE=[25-64))	*0.9450*
	P(SEV=KSI | LIG=I, ACT=HO, NOI=1, OI=>2, AGE=>64)	0.9558
6	P(SEV=KSI | LIG=I, ACT=HO, NOI=1, OI=>2, AGE=[18-25], GEN=F)	*0.9336*
	P(SEV=KSI | LIG=I, ACT=HO, NOI=1, OI=>2, AGE=[18-25], GEN=M)	**0.9629**
7	**P(SEV=KSI | LIG=I, ACT=HO, NOI=1, OI=>2, AGE=[18-25], GEN=M, ATF=GW)**	**0.9651**
	P(SEV=KSI | LIG=I, ACT=HO, NOI=1, OI=>2, AGE=[18-25], GEN=M, ATF=HR)	*0.9353*
	P(SEV=KSI | LIG=I, ACT=HO, NOI=1, OI=>2, AGE=[18-25], GEN=M, ATF=LR)	*0.9479*
	P(SEV=KSI | LIG=I, ACT=HO, NOI=1, OI=>2, AGE=[18-25], GEN=M, ATF=O)	*0.9283*