Universidade de Aveiro Departamento de Engenharia Mecânica
Ano 2013

GUILHERMINA CÂNDIDA ANTAS TORRÃO

Efeito das Características do Veículo na Segurança, Consumos e Emissões

Effect of Vehicle Characteristics on Safety, Fuel Use and Emissions

Universidade de Aveiro Departamento de Engenharia Mecânica
Ano 2013

GUILHERMINA CÂNDIDA ANTAS TORRÃO

Efeito das Características do Veículo na Segurança, Consumos e Emissões

Effect of Vehicle Characteristics on Safety, Fuel Use and Emissions

Tese apresentada à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Doutor em Engenharia Mecânica, realizada sob a orientação científica da Doutora Margarida Isabel Cabrita Marques Coelho, Professora Auxiliar do Departamento de Engenharia Mecânica da Universidade de Aveiro, e do Doutor Nagui Michael Rouphail, Professor no Departamento de Engenharia Civil e Diretor do Institute for Transportation Research and Education, North Carolina State University.

Apoio financeiro da FCT_Fundação para a Ciência e a Tecnologia, Ministério da Educação e Ciência_ através da Bolsa SFRH/BD/41530/2007 e do Projeto PTDC/SEN-TRA/113499/2009.
Apoio do PEst-C/EME/UI0481/2011.

To my mother who bravely survived a crash. For her endless support and unconditional love, I am for ever grateful.
To my father, who always encouraged me to pursue a higher education. I am forever proud of him.

To Fernando whose charisma I admire the most. His friendship and strong support are a driving force in my life.

Without your constant presence throughout my life, everything would be worth nothing.

To Oscar.

o júri

presidente

Prof. Doutor Vitor Brás de Sequeira Amaral
Professor Catedrático do Departamento de Física da Universidade de Aveiro

Prof. Doutora Ana Maria César Bastos Silva
Professora Auxiliar do Departamento de Engenharia Civil da Faculdade de Ciências e Tecnologia da Universidade de Coimbra

Doutora Elisabete Maria Mourinho Arsénio Guterres de Almeida
Investigadora Auxiliar do Laboratório Nacional de Engenharia Civil

Prof. Doutor José Paulo Oliveira Santos
Professor Auxiliar do Departamento de Engenharia Mecânica da Universidade de Aveiro

Prof. Doutora Margarida Isabel Cabrita Marques Coelho
Professora Auxiliar do Departamento de Engenharia Mecânica da Universidade de Aveiro

Prof. Doctor Nagui Michael Rouphail
Professor of Civil Engineering Department, Institute for Transportation Research and Education, North Carolina State University

agradecimentos

I would like to thank Prof. Dr. Margarida Coelho for presenting me the excellent opportunity to perform my PhD in the road safety area and sustainable mobility. I appreciate so much her constant support, dedication, trust and friendship. I am grateful for all the wonderful strength she kindly gave me.

To Prof. Dr. Nagui Rouphail for his outstanding supervision with research strategy and teaching. His powerful confidence and intelligent guidance were crucial during PhD development and he made me believe in this work. I deeply appreciate his exceptional supervision and friendship.

To Prof. David Dickey for his remarkable and intelligent support with data mining analysis. Although he is a distinguished Professor of the Department of Statistics of North Carolina State University, he used his time to teach me basic steps with Enterprise Miner. Then, he became an excellent adviser in crash modeling. I am thankful for his guidance and friendship.

To Prof. Dr. José Grácio for his continuous encouragement and Prof. Dr. Antonio Sousa for the remarkable support, at TEMA.

To Engineer Elson Filho and Dr. Margarida Alexandre for the valuable help with SAS modeling techniques. Dr. Sandra Simões and Dr. Jos Velden for all the technical support provided with SAS software, SAS Office at Lisbon.

To GNR Officers Leitão and Mota for their help with crash reports collection. Assistant Sergeant Joaquim Ferreira for his extraordinary help with crashes analysis. Captain Nuno Lopes for his great cooperation with data collection at the Traffic Division of the Territoral Command of Portuguese National Guard, at Porto.

To Engineer Gil Paulo and Mr. Francisco Lacerda, of the Institute for Mobility and Inland Transportation (IMTT), at Porto, for the friendly support in providing me information on vehicles registration IDs during the earlier stage of this research.

To Engineer Luís Paulo, Head of Certification and Registration of Road Vehicles and Engineer Jose Pinheiro, Director of Services, IMTT, at Lisbon, for the cooperation with vehicles registration plates information.

To Engineer Carlos Lopes, Director of the Road Safety Prevention Unit, Portuguese Road Safety Authority, for his excellent support with safety indicators and contacts that allowed me to ask for further helpful information.

To TT Group and team friends. In particular, to Sérgio Silva for the smart help and specially, to Tânia Fontes, who gave me an outstanding support with this Thesis.

To special friends: Ana Oliveira and Paula Marques, for the undisputed example of integrity, excellence and natural elegance.

To the exceptional friend, Tânia Barbosa, who is incredible powerful and always willing to help me.

To my friends at the University of Aveiro: Bernardete Coelho, Elisabete Ferreira, Gil Gonçalves, Joana Sousa, João Oliveira, Margarida Bola, Nuno Almeida, Raquel Silva, Patrícia Silva, Ricardo Pinto and Tiago Grilo for all the support provided and incredible moments. In particular, Marisa Henriques, who have changed my Aveiro life ever since.

To my friends at the North Carolina State University for the great help and happy moments: Anne LaPierre, Anxi Jia, Christie Vann, Hyejung Hu, Gurdas Singh, Katy Salamati, Linda Lancaster and Mike Lili.

To my all times life friends: Áida Pires, Birsen Ayaz-Maierhafer, Carina Rodrigues, Clarice Sandoval de Bausch, Joana Creissac, Sara Cameira, Lilian Domingues, Manuela Carvalho and Siew Chin.

Finally, I am forever thankful to my family, who have always supported me during my life!

palavras-chave

Acidentes com um veículo, acidentes com dois veículos, CART, consumo de combustível, CORINAIR, emissões, eventos raros, gravidade, regressão logistica, segurança, e veículo.

resumo

Nos últimos anos, o número de vítimas de acidentes de tráfego por milhões de habitantes em Portugal tem sido mais elevado do que a média da União Europeia. Ao nível nacional torna-se premente uma melhor compreensão dos dados de acidentes e sobre o efeito do veículo na gravidade do mesmo. O objetivo principal desta investigação consistiu no desenvolvimento de modelos de previsão da gravidade do acidente, para o caso de um único veículo envolvido e para caso de uma colisão, envolvendo dois veículos. Além disso, esta investigação compreendeu o desenvolvimento de uma análise integrada para avaliar o desempenho do veículo em termos de segurança, eficiência energética e emissões de poluentes. Os dados de acidentes foram recolhidos junto da Guarda Nacional Republicana Portuguesa, na área metropolitana do Porto para o período de 2006-2010. Um total de 1,374 acidentes foram recolhidos, 500 acidentes envolvendo um único veículo e 874 colisões.

Para a análise da segurança, foram utilizados modelos de regressão logística. Para os acidentes envolvendo um único veículo, o efeito das características do veículo no risco de feridos graves e/ou mortos (variável resposta definida como binária) foi explorado. Para as colisões envolvendo dois veículos foram criadas duas variáveis binárias adicionais: uma para prever a probabilidade de feridos graves e/ou mortos num dos veículos (designado como veículo V1) e outra para prever a probabilidade de feridos graves e/ou mortos no outro veículo envolvido (designado como veículo V2). Para ultrapassar o desafio e limitações relativas ao tamanho da amostra e desigualdade entre os casos analisados (apenas 5.1% de acidentes graves), foi desenvolvida uma metodologia com base numa estratégia de reamostragem e foram utilizadas 10 amostras geradas de forma aleatória e estratificada para a validação dos modelos. Durante a fase de modelação, foi analisado o efeito das características do veículo, como o peso, a cilindrada, a distância entre eixos e a idade do veículo.

Para a análise do consumo de combustível e das emissões, foi aplicada a metodologia CORINAIR. Posteriormente, os dados das emissões foram modelados de forma a serem ajustados a regressões lineares. Finalmente, foi desenvolvido um indicador de análise integrada (denominado “SEG”) que proporciona um método de classificação para avaliar o desempenho do veículo ao nível da segurança rodoviária, consumos e emissões de poluentes.

Resumo (cont.)

Face aos resultados obtidos, para os acidentes envolvendo um único veículo, o modelo de previsão do risco de gravidade identificou a idade e a cilindrada do veículo como estatisticamente significativas para a previsão de ocorrência de feridos graves e/ou mortos, ao nível de significância de 5%. A exatidão do modelo foi de 58.0% (desvio padrão (D.P.) 3.1). Para as colisões envolvendo dois veículos, ao prever a probabilidade de feridos graves e/ou mortos no veículo V1, a cilindrada do veículo oposto (veículo V2) aumentou o risco para os ocupantes do veículo V1, ao nível de significância de 10%. O modelo para prever o risco de gravidade no veículo V1 revelou um bom desempenho, com uma exatidão de 61.2% (D.P. 2.4). Ao prever a probabilidade de feridos graves e/ou mortos no veículo V2, a cilindrada do veículo V1 aumentou o risco para os ocupantes do veículo V2, ao nível de significância de 5%. O modelo para prever o risco de gravidade no veículo V2 também revelou um desempenho satisfatório, com uma exatidão de 40.5% (D.P. 2.1).

Os resultados do indicador integrado SEG revelaram que os veículos mais recentes apresentam uma melhor classificação para os três domínios: segurança, consumo e emissões. Esta investigação demonstra que não existe conflito entre a componente da segurança, a eficiência energética e emissões relativamente ao desempenho dos veículos.

keywords

CART, CORINAIR, emissions, fuel efficiency, logistic regression, rare events, safety, severity, single-vehicle crashes and two-vehicle collisions.

abstract

During the last years, the number of fatalities per million inhabitants in Portugal has always been higher than the average in the European Union. Therefore, at national level, there is a need for a more effective understanding of crash data and vehicles effects on crash severity. This research examined the effects of vehicle characteristics on severity risk, fuel use and emissions. The main goal of this research was to develop models for crash severity prediction in single vehicle-crashes and two-vehicle collisions. Furthermore, this research aimed at developing an integrated analysis to evaluate vehicle's safety, fuel efficiency and emission performances. Crash data were collected from the Portuguese Police Republican National Guard records for the Porto metropolitan area, for the period 2006-2010. A total of 1,374 crashes were collected, 500 single-vehicle crashes and 874 two-vehicle collisions. For the safety analysis, logistic regressions were used. For single-vehicle crashes, the effect of vehicle characteristics to predict the probability of a serious injury and/or killed in vehicle occupants (designed as binary target) was explored. For two-vehicle collisions, additional binary targets were designed: one target to predict the probability of a serious injury and/or killed in vehicle V1) and another target to predict the probability of a serious injury and/or killed in vehicle V2). To overcome the challenge imposed by sample size and high imbalanced data (only 5.1% were severe crashes), research methodology was developed based on a resampling strategy and 10 stratified random samples were used for validation. During the modeling stage, the effect of vehicle characteristics, such as weight, engine size, wheelbase and age of vehicle were analyzed.

For the vehicle's fuel efficiency and emissions analysis, pollutants were estimated using CORINAIR methodology. Following, emissions data were fit into linear regression models.

Finally, an integrated analysis indicator (entitled "SEG") that provides rating classification for the evaluation of vehicle's safety, fuel efficiency and emission performances, was developed.

Regarding these results, for single-vehicle crashes, injury severity prediction model identified age of the vehicle and engine size as statistically significant, at 5% level. Model performance accuracy rate was 58.0% (S.D. 3.1). For two-vehicle collisions, when predicting injury severity in vehicle V1, the engine size of the opponent vehicle (vehicle V2) increased the risk for the occupants of the subject vehicle (vehicle V1), at 10% level. Injury severity prediction model for vehicle V1 revealed a good performance with a mean prediction accuracy rate of 61.2% (S.D. 2.4). When predicting injury severity for the other vehicle involved (vehicle V2), the engine size of the opponent vehicle (vehicle V1) increased the risk for the occupants of vehicle V2, at 5% level. Injury severity prediction model for vehicle V2 achieved a mean prediction accuracy rate of 40.5% (S.D. 2.1).

abstract (cont.)

The results of the integrated analysis indicator, SEG, revealed that recent vehicle achieved better rating simultaneously for all the three domains: safety, fuel efficiency and emissions performances. Newer vehicles showed a better overall safety rating, were more fuel efficient (less CO_2 emissions) and reduced emissions (more environmental friendly). This research relevance showed that there is no trade-off between safety, fuel efficiency and emissions.

"*Excellence is an art won by training and habituation. We do not act rightly because we have virtue or excellence, but we rather have those because we have acted rightly. We are what we repeatedly do. Excellence, then, is not an act but a habit*."

Aristotle

INDEX

INDEX OF FIGURES

INDEX OF TABLES

LIST OF ABBREVIATIONS AND SYMBOLS

ACAP	Portuguese Automobile Association
AgeV1	Age of Vehicle V1
AgeV2	Age of Vehicle V2
AgeV2V1	Age differential between V2 and V1 in two-vehicle collisions
AIC	Akaike Information Criteria
AlcoholDrugs	Alcohol and/or Drugs
All	All the crashes including single-vehicle and two-vehicle collisions
AMLE	Analysis of the maximum likelihood estimates
ANSR	National Authority for Road Safety
AR	Accuracy rate
ASE	Average Square Error
BAC	Blood Alcohol Content
BEAV	Boletim Estatístico de Acidentes de Viação
BSR	Brand Severity Ratio
c.c.	Engine size
CAFE	Corporate Average Fuel Economy
CAR	Crash Analysis and Reporting
CARE	EU's road accident database
CART	Classification and Regression Trees Analysis
ccV1	Engine size of vehicle V1
ccV2	Engine size of vehicle V2
ccV2V1	Engine size differential between V2 and V1 in two-vehicle collisions
Chi-Sq	Chi-square test
CI	Confidence interval
CO	Carbon monoxide
CO_2	Carbon dioxide
Copert	Computer Program Emissions from Road Transport
CORINAIR	Emissions inventory guidebook
CrashCode	Crash Type
CSI	Crash Severity Index
DF	Degrees of freedom
DivisionCode	Divided/undivided

DUI	Driver under the influence
EC	European Commission
EM	Enterprise Miner
ENSR	National Road Safety Strategy
ER	Efficiency rating
ERSO	European Road Safety Observatory
ESC	Electronic Stability Control
ETSC	European Transport Safety Council
EU	European Union
EU27	European Union - 27 Members States
EuroNCAP	European New Car Assessment Program
EUROSTAT	European Statistics
FARS	Fatality Analysis Reporting System
FATALSIK	Overall crash severity; Probability of a serious injury and/or killed
FATALSIKV1	Probability of a serious injury and/or killed in vehicle V1
FATALSIKV2	Probability of a serious injury and/or killed in vehicle V2
FCT	Portuguese Science and Technology Foundation
FDOT	Florida Department of Transportation
FN	False negative
FP	False positive
GDP	Gross Domestic Product
GHG	Greenhouse Gases
GNP	Gross National Product
GNR	Portuguese National Guard
GR	Green rating
H	Hybrid
HeadOn	Crash type for collisions - Head-On
HEVs	Hybrid Electric Vehicles
hr	Hour
HLDI	Highway Loss Data Institute
IIHS	Insurance Institute for Highway Safety
IMT	Institute for Mobility Transportation
IMTT	Institute for Mobility and Inland Transportation
INE	National Statistics Institute
IRTAD	Traffic Safety Data and Analysis Group
IRTAD	International Traffic Safety Data and Analysis Group

ISS	Injury Severity Score
ITF	International Transport Forum
K	Killed
KDOT	Kansas Department of Transportation
LCL	Latent class logit
LDDV	Light Duty Diesel Vehicles
LDGV	Light Duty Gasoline Vehicles
LDV	Light-duty vehicles
LI	Light Injury
LOS	Level of service
LPDV	Light Passenger Diesel Vehicles
LPG	Liquefied Petroleum Gas
LPGV	Light Passenger Gasoline Vehicles
LPV	Light Passenger Vehicles
LR	Logistic Regression
LSS	Lane Support Systems
MARS	Multivariate Adaptive Regression Splines
MISC	Misclassification
MNL	Multinomial logit
MOVES	Motor Vehicle Emissions Simulator
N	Sample size
NASS CDS	National Automotive Sampling System Crashworthiness Data System
NASS GES	National Automotive Sampling System General Estimates System
NEDC	New European Driving Cycle
NHTSA	National Highway Traffic Safety Administration
NN	Neural networks
NO_x	Nitrogen oxides
NRSS	National Road Safety Strategy
NVehicles	Number of vehicles involved
OECD	Organisation for Economic Cooperation and Development
OR	Odds Ratio
OSI	Overall Severity Index
OSS	Overall Safety Score
P	Probability of the event
PCs	Passengers Cars

PEMS	Portable Emissions Monitoring System
PIN	Road Safety Performance Index
P_{leaf}	Probabilistic estimates of the cases correctly predicted in the leaf
PM_{10}	Particulate Matter with diameter less than 10 µm
$PM_{2.5}$	Particulate Matter with diameter less than 2.5 µm
pp	Page
PPPRA	Portuguese Plan for the Prevention of Road Accidents
PRSO	Portuguese Road Safety Observatory
PSP	Public Safety Police
RanOff	Crash type for single vehicles - Ran off road
RearEnd	Crash type for collisions - Rear End
RLP	Registration License Plate
RoadClass	Road Class
Rollover	Crash type for single vehicles - Rollover
S.D.	Standard deviation
SAS	Statistical Analysis Software
SC	Schwarz Criterion
SE	Standard error
SEG	Safety, Energy, Green indicator
SEMMA	Sampling, Exploring, Modifying, Modeling, and Assessing
SI	Serious Injury
Sideswipe	Sideswipe
SIK	Serious Injury and Killed
Single	Single-vehicle crashes
SpeedLevel	Speed Level
SPSS	Statistical Package for Social Sciences
SR	Safety rating
SRS	Severity risk score
SSE	Sum of square errors
SUV	Sport Utility Vehicle
TADS	Traffic Accident Database System
TEMA	Centre for Mechanical Technology and Automation
TN	True negative
TP	True positive
Two	Two-vehicle collisions
UN	United Nations

UNRSC	United Nations Road Safety Collaboration
US	United States
US	United States
V1	Vehicle V1
V2	Vehicle V2
VIM	Variable importance measure
VIN	Vehicle Identification Number
VPF	Value of preventing one road fatality
VSP	Vehicle Specific Power
WBGRSF	World Bank‘s Global Road Safety Facility
WBV1	Wheelbase of Vehicle V1
WBV2	Wheelbase of Vehicle V2
WBV2V1	Wheelbase differential between V2 and V1 in two-vehicle collisions
WeatherCode	Weather Conditions
WF_E	Weighting factor for efficiency rating
WF_G	Weighting factor for green rating
WF_S	Weighting factor attributed to the safety rating
WHO	World Health Organization
WTV1	Weight of Vehicle V1
WTV1V2	Weight differential between V2 and V1 in two-vehicle collisions
WTV2	Weight of Vehicle V2
Y	Response variable (target being modelled)
β0	Intercept
β1	Estimate for the parameter x1

CHAPTER 1

INTRODUCTION

Worldwide, 1.3 million people die annually as a result of a road traffic accidents, leading to more than 3,000 deaths each day [1]. Between 20 to 50 million more people suffer non-fatal injuries, with many suffering a disability as a result of their injury level [2]. The World Health Organization (WHO) has estimated around the same rate, 1.3 million deaths per year, caused by urban air pollution [3]. During the last years, passenger vehicles have shifted towards two extremes: small and light vehicles and large and heavy vehicles [4]. As a result, vehicle fleet is now highly variable in terms of mass, engine power and vehicle size. The main goal of this Doctoral Thesis was to investigate the effect of vehicle characteristics in injury severity risk, fuel consumption and emissions. It considers if lighter and smaller vehicles represent a higher risk to its occupants. On the other hand, it explores if larger and heavier passengers' vehicles decrease the risk towards its occupants, imposing at the same time, higher risk towards the occupants of a lighter and smaller vehicle involved in the collision. The research then combines those findings with vehicles emission estimations to address the important question if there is a trade-off between vehicle's safety performance and its fuel efficiency and emissions performance.

An introduction to the present work is carried out in this Chapter, which comprises: background for road safety and vehicles emissions, research motivation and main objectives. Finally, a structured reading guide for this Thesis is provided.

1.1 Background

During the last two decades, the number of registered vehicles has increased exponentially worldwide leading to a significant increase in road emissions, as well fuel used by the transportation sector. For passengers travel, road transport dominates as it carries 79% of passenger traffic [5]. Between 1970 and 2000, the number of cars in the European Union (EU) increased from 62.5 million to nearly 175 million [5]. Since motor vehicles become a common means for transportation, traffic injuries are not the only major concern. Reduction of greenhouse gases (GHG) emissions and fuel consumption have also become a main issue to health, environmental and transportation authorities. As traffic volume is increasing, road transport alone accounts for 84% carbon dioxide (CO_2) emissions attributable to transport [5].

Road safety progress depends to some extent on what one uses as a measure of exposure to risk (for example, population, registered vehicles, distance travelled). More than 90% of the world's fatalities on the roads occur in low-income and middle-income countries, even though these countries have approximately half of the world's vehicles [2]. In 1998 the ratio of the number of road deaths in Sweden and Portugal, two countries with comparable population, was 1 to 4.5 [6]. As the health and transport sectors developed their level of co-operation, fatalities per 100 000 population is becoming more widely used [7, 8]. Fatalities over distance travelled have traditionally been preferred by road transport authorities as this implicitly discounts fatality rates if travel is increased [7].

Along with the human suffering described above, road crashes have economic costs. In 2010, the United Nations (UN) and World Health Organization (WHO) reference the economic consequences of motor vehicle crashes as representing 1 to 3% of the gross national product (GNP) of the world countries, reaching over $500 billion [1]. The value of preventing one road fatality (VPF) has been estimated in 1.84 million Euros [9]. At the National level, in 2010, the economic and social cost of road accidents has been estimated at 1,890 thousand million Euros, representing 1.17% of the Portuguese GNP [10].

1.1.1 Road accidents- contributing factors

Road traffic accidents result from a combination of factors related to the elements of the system involving roads, environment, vehicles and road users, and the way they interact [11]. Some factors contribute to the occurrence of an accident and they could be part of crash causation as well. Other factors magnify the effects of the collision and thus contribute to severe outcomes.

The risk factors involved in road crashes injuries are grouped into two categories [11, 12]:

1) Risk factors influencing crash involvement: a) Inappropriate and excessive speed; b) Presence of alcohol and/or drugs; c) Fatigue; d) Being a young male; e) Inadequate visibility or poor weather

conditions; f) Vehicle factors (such as braking and maintenance); g) And defects in road design and inefficient maintenance.

2) Risk factors influencing crash severity: a) Human tolerance factors (such as age, sex and health conditions); b) Excessive speed; c) Seat-belts and child restraints not used; d) Roadside objects not crash-protective; e) Presence of alcohol and other drugs; f) And insufficient vehicle protection for occupants and for those hit by the vehicle.

In addition, there are also factors influencing the exposure to risk, such as economic factors and social deprivation, and risk factors influencing post-crash outcomes of injuries as difficulty in rescuing and delay in transport of those injured to the hospital. More information on popular analytical approaches to identify risk factors involved in road traffic injuries are provided at Appendix 1.

1.1.2 Road safety in Europe

Despite the improvement in road safety, road accidents and their consequences remain a serious social problem: on average 75 people lose their lives every day on European roads and 750 are seriously injured [13]. Road safety statistics for the EU and Portugal are presented.

1.1.2.1 Road safety performance in the EU

The number of road fatalities in the EU-27 fell during the decade between 1999 and 2009, from 57,691 deaths to an estimated value of 34,500 deaths [14]. The year of 2001 was a reference year since the European Commission (EC) published the White Paper- *"*European transport policy for 2010: time to decide*"*, which aimed to set an ambitious target of reducing the yearly number of road deaths by 50% by 2010 compared to 2001 [6]. Subsequently, the EU set an ambitious goal to halve the number of road deaths by 2010, expecting to save 25 000 lives [15]. As illustrated in Figure 1.1, the proposed target of halving road deaths between 2001 and 2010 was not achieved in the EU (30,500 deaths were above the target).

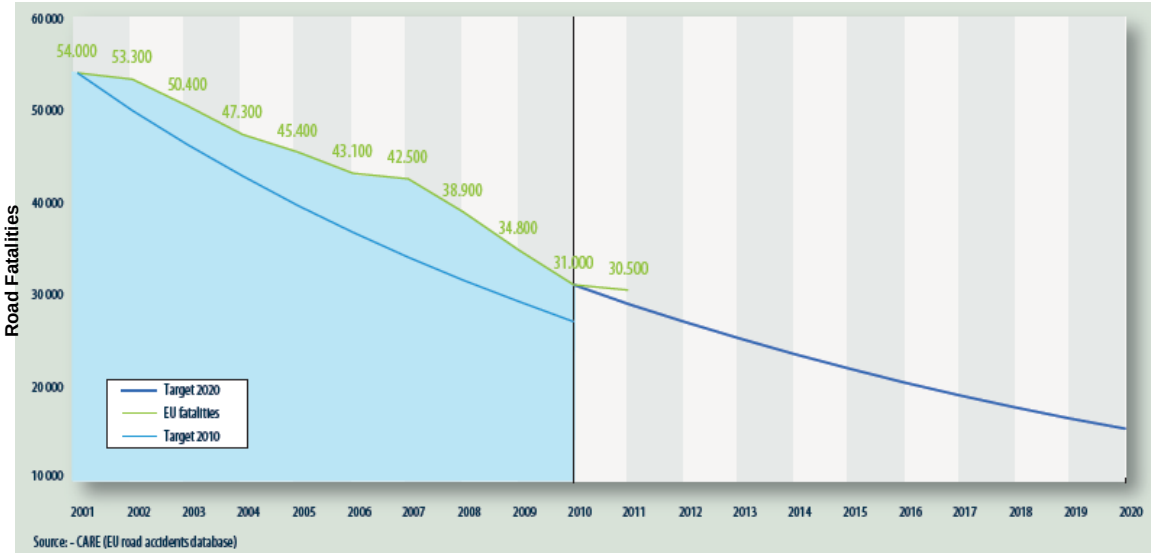


Figure 1.1 - Road fatalities in the EU since 2001 and targets objective from 2010 to 2020 [16].

Table 1.1 shows police-recorded road fatalities on the basis of death within 30 days for selected members of the International Traffic Safety Data and Analysis Group (IRTAD) [7]. Sweden was the country that have achieved the highest reduction in road fatalities (-52.0%) for the long-term (2010-2001). IRTAD data, showed a reduction of 49.3% in road fatalities, for Portugal during the same long-term period.

Table 1.1 – Selected European Countries Road Fatalities on the 30 days basis [7].

Recent data			Change trend	
Country	**2010**	**2009**	**Annual change 2010-2009**	**Long-term change 2010-2001**
France	3992	4273	-6.58%	-51.1%
Germany	3648	4152	-12.1%	-47.7%
Portugal	**937***	**929***	**0.9%**	**-49.3%**
Sweden	266	358	-25.7%	-52.0%
United Kingdom	1905	2337	-18.5%	-47.1%

***Data for 2010 was previous to the National Road Safety Strategy 2008-2015 Midterm Review

Comparison of road safety progress between 2001 and 2010 shows that EU achieved a reduction of 43% of road fatalities, from 54,302 to 30,900 road fatalities, respectively [17]. For the same period, Portugal have achieved a reduction of 50%, from 1670 in 2001 to 845 in 2010, using a basis of 24 hrs [17]. The results achieved for this period were published under the 5th Road Safety Performance Index (PIN) Annual Report [18]. Latvia, Estonia, Lithuania, Spain, Luxembourg, Sweden, France and Slovenia all reached the EU 2010 target.

Following the EU target between 2001 and 2010, EU has renewed its commitment to improving road safety by setting a target of reducing road deaths by another 50% by 2020, compared to 2010

levels. In 2011 more than 30,000 people died on the EU roads [19]. The current 6th PIN Annual Report presents in the results of the first year of progress towards the EU target of halving road deaths between 2011 and 2020. Norway leads this ranking with a 20% reduction in road deaths. On the other hand, Portugal reached the 2010 target with just one year of delay [9]. The 3% reduction in road deaths in the EU in 2011 compared with 2010 is below the 5.7% average annual reduction observed for the 2001-2010 decade and also below the 6.7% annual reduction that would have been needed from 2010 to reach the EU 2020 target [9].

1.1.2.2 Road safety performance in Portugal

Portugal has adopted directives that aim safer roads, compulsory use of seatbelts, standardized driving licenses and roadworthiness testing of vehicles [6]. In 2003, the Portuguese Plan for the Prevention of Road Accidents (PPPRA) was approved in order to control the high level of road accidents [20]. The target adopted by PPPRA was a 50% reduction in the number of fatalities and serious injuries by 2009 in comparison to the average for 1998-2000 [20]. In 2007 the National Road Safety Authority (ANSR) was created under the Ministry of Internal Affairs. In 2009 the National Road Safety Strategy (NRSS) for 2008-2015 it was presented with the purpose to define 10 strategic objectives, monitoring and assessing further actions [20]. The two major targets of NRSS for 2008-2015 are presented next. The first target, aims the reduction in the road mortality rate (expressed by the number of road deaths per population) [20]:

- 78 deaths per million inhabitants by 2011;
- 62 deaths per million inhabitants by 2015.

The second target, intends to control the road deaths to 579 until 2015 [15]. Prior to 2009, fatalities were reported on the 24 hrs basis. Working groups have defined correction factors as a conversion coefficient to estimate the fatalities, so that comparisons on the basis of the 30 day-definition could be made with other countries. Until 1997 Portugal applied a conversion factor of 1.30 (shadow area in Figure 1.2) and starting in 1998, this value was updated by a working group to 1.14 [15]. In 2009, to meet international agreed definitions, the NRSS established a methodology to account the road deaths within 30 days, based on the government document “Despacho n.º 27808/2009” [21]. Between 1970 and 2010, the number of fatalities decreased by 48% while the number of vehicles was multiplied by seven [7]. Figure 1.2 illustrates that despite of the overall progress, after 1970 (when motorization become more visible) the number of fatalities per million inhabitants have always been higher in Portugal, than the average in the European Union.

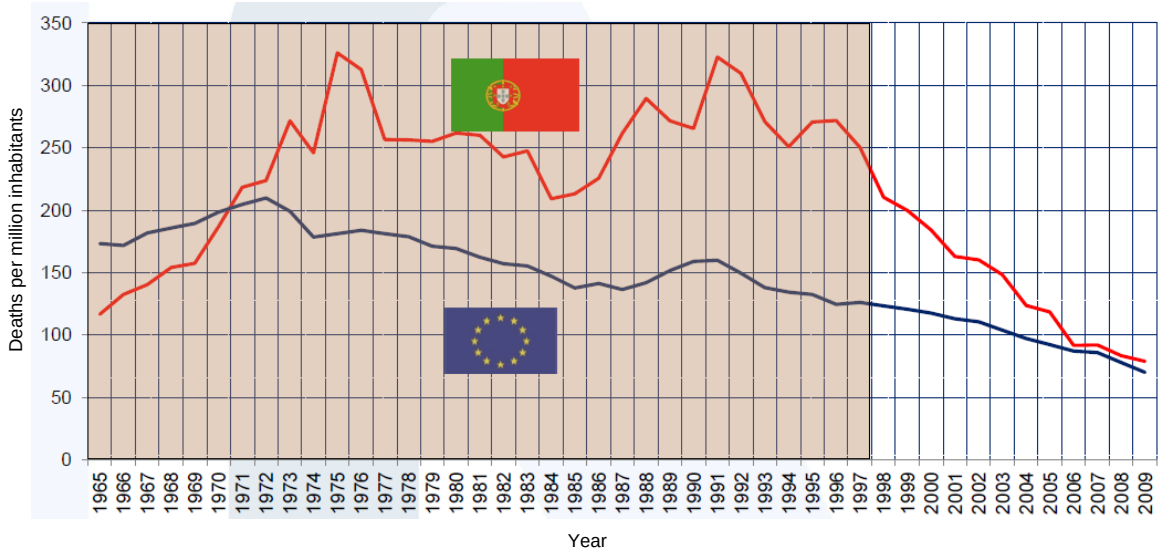


Figure 1.2 - Road fatalities in Portugal and the UE per million inhabitants: 1965 to 2009 [15].

Since 2000, the rate of decline has accelerated, with an average annual decrease of 7.3% between 2000 and 2010 [7]. For the decade, 2000 to 2010, the decrease in fatalities was reduced by -54%, as shown in Table 1.2 [7].

Table 1.2 - National reported road fatalities, injury crashes and rates in Portugal: 1970-2010* [7].

Indicator	1970	1980	1990	2000	2009	2010	2010 change over		
							2009	2000	1990
Fatalities	1785	2850	2924	2053	929	937	0.9%	-54%	-68%
Injury crashes	22662	33886	45110	44159	35484	35426	-0.2%	-20%	-21%
Deaths/100000 population	20.6	30.6	31.2	20.0	8.7	8.8	0.8%	-56%	-72%
Deaths/10000 registered vehicle	22.7	14.7	13.4	4.3	1.6	1.6	0.0%	-63%	-88%
Motorize vehicles/1000 inhabitants	91	208	234	462	543	545	0.4%	18%	133%

***Data for 2010 was previous to the National Road Safety Strategy 2008-2015 Midterm Review

ANSR has available road fatalities on the 30 days basis since 2010. In 2010 there were 35,426 injury crashes, which had result in 2,475 serious injured and 937 fatalities [22]. The latest ANSR annual report on road safety on the 30 days basis showed that during the year 2011, there has been a total of 32,541 crashes involving injuries and those resulted in 2,265 serious injured and 891 fatalities [23]. In 2012, ANSR has released a term review document of the National Road Safety Strategy for 2012-2015 in order to improve statistics accuracy [24]. During this review process, the impact of the new methodology on assessing fatalities was visible in comparison with the records on the 24 hours following the road crash, for which an increment of 14% was been

applied pos 1998 [24].The real number of road deaths, within the 30 days, was 26% and 29% higher for 2010, and 2011, respectively [24]. Following this revision, for 2010, ANSR has updated the previous indicator of 88 deaths/(million inhabitant), in Table 1.2, to 92 deaths/(million inhabitants), much higher than 62 deaths/(million inhabitant) for the average in the EU-27 [23, 25]. Regarding to the strategic target set for 2011, 78 deaths.(million inhabitant)$^{-1}$ was not reached, since there were 89 deaths.(million inhabitant)$^{-1}$ [26].

Previously to close the section, Road Safety in Europe, the economic crisis may had an impact in the positive road safety progress in the EU through a variety of effects in the society: a decrease in mobility, less inexperienced drivers with relatively higher risks, a reduction in leisure driving, and a safer driving behavior intended to save fuel [7]. However this relationship is not fully explained. If cost concerns may reduce individuals trips, hence reducing the risk of a crash exposure, on the other hand, vehicles owners tended to avoid spending money with vehicle maintenance. In addition, the increase of the unemployment rate and purchasing loss power force consumers to drive older cars.

1.1.3 Trends in vehicle's emissions and fuel use

Transportation systems are vital to world's prosperity, having significant impacts on economic growth, social development and the environment. Although the transportation sector accounts for about 7% of European GDP, its environmental cost offset 1.1% of GDP [27]. In 2010, transport sector account for 31.7% of the energy consumption in the EU-27, and road transportation in particular represented 82.1% of the total transportation consumption [25]. In 2010, gasoline/diesel oil accounted for 53% of the total consumption, reflecting an increase of 9% compared to 2000 [28].

Transport greenhouse gases (GHG) emissions accounted for 24% of GHG emissions from all sectors in the EU-27, in 2010 [29]. In particular, road transport contributed to 71.1% of the 24% share in GHG emissions from the transportation sector. Transport GHG emissions (including from international aviation) as the target defined in the White Paper, were 26% above 1990 levels [29]. In 2010, transport emissions decreased by 0.4% compared to 2009 [29]. For 2011, a similar reduction of 0.4% was estimated [29]. The decline in GHG emissions from road transport since 2009, can be mainly attributed to the decline in freight transport demand related to the economic recession and higher fuel prices [29].

In the analysis of CO_2 emissions among the EU car fleet, vehicle weight is a very important factor as more weight needs more energy to move the vehicle, thus, it increases the fuel needed for the same driving distance. During 1995 to 2003, diesel vehicles weight increased by 11.6% (140 kg), while the average gasoline vehicle by 15% (160 kg) [4]. Even if weight generally increases during

those years, CO_2 emissions decrease was due to the increased combustion efficiency, leading to lower fuel consumption and thus, lower CO_2 emissions. Generally, diesel and gasoline light passenger vehicles are shifted to the two extremes in the passenger vehicle fleet: very light and very heavy vehicles. As for vehicle weight, there was a general shift to smaller and bigger engines for both diesel and gasoline light passenger vehicles [4].

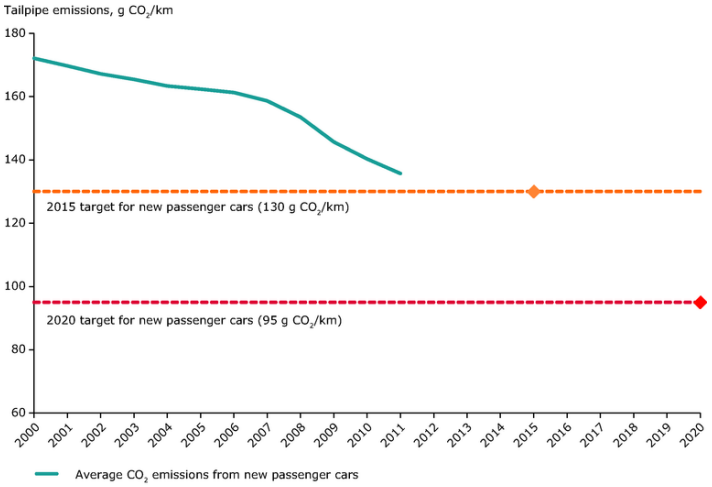


Figure 1.3 - Average CO_2 emissions for new cars ($gCO_2.km^{-1}$) in EU-27 and targets for 2015 and 2020 [29].

In general, CO_2 emissions for passenger cars have been decreasing since 2000, as illustrated in Figure 1.3. The average passenger car emissions target of 130 g $CO_2.km^{-1}$ for the new car fleet by 2015, and a target of 95 g $CO_2.km^{-1}$ from 2020 onwards are marked on orange and red colors respectively, in Figure 1.3. CO_2 emissions from the new passenger car fleet in the EU-27 decreased from 140.2 g $CO_2.km^{-1}$ in 2010 to 135.7 g $CO_2.km^{-1}$ in 2011 [29]. In 2011, average CO_2 vehicle emissions for most carmakers were below target levels estimated for 2012. New cars in 2011 were on average 3.3 % more efficient than those vehicles models registered in 2010 [30].

The progress done with EU regulations and emissions targets has been decreasing the average vehicle CO_2 emissions. In 2009 the European Union adopted a Regulation [EC] No. 443/2009 to impose the CO_2 emissions of 130 $g.km^{-1}$ on the fleet average, by 2012 [31]. However, due to the economic recession worldwide and its effect on the automotive industry, the EU has shifted the CO_2 emissions of 130$g.km^{-1}$ target to be achieved by 2015. The target of 130 $g.km^{-1}$ (5.6 $L.(100km)^{-1}$) for the average emissions of new cars was also phased-in by 2015 and 95 $g.km^{-1}$ (4.1 $L.(100km)^{-1}$) by 2020 [31, 32]. Then, the EU is expecting that 2015 and 2020 targets will represent a reduction of 18% and 40% respectively compared with the year 2007 fleet average of 158.7 $gCO_2.km^{-1}$ [32]. CO_2 emissions and fuel consumption are closely related. To achieve Europe's targeted 80% CO_2 reduction by 2050 compared to 1990, oil consumption in the transport sector must drop by around 70% from nowadays [28].

Actions to reduce GHG emissions, pollutants and noise from vehicles will benefit from shifting from conventional modes to hybrid and electric vehicles, cleaner fuels and improved vehicle technology. This form should be complemented by better managing transport demand. Also, reduction of motorway speed limits from 120 to 110 $km.h^{-1}$ would reduce fuel consumption by 12 % for diesel cars and 18 % for gasoline cars [29].

During the last years, goals have been set for safer and more sustainable mobility. In 2010, the United Nations Road Safety Collaboration and the World Health Organization lunched the Global Plan for the Decade of Action for Road Safety 2011-2020 in more than 100 countries, with one goal: to prevent five million road traffic deaths globally by 2020 [1]. In 2011, the White Paper "Roadmap to a Single European Transport Area – Towards a competitive and resource efficient transport system" was published [33]. Concerning to road safety, the framework established the goal that by 2050, the EU must move closer to zero fatalities in road transport. This document defined ten goals for a competitive and resource efficient transport system benchmarks for achieving the 60% GHGs emission reduction target [33]. It sets the 'Europe 2020' strategy to achieve CO_2 emissions reductions by 60 % by 2050 compared to 1990 levels [29, 33]. Hence it is required to cut the emissions in 68 % from 2010 to 2050 to meet this target. Concerning to road safety, the framework established the goal that by 2050, the EU must move closer to zero fatalities in road transport. On the other hand, the Horizon 2020 Transport challenge work program encourages research in areas such as: power train technology for law CO_2 and polluting emissions, and traffic safety [13]. Is that possible an integrated approach towards vehicle safety and emissions? Thinking about an answer to this question leads to the motivation of this Doctoral research, stated in the next section.

1.2 Motivation

A study amongst 21 European Countries has indicated that Portugal had the lowest road safety performance score, and suggested that Portugal should invest more in vehicle safety technology and in promoting new(er) cars [34]. During the last decennia there has been an increase in the amount of consumer interest in the vehicle safety performance and fuel economy. Consumers tend to equate vehicle safety with the presence of specific features or technologies rather than with the outcomes of vehicle crash safety/test or crashworthiness [35]. Crash testing is a valuable source for consumer regarding vehicle crash safety and credits a car manufacturer for focusing on safety. Under the EuroNCAP, the frontal impact takes place at 64 $km.h^{-1}$, meanwhile the car strikes deformable barrier that is offset [36]. It simulates one car having a frontal impact with another car of similar weight. Hence, it can only be compared with vehicles in the same class and within a 113 kg weight range [37]. EuroNCAP discourage consumers from comparing ratings of cars from different segments, and in real crashes, there is obviously no control on the vehicle categories involved. Despite the scientific conditions under which crash tests are conducted, they have limitations as follows. First, they do not account for weight differential between the vehicles involved within the collision. Second, the speed of the crash impact frequently is higher than 64 $km.h^{-1}$, which is the speed at the frontal impact takes place in crash testing [38-40]. Third, crash testing is only performed for selected models, whereas in real roads there is no control neither in vehicle body type, neither in the age of vehicles model year. EuroNCAP recognizes there is no capability to determine what would happen if cars of widely different masses impact each other [41]. Crash testing programs do not attempt to predict the real crash outcome, rather than provide an indication of safety best practices that had been implemented in individual vehicle models. During the last years, due to fuel economy and CO_2 emissions targets, and global recession, manufacturers have increase the sales of smaller, lighter cars to offset the fuel economy by their bigger, heavier models. Minicars are more affordable, and they use less fuel and emit less pollutants, however the safety tradeoffs are a challenge. In a collision involving two vehicles that differ in size and weight, the occupants of the sampler lighter car will be in disadvantage? Would a consumer have to choose the heaviest on the road to gain safety benefits? But if it does, other road users could be at higher risk specially the ones travelling in a lighter car. On the other hand, if all new passenger cars would shift towards larger and heavier vehicles, then what would be the cost in fuel consumption and emissions? Addressing these questions yield to fourth main motivations for this research:

1. In Portugal, there is a gap in incorporating vehicle characteristics in road safety analysis.
2. Crash testing has limitations in prediction crash compatibility amongst vehicles of different segments.
3. It is unclear if more environmental friendly vehicles impose a trade-off on its vehicles' occupants.
4. An integrated approach towards vehicle safety, energy and emissions should be available not only to policymakers but also to consumers.

In road safety analysis three key elements are fundamental: vehicle, infrastructure and driver. Infrastructures design has been significantly improved over the decades. Driver behavior is complex, subjective and often unpredictable. Therefore the analysis of vehicles effects on severe crash outcomes plays a central role. Police records data is a valuable source for crash analysis. A better understanding of the severe crashes outcomes demands the analysis of complex data, which events are significantly less frequent compared with minor severe crashes resulting in light injuries and/or property damage only. Rare events are part of the nature of crash injury data: injury severity level has been estimated by the following distribution: 61.0%, 15.3%, and 2.8% for no injury, possible injury, evident injury, and severe/fatal injury, respectively [42]. Other sources have estimated the overall probability of injury cases at about 2.8%, hence there would be about 35 times more probability for classifying a case as non-injury, than injury [43]. Data from the United States during the year 2010 reflects the imbalance between non-fatal crashes and fatal crashes; 99% to 1%, respectively [44]. During 2010 and 2011, the ratio of fatal crashes has been estimated around 2.7% [23]. With regard to binary data classification (severe crash vs. non severe crash), analysis of data containing rare events, poses a great challenge to the machine learning community. When probabilistic statistical methods are used, such as logistic regression, they underestimate the probability of the rare events because they tend to be biased toward the majority class (non severe crashes), which has significantly higher frequency compared to the minority class (severe crashes). When modeling a rare event, which happens in a very low frequency, it is difficult for the algorithm to find a valuable split, because the model is already predicting right the common event. The topics of imbalance datasets and sample balance have not been subject to a formal study in crash analysis. The overall crash severity at the crash sample explored in this study was 5.1%. The greatest challenge faced by this study was due to the disproportionate class distributions of the non-severe and severe events being predicted. To overcome this challenge, a balanced sample was derived from the original crash sample and it was modeling using binary classification methodologies.

This Doctoral research is part of the “SAFENV: Predicting the Trade-offs between Safety and Emissions for Road Traffic” (PTDC/SEN-TRA/113499/2009), project funded by the Portuguese Foundation for the Science and Technology (FCT). This is the first study conducted in Portugal which links vehicle specific characteristics with the crash outcomes. The analysis of crashes reports sample from the Oporto metropolitan area for the time 2006 to 2010, leads to important findings to address the contribution of the national car fleet in road safety progress. This research is intended to support decision-making for safe and sustainable transportation policy and mobility in Portugal.

1.3 Research Objectives

The main goal of this research was to develop safety prediction models based on real world crash data, which was collected from the Portuguese Police crash reports records. The effect of vehicle characteristics, such as make and model, engine size, weight, wheelbase, registration year (age of vehicle), and fuel type, on crash outcomes, expressed by the number of injuries and fatalities among the passengers, is analyzed. It is important to notice that the study focuses on post-crash consequences rather than on pre-crash contributing factors to the event. In addition to the safety analysis, vehicles technical information was also used to quantify their impact on fuel consumption and emissions.

The major objectives of this Doctoral Thesis are:

1. Determine if vehicles characteristics affect crash outcomes and identify which factors are more significant to predict crash injury severity.
2. Develop decision models to predict the probability of a serious injury and/or fatality in single-vehicle crashes and in two-vehicle collisions.
3. Develop logistic regression models to predict the probability of a serious injury and/or fatality in single-vehicle crashes and in two-vehicle collisions.
4. Identify which vehicles auto brands are more frequently involved in severe crashes and evaluate brand severity ratio involvement in the sample with the overall severity at Portuguese fleet.
5. Develop an integrated analysis score to evaluate vehicle's safety, fuel efficiency and green performances.

This study addressed the following questions: is there any vehicle dimension important for the crashworthiness? Is vehicle size or size differential between the two vehicles involved fundamental to safety? Is it possible for designers of new vehicles to cut carbon emissions without negatively affecting their safety performance? Is there a trade-off between vehicle safety, fuel efficiency and emissions performance? Can manufacturers accomplish the European Commission goal to decrease CO_2 emissions to 130 $g.km^{-1}$ by 2015, and still achieve a better management of crash forces?

In summary, this research is intended to support the decision-making process for transportation policy for safe and sustainable mobility in Portugal. The findings discussed in this Thesis will provide meaningful interpretations that can be used to identify potential correlations amongst crash analysis and vehicle characteristics effects in road severity risk. Further, the conclusions will provide a new assessment of the trade-off between safety and environment in the transportation research. It will also provide important information for automotive industry to produce low emission vehicles without compromising many of the basic vehicle functions of performance and safety.

1.4. Thesis Organization

The present Thesis is organized in 10 chapters, including: introduction, literature review, safety methodology, descriptive statistics, safety analysis results for single-vehicle crashes and two-vehicle collisions, emissions estimation and modeling, integrated analysis for vehicle's safety, energy and environmental performance, and conclusions. A Thesis reading guide is presented in Figure 1.4.

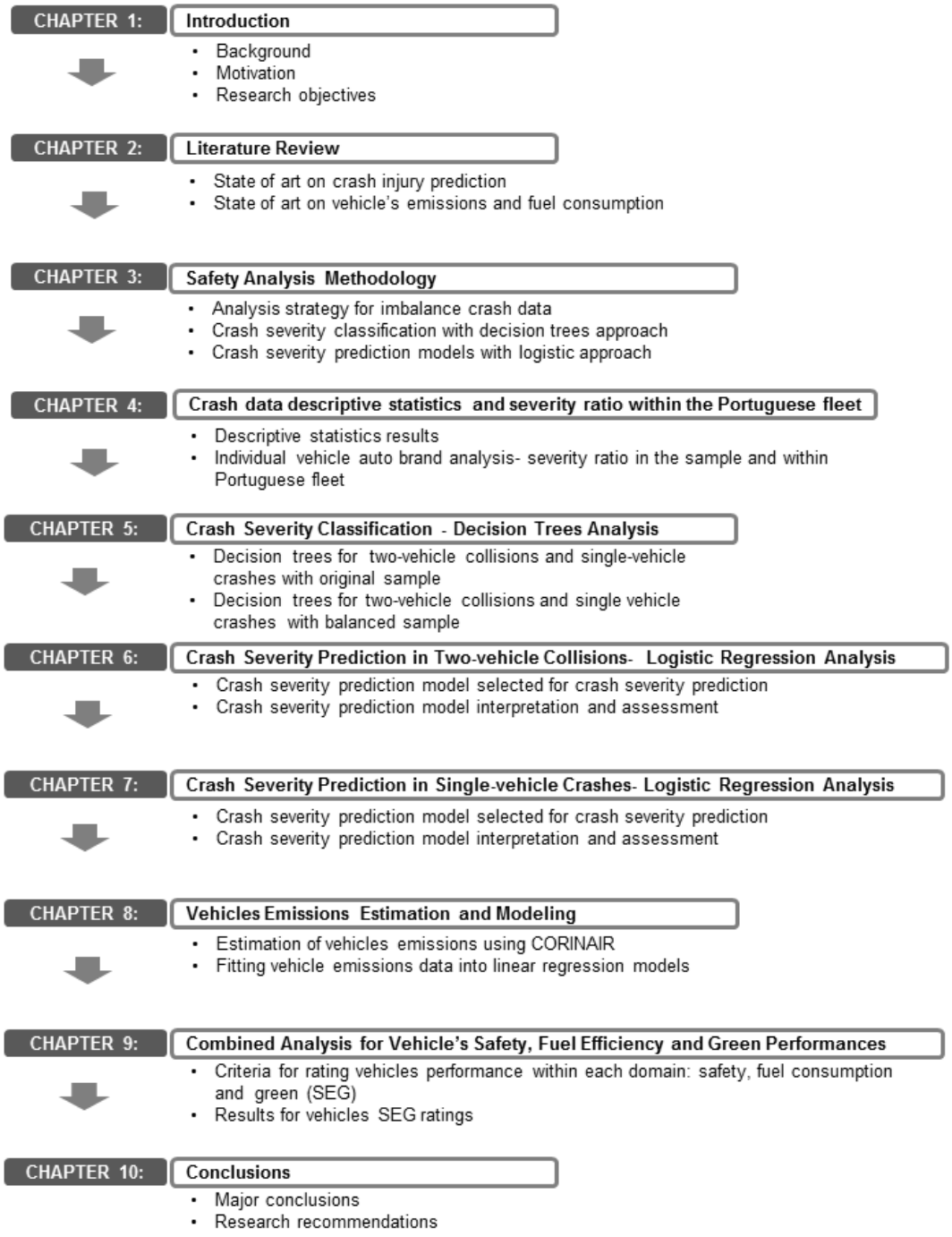


Figure 1.4 - Thesis reading guide fluxogram.

CHAPTER 2

LITERATURE REVIEW

Since safety and environment are "*transported"* together through this Doctoral Dissertation, this Chapter highlights previous studies for crash injuries analysis and vehicle's safety and environmental performance analysis. First, research in crash injury severity prediction modeling is presented. Second, it discusses statistical approach to deal with crash data complexity and unbalanced classes (among severe and non-severe observations). Third, it discusses the correlation of crash testing with real life crash outcomes. Fourth, studies addressing the trade-off between vehicle's safety and environmental performance are presented. Finally, main remarks of the existing studies in this literature review are emphasized.

2.1 Road Safety Main Risk Domains and Drivers' Behavior

Traffic safety is a subject with complex interactions amongst these three main factors: human behaviour, road and vehicle. Delen et al. had identified the factors that affect the risk of increased injury of occupants during a crash: demographic, behavioral characteristics of person, environment, roadway conditions and technical characteristics of the vehicle, among others [45]. Hermans et al. identified the following risks for road safety outcomes: alcohol and drugs, speed, protective systems, infrastructure, vehicle, and trauma management [34]. Driver behavior and driver characteristics not only affect the probability to be involved in crash event, but also, how his body will sustain the impact and his condition following the crash [15, 46].

Multiple socio-physiological factors may influence the injury and fatality outcomes in motor vehicle crashes. Awadazi et al. had investigated main risk factors for motor vehicle injuries and fatalities among younger and drivers 65 years of age or older [47]. The point of impact on a vehicle during a crash had increased risk of both injury and fatality for older drivers. Behavioral factors, such as alcohol involvement and lack of seatbelts, were likely to place all drivers at increased risk, with higher likelihood for crash fatalities [47]. The evidence shows major gender differences on the impacts of driver condition, seatbelt use and airbag deployment on injury severity risks. "Male drivers, older drivers, drivers who are not wearing safety belts, collisions occurring in a higher speed zone and head-on collisions significantly increase the risk of death" [48]. Airbag deployment, may impose a higher risk for female than for males [47]. Women and older drivers are more frequently killed than other groups under equivalent impact conditions [47-49]. As far as driver's age effect, 16 and 17-year-old drivers pose more than twice as much risk to occupants of other vehicles as do drivers aged 85 and older [50]. In addition to vehicle mass and vehicle type, drivers characteristics, as well as the circumstances of the collision affected the drivers' condition post-crash [48]. Despite of the drivers' conscious and/or unconscious behavior, Pompili et al. suggested that above 2% of the traffic accidents are suicide behaviors [51].

Human factor comprehension on crash injuries and fatalities requires further research and more efficient cooperation between police makers and auto-industry. The injury prevention measures for fatal crashes may potentially benefit younger and older drivers alike [47]. Eleven thought younger drivers were linked to the highest risk of collision (manly younger males), individuals aged 50 and over become the largest segment of potential buyers of automobiles in the marketplace, accounting for more than 40% of all new cars purchase [52]. If the automotive manufactures want to remain competitive, particularly given the recent economic downtown in this sector, "understanding the needs of older consumers and incorporating them into the design of the automobile is important" [52]. The development, design, and marketing of crash avoidance and safety-related vehicular technology to consumers are critical to ensure the vehicle purchased is the best fit with their safety and driving needs.

2.2 Vehicles' Size and Weight Effects on Occupants' Injury Risk

Evans explored vehicle mass and size basing his study on Newtonian mechanics. In this study, for crash between two cars of different masses, the fatality risk ratio of a lighter to a heavier car increases as a power function of mass ratio of the heavier to the lighter car [53]. Based on the law of the conservation of momentum, the change in the velocity for the individual vehicle is subject to the relative speed and mass proportions between the two vehicles involved in the collision [54]. Hence the mass influence the impact yielding to injury severity [54]. Vehicle mass and size variables are strongly correlated, which makes it difficult to determine the separate contribution of mass and size on crash risk [53]. Wood showed that in collisions between cars of similar size and in single vehicle crashes the fundamental parameters which determine the injury risk are associated to the size, i.e. the length of the vehicle [55]. However, in collisions between dissimilar sized cars the fundamental parameters are the weight and the structural energy absorption of the vehicle[55]. Wenzel and Ross found that mass alone is not an effective predictor of risk, on the basis of driver deaths per year per million registered vehicles for a given car model [56]. These authors suggested the quality of cars may be more correlated to risk than weight, but this correlation was not strong [56]. Robertson analyzed vehicles models from 2000-2005 and stated that although excess weight and horse power are adverse to other road users (cyclist and pedestrians), larger vehicle size is related to lower risk because "it gives occupants more room to decelerate in a crash" [57]. During 2007, the death rate in 1-3 year old minicars involved in multiple-vehicle crashes was nearly twice as high as the rate in very large cars [58]. Also for single-vehicle crashes, the fatality risk in minicars was found high as well as in multiple-vehicle crashes [58]. Broughton showed that the driver casualty rate decreases with the size of his/her car, however the driver casualty increased with the size of the other car involved in the collision [59]. Newer cars are safer for their occupants and more aggressive to occupants of cars with which they collide [59]. The author claimed that these effects are partly due to an increase in the mass of new cars [59]. A further update to this work, showed that the mean risk of death for a car driver in a collision with a car registered in 2004–2007 is about 23% greater than in collision with a car registered in 1988–1991 [60]. In car-car collisions when modern cars are involved, it was found fewer casualties, suggesting that the overall benefits of improved secondary safety have clearly outweighed the disbenefits of increasing aggressivity [60]. A more modern car provides better protection to its occupants, mainly achieved by the design efforts that have been made to improve secondary safety (crashworthiness), also the tendency to greater mass [60]. Méndez et al. advised that vehicles aggressivity and crashworthiness were influenced by vehicles mass, size and structural properties [61]. Improvements on vehicles safety increased the injury risk on the occupants of the older vehicles [61]. Zachariadis suggested that mass seemed to play an important role in frontal crash tests only [62]. Distribution of mass among vehicles, and not mass per se, is largely responsible for injury risks [62]. Huang et al. suggested that crashworthiness and crash aggressivity significantly vary by vehicle type with the dominating effect of vehicle mass [63].

Tolouei and Titheridge showed that increasing vehicle mass generally decreases the risk of injury to the driver [64]. The injury risk of occupants in the lighter car is higher than for heavier car, due to the greater velocity change during the collision [64]. More recently, Tolouei et al. confirmed that the probability of injury of the driver of vehicle 1 increases with speed limit and with increasing mass ratio ($mass_2/mass_1$), whereas the probability of injury of the driver of vehicle 2 increases with speed limit and with decreasing mass ratio [65]. Also, this study evokes that there is a protective effect of vehicle size above and beyond that of vehicle mass for frontal collisions [65].

Fredette et al. analysis showed that drivers of pickup trucks, minivans and sport utility vehicles were more aggressive than the drivers of others vehicles involved, while their vehicle provided ahead protection [48]. Keall and Newstead found that in single-vehicle crashes, SUVs are potentially harmful to their own occupants due to its high centre of gravity compared to the width of the wheel track, leading to greater instability and a higher risk of rollover [66]. When considering fatality rates by vehicle type, SUVs showed the highest rate per licensed vehicles [66]. However driver risk behavior was suggested as a strong contributor of this elevate risk [66]. Regarding to vehicles incompatibility between passenger cars and light trucks, motor-vehicle manufacturers have taken voluntary measures to reduce light truck aggressivity by adding crumple zones and reducing vehicle height [67]. When subject to a frontal crash, passenger vehicles are designed to absorb crash energy through deformation or crush of energy-absorbing structures forward of the occupant compartment. However, in collisions between vehicles of different body type, such as cars and light pickups or SUVs, the capacity of energy-absorption structures would not be fully utilized because mismatches often exist between the heights of these structures. Therefore, in 2009 new light trucks were required to have the front structure (frame rails) low enough to interact with the primary structures in cars, which for most cars is about the height of the front bumper [68]. Baker et al. study suggested that the lower front energy-absorbing structure showed a benefit of 19 % reduction in fatality risk to belted car drivers in front-to-front crashes crashworthiness has been a constant concern for road safety and vehicle design [68].

2.3 Crash Testing and Vehicle Safety Performance in Roads

The improvement in vehicles secondary safety (or crashworthiness) over the years has been proven by several studies [36, 49, 61, 69-71]. However, debating has been arising if the crash test results indicate the risk of fatality or injury in serious crashes. This section highlights studies on vehicles' safety and crash tests reliability with real crashes.

2.3.1 Perception of vehicles' safety

Once introducing model variations on the market, car manufacturers face trade-offs when choosing between interior volume, length*width, mass, maximum engine power, power-to-weight ratio,

acceleration, and fuel efficiency [72]. Understanding consumers' preferences for safety is essential for designing safer vehicles, for encouraging safe driving behavior and to improve overall road safety. Several studies have been conducted in order to gain a better understanding about consumer's perception of vehicle's safety [35, 52, 67, 73]. Koppel et al. investigated the key parameters associated with ranking 'vehicle safety' as the most important consideration in the new vehicle purchase [35]. Safety-related factors (e.g., EuroNCAP ratings) were more important in the new vehicle purchase than other vehicle factors (e.g., price, reliability) [35]. Likewise, safety-related features (e.g., advanced braking systems, front passenger airbags) were considered as more important than non-safety-related features (e.g., route navigation systems) [35]. Vrkljan and Anaby found that consumer's vehicle purchase is influenced by: crash test rating, cost, design, and reliability. In this study, safety, along with reliability, were considered most important if purchasing a vehicle amongst overall consumers [52]. Thus, studies have recommend a better understand of consumers' perceptions of safety to make easier to plan more effective safety policies and safety campaigns" [67, 73]. Consumers need to understand the importance of seeking low aggressivity in the vehicles they are purchasing to minimize harm to other road users with whom they may crash [66].

2.3.2 Correlation of crash testing with real crashes

There has been a long-standing debate about whether vehicle secondary safety is superior measured through real world crash analysis or controlled during laboratory testing. Lie and Tingvall focused on how do EuroNCAP results correlate with real-life injury risks, based on police reports crashes [54]. These authors claimed that Euro NCAP is not able to predict crash outcomes because start rakings system does not reflect the mass of the vehicles involved in the collisions, and mass has an important role in the impact severity distribution [54]. The results suggested that four-star cars seem to reduce the risk of a serious and fatal injury by more than 30% [54]. The importance of vehicle's weight (mass) should not be underestimated, and while this factor is not taken into account in crash tests into fixed barriers, in a car-to-car impact a 100 kg more weight difference will decrease the risk of any level of injury by 7% [65]. On the other hand, in single vehicle crashes, the mass should not have any significant influence on safety [65]. Mendez et al. showed that the average score of EuroNCAP test of new cars sold in Europe rose from 2 starts in 1988 to 4 stars in 2005. However this improvement on new cars safety rating did not translate into reductions of the risk of injuries faced by drivers in real traffic situations, because of the evolution of the car's mass fleet [61]. Kullgren et al. compared injury risk measures between Euro NCAP 2 and 5 Star cars with real-world injury outcomes using police and insurance injury data [74]. The 5-star rated cars were found to offer a superior safety performance over 2-star rated vehicles in the crash tests and real-world crash and injury performance. Contrary to the work of Lie and Tingvall, mentioned above, Kullgren et al. claimed that Euro NCAP crash tests were highly correlated with serious crash outcomes. These authors stated "though weights of new cars have gone up

substantially in recent years, the results of this study confirm that improved crashworthiness has been the primary factor in enhanced vehicle safety, rather than the increase in mass" [74]. Newstead et al. maintained that crash tests "do not account for vehicle mass effects in the real world and they only cover a limited range of crash types" [75].

Based on the literature review, two major limitations are pointed to crash testing.

First limitation is related to the crash testing impact speed. The speed of collision, the delta-v, has been identified by several authors as the most important variable to access crash severity outcomes [43, 53, 54, 61, 76]. For vehicle's occupants involved in impacts with a delta-v ≥50 $km.h^{-1}$, the risk of severe injury is more than five times greater than for those in the lower delta-v [69]. However, EURO NCAP frontal impact testing protocol version 6.0 included a car impact speed of 64 $km.h^{-1}$resulting in a delta-v of approximately 32 $km.\ h^{-1}$ for the occupants [69].

Second limitation is related to the difficulty to compare vehicles safety ratings amongst different segments. IIHS endorse the consumers to not compare ratings across vehicle size groups because size and weight influence occupant protection in serious crashes [77]. "Larger, heavier vehicles generally afford more protection than smaller, lighter ones" [77]. On the other hand, Euro NCAP recommend that crash testing only can be compared with vehicles in the same class and within a 113 kg weight range [37].

2.3.3 Vehicles' improvements: primary safety and secondary safety

Some authors have study vehicles improvements over the years, others have discussed the benefits of improved car primary safety.

Regarding vehicles improvements over the years, and following the studies presented in section 2.2, Broughton demonstrated that the proportion of injured car drivers who were serious injured or killed in modern cars was clearly less than in older cars [70]. The author suggested that the benefits have been proportionately greater in accidents occurring on roads with speed limits of at most 40 mph [70]. However, it was not conclusive if those severe injuries were due to the efforts of regulators and manufacturers to produce safer vehicles, or weather independent factors had contributed to the observed reductions. Ritcher et al. results reveal a decrease in crash severity (based on collision speed) and injury severity during the 1990s compared to the 1970s. It would appear that improvements in vehicle design lead to a greater reduction in injury severity from decreased crash severity alone [69]. Lund stated that, whereas vehicle safety has continuously improved for vehicle occupants as a whole, it has worsened for many individual drivers who are not driving the newest vehicles [78]. The author recognized that improvements in occupants protection from vehicle design have been offset by an increasing risky environment, such as driving behaviors and higher aggressively of the opponent vehicle [78].

Regarding to advanced safety technologies, some examples are highlighted in Appendix 2. Electronic Stability Control (ESC) “was the most important innovation in reducing of vehicle-related mortality in decades, perhaps the single most effect innovation since the invention of seat belts” [57]. Farmer focused on the potential of five crash avoidance technologies: blind spot detection/warning, forward collision warning, emergency brake assist, lane departure warning/prevention, and adaptive headlights [79]. Of those technologies, the one with the greatest potential was the forward collision warning system could prevent 2.3 million crashes in the United States each year [79]. Similarly to Farmer’s research, Jermakian suggested that a combination of four current technologies (side view assists, forward collisions warning/mitigation, lane departure warning and adaptive headlights) could mitigate 149, 000 serious and moderate injury crashes and 10,238 fatal crashes each year [80]. Also, forward collision warning was found by the author as having the greatest potential for preventing crashes of any severity.

A report from IIHS published the results of the Highway Loss Data Institute (HLDI) that analyzed five existing features: antilock brakes, electronic stability control (ESC), driver frontal airbags, side airbags, and forward collision warning, introduction in the vehicles fleet [81]. The IIHS reports states that it takes typically three decades for a promising safety feature first introduced in few luxury cars to spread through the fleet [81]. Although US government began requiring frontal airbags installation in some vehicles in 1996, it won’t be until 2016 that 95% of all registered vehicles will have frontal airbags [81]. ESC was introduced in 1995 models and was standard on 10 percent of 2000 models [81]. It is predicted that 95 percent of registered vehicles in 2029 will have ESC [81]. A newer report from IIHS stated that forward collision avoidance systems, particularly those that can brake autonomously, along with adaptive headlights, which shift direction as the driver steers, show the biggest crash reductions [82].

Regardless of all their potential benefits, the success of crash avoidance technologies in preventing crashes depends on several factors, including driver acceptance as well as drivers understanding which could make them to inappropriately respond to the alerts [80, 81]. On the other hand, drivers with too much confidence in the vehicle safety features may be less observant or drive more aggressively, thus offsetting the potential benefits of those systems [80].

2.4 Statistical Approaches on Crash Severity Analysis

The development of effective countermeasures for road safety requires a thorough understanding of the factors that affect the likelihood of a severe injured given any injury level sustained by vehicles’ occupants following a crash. To gain such understanding, a wide variety of methodologies have been applied over the years, as discussed in this section.

2.4.1 Crash analysis- General review

In crash severity prediction, the analysis focuses on the contribution of several factors and its relationship to the crash outcome. Logistic regression provides important information to discuss the correlation effect between the factors and response variable [83-85]. These factors are called independent variables or predictors variables, which may explain the response variable (also called dependent variable or target) [84]. Logistic methodology provides information on the parameters estimates (input factors), their standard error and their significance level and their confidence intervals and assumes independence among observations [84]. However, regression models have many assumptions and implicit underlie relationships between the dependent and independent variables [86, 87]. An advanced and powerful data mining technique is the Classification and Regression Trees Analysis (CART) [88]. CART methods do not require predefined causal relationship between the target and predictors. Decision trees provide an excellent starting point to predictive modeling and are useful to predict new cases, select useful inputs and optimizing complexity [84, 89]. CART is a flexible non-parametric technique which can provide more informative and smart set of models, and its application is a valuable precursor to a more detailed logistic regression analysis in crash injury data [86]. CART can provide higher prediction accuracy than the conventional binary logit model [88]. Due to the nature of CART, p-values and hence significance cannot be explicitly as in logistic regression. However, CART is based on a surrogate approach for selecting sets of significance variables, and the variable importance rankings could also act as a surrogate for significance [84, 89]. Thus, logistic regression remains the most popular method applied by practitioners working within financial services, industry, medicine, marketing and crash analysis, and it offers a suitable balance of accuracy, efficiency and interpretability [43, 47, 83]. On the other hand, CART is also popular, due to the relative easy way in which models can be developed, their limited operational requirements, and particularly their interpretability [43, 46, 47, 83].

Savolainen et al. had reviewed statistical methods for motor-vehicle injury severities and the challenges that complex data imposes, such as endogeneity, when explanatory (predictor) variables are potentially influenced by injury-severity outcomes [90]. The authors give the example of a model that would use the presence of airbags as an explanatory variable in a model of injury outcome. In that case, drivers owing vehicles with airbags may also tend to be more risk homeostasis. Simply stated, the presence of the airbag releases the perceived risk by the driver, thus allowing him/her to be more aggressive and/or taking dangerous maneuver when overtaking on the road. Other example of endogeneity was identified by Méndez et al. when drivers may take advantage of design improvements and travelling at higher speeds, which translates in higher impact speeds and therefore, higher injury severity [61].

In crash severity prediction modeling, usually researchers look to several classes of targets, which sometimes make difficult the comparison results among different studies. Some researchers have inspected the injury severity of crashes by considering the injury level of the driver only [42, 61, 91-

93]. Others have included in the analysis the injury-severity of the most severely injured occupant, whereas, others have included multiple injury levels per crash event [42, 92, 93]. Therefore, comparison of results among crash severity prediction studies must be made with prudence because those results are significantly influenced not only by the class of target being modeling, but also by the data source. Police accident reports are used worldwide for crash analysis and road safety. However several authors have claimed the misclassification of injury severity among road casualties in police reports. Hauer claimed that Police miss near to 20% of injuries that require hospitalization [94]. Tsui et al. study claimed that police reports overestimate injury severity significantly [95]. This study identified that victims' age, the Injury Severity Score (ISS), and the position of the victim significantly determine the likelihood of police injury misclassification [95]. Amoros et al. stated "Police crash data, which are the basis for safety research in most countries, are incomplete and biased" [96]. Whereas fatal casualties are quite clearly defined and well reported, non-fatal casualties could be biased [96, 97]. Al-Ghamdi stated that police reports "do not describe injuries in much detail because of the lack of police qualifications and training as well as facilities needed to perform complex examinations", and "medical reports are hard to obtain because police accident data and medical data are not kept together" [85]. Despite of the above, Police accident reports are the main source for crash analysis and prevention.

As far as crash data access worldwide, examples are provided for some of the studies under discussion in this Chapter. In Austria, Boufous et al. data was obtained from a Traffic Accident Database System (TADS) [98]. In U.S., Bédard et al. and Jermakian used data from Fatality Analysis Reporting System (FARS) [80, 91]. Also in US, Kockelman and Kweon and Chen and Kockelman had access to the National Automotive Sampling System General Estimates System (NASS GES) [92, 99]. Kononen el. al. examined data from National Automotive Sampling System Crashworthiness Data System (NASS-CDS) [43]. Das et al. used crash data from the Crash Analysis and Reporting (CAR) system [87].

2.4.2 Crash severity prediction models- A review of previous studies

Previous studies related to crash analyses have used a broad spectrum of statistical models to reach conclusions. For example, statistical regression models have been widely used for analyzing contributing factors to injury severity [37, 85, 91, 98-101]. Often, researchers combine different methods in order to extract partner relationships between variables and to overcome data complexity [83, 92, 102, 103].

Boufous et al. used multiple linear regression analysis to evaluate factors affecting injury severity [98]. Results showed that road type, the presence of complex intersections, road speed limit as well as driver's error, speeding, and use of seat belt were significant predictors of injury severity

[98]. Bédard et al. used a multivariate logistic regression to identify the independent contribution of driver, crash, and vehicle characteristics to fatal injuries sustained by drivers [91]. Older drivers, female gender, blood alcohol concentration (greater than 0.30), driver-side impacts, speeds in excess of 111 $km.h^{-1}$ prior to the crash, and no seat belt use were related to higher fatality ratios [91]. Al-Ghamdi applied logistic regression to accident data to examine the contribution of several variables to accident severity [85]. Accident location (intersection or not) and cause were found significant to predict a fatal crash [85]. Kockelman and Kweon applied ordered probit models to examine the risk of different injury levels sustained under all crash types, two-vehicle crashes, and single-vehicle crashes [92]. Pickup-trucks and SUVs were less safe than passenger cars under single-vehicle crash conditions. However, in two-vehicle crashes, these vehicle types were associated with less severe injuries for their drivers and more severe injuries for drivers of their collision partners [92]. Abdel-Aty studied driver injury severity levels using the ordered probit models [42]. Models results showed the significance of driver's age, gender, seat belt use, point of impact, speed, and vehicle type on the injury severity level [42]. Das et al. used random forests, which were ensembles of individual trees grown by CART algorithm [87]. This methodology has identified alcohol/drug use and higher posted speed limits as contributing factors to severe crashes outcomes [87].

Kuhnert et al. combined non-parametric models (such as CART) with logistic regression to determine if “risk-taking” was a significant contributor to crashes resulting in serious injury or death [102]. These combined techniques had identify age, driving experience, sex, and seatbelts as the major contributors to serious injury resulting in hospitalization from motor vehicle accident [102]. Kashani et al. used CART methodology to identify the most important factors which affect injury severity of vehicles drivers [104]. The results reveled that seat belt use, improper overtaking and speeding were the most important factors associated with drivers injury severity [104]. Sobhani et al. developed a kinetic model of two-vehicle crash injury severity using generalized linear regression model. Mass ratio and speed limit had positive effect on the injury severity score of the crash [105]. Martin and Lenguerrand estimated driver protection provided by passenger cars for French vehicles fleet using a conditional Poisson regression [49]. “Recent cars protect their drivers better than older cars in the event of a collision” [49]. However, for the single-car crashes the advances in secondary safety were not apparent, “probably because of higher impact speeds” [49]. Méndez et al. evaluated the crashworthiness and the aggressiveness of the Spanish car based on car’s year of registration by applying two types of regressions: logistic models for single-crashes and generalized estimation equation (GEE) models in tow-crash crashes [61]. Crashworthiness had improved in two-car crashes, and drivers of cars registered before 1985 had a significantly higher probability of being killed or seriously injured than drivers of cars registered in 2000–2005 [61]. Also, for single-car crashes, the improvements in crashworthiness were also very slight [61]. Chen and Kockelman used a heteroscedastic ordered prohibit model to differentiate the effects of vehicle weight, footprint (defined as the product of wheelbase and width) on the severity of injuries

of vehicle occupants [99]. The impact of vehicle's attributes was also found more significant in one-car crashes than in two-car crashes. For single-vehicle crashes, larger footprint vehicles seemed to reduce the risk of serious injuries; while in a two vehicle collision those same vehicles attributes seemed less crashworthy [99]. Also, heavier vehicles were expected to be more crashworthy regardless of crash type [99]. Kononen et al. used logistic regression model for predicting serious injuries associated with motor vehicle crashes [43]. Delta-V, seat belt use and crash impact direction were found the most important predictors of serious injury [43]. Xie et al. focused on the analysis of driver injury severity in rural single vehicle crashes using both the multinomial logit (MNL) model and the latent class logit model (LCL) to find out the relationship between injury severities and related traffic factors [93]. Driver age, DUI, seat belt usage, points of impact, lighting condition, speed, which were found to be closely related to driver injury severity levels [93].

Newstead et al. estimated the risk of death or serious injury based on a total secondary safety index developed with logistic regression model [75]. Crashworthiness and risk impose to another vehicle were largely independent, with a slight correlation with vehicle mass, which tends to improve crashworthiness but increases "agressivity" [75]. Total secondary safety rating was found to be the best for medium vehicles size, whereas, light cars showed the poorest [75].

Table 2.1 highlights the main important studies in the technical literature for crash severity risk factors modeling and injury severity prediction. For each study, data source, sample description, selected statistical methods, key findings and research limitations are outlined.

Table 2.1 – Studies for risk factors analysis and crash injury severity prediction.

AUTHOR'S STUDY	DATA SAMPLE	STATISTICAL METHODS	KEY FINDINGS	LIMITATIONS
Abdel-Aty (2003)	Crash data for the Central Florida, from 1996 and 1997.	Ordered probit models for multinomial variables (injury severity levels) which were inherently ordered.	Driver's age, gender, seat belt use, point of impact, speed, and vehicle type were significant on the injury severity level.	Only focus on driver's injury risk. For vehicle information only use the vehicle type, as being PC or not.
Al-Ghamdi (2002)	Traffic police records from 1997 to 1998 Total of 560 crashes selected in a systematic random process from all accident records in Saudi Arabia.	Logistic regression was used to classify accidents being fatal or non-fatal. During the modeling phase some variables were dropped from the model, those that were not adding useful information to the variability of the response variable.	After dropping some variables location and cause of the crash were found significant. For two-vehicle collisions, driving a heavy duty trunk seemed to offer better protection. For two and single-vehicle crashes, vehicle age and alcohol were positively associated with injury level. The odds of being in a fatal accident at a non-intersection location are 2.64 higher than those at an intersection.	Vehicle information only relied on vehicle body type classification. Only 560 serious crashes were examined and this sample mix pedestrians, cyclist and vehicle collisions.
Baker et al. (2008)	FARS was used for two-vehicle crashes between 2000 and 2003.	Driver fatalities in struck passenger cars were grouped by crash configuration (front-to-front or front-to-driver-side), reported driver belt use, light truck body type (pickup or SUV), and whether or not the height-matching criteria were met. Driver fatalities per million light truck vehicle registration-years then were calculated for each of these groups.	The estimated benefits of lower front energy-absorbing structure were a 19 % reduction in fatality risk to belted car drivers in front-to-front crashes with light trucks and a 19 % reduction in fatality risk to car drivers in front-to-driver-side crashes with light trucks.	Focus on the risk to the driver only. The vehicle characteristics being analyzed were limited to matching of primary energy-absorbing structures that affect the agressivity of light trucks with cars.
Bédard et al. (2002)	FARS data for US traffic fatalities from 1975–1998.	Multivariate logistic regression.	Odds ratio (OR) of a fatal injury increased with age, 4.98 (for drivers aged 80+ compared with drivers aged 40–49 years. Female gender (OR=1.54) and blood alcohol concentration greater than 0.30 (OR=3.16) were associated with higher fatality odds. In comparison with front impacts, driver-side impacts doubled the odds of a fatality (OR=2.26), and speeds in excess of 111 kilometers per hour (were related to higher fatality odds (OR=2.64) compared with speeds of less than 56 kph.	Only focus in single -vehicle crashes. Risk to the drivers only.
Boufous et al.(2008)	Database linking hospital from the Inpatient Statistics Collection (ISC) to police crash casualty records from the Traffic Accident Database System (TADS), in Australia.	Injury resulting from traffic crashes was measured using the International Classification of Diseases, 10th revision (ICD-10) Injury Severity Score (ICISS). Univariate and Multiple linear regression analysis. Different Models were developed: for analysis impact driver characteristics, for analysis impact of environment and road and for analysis of the impact of vehicle and crash information on injury severity.	Road type, presence of complex intersections, road speed limit as well as driver's error, speeding, and use of seat belt were significant predictors of injury severity in older people hospitalized as a result of a traffic crash	Only focus in older diver risk. Vehicle information is only limited to being a car or "other vehicle" and year of manufacture. It includes the number of vehicles involved in the crash but not the effect of the opponent vehicle.

AUTHOR'S STUDY	DATA SAMPLE	STATISTICAL METHODS	KEY FINDINGS	LIMITATIONS
Broughton (2008)	British SATS19 national road accident reporting system. Data crashes from 2001 to 2005	Car models were grouped into six types, ranging from 'Minis and Superminis' to '4 × 4s and PCs. Statistical models were fitted to identify the influence on a driver's risk of injury in a car-car collision based on type and registration year of the driver's car and the type and registration year of the other car in the collision. Risk has been calculated as driver casualty rate per million. Generalized linear model was used to fit the driver casualty rate data Analysis focus on secondary risk estimated from two-vehicle collisions.	Driver casualty rates falls with size of car, except for sport cars. In car-car collision driver fatality rise with size of the other car. In multi-vehicle accident, occupants of smaller vehicle face greater risk and the asymmetry of risk increases with mass ratio. In car-car collisions, driver of the earlier car tends to face greater risk than the driver of the later car. The risk of death for the driver of the smallest type of car was 4 times the risk for the largest type. The risk of death for a driver in collision with the largest type of car was over twice the risk when in collision with the smallest type, The risk of death for the driver of a car registered in 2000-2003 is less than half the risk for the driver of a car registered in 1988–1991. The risk of death for a car driver in collision with a car registered in 2000–2003 is about 46% greater than the risk when in collision with a car registered in 1988-1991. The risk of being killed or seriously injured varies less with car type and registration year than the mean risk of being killed. Nature and severity of an accident tend to vary with the local speed limit.	Focus on the risk to the driver only. Car information was limited to type and registration year).
Broughton (2012)	Crashes extracted from the British National Road Accident Reporting System. Accident data from 2003 to 2007.	Two models were fitted to the accident data and the dependent variable for each model was proportion of injured drivers who were killed or seriously injured. One model comprise comprised the driver's age and sex and the registration year of the driver's car. Other model had added type of car and registration year. Separated models were fitted for type of road.	The mean risk of death for a car driver in collision with a car registered in 2004-2007 was 23% greater than in collision with a car registered in 1988-1991. Newer cars are associated with lower risk of injury than older cars, namely protection of occupants in fatal and serious accidents and aggressivity in serious accidents. Fewer casualties in car-car collisions were registered when more modern cars are involved. So the casualty benefits of improved secondary safety have clearly outweighed the disbenefits of increased aggressivity.	Only focus on risk to the drivers. Predictors were based on type of car and registration year. It does not take into account vehicles' differential size and mass.
Chen and Kockelman (2012)	Data was used from 2007 through 2009 NASS GES. 26,421 occupant observations for one190 vehicle crashes and 72,139 occupant observations for two-vehicle were analyzed. 1V and 2V	Data from NASS GES was matched with additional vehicle-specific characteristics (obtained using HLDI's database) based on abbreviated vehicle identification numbers (VINs). Heteroscedastic ordered probit model to distinguish the effects of vehicle weight, footprint (wheelbase*width) and height on the severity of injuries sustained by vehicle occupants.	Larger-footprint vehicles and shorter vehicles are estimated to reduce the risk of serious injury. In single-vehicle crashes, they appear to be less crashworthy in two-vehicle collisions. Heavier vehicles are anticipated to be more crashworthy regardless of crash type. Moderate changes in vehicle weights, footprints are estimated to have small impacts, while other factors, such as seat belt use, driver intoxication, and the presence of roadway curvature and grade influence crash outcomes much more noticeably.	The methodology does not explain if the effect of the opponent vehicle was on the case vehicle injury outcomes. Vehicles differential characteristics such as weight differential between the two vehicles involved in the collisions were not shown.
Das et al.(2009)	Crash data from the Crash Analysis and Reporting (CAR) System , Florida Department of Transportation (FDOT), for the years 2004 through 2006.	Random Forests, which are ensembles of individual trees grown by CART algorithm, were used to classify crash severity. Severity level was defined as Binary (1 = incapacitating injuries/ fatalities; 2 = possible/ non-incapacitating injuries).	Alcohol/ drug use was associated with increased severity of crashes irrespective of the length of the corridors or the type of crashes. Failure to use safety equipment by all passengers and presence of driver/passenger in the vulnerable age group (> 55 yr or <3 yr) increased the severity of injuries.	Only consider crashes occurring in urban arterials. For vehicle information only consider vehicle type category: light trucks; heavy vehicles and light slow moving vehicles.

AUTHOR'S STUDY	DATA SAMPLE	STATISTICAL METHODS	KEY FINDINGS	LIMITATIONS
Evans (2004)	Crashes cases were extracted from FARS data, for 1975-1998	Analysis the quantitative relationship to to explore what length increases was required to offset the risk increases from reducing vehicle mass. Analysis derived from frontal two-car crashes.	If a car is heavier, it reduces risk to its driver but increases risk to other drivers. If a car is larger (without being heavier) it reduces the risk to its driver and also reduces risk to other drivers. Increased dimensions in a car provide increased occupant comfort. To reduce fatality risk in crashes between large and small cars requires increasing vehicle length while reducing mass.	Focus only on two-car crashes.
Fredette (2008)	Data from 2 vehicle collisions occurring between 1993 and 2001, from National Collision Database, in Canada. Data for 2,999,395 drivers.	Logistic regression was used to model the risk of driver death or major injury (defined has being hospitalized).	Pickup trucks, minivans and sport utility vehicles (SUVs) are more aggressive than cars for the driver of the other vehicle and more protective for their own drivers. Like vehicle mass and type, characteristics of drivers and circumstances of the collision influence the driver's condition after impact. Male drivers, older drivers, drivers who are not wearing safety belts, collisions occurring in a higher speed zone and head-on collisions significantly increase the risk of death.	Only focus in older diver risk. It classifies six vehicle types: passenger car, SUV, pickup truck, minivan, heavy truck and bus. Mass ratio for driver car and impact were included. The study did not include vehicle's technical data information.
Kashani et al.(2011)	Dataset include 213,569 drivers that were involved in rural road crashes from 2006 to 2008, in Iran.	CART was applied to model 13 independent variables, and the target variable injury severity, which includes 3 classes: no-injury, injury and fatality.	Seat belt use, cause of crash and collision type as the most important variables influencing the injury severity of traffic crashes.	Vehicle information is only in respect of vehicle type classification. Only focus on drivers' risk.
Keall (2008)	Crash data in the years 2005-2006, New Zealand. Population with 2,996,000 vehicles of which 17,245 were involved in an injury crash.	Vehicles grouped by category. Poisson regression was used to estimate the number of casualties resulting from crashes involving the vehicle marker group. Multivariate logistic regression models were used to estimate crash risk.	Sport cars high crash involvement rate and injury rate is likely to be largely due to the way they are driven rather than to inherent characteristics of the vehicles themselves. SUVs are dangerous when in the hands of young drivers. Safety conscious vehicle purchaser should also avoid sports cars because of the tendency for drivers to take additional risks when provided with high levels of acceleration and performance.	Only two continuous variables available for the analysis, vehicle age and annual distance driven. The logistic model for injury crash involvement had a non-significant Hosmer-Lemeshow Goodness-of-fit test, providing evidence of a poor fit.
Kockelman and Kweon (2002)	Data from National Automotive Sampling System GES, which has all police-reported crashes in the US for 1998 year.	Ordered probit regression was applied to model four levels categories: no injury (0), minor injury (1), severe injury (2), and fatal injury sustained by driver (3).	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.	Only considered the risk to the driver. For vehicle information only consider vehicle age (model year) and vehicle type category, such as motorcycle, SUVs, van, pick-up, heavy duty vehicle.
Kononen et al. (2011)	National Automotive Sampling System Crashworthiness Data System (NASS-CDS) for 1999-2008. Sample had 14,673 vehicles, 1212 (8.3%) contained one or more occupants with ISS 15+ injuries.	Injury Severity Score (ISS) was considered for crash outcomes injury level analysis. Logistic regression was conducted using SAS 9.2. The target was the percent of vehicles with seriously injured occupant(s).	Delta-V (mph), seat belt use and crash direction were the most important predictors of serious injury.	Lack of vehicles characteristics for models inputs. The only information used was vehicle type (utility, van, pickup and car).
Kuhnert et al. (2000)	Survey from 1997 to 1998, in Australia. 2000 people were inquired.	Participants were stratified by sex, vehicle type and postcode areas). Combined non-parametric modeling procedures (CART) and multivariate adaptive regression splines (MARS) with logistic regression.	MARS and CART are not only modeling tools but exploratory tools for a more detailed analysis. Models have identified age, experience, sex and seatbelts are major contributors to serious injury.	Vehicle information was limited to vehicle type classification. It center in the analysis of driver characteristics rather than other contributor factors to injury outcomes.

AUTHOR'S STUDY	DATA SAMPLE	STATISTICAL METHODS	KEY FINDINGS	LIMITATIONS
Li (2008)	The crash data were originally obtained from the Kansas Department of Transportation (KDOT) database. Data includes **85 fatal crashes** and 604 injury crashes between 1998 and 2004.	Crash severity index (CSI) for work zone safety evaluation was proposed and a set of CSI models were developed through the modeling of work zone crash severity outcomes. Chi-square statistics and Cochran–Mantel–Haenszel (CMH) statistics were employed to ensure the accuracy of risk factor identification. First, a wide range of crash variables were examined in a comprehensive manner and the significant risk factors that had impact on crash severity were selected. Second, the CSI models were developed using logistic regression technique by incorporating the selected risk factors. Finally, the developed models were validated using the recent crash data and their ability in assessing work zone risk levels were analyzed.	CSI models can provide straightforward measurements of work zone risk levels.	Training model developed with 267 injury crashes and 67 fatal crashes. The crash data used for model validation had only 18 fatal crash cases. The size of the fatal crash sample might not be large enough to validate the developed models under typical fatal conditions.
Martin and Lenguerrand (2008)	Crashes by the police in France between 1996 and 2005. The risk of the driver being killed has been evaluated for a sample with 144,034 drivers. Single and two-vehicle crashes.	Poisson regression was used to assess the relative risks. With this regression the relative risks for drivers within the same crash are estimated by conditioning the Poisson likelihood on the number of deaths in each matched set (single and two-vehicle crashes)	When a recent car is in collision with an older car, the driver of the former is better protected than the driver of the latter. Improvements in secondary safety are not observed in the case of single-car crashes, very probably because of higher impact speeds.	Data which would have allowed a good estimate of impact conditions in terms of Delta-V was not available. Lack of precision concerning vehicle characteristics, mainly registration year, mass and power.
Mendez et. Al (2010)	Data extracted from the Spanish Road Accident Database, for cars registered before 1985 and cars registered, in 2000-2005	Two types of regression models have been used: logistic regression models in single-car crashes, and generalized estimating equations (GEE) models in two-car crashes. Dependent variables have been defined as proportion of injured drivers who were killed or serious injured in the Spanish car fleet.	Crashworthiness improved in two-car crashes: when crashing into the average opponent car, drivers of cars registered before 1985 have a significantly higher probability of being killed or seriously injured than drivers of cars registered in 2000-2005. In single-car crashes, the improvement in crashworthiness was very slight. Increase in the aggressivity of newer cars.	Only focus on analysis of the drivers risk. Vehicle information is only limited to the registration year.
Pakgohar (2011)	Database extracted from Traffic Accidents of Iran's Police. The size of the target population was 347,285 road crashes during 2006.	Descriptive analysis, Logistic Regression, and CART were employed. The dependent variable (Accident Severity) had three levels: "Fatal", "Injury", and "No Injury". During running CART and LR algorithms through SPSS, the software's defaults were adopted.	After executing algorithms, the accuracies of 81% and 78.57% were achieved for CART and LR, respectively. Thus CART had higher accuracy than LR method.	Accident severity did not take into account vehicle effect but driver´s age and gender, seat belt use, and driving license.
Tolouei (2009)	UK data from two-car accidents where at least one driver was injured, from 2000-2004.	Logistic regression models were used to represent the independent influence of speed limit (proxy for accident severity), first point of impact, driver sex and driver age. Linear model was estimated using ordinary least square to investigate the effect of vehicle mass on its adjusted crash injury risk to the driver.	A 100 kg increase in mass decreases risk of injury to the driver in a two-car injury accident between 2.6% and 3.2%. Characteristics of the fleet, and in particular the distribution of mass within the fleet, it is an important factor in determining the relationship between mass and secondary safety performance of individual vehicles.	Only focus in risk to the driver. It seems that uses an average of vehicle's mass and engine size for auto-brands rather than using vehicles individual's mass and engine size.

AUTHOR'S STUDY	DATA SAMPLE	STATISTICAL METHODS	KEY FINDINGS	LIMITATIONS
Tolouei (2013)	UK STATS19 Police reported data from 2000 to 2006. Sample dataset included two-car crashes where at least one of the drivers was either killed or seriously injured (KSI); this included a total of 2485 two-car crashes. Two vehicle collisions.	Disaggregate analysis of two-car crash data to estimate the partial effects of mass, through the velocity change, on absolute driver injury risk in each of the vehicles involved in the crash. Absolute injury risk is defined as the probability of injury when the vehicle is involved in a two-car crash. It separates the effect of vehicle mass from size (length x width). The driver injury probability is described by a logistic function that includes, for each vehicle involved in the crash, the velocity change (defined as a function of mass ratio and closing speed) as well as various driver and vehicle characteristics.	The probability of injury of the driver of vehicle 1 increases with speed limit and with increasing mass ratio (μ = m2/m1) while the probability of injury of the driver of vehicle 2 increases with speed limit and with decreasing mass ratio; that is, in a two-car collision vehicle mass has a protective effect on its own driver injury risk and an aggressive effect on the driver injury risk of the colliding vehicle. There is a protective effect of vehicle size above and beyond that of vehicle mass for frontal collisions. Mass might not necessarily impose a trade-off between safety and environmental goals in the vehicle fleet as a whole. This is because the secondary safety performance of a vehicle depends on both its own mass and the mass of the other vehicles in the fleet.	Only estimate risk to the driver. Crash analysis focus only frontal two-car crashes.
Wenzel (2005)	Crashes from fatality analysis reporting system FARS, for 1997-2001.	Used the number of driver fatalities during the for selected vehicle types/models from model years 1997–2001 and divide the number of fatalities for a given vehicle type or model by the number of "registration-years". Risk defined as drivers deaths per million registered vehicles for a given car model. Use both primary risk (crash involvement) and secondary risk (injury risk) during the analysis.	Range in cars' risk must be attributed to vehicle design (which encompasses mass and size) and to difficulty to driver characteristics and/or behavior. Mass alone is a "modest" predictor for risk. Mass and size correlates inversely with risk; large and mid-size cars have safer records than average subcompact, but the correlation is not strong. Better correlation was found between vehicles quality and safer records. It remains inconclusive whether design features or driver characteristics and/or behavior are more important to risk.	Focus on the risk to the driver only. The "other vehicle" could be any model, including motorcycles, buses and heavy vehicles. Study the dependence of risk on vehicle type and especially on vehicle model, but not took into consideration vehicles technical information. In the risk to the driver did not consider the effect of vehicles characteristic differentials.
Xie et al. (2012)	Total number of crashes with valid data was 4,285 obtained from Florida Traffic Crash Records Database, in 2005. Single-vehicle crashes.	Multinomial logit (MNL) model and latent class logit (LCL) model were used. Five crash injury outcomes were considered in this research: "no injury", "possible injury", "non-incapacitated injury", "incapacitated injury", and "fatal injury". To further assess the performance of the LCL model, a prediction experiment was conducted to evaluate the goodness-of-fits of the two models. From the collected data, 3,000 observations were randomly drawn for model fitting, and the remaining data are used for evaluation. This process is repeated 10 times.	Compared to the MNL model, the LCL model improves the prediction accuracy for the possible injury category by around 37%. For other injury outcomes, the improvements from the LCL model range between 10% and 20%, which are quite significant considering that this is the average result based on 10 randomly generated samples. Model´s significant risk factors were: driver age, DUI, seat belt usage, points of impact, lighting condition and speed. Vehicle age and surface condition were not significant.	Focused on rural single-vehicle traffic crash and only in crash driver injury severity risk. Vehicle information was only limited to vehicle age and being an automobile or a van.
Zhang (2000)	Crashes obtained from Canadian Traffics Accident Information Databank from 1988-1993. 17,367 crashes including 711 fatal observations.	Multivariate logistic regression was used to calculate the estimated relative risk based on odds ratios (OR).	Factors significantly related to the increased risk of fatal-injury in crashes were: age (OR=1.4 for 70–79), sex (OR=1.4 for males), failing to yield right-of-way/disobeying traffic signs (OR=1.7), non-use of seat belts (OR=4.0), ejection from vehicle (OR=11.3), intersection without traffic controls (OR=1.7), roads with higher speed limits (OR=7.9 for 70–90 $km.hr^{-1}$; OR=5.8 for 100 $km.hr^{-1}$), head-on collisions (OR=55.1), two-vehicle turning collisions (OR=3.1 for left-turn, OR=8.7 for right-turn), overtaking (OR=5.6), and changing lanes (OR=2.1).	Vehicle information was limited to automobile or van. Only focus on risk injuries to the elderly drivers.

2.5 Modeling Rare Events- Imbalanced Data

Problems of classification and prediction models with imbalanced classes are common in several domains. This section discussed the challenges imposed by rare events and summarizes authors' findings dealing with this topic.

2.5.1 Why are rare events a problem?

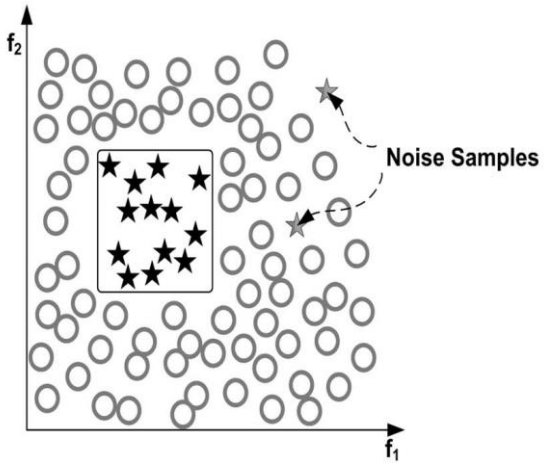


Figure 2.1 - A scheme illustrating a dataset with imbalance classes (used with permission [106].

Imbalanced data sets exists in many real-world domains, such as spotting unreliable telecommunications customers, detection of oil spills in satellite radars images, detection of fraudulent telephones calls and credit card frauds [83, 106, 107]. High imbalance events occurs in real-world where the decision is aimed to detect a rare but important case [107]. Imbalance data correspond to data exhibiting significant and sometimes extreme imbalances between the classes. A dataset is imbalanced if the classes are not approximately equally distributed. Some authors claim that natural distributions are not the bets distribution for learning a classifier [107-111].

Figure 2.1 illustrates in a simpler manner an imbalanced classes distribution [106]. The stars represent the minority class and the circles represent the majority class. In some domains of civilian life to national security, between class imbalances are on the order of 100:1, 1000:1 and 10000:1, where for each case, one class severely outrepresents another [106]. Classifiers (or algorithm method) tend to provide a severely imbalance degree of accuracy: with the majority class having close 100% accuracy, and the minority class having accuracies in the interval of 0-10% [106]. There is the need to have an algorithm method that it will provide high accuracy for the minority class, without making vulnerable the accuracy of the majority class.

In literatures, rare events have proven difficulty to explain and predict [110]. The importance of addressing the imbalance distribution between the majority and minority classes in modeling is

often derived from the underlying decision context and the costs associated with it [83]. The nature of the class imbalanced was defined as “a relative problem depending on both the complexity of the concept represented by the data in which the imbalanced occurs and the overall size of the training set, in addition to the degree of class imbalance present in the data and the classifier involved” [110]. High complexity and imbalance classes, as well as small training set sizes, lead to very small subclusters that cannot be classified accurately [110]. Contrariwise, the class imbalance problem causes no harm when all subclusters have a reasonable size, thus dismissing the belief that classification errors will necessarily occur if one class is represented by a large data set and the other, by a small one [110].

As far as the answer to the question “Why are rare events a problem?” there are several reasons as explained next.

Explanation 1: Some small disjunctions may not indicate a rare case or exceptional observation, but rather noisy data [107]. Hence, just small disjunctions that area meaningful should be reserved for the analysis [107]. In logistic regression modeling to predict a binary target outcome (Y=”0” or “Y=1”) with unequal sample frequencies of the two outcomes (“0” and “1”), the less frequent outcome (“1”) always has lower estimated prediction probabilities than the other outcome [112]. Thus, the logit model would estimate high prediction probabilities for the most common event and very low for the less frequent event. Hence the inequality of sample proportions of the outcomes leads to a high overall estimated prediction probabilities and to high log-likelihood [112]. Cramer stated that a good prediction would be simply a matter of choosing the right predictors[112]. Whatever value the rare outcomes can attain, on average the prevalent outcome will always be predicted even better [112]. The extent of this asymmetry differs with the fit of the model, which is usually mediocre, as a “rule” there is a great contrast between the poor prediction of the rare event and the good prediction of the common event [112].

Explanation 2: The problems of logistic regression in rare events are mainly related to two sources: statistical procedure can sharply underestimate probability of rare events and commonly data collection strategies are inefficient for unbalanced data [83, 109]. The first source of problems of rare events data analysis with binary dependent variables is related to the mean of the binary variable which is the relative frequency of the events in the data. For instance, logit coefficients are biased in small samples (under 200) and this problem have been well documented in the literature [109]. However, it is not widely understood why in rare events data, the biases in probabilities can be substantively significant for large sample sizes (above thousands) [109]. In addition, the probabilities of events in the logit analysis are suboptimal in samples containing rare events, leading to errors in the same direction as biases in the coefficients [83, 109]. The second source of problems with rare events data is derived from the data collection. Collecting data sets with no events (and thus no variation on the dependent variable “Y”) led to choice of very large number of observations with poorly measure explanatory variables [109]. King and Zeng stated that “a trade-off always exist between gathering more observations and including better or additional variables” [109].

Explanation 3: In many applications and domains of data mining, the costs of type I and type II errors is dramatically asymmetrical, making an invalid prediction of the minority class more costly than an accurate prediction of the majority class [83]. Traditional algorithms usually have a bias towards the majority class, which provides more error signals [83]. Moreover, the error signals derived from different numbers of events “1” and events “0” may shift the decision surface in feature space for those methods estimating decision boundaries using fundamentally different approaches to classifier design, depending on their statistical efficiency. Hence, there is the need to avoid collecting the vast majority of observations without efficiency loss. Some approaches designed to handle with this problem rely on selecting the events (Y=”1”) which are relevant, however those approaches might lead to alter the population to which are inferring or requires conditional analysis [109].

The above explanations prove why handling imbalance data requires either the development of distribution insensitive algorithms or an artificial rebalancing of the datasets through sampling [83]. The computer time and memory required for the statistical analysis depend on the number of cases, the number of variables, the complexity of the model, and the algorithm. Therefore, for many modeling situations, there is a trade-off between time and memory.

2.5.2 Strategies and methodologies to handle imbalanced data

Solutions to handle imbalanced data sets include: sampling techniques, cost-sensitive methods and kernel-based methodology [83, 106, 107, 109, 111, 113-115]. The sampling methods comprise different forms of re-sambling, such as: oversampling, undersampling, cluster-based sampling and boosting [83, 106-108, 111, 113, 114]. Balancing methods attempt to balance the distributions by taking into account the proportions of the classes. Whereas, cost-sensitive methods target the imbalanced data problem by using cost matrices that address the cost of misclassifying any data. Attention is given to oversampling and undersampling which are among the most common re-sampling methods.

Crone and Finlay defined undersampling as ”instances of the minority and majority classes are selected randomly in order to achieve a balanced stratified sample with equal class distributions, often using all instances of the minority class and only a sub-set of the majority class” [83]. Whereas, oversampling have been defined as “the cases of the under-represented class are replicated a number of times, so that the class distributions are more equal” [83]. These authors alerted for the inconsistencies in this terminology, which are frequent. Anderson had referred to oversampling, but essentially described it as undersampling by removing instances of the majority class[83, 116]. Also, Sarma defined oversampling as including all the cases of the “responders” and only a fraction of the “non-responders” [117]. Japkowicz and Stephen defined random oversampling as “oversampling the small class at random until it contains as many examples as the other class” [110]. On the other hand, random undersampling was defined as “eliminating at random elements of the over-sized class until it matches the size of the other class” [110]. Nisbet et

al. defined “oversampling” as increasing the sample rate category, and “undersampling” as reducing the sample of the common category [118].

King and Zeng strategy was to select on Y by collecting observations (randomly or all those available) for which Y=1 and a random selection of observations for which Y = 0 [109]. In fields were the number of observable ones is strictly limited (such as in crash injury severity events) the authors recommend collecting all available or large number of ones. Subsequently, the decision how many numbers of zeros must be collected depends if that collection is not costless, the analysis must collect more zeros than ones [109]. A useful practice is sequential, involving first the collection of all ones and an equal number of zeros [109]. “Real information in the data lies much more with the ones than the zeros, but researchers must be careful to avoid selection bias [109].

Japkowicz and Stephen compared various strategies to handle class imbalanced: two re-sampling methods (random oversampling and random undersampling) and cost-modifying [110]. These authors found random oversampling more useful than random undersampling [110]. In some applications, cost sensitive methods perform better than sampling methods [106, 113, 119]. Cost-sensitive learning outperforms random resampling [110]. However the cost of misclassification is generally unknown in real cases [107].

Each of the above methods has advantageous and disadvantageous and they have been subject of several discussions in the literature. The major drawback of undersampling is that can discard potentially useful data [107]. On the other hand, random oversampling can increase the likelihood of occurring overfitting, when this methodology relives exact copies of the minority class [107]. Chawla suggested that undersampling is usually better than oversampling with replications [111]. Nisbet et al. recommend that if the data set is not large it is better to oversampling the rare category [118]. In the case of oversampling selection, overfitting may occurs when classifiers produce multiple copies of the same example; although the training accuracy will be high the classification performance on the unseen testing data is worse [106].

Sampling methodologies (under and oversampling) generally lead to models with an enhanced discriminatory power, but both random oversampling and random undersampling methods have their shortcomings: random undersampling can discard potentially important cases from the majority class, thus impairing an algorithm’s ability to learn the decision boundary, while random oversampling duplicates records and can lead to the overfitting of similar instances [83]. Therefore, undersampling tends to overestimate the probability of cases belonging to the minority class, while oversampling tends to underestimate the likelihood of observations belonging to the minority [119]. As both over and under-sampling can potentially reduce the accuracy in generalization for unseen data, a number of studies have compared variants of over- and under-sampling, and have presented (often conflicting) viewpoints on the accuracy gains derived from oversampling versus undersampling [111, 113]. The presence of irrelevant data it would make undersampling more effective than oversampling or even cost-modifying on fields presenting a large variance in the

distribution of the larger class [110]. However, undersampling, removing examples from the majority class, may cause the classifier to missing important information [111, 113].

2.5.3 Effect of sample size in predictive modeling

With regard to binary data classification, analysis of data containing rare events or imbalance class distributions poses a great challenge to industry and to the machine learning community [114]. Sample size and balance may affect not only the accuracy but also the interpretability and efficiency of the algorithms [83]. Larger sample sizes raise the probability that a sample will be representative of the entire population, and therefore guarantee similar predictive accuracy, however increases computation times and data acquisition costs. On the other hand, smaller samples, the patterns contained in the data may be missed or erroneous patterns may be detected, thus enhancing efficiency at the cost of limiting accuracy [83]. The ratio of events to variables tends to be a less important factor for larger samples, hence decreasing the probability of overfiting.

Harrell claimed that amount of information in a data set with a categorical outcome is determined not by the total number of cases in the data set itself, but instead by the number of cases in the rarest outcome category (for binary target data sets) [103]. Therefore, this author recommended separating sampling as an effective resampling strategy for productive modeling [103]. Crone and Finlay suggested that logistic regression had a near optimal performance using far fewer observations than methods such as CART, when there is a concern with sample size on the efficiency of the algorithm [83]. Also, this work stated that oversampling significantly increases the accuracy relatively to undersampling, across all algorithms. For logistic regression, the balancing applied to datasets appears to be of minor importance. However, the other methods demonstrate a greater sensitivity to balancing, particularly CART [83].

As final remarks regarding to re-sampling strategies, it should be noted that over and undersampling will impact not only the predictive accuracy, depending on the statistical efficiency, but also the resource efficiency in model construction and application. Balancing (re-sampling) has an impact on the total sample size by omitting or replicating good and/or bad instances, thereby decreasing or increasing the total number of instances in the dataset, which impacts the time taken for model parameterisation [83]. “It is still unclear which sampling procedure performs best, what sampling rate should be used and that the proper choice is probably domain specific” [107]. Although, algorithms presented in this literature review (section 2.4) claimed to improve classification accuracy, there are certain situations in which learning from original data sets may provide better performance [106]. Thus, it would be desirable a uniform benchmark platform to provide assessment between existing and future methodologies. Henceforth, the results are not universal and depend on the dataset properties and the application domain [83].

2.5.4 Severe crashes as rare events- Predictive challenges

In contrast to the domain of credit card fraud detection, injury severity prediction on road safety analysis is missing an approach to deal with the rare events (severe crashes). The imbalance between severe crashes and non-severe crashes highlighted with the following road safety indicators. A study in US, with crash data from 1996 to 1997 have shown the following distribution driver injury: property damage only (no injury; 58.8%); possible injuries (20.7%); evident injuries (9.0%); and severe/fatal injuries (4.8%) [42]. In 2009, crash data provided by the Fatality Analysis Reporting System (FARS) showed that from total police reported motor vehicle crashes (5,500,000), fewer (30,797) than one percent resulted in death (1%) [44].

Unfortunately, the best practices for resampling have not been explored in crash severity injury prediction. The evidence of the existing gap in resampling strategies to deal with rare events among crash data is illustrated by the next four studies.

1. Xie et al. analyzed driver injury for data obtained in Florida for the time period 2002 to 2006 using logit regression methods. In this study, the percent of fatal crashes was 1.71% and 0.78% for rural and urban roads respectively [93].
2. Pakgohar et al. applied CART and logistic regression for the analysis of crash severity in a data set where injuries were 8% and fatalities were 1% among the data [46].
3. Li developed a crash severity index comprehensive models using for 267 injury crashes and 67 fatal crashes [100]. Models validation was performed with new crash data, 337 injury crashes and 18 fatal crashes. The author recognized that the size of the sample could not be large enough to validate developed model [100].
4. Only Kononen et al. had shown concern with the imbalance classes between the non-injury cases and the injury cases [43]. This study using National Automotive Sampling System Crashworthiness Data System (NASS-CDS) data the overall probability of injury cases was 2.8% [43].

Despite of above highlighted imbalance data sets with disproportion between severe and non-severe observations, those authors have not shown resampling strategies. As a consequence, issues of sample size and balancing have been neglected within road safety expertise as a topic of study. Thus, the gap in sample balancing for crash severe events studies, lead the development of an own strategy in this study to overcome the challenge imposed by imbalance between severe crashes and non-severe crashes in the Portuguese collected data [120]. Also, this is the first research conducted in Portugal that integrates vehicles technical characteristics with crash data analysis [120-122]. Chapter 3, dedicated to Safety Analysis Methodology, will present the balancing approach developed for the analysis of the Portuguese crash sample.

2.6 Trade-off of Vehicle Safety, Fuel Efficiency and Emissions

The trade-off between vehicle´s safety and fuel economy has been a controversial issue since the energy crisis of the 1970s. In 2007 and 2009, the EU regulations set CO_2 emissions performances targets for manufacturer's new car sales moved the technological trade-off in favor of increased fuel efficiency. First, CO_2 emissions limits measures are discussed. Second, research on the vehicle's safety and fuel efficiency trade-off analysis is outlined.

2.6.1 CO_2 emissions measurements

Fontaras and Dilara investigated how vehicles characteristic affects real world emissions performance [123]. The difference between real world performance and the certified test was estimated in 15-20% [123]. The authors claimed that NEDC does not take into account other important factors affecting vehicles' emissions such as: use of air conditioning, vehicles accessories, and reduction of tyre pressure [123]. Leduc et al. compared CO_2 emissions and energy use under real world conditions with those under the NECD and found that NECD had lower emissions by 14% [124]. Zervas recommended that NEDC CO_2 emissions should account for annual mileage [4].

Franco et al. revised emission measurements techniques for road vehicle emissions [125]. There are models that only required mean travelling speed to estimate emissions (e.g. COPERT), models that need traffic situations to express emissions (e.g. HBEFA), and others that require second-by-second engine data (e.g. PEMS, MOVES, VSP) to originate emission information for the driving profile [125]. The author argued that emissions measures under real-world conditions (such as in tunnel or on-board measurements (PEMS)) are usually less precise and repeatable than those performed in an engine and chassis dynamometer studies, due to the absence of a standard test cycle and the presence of additional sources of variability such as environmental or traffic conditions, driver behaviour or highly transient operation [125]. This study suggested that the selection of the appropriate emission method depends on the application considered [125]. Bampatsou and Zervas claimed that specific CO_2 emissions are measured on the NEDC for all PCs, but all PCs do not have the same annual traveling distance. The authors have shown the average annual mileage of new gasoline and diesel passenger cars, is a function of segment and model year of the vehicle [126].

2.6.2 Are CO_2 emissions standards compromising the trade-off analysis between fuel efficiency and vehicle safety?

Thought automakers must comply with emissions regulations, consumers' preferences influence the market share by selecting vehicle attributes, such as car segment, fuel type, mass/size, and engine power. Kok has assessed the effects of consumer preferences and technological advances

on sales-weighted average CO_2 emissions of new passenger cars [72]. Until 2007, the results showed that consumers preferences shifts towards larger and less fuel-efficient car segments and also towards larger, heavier and more powerful cars within the same car segment [72]. From 2007 to 2011, this trend decline reflecting consumer preferences shifts toward smaller car segments [72]. Between 2000 and 2007, 56% CO_2 reduction from technological advances had been covered by an increase in larger vehicles sizes. Though from 2008 to 2011 purchasing trends reduced CO_2 by 31% over those from technological advances [72].

Despite of the air emissions regulations, some criticism have been addressing the standards for CO_2 emissions and fuel economy, which are based on vehicles attributes. The mass-based vehicle, (almost half of the world automobile market), apparently seem to be logical choice for the regulatory structure, because vehicle mass is a fundamental determinant of vehicle efficiency. In addition to mass, rolling resistance, powertrain efficiency, and aerodynamics have been improved during the last decade and they have been contributing to expressively reduce emissions.

However, Lutsey argued that “vehicle mass reduction technology (advanced materials, mass-optimized designs) is a major technology strategy for increasing vehicle efficiency” [127]. Thus, “by using a mass-based standard structure, the core efficiency technology of mass-reduction is essentially neutralized” [127]. Bampatsou and Zervas criticized the regulatory emissions of the exhaust CO_2 exhaust emissions from PCs in the EU by the Regulation [EC] No. 443/2009, previously introduced in section 1.1.3, [126]. This study highlighted four critical points.

First: “the regulation proposes a limit on exhaust CO_2 emissions based on the average emissions of each manufacturer sales and not a limit for each passenger car” [126]. Thus if a car manufacturer sells a number of PCs with CO_2 emissions higher than the limit, it must sell a number of passenger cars with CO_2 emissions lower than the limit to compensate the difference [126].

Second: the regulation allows the manufactures to create groups of car makers which applied an average value of CO_2 for the entire group [126]. This “transference of CO_2 emissions limit through car groups” is based on the principle of “flexibility to the compliance” [126]. However, other regulations/directives concerning emissions and safety of passenger cars are not flexible but they have specific targets such as the Euro5/Euro 6 limits imposed by the EU Regulation (EC) No. 715/2007 and cars safety features established by the EU (EC) Regulation No. 19/2011, [126, 128]. This “flexibility” has implications in the ethical point of view, “as the “polluter-pays” principle becomes “someone who can pay, can pollute” principle” [126].

Third: the Regulation [EC] No. 443/2009 proposed a penalty proposed for CO_2 emissions exceeding the average upper limit (95 euro per exceeding gram of CO_2 $g.km^{-1}$ per vehicle) [126]. This penalty will be included in the final price of the vehicle rather than for the car maker.

Forth and last: the critical point is related to the proposed $95g.km^{-1}$ for 2020 which could be a very ambitious target. Average EU15 CO_2 emissions decreased from 186.6 $g.km^{-1}$ to 153 $g.km^{-1}$, between 1995 and 2008, which corresponds to a decrease of 17.67% during 13 years [126].

Therefore necessary change to reach CO_2 95g.km^{-1} would be 49% [126]. Thus, from 2008 to 2020, CO_2 emissions would require a decrease of 31.4% [126].

Zervas shown that the average CO_2 emissions by car firm selling volumes in the European market [4]. Lamborghini, Ferrari, Porsche, they are required to overcome a higher challenge to reach the proposed Regulation [EC] No. 443/2009 target [126]. Seat, Citroen, Renault, Peugeot and Fiat have lower effort to reach target of 95 g.km^{-1} by 2020. As proposition for the CO_2 regulations by 2020, the previous work study had suggested the same CO_2 limit of all new passenger cars without derogations and penalties [126]. In the US, the problem with the current structure of fuel economy standards for cars is that the target of 27.5 miles per gallon is applied to an automaker's whole fleet, no matter the mix of cars an individual automaker sells [58]. A cross-disciplinary cooperation between different industry segments and political institutions is recommended for improvements towards sustainable mobility.

2.6.3 The trade-off between fuel efficiency, emissions and vehicle safety really exists?

Some studies have intended to discuss if there is a trade-off or not between fuel efficiency and vehicles safety, as summarized below.

Wenzel suggested that the relationship between footprint (wheelbase x width) and casualty risk to the drivers of individual vehicle models, including cars and light trucks is very week [129]. Vehicle design, which can be improved by safety regulations, would be more effective on occupant safety than fuel economy standards that are structured to maintain vehicle size and weight [129]. On the other hand, Tolouei and Titheridge stated that in vehicle design, there is a trade-off between fuel economy and secondary safety performance imposed by mass [64]. Even though mass imposes a trade-off in vehicle design, between safety and fuel use, this do not mean that it imposes a trade-off between safety and environmental goals in the vehicle fleet as a whole" [64]. The "secondary safety performance of a vehicle depends on both its own mass and the mass of the other vehicles with which it collides" [64].

Chen and Ren analyzed the relationship between vehicle safety ratings and fuel efficiency for 45 new vehicles models [37]. From 2002 and 2007, the relationship between vehicle safety ratings and fuel efficiencies seem to have been mostly positive [37]. Zachariadis examined 192 car models to investigate whether a safer car consumes more fuel than its less safe counterparts [62]. Enhanced safety of modern cars has a very small effect on vehicle mass and does not significantly affect fuel consumption [62]. Safer cars are heavier by only a few kilograms and do not consume more fuel than their counterparts with lower safety scores [62]. The author suggested that there is almost no trade-off between better car safety and CO_2 emission reduction [62].

While the advocates of the new standards claim the benefits of energy and environment, opponents argue that vehicle safety will be compromised with the new fuel standards. The current

structure of fuel economy standards could encourage manufacturers to sell more smaller, lighter cars to offset the fuel consumed by their bigger, heavier models [58]. “Automakers even sell the smaller and less safe cars at a loss to ensure compliance with fleetwide requirements” [58]. Bampatsou and Zervas claimed that there are two ways to decrease real CO_2 emissions: to decrease the mileage and to decrease the emissions per kilometer [126]. However, other study argued that the main way to reduce CO_2 emissions is by reducing car weights, which means downsizing vehicles, but this would cause conflit with occupants safety goals [58].

The application of lightweight design with thermoplastics offers a possibility to reduce the CO2 emission and fuel consumption [130]. The use of nanocomposites in vehicle parts and systems potentially can to improve manufacturing speed, enhance environmental and thermal stability, promote recycling, and reduce weight. Substituting reinforced polymers in vehicle body components is a promising approach to weight reduction and fuel savings. An estimated 30% improvement of roll-resistance, air-resistance, car-weight and powertrain might reduce the fossil fuel consumption by 4%, 6%, 15%, or 28%, respectively [131]. Nanotechnology application into the automotive industry leads to lighter car bodies without compromises stiffness and crash resistance and results in less fuel consumption. General Motors (GM) produced the electric Chevy Volt that uses 45.4 Kg of thermoplastics, including composites in the hood and doors, plus unreinforced polymeric materials in the rear deck lid, roof and fenders [131, 132]. Volt model also incorporates glass fiber reinforced composite for lightweight horizontal body panels. Tesla Roadsport electric model uses innovative lightweight body panels of carbon fiber/epoxy composite [131, 132].

2.7 Concluding Remarks

In technical literature, much attention is paid to vehicle type and its risk to drivers, but not to its relation to crashworthiness. Vehicles' speed collision was identified as one of the most important fact influencing crash severity outcomes. During a crash, the change of velocity distribution depends on the mass of the vehicle; hence the mass influences the impact of a severity. Thus, vehicle mass was found to be a significant factor of crash severity, that not only influences the vehicle crashworthiness and "agressivity", but also impacts vehicle fuel use and air emissions. However, vehicle's mass alone is a "modest" predictor for injury risk. There is a lack of a methodology to estimate the effect of vehicles characteristics on crash severity following vehicles collisions.

Crash testing protocols provide a valuable tool in consumer guidance, but they cannot predict real-life crash outcomes. During the last few years, improvements in vehicle's safety have been significant, and advanced safety technologies have been recognized to save lives. However, despite of the potential safety features benefits, how the drivers will interact with those technologies will influence the effectiveness of these avoidance systems.

A number of studies have attempted to correlate safety and vehicle design features. However this relation is not fully understood. In addition, crash samples are highly imbalanced for minor injury vs. serious injury and/or killed. Therefore, crash analysis faces a challenge when investigating crash severe events, and no attempted has been made in the literature on how to approach the imbalanced classes in real crash data.

Larger vehicles usually show an extra size and weight that enhance occupant protection in collisions. Nevertheless small cars are more affordable; they use less gas and emit fewer pollutants. The safety and environmental tradeoffs are still not fully explained and they impose a challenge for the transportation and environmental authorities. The trade-off between vehicle's safety performance and environmental performance has been raising some debate. The few existing studies on this trade-off analysis usually focus on the relationship between vehicle's safety and fuel consumption, targeting CO_2 emissions but other exhaust air pollutants are not covered. Furthermore, previous research analyzed vehicle's safety performance based on the individual vehicle only, and they have not considered the risk of exposure in the fleet.

CHAPTER 3

SAFETY ANALYSIS METHODOLOGY

This Chapter describes the methodology for the safety analysis of the Portuguese crash sample. The motivation for this research was to focus on the light vehicle fleet (passenger cars and light duty vehicles) technical characteristics and analyze which one, if any, has a stronger impact on crash severity, expressed by the risk to drivers and passengers, based on real crash data. As an outline of the designed methodology, first data preparation and variables definition are presented. To overcome the challenge imposed by few rare events (severe crashes) in the sample, an advanced strategy was developed to balance the distribution between severe and non-severe events. In conclusion, CART and logistic regression modeling techniques are explained for the crash severity classification and prediction.

3.1 Research Domain

Crash severity was analyzed by exploring the contribution of vehicle related variables: auto brand make, weight (mass), engine size (power), wheelbase, year of registration (age) and fuel type. Crash severity is related to the occurrence of severe injuries and/or fatalities among vehicle's occupants, during the event of a crash involving light passenger vehicles and light duty vehicles. As stated in Chapter 2, factors affecting the risk of increased injury level of occupants during a crash include: demographic and behavioral characteristics of person, environmental factors, roadway conditions and vehicle [45]. This research was not designed to understand the circumstances under which the crash had occurred, such as presence of roadside obstacles, inattentive driver, failure to press the braking system, and traffic volume among other causes. Further, this research focused exclusively on post-crash consequences centered on the injury level outcomes, rather than on pre-crash contributing factors to the event. It focused on the understanding of how technical characteristics of the vehicle may affect the risk of severe injury and/or fatality among its occupants. It is important to point out that, drivers' characteristics, such as age, gender, and agressivity, as well as socio-demographic factors were beyond the scope of this study. Although vehicle's speed at the moment of the crash had been identified as one of the most important factors of injury risk [34, 42, 43, 74, 91, 92, 99], this information is usually not accessible. Information on occupant's seat belt use, airbag data, and vehicle protective systems, as well as trauma management were not available at the Portuguese police crash reports. Figure 3.1 summarizes the steps undertaken to execute the general methodology followed in this study, although this chapter focused the safety analysis methodology.

As discussed in the literature review, previous research generally has attempted to model overall crash severity without taking into account the effect of the opponent vehicle [43, 85, 86, 91, 93, 98, 102, 104]. However, in multi-vehicle collisions the injury severity outcomes depends on the attacking ability of striking vehicle as well as the protective ability of struck vehicle [63]. Some studies have analyzed the effect of vehicle on crashworthiness (ability to protect its own occupants) and "agressivity", hazardousness that the subject vehicle imposes to the opponent vehicle [48, 49, 59, 61, 63]. However these studies focused only in risk to the drivers and largely they only have analyzed the effect of vehicle type (category). In addition, those studies have not clarified how the effect of the opponent vehicle was taken into account on the injuries prediction for the occupants of the vehicle being analyzed. This gap in the previous research work, lead to the development specific target variables to model not only the overall crash severity, but also to model crash severity exclusively for the each vehicle involved in the collision.

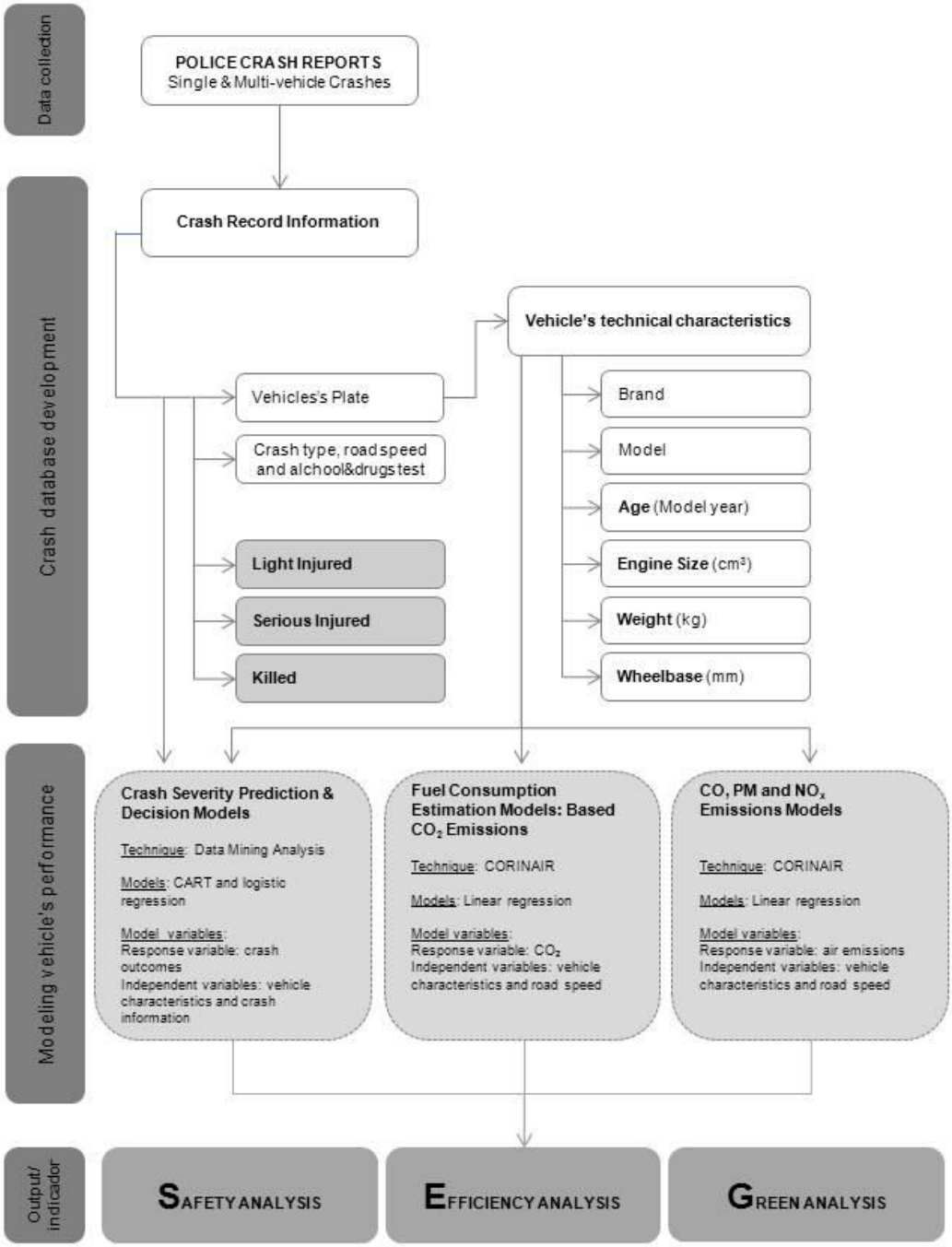


Figure 3.1 - Methodology overview.

For single-vehicle crashes, vehicle individual technical data, such as brand, model, age (vehicle model year), engine size, weight and wheelbase were analyzed for their contribution to crash outcomes. For two-vehicle collisions, in addition of vehicles individual technical data, differential variables were created to express the quantitative difference between the characteristics of the vehicles involved in the collisions (such as: age difference, engine size difference, weight

difference and wheelbase difference) in order to analyze their contribution to crash severity. A detailed explanation of response variables (derived from crash outcomes) and independent variables (crash information and vehicle technical characteristics) is given in section 3.3.

3.2 Data Collection

This section describes the data collection process, limitations within the crash reports and development of the crash database for the safety and environmental analyses.

3.2.1 Site description

During 2010, in Portugal the highest rates of crash fatalities occur for the districts of: Lisbon, Porto and Aveiro, with 123, 108 and 88 road deaths, respectively [23]. The districts of Aveiro and Porto were selected for this study because of two reasons: first, due to the higher rates of fatalities, second, for data collection convenience, that would be performed in the proximity of TEMA/UA where this study was developed. Figure 3.2a) signalizes the crash data collection area in Portugal.

For the selected region above, the accessed crash reports records involved accidents on roads which are included on the 2000 National Roadway Plan, in the Northeast side of Portugal. The reported crash records included the following road classes:

- Main Road (speed limit is 90 $km.h^{-1}$);
- Principal Itinerary (speed limit is 100 $km.h^{-1}$);
- Complementary Routes (speed limit is 100 $km.h^{-1}$);
- And Motorways/freeways (speed limit is 120 $km.h^{-1}$).

Figure 3.2b) signalizes some examples of those road classes for Porto metropolitan area, as follows:

- Main Road (such as EN1, EN14);
- Principal Itinerary (such as IP1, IP4);
- Complementary Routes (such as IC1, IC24);
- And Motorways/freeways (such as A1, A29).

It must be clarify that crash data collection was not controlled for those road classes. However, for each crash observation, the road name ID was recorded, as explained in section 3.2.4.

a) b)

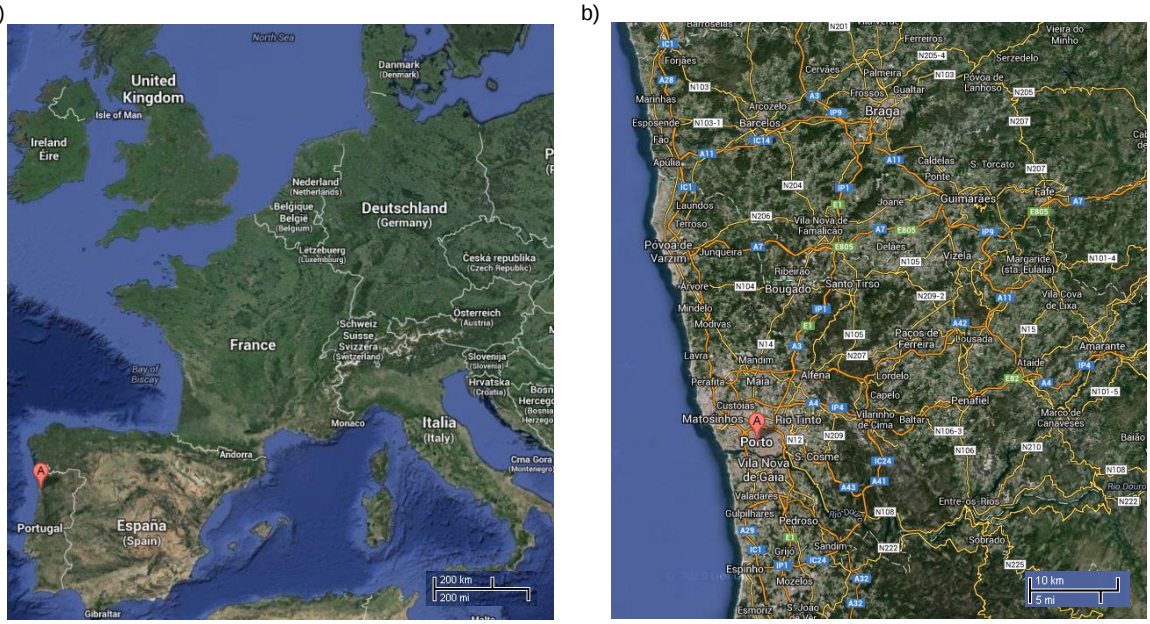


Figure 3.2 - Crash Site location for crash data collection: a) in Porto, Portugal, Europe; and b) Porto metropolitan area.

Among the several road classes identified in the crash records and illustrated in Figure 3.2b), there is A29 motorway, which is a toll road with high volume of traffic, selected often by drivers travelling between Aveiro and Porto, and vice-versa. A29 is among the Portuguese roads with more black spots (five or more severe crashes in 200 meters of the road length in question) [133].

3.2.2 Crash reports selection

Data for the crash severity models development were collected from the Road Traffic Division (RTD) of the Portuguese Road Safety National Republican Guard (GNR) located in Porto and the Portuguese Public Safety Police (PSP) located in Oporto and in Aveiro. From extensive crash reports records data gathered by GNR and PSP, reports were selected based on the following criteria.

1. Recorded crash reports involving property damage only were excluded because this research was focused exclusively on crashes involving any level of injury.
2. Crashes with injuries and/or fatalities and involving light passenger vehicles and light duty vehicles (such as passengers' cars, sport utility vehicles (SUVs) vans and pick-up trucks) were selected.
3. From those, crashes involving pedestrians and/or cyclists were excluded because the designed analysis aimed to explore the relationship between vehicle characteristics and occupants injury risk only.
4. Crash reports data were gathered for the time period of five years, 2006 to 2010.

Following the selection criteria, a total of 2,270 reports were personally collected, as summarized in Table 3.1. Initially crash data were gathered at PSP. Secondly, data were collected at the RTD of GNR. In the first phase of this research, single-vehicle crashes and multi-vehicle collisions were gathered from 2006 to 2008. On the second phase, additional crash data was gathered with focus on two-vehicle collisions from 2008 to 2010.

Table 3.1 - Relevant crash frequencies gathered in the study for the time period 2006 to 2010.

Data Source	2006	2007	2008	2009	2010	Total by Data Source
GNR Porto, PT	298	548	508	161	184	**1699**
PSP Aveiro, PT	-	65	65	-	-	**130**
PSP Porto, PT	-	166	275	-	-	**441**
Total	**298**	**779**	**848**	**161**	**184**	**2270**

3.2.3 Challenges faced to developed the crash database

Several difficulties were faced previously to accomplish the full develop the crash database investigated in this research, as presented in the next section. Following data collection, the extracted information from each crash report was analyzed in more detail and criteria selection was followed in order to develop a database adjusted to the objectives of this research. **Contrary to simple-easier researcher access to crash databases as exemplified in section 2.4.1, in Portugal crash data access is quite different, since crash records are not available in digital files and crash information is not centralized.**

At national level, the crash database is managed by the ANSR. The Police Officers are responsible for submitting selected information from the crash records reports on the 15 days basis to this Authority using a form called “Boletim Estatístico de Acidentes de Viação (BEAV)”. However, the extracted information in the BEAV is brief and standardize, usually indicating the cause of the crash, the outcomes, information on the day, and hour. The ANSR crash database does not include any information on the vehicles involved in the crash, rather than vehicle category (such as light vehicle or heavy vehicle). Thus, the strategy for this study was to personally collect the data at the Police Road Traffic Divisions.

Some published studies in the U.S. had matched crash data with the vehicle identification number (VIN), obtained from the Highway Loss Data Institute [43, 99]. In Portugal this procedure was different. The one of the most challenging tasks of completing the Portuguese crash database based it was to obtain legal permission to access the vehicle technical information derived from vehicles’ registration license plate (RLP). Due to this difficulty, it was decided to focus on the GNR records (1,699 observations) rather than the total 2,270 gathered crash observations. The reason why the “priority” was given to GNR crashes GNR was related to the fact that this Police Force in general is responsible for patrolling roads with higher speed levels, whereas PSP usually operates in urban areas, where the legal speed limits are lower. Therefore, severe crashes are at higher frequency at the GNR records.

From the selected GNR 1,699 crashes, only 1,374 were manageable for further analysis. The reasons why this research was centered on 60% (1374/2270) of the original collected data are presented underneath.

1. As result of the constraints to have legal access to vehicle specific technical data, from the total of 2,270 gathered reports, priority was given to 1,699 collected reports at GNR.
2. From those 1,699 observations, multi-vehicle collisions with more than 2 vehicles were eliminated because the individual vehicle contribution to the overall crash severity would be masked by the interaction with other vehicles involved in the collision.
3. Observations including vehicles which the RLP did not match the Portuguese standard label were excluded because no further information could be gleaned about its technical attributes from international entities. In general the vehicles Portuguese plates follow the partner: four numbers plus two letters, for a total of six digits. For vehicles from abroad is not possible to request vehicles specifications.
4. Each vehicle's information in the crash dataset was recorded following the order stated at the Police record. As an example, the first vehicle (V1) in a collision report tends to be related with the one that initially collided with the second vehicle (V2) and/or caused the crash collision. However the vehicle order in the police records does not follow this protocol uniformly and there was limited information to assume that vehicle V1 always hits vehicle V2 or that vehicle V2 is always struck by vehicle V1. While for rear-end collisions scenarios, it would be easier to identify the vehicle that hits the car in front of it, for the general collisions this identification is more complex.
5. Report content may be unclear; sometimes information was missing or could show inconsistent information and also human errors. For instance, crash reports identify the vehicle type/category information as: light duty vehicle or light passenger vehicle. However, when developing the database and matching vehicles' RLP with technical information it was noticed that the label light duty vehicle was either a heavy duty vehicle, or a non-road vehicle (agricultural tractor). Other reports mislabel "light passenger vehicle" to designate a scooter, or a motorbike. When such errors were detected, the crash record was eliminated from further consideration.

Regarding to crash outcomes information accuracy, it is relevant to clarify the uncertainty which could be associated with some injury levels reported at the crashes records. The Portuguese Police Forces consider three level of injury risk: Light Injury (LI), Serious Injury (SI) and Fatality (F). In Portugal, a serious injured is reported if following the crash the individual required hospitalization at least during 24 hrs. Status of seriously injured was not traced overtime, with the possibility that the harm injuries would result in a death. As explained in Chapter 1, (section 1.1.2.2), until 2010, Portuguese methodology did not apply the threshold of 30 days [7, 15, 23]. Also, in 2011, during the second stage of crash data collection, which took place the Oporto GNR headquarters, a Police

Officer inquired claimed that new methodology procedure was being implemented with difficulties, and the required collaboration between hospital services and Police Forces was yet imperfect to ensure the monitoring of the victims' status in the 30 days basis. As far as the crash sample used in this thesis analysis, it must be said that data collected for the years 2006 to 2009 did not follow the 30 days methodology. For the 2010 crash data collection, fatalities were also recorded in the 24 hours basis, as they were registered in the crash reports by GNR Officers and to ensure consistency with previous data on the crash sample.

3.2.4 Development of the crash database

The 1,374 records selected from the GNR source included single-vehicle crashes and two-vehicle collisions resulting in injuries and fatalities. For each crash event, information extracted from each report was as follows: a) road name and location, b) weather conditions, c) driver's alcohol and/or drugs test results, d) crash type, vehicles' registration plate and registration year, and f) crash outcomes, namely vehicle occupant's injuries and/or fatalities. Appendix 3 shows a copy of a severe crash report record, which outcomes resulted in a fatality, driver of vehicle V2.

At the crash reports, the technical information related to the registered vehicles was minimal, mainly restricted to vehicle's registration plate and vehicle's registration year. Since one of the major goals of this research was to analyze vehicle characteristics effects on the crash severity outcomes, it was obligatory to fulfill vehicle information with exact technical data for each individual vehicle, such as its specific weight, engine size (engine displacement) and wheelbase dimensions. The vehicle technical features were obtained from the former Institute for Mobility and Inland Transportation (IMTT), which is currently the new Institute for Mobility Transportation (IMT). IMTT database allowed to match vehicle registration plate (VRP) (extracted from crash reports) to be augmented with details such as the date of the first registration and specific vehicle's make and model technical data.

For each crash observation, vehicle registration plate was matched with the correspondent VIN, which is equivalent to the "*N.º Homologação Nacional*" at "*Folha de Aprovação do Modelo*", IMTT sheet. As an example, a copy of this document is presented in Appendix 4 for a Toyota Corolla E12T, 2005 vehicle model year, (vehicle's registration plate was deleted on purpose for safeguarding owner privacy). For this vehicle in particular, the characteristics acquired from IMTT databases are listed below:

- **Brand Name** (Toyota),
- **Model** (Corolla E12T),
- **Wheelbase** (2600 mm),
- **Length Size** (Not available for this model),
- **Curb Weight** (1360 kg),
- **Engine Size** (1364 cm^3),

- **Fuel** (Diesel),
- **Vehicle Registration Year** (2005).

Successively, vehicles registration plates were matched with the "*Folha de Aprovação do Modelo"* in an Excel spreadsheet, which dataset contained the record of each crash observation. An integrated database was developed where each crash record and technical characteristics for the vehicles involved in the collision were combined into a unique crash event observation. Appendix 6 shows the code applied to the statistical analysis software (SAS) for reading all the data imported from the Excel spreadsheet crash database (explained in section 3.2.4) and converting it into SAS data source. Following, SAS crash data source was subject to data mining analysis with Enterprise Miner (EM) software.

3.3 Structure of the Database and Variables Definition

This section explains the crash database subdivided by datasets and it defines the variables used in the crash sample. For simplicity, three crash datasets were defined based on the number of vehicles involved:

- **All** represents the total of the crashes observations including single-vehicle crashes and two-vehicle collisions (N=1,374),
- **Two** represents the two-vehicle collisions (N=874),
- And **Single** represents the single-vehicle crashes (N=500).

The crash dataset includes two types of variables and three classes of variables. The two types of variables including in this analysis were: categorical and continuous. The categorical has values that function as labels rather than numerical information, and in some programs are called as "nominal" variables, such as in data mining software. On the other hand, the continuous variables have numeric values. In the crash dataset, examples of those types of variables are presented next.

a) Categorical variables: crash type, speed level and weather conditions.
b) Continuous variables: vehicle weight, vehicle engine size, vehicle age and vehicle wheelbase.

The two classes of variables used during the crash data modeling are presented as follows:

a) **Target variable/dependent variable or response variable** is the variable whose values are modeled and predicted by other variables. An example is crash severity.
b) **Predictor variable/independent variable or explanatory variable** is a variable whose values are used to predict the target variable. An example is vehicle weight.

The most widely adopted approach for predictive modeling of crash severity is to categorize the data using dummy variables (which are an artificial variable created to represent an attribute with

two or more distinct categories/levels). For example, alcohol and/or drugs test results, which originally was recorded as a continuous variable was converted into a dummy variable having two levels: “1” if the driver had alcohol content in the blood >0.5 $g.L^{-1}$ and/or the test for drugs were positive, “0” if the driver was legal. Dummy variables provide a good linear approximation of the non-linear features of the data. In this Thesis, binary targets were used to predict the crash severity as explained next.

Regarding two-vehicle collisions, it is convenient to explain vehicle identification/order: vehicle V1 and vehicle V2, previously mentioned in section 3.2.3. Hard copy reports usually warn that their contents includes crash witness’s description (if there is any), rather than providing much technical and/or official explanation. Also, for a collision involving two vehicles usually is its unknown what vehicle was responsible for the crash. Thus to avoid judgments, Officers just identified vehicles as vehicle V1 and Vehicle V2. The order in the crash records does not obey a restricted and predefined procedure. Similarly to Tolouei et al., vehicle V1 and vehicle V2 keep the same labels as those in the original police crash reports and this order are believed to be arbitrary [65].

Table 3.2 identifies the independent (explanatory) variables that were analyzed to estimate and/or predict their impact on crash severity outcomes. Table 3.2 also presents the derivative variables for vehicles V1 and V2 differential characteristics. For instance, in a two-vehicle collision, the weight differential between V2 and V1 as expressed by WTV2V1 (kg), which was obtained by subtracting the weight of vehicle V1 from vehicle V2. The same procedure was applied for the vehicle’s engine size, wheelbase, and age, leading to the following derived variables: ccV2V1, WBV2V1, and AgeV2V1, respectively.

Table 3.2 - Description of independent variables used in the analysis of crash database.

Variable	Description	Symbol
Age, Vehicle 1	AgeV1 (yr): year of the crash event -year of the first vehicle registration.	AgeV1
Age, Vehicle 2	AgeV2 (yr): year of the crash event - year of the first vehicle registration.	AgeV2
Age Difference vehicle between (V2) and (V1)	AgeV2V1 (yr): age of vehicle V2 - age of vehicle V1.	AgeV2V1
Alcohol and/or Drugs	The Driver´s test for alcohol and or drugs is presented as: Code=0, legal; Code=1, illegal	AlcoholDrugs
Number of vehicles involved	The number of vehicles involved distinguish between single vehicle crash and multi-vehicles collisions and it is coded as follows: NVehicles=1, if a single vehicle is involved in the crash NVehicles=2, if two vehicles are involved in the crash	NVehicles
Crash Type	RanOff=1, if crash type is RanOff Road, else RanOff =0 Rollover=1, if crash type is Rollover, else Rollover=0 RearEnd=1, if crash type is Rear End, else RearEnd=0 HeadOn=1, if crash type is Head-on, else HeadOn=0 Sideswipe=1, if crash type is Sideswipe, else Sideswipe=0 Other=1, if crash type is Other, else Other=0	CrashCode
Divided/ undivided	Existence or absence of physical median/barrier: Code=0, undivided Code=1, divided	DivisionCode
Engine Size Vehicle 1	Engine size of vehicle (V1) (cm^3)	ccV1
Engine Size Vehicle 2	Engine size of vehicle (V2) (cm^3)	ccV2
Engine Size Difference between vehicles (V2) and (V1)	ccV2V1: engine size of vehicle V2 - engine size of vehicle V1, at crash observation (cm^3).	ccV2V1
Road Class	Based in the number of lanes and coded as follows: Code=0, two lanes Code=1, multi-lanes Code=2, motorway	RoadClass
Speed Level	SpeedLevel=1, if Speed Limit > 90 $km.h^{-1}$, else, SpeedLevel=0	SpeedLevel
Wheelbase Vehicle 1	Wheelbase of vehicle (V1) (mm)	WBV1
Wheelbase Vehicle 2	Wheelbase of vehicle (V2) (mm)	WBV2
Wheelbase Difference between vehicles (V2) and (V1)	WBV2V1: wheelbase of vehicle V2 - wheelbase of vehicle V1, at crash observation (mm).	WBV2V1
Weight Vehicle 1	Weight of vehicle 1 (V1) (kg)	WTV1
Weight Vehicle 2	Weight of vehicle 2 (V2) (kg)	WTV2
Weight Difference between vehicles (V2) and (V1)	WTV2V1 stands for weight of vehicle V2 minus the engine size of vehicle V1, at crash observation (kg).	WTV2V1
Weather Conditions	Weather conditions at the moment of the crash: Code=0, Clear and/or dry pavement Code=1, rain and/or wet pavement	WeatherCode

Table 3.3 identifies four categories for the dependent variables (response variables or targets) used during the statistical modeling. The dependent variables categories were defined by performing calculations and aggregations with the original crash outcomes, namely the number of light injuries (LI), serious injuries (SI) and killed (K) in a crash record. As an example, the dependent variable labeled “SIK” was created to signify the sum of the number of serious injuries and fatalities in a crash.

For the single-vehicle crashes, the response variable was crash severity expressed by the variable FatalSIK, which represents the probability of serious injuries and/or fatalities among the occupants of the vehicle being studied following the crash.

For two-vehicle collisions, the safety analysis included not only the contribution of each individual vehicle in the overall crash severity, but also explores the individual impact of each vehicle in the protection of its occupants and risk imposed to the occupants of the opponent vehicle. **Thus, for the two-vehicle collisions, three response variables have been defined as follows.**

1. **The overall crash severity is expressed by the variable FatalSIK, which represents the probability of serious injuries and/or fatalities among the occupants of the two vehicles involved in the collision, regardless of the vehicle's identification.**
2. **Crash severity for the studied vehicle is defined by FatalSIKV1, which represents the probability of serious injuries and/or fatalities among the occupants of the studied vehicle, vehicle V1.**
3. **Crash severity for the opponent vehicle is defined by FatalSIKV2, which represents the probability of serious injuries and/or fatalities among the occupants of the vehicle V2.**

Thus, FatalSIKV1 takes into account the protective effect of vehicle V1 and the risk imposed by the vehicle V2 into the severity sustained by the occupants of V1. On the other hand, FatalSIKV2 takes into account the protective effect of vehicle V2 and the risk imposed by vehicle V1 into the severity sustained by the occupants of V2.

Table 3.3 - Description of dependent variables for crash data set modeling.

Variable	Description	Symbol
Number of Killed (K) plus Serious Injured (SI)	SIK: sum of occupants serious injured (SI) + sum of occupants killed (K) in a crash event.	SIK
Serious injured and/or killed in the crash (with one vehicle or two vehicles involved)	FatalSIK: categorical response for a crash outcome used to predict either a serious injury, or fatality in a crash event. FatalSIK=1, if SI>0 and/or K>0, else, FatalSIK=0	FatalSIK
Serious injured and/or killed in vehicle V1 occupants	FatalSIKV1: categorical response for a crash outcome used to predict either a serious injury, or fatality or both for occupants in vehicle 1 in a crash event. FatalSIKV1=1, if SI>0 and/or K>0, else, FatalSIKV1=0	FatalSIKV1
Serious injured and/or killed in vehicle V2 occupants	FatalSIKV2: categorical response for crash outcome for a crash outcome used to predict either a serious injury, or fatality or to both for occupants in vehicle 2 in a crash event. FatalSIKV2=1, if SI>0 and/or K>0, else, FatalSIKV2=0	FatalSIKV2

Following data description and variables definition, the next sections of this Chapter explain the approach developed for the crash data analysis.

3.4 Vehicle Brand Severity Ratio Analysis

The individual vehicle analysis aims to infer severity index at the crash sample with the overall severity index at national fleet. Also, it gives attention to the vehicle brand representatively the in sample and the severity ratio for the crashes involving the vehicle's auto brand being analyzed.

The Portuguese crash database covers all the police injuries and fatalities registration records segregated by light injured, serious injured and killed, for crashes involving one single vehicle or two vehicles involved. For these crashes matching the criteria established in this study (in section 3.2.2) the overall severity index (OSI) was defined by the equation above:

$$OSI\,(\%) = \frac{SIK_{\ PT}}{LI_{PT} + SI_{PT} + K_{PT}} * 100 \qquad \text{Equation 3.1}$$

Where "OSI" is the overall severity index for the national fleet, "SIK_{PT}" is the sum of the number of serious injured and killed, and "LI_{PT}+SI_{PT}+K_{PT}" is the sum of all the injuries and killed for the national fleet. The OSI was estimated for the time period 2006-2010 and individually for single vehicle crashes and two-vehicle collisions, in order to allow the comparison with the crash data sample used in this study. Following, a crash severity index (CSI) was calculated for each crash dataset: Single and Two, as established on Equation 3.2:

$$CSI(\%) = \frac{SIK}{LI + SI + K} * 100 \qquad \text{Equation 3.2}$$

Subsequently, for each crash dataset, the vehicle brands that showed a higher frequency in crash involvement were investigated for the numbers of occupants distributed amongst the injury level. A brand severity ratio (BSR) was defined as follows:

$$BSR_i(\%) = \frac{SIK_i}{LI_i + SI_i + K_i} * 100 \qquad \text{Equation 3.3}$$

Where "i" is the Auto Brand, "BSR_i" is the brand severity ratio, "SIK_i" represents the sum of number of serious injured and killed for crashes involving that brand, and "LI_i+SI_i+K_i" is the total number of injured and killed in the crashes where that brand was involved. Firstly, for Single and Two datasets, BSR_i for the most frequent brands was compared with the corresponding CSI. Secondly, each BSR_i was evaluated by comparing with OSI.

For the inference of individual vehicle brand injury severity ratio with the injury severity level at the Portuguese fleet, specific road safety data was requested to ANSR in order to estimate the OSI. Then, those brands were analyzed base on their share in the Portuguese fleet. Brands sales information and annual number of vehicles register at the National fleet were obtained from the Portuguese Automobile Association (ACAP) [134, 135]. Then, BSR_i was discussed taking into account brands exposure on the national fleet based on brans sales annual percentage by numbers of vehicles annually registered.

3.5 Analysis Strategy for Imbalance Crash Data

In this study, the main constrains of the crash dataset modeling were related to small sample size, and disproportion between severe and non-severe events. The safety analysis methodology identifies which factors are determinant for crash severity prediction. With regard to binary data classification (such as severe or non-severe crashes), analysis of data containing rare events or imbalance class distributions poses a great challenge to the machine learning community [114]. There is the need to have an algorithm method that would provide high accuracy for the minority class, without making vulnerable the accuracy of the majority class [106]. Previous authors (see section 2.5.4) have not shown any strategy to deal with the problem of imbalanced classes in crash analysis. This gap in previous research leads to the greatest challenge of this work: design an approach to resampling crash events in order to allow further modeling analysis with adequate degree of accuracy. First, proof of original crash imbalanced data is presented. Second, the strategy to balance the original crash data is explained.

3.5.1 Imbalance data within the original crash sample

From a total of 1,374 crashes selected for this study, only 5.1% had resulted in serious and/or fatal crashes. Thus, for a binary target classification, this means that there were 70 severe crashes (events being “1”) and 1,304 non-severe crashes (events being “0”). The overall sample crash severity proportion of 5.1% proves a clear imbalance distribution between severe and non-severe events. Consequently modeling the original imbalance sample would lead to high accurate predictions for non-severe crashes, but poor predictions for the severe crashes, since they represent the minority class. As a result, there was the need to have an algorithm method that will provide high accuracy for the minority class, without making vulnerable the accuracy of the majority class.

3.5.2 Balancing strategy- Stratified random sample

This section explains the balanced strategy which was applied to both predictive methods: CART and logistic regression. Random sampling often does not provide enough targets to train a predictive model for rare events. Since the response rate was very low it was necessary to include all the responders available and only a random fraction of non-responders [117]. Studies have shown that for several classifiers, a balanced data set provides improved overall classification performance when compared to an imbalanced data set [83, 106]. However studies do not imply that classifiers cannot learn from imbalanced dataset [83]. As a matter of fact, some studies have shown that classifiers applied to certain imbalanced dataset are comparable to classifiers induced from balanced datasets [106, 110]. In balanced sampling, the attempt is to draw samples from a

population but with the composition of the dependent variable in the sample being different from that in the original population [136].

For balancing the crash data for predictive modeling a resampling strategy was applied [89, 103, 117, 118]. To deal with the overrepresentation of non-severe crashes (target with outcome being “0”) a resampling approach was applied. Instead of randomly sampling cases from the modeling sample, cases from each outcome level were separately sampled. Since the number of the cases of interest (target “1”) was especially small, all available severe cases were selected, and then, they were matched with one non-severe case (target “0”), which was randomly selected.

To model rare events with SAS® Enterprise Miner™, all the observations having the rare event (severe crash) were included, but only a fraction of the non-event (non-severe crash) was included [103, 117]. The fraction of the non-event (or majority class) was randomly selected. At the EM interface, the sample was configured for stratified random sampling properties, by omitting cases of the common classes in the trading dataset.

Each crash dataset, (All, Two and Single), was stratified to the target proportion 0.5, leading to training samples were the proportion of target level “1” (severe crash) was equal to the target level “0” (non-severe crash). However this procedure biases the sampling to provide enough target events to effectively train a predictive model, leading to overrepresentation of target level “1” (severe crashes), which is the response level of interest for this research. Thus, the models developed from the balanced sampling would be biased unless a correction is made for the bias caused by over-representation of the target “1”. The approach followed to correct this bias was different for each predictive modeling technique, since the algorithms sensitivity to the balancing samplings is different. For logistic regression, the balancing applied to datasets appears to be of minor importance. However, the other methods demonstrate a greater sensitivity to balancing, particularly CART [83]. For logistic regression the solution include adjusting the decision threshold by adding a cutoff node function, as going to be explained in section 3.7.2. For the decision prediction modeling (decision trees) the approach used to correct the bias introduce by balancing by adjusting prior probabilities, as explain in the next section. The predictive models were developed using SAS® Enterprise MinerTM 7.1 [84, 89, 117].

3.6 CART Methodology

Decision trees provide an excellent introduction to predictive modeling and are useful to predict new cases, select useful inputs and optimizing complexity [84, 89, 118]. Tree prediction algorithms can be applied for distinct predictions types, namely decisions, rankings and estimates. This section explains the modeling approach with CART methodology. First, the reasons why decision trees are sensitive to relative high imbalanced classes are presented. Then, the strategy

implemented to correct the bias introduced by balancing the crash data is explained. Following decisions trees development and assessment are explained.

3.6.1 CART methodology selection

Trees as predictive algorithms do not assume any association structure, they simply isolate concentrations of cases with like-valued target measurements [89].

CART methodology was selected for the following reasons.

1. Traditional statistics have limited utility in the task of variable selection for multiple variable comparisons. Apart from identifying the variables that improve classification accuracy, the methodology also identifies clearly the variables that are neutral to accuracy, and also those that decrease it [137].
2. Predictor variables are rarely satisfactorily distributed and decisions trees can deal with missing data [46, 86, 138]. Fortunately, at the crash data set, there were no missing inputs for any of the variables included in this analysis.
3. Complex interactions may exist amongst the explanatory variables, such as vehicle engine size, vehicle weight, crash type and weather conditions. CART has the potential to "uncover complex interaction between predictors which may be impossible to uncover using traditional multivariate techniques" [86].
4. It is a powerful method to deal with prediction and classification problems, mainly when there is a large amount of data with many independent variables [104].
5. CART output is almost intuitive and offers an easier comprehension between the target and the explanatory variables.

3.6.2 Decision trees structure

The decision tree represents a segmentation of the data that is created by applying a series of rules, resulting in a hierarchy of segments within segments. The hierarchy is called a tree, and each segment is called a node (or a leaf). A simplified decision tree is illustrated in Figure 3.3. The original segment contains the entire data set and is called the root node of the tree. Then, the root node is divided into child nodes (also called tree leafs) on the basis of an independent variable (splitter in Figure 3.3), which creates the best purity in the way that the data in the child note is more homogeneous than in the upper parents node [104]. For each leaf, a decision is made and applied to all observations in the leaf. This process will last until all data in each node have as much as possible homogeneity, leading to the terminal nodes or terminal leafs.

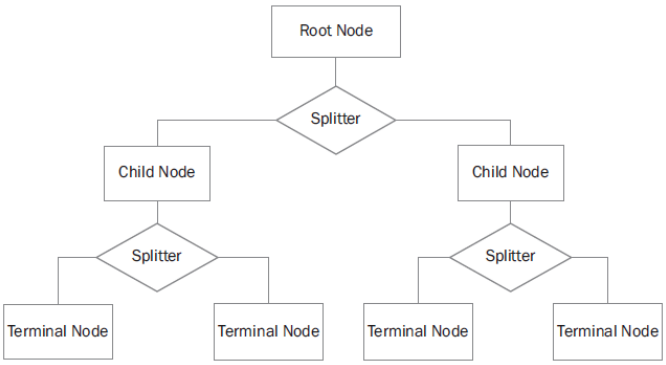


Figure 3.3 - General structure of a decision tree [104].

Therefore, a decision tree split, for a binary class, can be expressed by a confusion matrix. The parent node contains positive and negative examples, by splitting, one node will carry the true and false positive observations and the other node will carry the true and false negative observations [139]. CART provides an advance methodology for predictive modeling, the decision is simply the predicted value [89]. To select useful inputs, trees employ a split search algorithm. The split search selects an input for partitioning the training data. If the input was coded as an interval variable (for instance, vehicle weight), each unique value serves as a potential split point for the data. If the input is categorical (for instance, speed level), the average value of the target is used [89]. For a selected input, two groups are generated, resulting in two leafs (or child node).

If input values are less than the split point are said to branch left. If input values greater than the split point are said to branch right. The groups, combined with the target outcomes, form a 2x2 contingency table with columns specifying branch direction (left or right) and rows specifying target value (0 or 1). For the slipping rules, the criterion is based on either a statistical significance test, namely a F test or a Chi-square test, or on the reduction in variance, Gini index [89]. The significance level specifies the maximum acceptable p-value for the worth of a candidate splitting rule, and by default was configured for 0.2 [84, 140]. The F test and Chi-square test accept a p-value input as a stopping rule [89]. A Pearson chi-square statistic is used to quantify the independence of counts in the table's columns. Large values for the Chi-square statistic suggest that the proportion of zeros and ones in the left branch is different than the proportion in the right branch [89, 117]. A large difference in outcome proportions indicates a good split. The p-value indicates the likelihood of obtaining the observed value assuming identical target proportions in each branch [89]. For large data sets, these p-values can be very close to zero. For this reason, the quality of a split is reported by logworth = -log(chi-squared p-value). At least one logworth must exceed a threshold for a split to occur with that input. A threshold corresponds to a chi-squared p-value of 0.20 or a logworth of approximately 0.7 [89]. Hence, the best split for an input is the split that has the highest logworth. For more details of the tree algorithm, the paper by Das is recommended [87].

The Decision Tree analysis provides information for the output selected variables based on their relative importance. The relative importance of an input variable in subtree denotes the primary or surrogate splitting rule using that input in a way the node assures the reduction in sum of squares errors (SSE) from the predicted values. It must be noticed that the variables relative importance may or may not follow the order of the variables selected by the tree for the split. The split is based on logworth. Hence, input variables that have a larger -log(chi-squared p-value) are selected first. On the other hand, the variable importance choses as most important the variable that will minimize the SSE associated with the other independent variables. The variable importance measure (VIM) is one of the CART method output that is helpful for the analysis of which variables are more important to classify or predict the target [87, 89, 104]. More information on the variable importance score algorithm can be found at Das and Kashani [87, 104]. VIM is very helpful for variables selection and will be used in the discussion of the decision trees modeling results (Chapter 5).

3.6.3 Decision trees- Strategy to handle the imbalanced data

CART is one of the most popular algorithms in decision tree induction, however splitting criteria is considered to be skew sensitive, because splitting criteria as the skewness increases, the information gain will become poorer [139, 141]. This occurs because the sampling methods prior to the decision tree induction alter the class distribution driving the bias towards the majority or positive class [139]. The objective functions used by the classifiers methods typically tend to favor the larger, less important class in the analysis of imbalanced datasets [139]. Thus, the predictive accuracy might not be appropriate when the data is imbalanced and /or the cost of different errors vary significantly [111]. With imbalanced datasets it is useful to incorporate the prior of the positive class to smooth the probabilities so that the estimates are shifted toward the minority class base rate [106].

Following the balanced strategy (section 3.5.2), the solution to correct the bias imposed by the imbalanced crash data was to adjust the probabilistic estimates at the tree leaf [109, 117, 140]. The bias introduced by over representing level “1” was corrected by adjusting the predicted probabilities with prior probabilities, allowing the model to predict the original distribution of target “1” for the original crash data. The adjustment of prior probabilities was performed with a decision node. As explained, the original portability of a severe crash was: 0.051, 0.037 and 0,073 for All, Two and Single crash datasets, respectively. To balance the bias by stratified 0.5 level training samples generation, the prior probabilities were adjusted for the original proportion of target level “1” and “0”. For instance, for the two-vehicle collisions dataset, the stratified sample procedure has generated a training sample including all the severe crashes (32 events) and equal proportion of non-severe crashes (which were randomly selected). Then the prior probability of 0.5 was adjusted for the original probability of 0.963, and 0.037, for targets levels “0” and “1, respectively. Table 3.4

summarizes the adjusting prior probabilities for the stratified training samples used in the trees model development.

Table 3.4 – Stratified Training Samples adjust prior probabilities for the original crash dataset.

Data set	Stratified Levels		Prior Probabilities	Adjusted Prior
	Level	Count		
All	1	70	0.5	0.051
	0	70	0.5	0.949
Two	1	32	0.5	0.037
	0	32	0.5	0.963
Single	1	38	0.5	0.076
	0	38	0.5	0.924

It must be pointed out that in this study the 0.5 stratified level was chosen under the constrain of the available observations for the minority class (rare event), so that all random samples would contain all the rare events (severe crashes), since the sample size was small and imbalanced for the crash severity distribution. In addition, with small or moderate data sets, data splitting would be inefficient; the reduced sample size can reduce the fit of the model training and validation [43, 89, 103]. However, the conventional split between training data and testing data was not applied in this study, due to sample constrains. Thus, for decision trees assessment significance test analysis (to be explained in section 3.6.5) was applied.

3.6.4 Decision trees development

The process flow diagram for the decisions trees was created as follows. Each input dataset, All, Two and Single, were imported into the software interface. The sample node allowed to extract a sample from crash input data source. Then each tree node was connected to the decision node. The trees were created with the assessment method and assessment measure set for decisions because decision trees were applied to produce only a class decision, such as severe crash or non-severe crash, in this study. Table 3.5 shows the variables that were used as inputs for each tree development, as well as, the dependent variable used as a target.

Table 3.5 – Description of input variables and targets in CART modeling.

Variable	Description	Abbreviation	Variable role	
			Input	Target
Age of Vehicle 1	AgeV1 (yr) was calculated based on the year of the crash event minus the year of the first vehicle registration.	AgeV1	Figures 5.1 to 5.8	-
Age of Vehicle 2	AgeV2 (yr) was calculated based on the year of the crash event minus the year of the first vehicle registration.	AgeV2	Figures 5.1, 5.2, 5.4, 5.5, 5.7 and 5.8	-
Age Difference between vehicles (V2) and (V1)	AgeV2V1 (yr) stands for age of vehicle V2 minus the age of vehicle V1, crash observation.	AgeV2V1	Figures 5.1, 5.2, 5.4, 5.5, 5.7 and 5.8	-
Alcohol and/or Drugs	The Driver´s test for alcohol and or drugs is presented as: Code=0, legal; Code=1, illegal	AlcoholDrugs	Figures 5.1 to 5.8	-
Crash type for single vehicles	Ran off road Rollover	RanOff Rollover	Figures 5.3 and 5.6 -	-
Crash type for collisions	Rear End Head-On Sideswipe Other	RearEnd HeadOn Sideswipe Other	Figures 5.1, 5.2, 5.4, 5.5, 5.7 and 5.8 -	-
Divided/undivided	Existence or absence of physical median: Code=0, undivided Code=1, divided	DivisionCode	Figures 5.1 to 5.6	-
Number of vehicles	Number of vehicles involved in the crash: Code=1, if only one vehicle was involved Code=2, if two vehicles were involved	NVehicles	Figure 5.1	-
Serious and/or killed in the crash (involving one vehicle or involving two vehicles)	FatalSIK is a categorical response for a crash outcome used to predict either a serious injury, or fatality in a crash event. FatalSIK=1, if SI>0 and/or K>0, else, FatalSIK=0	FatalSIK	-	Figures 5.1 to 5.6
Serious and/or killed at vehicle 1 (V_1) occupants	FatalSIKV1 is a categorical response for a crash outcome used to predict either a serious injury, or fatality or both for occupants in vehicle 1 in a crash event. FatalSIKV1=1, if SI>0 and/or K>0, else, FatalSIKV1=0	FatalSIKV1	-	Figure 5.7
Serious and/or killed at vehicle 2 (V_2) occupants	FatalSIKV2 is a categorical response for crash outcome for a crash outcome used to predict either a serious injury, or fatality or to both for occupants in vehicle 2 in a crash event. FatalSIKV2=1, if SI>0 and/or K>0, else, FatalSIKV2=0	FatalSIKV2	-	Figure 5.8
Speed Level	The speed level was coded as follow: If Speed limit≤90 km.h^{-1}, then code=0 If Speed limit>90 km.h^{-1}, then code=1	SpeedLevel	Figures 5.1 to 5.7	-
Wheelbase of Vehicle 1	Wheelbase of vehicle (V1) (mm).	WBV1	Figures 5.1 to 5.8	-
Wheelbase of Vehicle 2	Wheelbase of vehicle (V2) (mm).	WBV2	Figures 5.1, 5.2, 5.4, 5.5, 5.7 and 5.8	-
Wheelbase Difference between vehicles (V2) and (V1)	WBV2V1 stands for wheelbase of vehicle V2 minus the wheelbase of vehicle V1, at crash observation, (mm).	WBV2V1	Figures 5.1, 5.2, 5.4, 5.5, 5.7 and 5.8	-
Weight of Vehicle 1	Weight of vehicle 1 (V1) (kg).	WTV1	Figures 5.1 to 5.8	-
Weight of Vehicle 2	Weight of vehicle 2 (V2) (kg).	WTV2	Figures 5.1, 5.2, 5.4, 5.5, 5.7 and 5.8	-
Weight Difference between vehicles (V2) and (V1)	WTV2V1 stands for weight of vehicle V2 minus the engine size of vehicle V1, at crash observation (kg).	WTV2V1	Figures 5.1, 5.2, 5.4, 5.5, 5.7 and 5.8	-
Weather Conditions	Weather conditions at the moment of the crash: Code=0, Clear and/or dry pavement Code=1, rain and/or wet pavement	WeatherCode	Figures 5.1 to 5.7	-
Engine Size of Vehicle 1	Engine size of vehicle (V1) (cm^3).	ccV1	Figures 5.1 to 5.8	-
Engine Size of Vehicle 2	Engine size of vehicle (V2) (cm^3).	ccV2	Figures 5.1, 5.2, 5.4, 5.5, 5.7 and 5.8	-
Engine Size Difference between vehicles (V2) and (V1)	ccV2V1 stands for engine size of vehicle V2 minus the engine size of vehicle V1, at crash observation, (cm^3).	ccV2V1	Figures 5.1, 5.2, 5.4, 5.5, 5.7 and 5.8	-

The CART methodology for decisions classification of target FatalSIK was performed for each crash dataset based in two procedures: imbalance sample (original sample distribution of severe and non-severe crashes) and balance sample (stratified sample with equal proportion of severe and non-severe crashes). For an advanced analysis of the vehicles' effect on crashworthiness and

risk imposed to the other car involved in the collision, two additional target variables were explored: FatalSIKV1 and FatalSIKV2 (as explained in section 3.3). These response variables have few observations for the target level “1”: 21 and 14 for FatalSIKV1”1”, and FatalSIKV2”1”, respectively. Due to the limited number of the target with the level of interest, the resampling approach was not performed, otherwise the stratified random sample procedure (randomly removing the majority class to a balanced proportion) would lead to small training samples: 42 observations to model FatalSIKV1 and 28 observations to model FatalSIKV2. As it was explained in section 3.4.2.1, decision trees are very sensitive to the sample size and small leafs (i.g, small number of observations in the tree node). Therefore, for the two-vehicle collisions, decision trees modeling were developed with the distribution (0.037, and 0.963 for severe and non-severe collisions, respectively).

3.6.5 Decision trees significant test analysis

Chi-square statistics is widely employed to ensure the accuracy of risk factor identification [87, 100]. In this study, to examine whether there is an association between the predictor variables selected at the trees’ leafs and the target, Chi-square test (Chi-Sq) was conducted. Chi-Sq test measure the difference between the observed cell frequencies and the cell frequencies that are expected if there is no association between the variables. If the p-value is small (less than 0.05) there is enough evidence at 5% significance level to reject the null hypothesis. If the association test results in a significant Chi-Sq statistic, there is strong evidence that an association exists between the variables. The value of the Chi-Sq statistic only indicates how confident the researcher can be to reject the null hypothesis. This test does not show the magnitude between the variables being analyzed. When more than 20% of the cells (nodes at the tree) have expected frequencies of less than 5, the Chi-Sq test might not be valid [89]. This happens with the crash data sample used in this study, since there are a limited number of observations. For small samples, exact p-value is useful, however sometimes it might requires a prohibit augment of time and computing memory for the EXACT statement in SAS® v9.2. The exact p-value reflects the probability of observing a table with at least the same evidence of an association as the one actually observed, given there is no association between the variables. Therefore, Fisher’s exact test was used to ensure the accuracy of severe crash factors identification, for the situation where the Chi-Square test was not valid at the 5 % significance level (for those cells that had expected counts less than 5) [89].

3.7 Logistic Regression Methodology

Regression offers a different approach to prediction modeling compared to decision trees [61, 89]. Regressions, as parametric models, assume a specific structure between inputs (predictors) and target. Whereas trees as predictive algorithm, do not assume any association structure, they simply isolate concentrations of cases with like-valued target measurements. A great advantage of logistic

regression technique comparing to CART technique is that regression provides valuable information on the parameters estimates, their standard error and their significance. Logistic regression method was selected to predict the probability that the binary target will acquire the event of interest as a function of the independent inputs. First, this section provides a background of logistic regression analysis. Second, the developments of logistic models are explained. Third, logit regression models validation approach is presented.

3.7.1 Logistic regression background

The logistic regression is widely used for predictive modeling of binary targets. The binary logistic regression model was developed primarily by Cox and Walter and Duncan [103]. The odds of an event can be expressed by the probability of that event as Equation 3.4:

$$Odds = (\frac{P}{1-P}) \quad \text{Equation 3.4}$$

Where “P” is the probability of the event. In logistic regression, the dependent variable responds to a logit, which is the natural log of the odds, (Equation 3.5), that is:

$$\log(odds) = logit\ (P)\ = \log\left(\frac{P}{1-P}\right) \quad \text{Equation 3.5}$$

The logit transformation in the logistic regression model is described by the following equation:

$$logit(P) = \log\left(\frac{P}{1-P}\right) = \beta 0\ +\ \beta 1 * \mathbf{x1} + \ldots + \beta k * \mathbf{xk} \quad \text{Equation 3.6}$$

Equation 3.6 expresses a linear relation between the odds and X in terms of probability. The logistic function is the inverse of logit function. A logistic regression applies a logit transformation (a natural log of the odds) to the probabilities and ensures that the model generates estimated probabilities between 0 and 1. At this function, x has an unlimited range while P (Probability) is restricted to range from 0 to 1. The preceding Equation 3.6 could be transform to probabilities by applying the natural log by sides of the above equation and solving for “P”. Subsequently the above equation in terms of probability it can be rewritten as Equation 3.7:

$$P = \frac{\exp(\beta 0\ +\ \beta 1 * \mathbf{x1} + \ldots + \beta k * \mathbf{xk})}{(1 + \exp(\beta 0\ +\ \beta 1 * \mathbf{x1} + \ldots + \beta k * \mathbf{xk}))} \quad \text{Equation 3.7}$$

Where “$\beta 0$" is the intercept, “$\beta 1$" is the estimated for the parameter “$\mathbf{x1}$”, and the same for “$\beta k * \mathbf{xk}$”. The logistic mathematical model assumes a linear relationship between predictors and the logit for the response variable. The slope coefficient in the logistic regression model represents the change in the logit for a change of one unit in the independent variable “x” [85]. Unlike linear regression, the logit is not normally distributed and the variance is not constant. Hence, the least squares estimation is abandoned in favor of maximum likelihood estimation. The logistic regression requires a more complex estimation method than the linear regression, called maximum likelihood to

estimate the parameters. For the logistic regression analysis the function chosen to measure the fit of the model is the maximum likelihood. The likelihood function is the joint probability density of the data related as function of the parameters. The Likelihood is a conditional probability (P|X), the probability of Y given X. Hence in model selection the parameters that will be chosen are the ones that yield to the greatest likelihood computed. The estimates are called maximum likelihood because the parameters are chosen to maximize the likelihood of the sample data. The logistic regression finds the parameters estimates that are most likely to occur given the data [89]. This procedure is achieved by maximizing the likelihood function that expresses the probability of the observed data as function of the unknown parameters [84, 89].

3.7.2 Logistic regression modeling

The modeling process flow was developed with application of specific functions of the EM program: such as drop, transformation, regression and cutoff nodes. Appendix 5 provides detailed information for the logistic models development using SAS®Enterprise MinerTM 7.1 [84, 89, 117]. During the training process, four selection methods for variables input in the model were used:

- *Backward* - begins with all candidate effects (inputs) in the model and removes effects until the stay significance level is met. It creates a sequence of models decreasing complexity.
- *Forward*- begins with no candidate inputs in the model and adds inputs until the entry significance level is met. In contrast with backward selection creates a sequence of models of increasing complexity.
- *Stepwise*- begins as in the forward selection but may remove inputs already in the model. This procedure sequentially adds inputs with the smallest p-value below the entry cutoff. As each input is added, the algorithm re-evaluates the statistical significance of all included inputs in the model. If p-value of the selected inputs exceeds a stay cutoff, the input is removed from the model.
- *None*- When none of the above selections methods are selected, the regressions use all the available inputs to fit the model. Usually, it generates models with higher complexity since all the predictor variables stay in the model.

As result of the input selection methods, several candidate models were developed, some incorporating all the input variables (when “none” method was selected), others candidate models with several or few inputs.

Following the development of several models candidates, the best model to predict the target was selected based on the goodness of fit of the model to the crash data. Following the selection of the best model, cutoff, score and SAS code nodes were added to the diagram for further assessment of the prediction accuracy.

The cutoff function provides graphical information to determine the appropriate probability cutoff point for decision making with binary target models. The establishment of a cutoff decision point entails the risk of generating false positives and false negatives, but an appropriate use of the cutoff node can help minimize those risks. During the models training, the optimal cutoff value was obtained for 0.69. This value was found by taking into account which cutoff would result in a higher overall classification rate and the prior probabilities for the severe crashes in the data set.

The score function creates predictions using the best model selected based on the model comparison node, described above. To evaluate the performance of the selected model from the training procedure, a new data source must be dragged into to diagram workspace. While for the training models development the data set's role was set to "raw", for the score stage, the data set was set to "score" role. This attribute allows the score node to use the data set to generate predicted values for a data set that might not contain a target.

Finally, at the end of the models development process, sas score code function was linked to the score node. This function allows to programing code to generate an output for the model performance when evaluating its prediction accuracy with the original data. The generated report output creates the scores results for the classification assessment, (that will be discuss in the next section).

3.7.3 Models assessment and validation

The most frequent metrics for models assessment are accuracy and error rate [106]. By convention the class label of the minority class is positive, and the class label of the majority class is negative [107, 108]. Given a classification model (also called classifier) and a response, there are four possible outcomes. If the response is positive and it is classified as positive, it is counted as a true positive; if it is classified as negative, it is counted as a false negative [142]. If the response is negative and it is classified as negative, it is counted as a true negative; if it is classified as positive, it is counted as a false positive [142]. Given a classifier and a set of responses, a two-by-two confusion matrix (also called a contingency table) can be constructed representing the dispositions of the set of responses, with the true class on the columns and the predicted class on the lines. This matrix forms the basis for many common metrics and provides information on the performance of the model [106-108, 142].

In this study, the event classification table (metric provided by Enterprise Miner) is used to measure the assessment score rankings for the model, showing the predicted probabilities of the observed response (target being modeled). Binary targets can be classified as event or non-event. Predicted and observed targets results follow into four classification categories: False Negative, True Negative, False Positive, and True Positive.

Thus, the event classification analysis classifies the response output accuracy for the target being modeled as:

- False Negative (FN), which means that the target was predicted as “0” when it was “1” in reality.
- True Negative (TN), which indicates that the target was correctly predicted as “0”.
- False Positive (FP), which means that the target was incorrectly predicted as “1” when it was in reality “0”.
- And True Positive (TP), which means that the target was correctly predicted as “1”.

Table 3.6 shows the measures classification approach developed based on the Enterprise Miner software for the assessment score and confusion matrix for a binary classification; in this case FatalSIK. Table 3.6a) shows the assessment of the training model evaluation. The TN category refers to the observations where a crash was non severe (actual value was FatalSIK”0”) and it was predicted as non-severe (FatalSIK”0”). When a crash was severe (actual value FatalSIK”1”) and it was predicted as severe crash (FatalSIK”1”), this observation follows into the category TP.

Table 3.6 - Assessment of FatalSIK prediction based on event classification table.

Model Assessment Score				
a) Assessment of selected model with the training sample/balanced sample				
Target	**False Negative (FN)**	**True Negative (TN)**	**False Positive (FP)**	**True Positive (TP)**
Predicted Actual	FatalSIK”0” FatalSIK”1”	FatalSIK”0” FatalSIK”0”	FatalSIK”1” FatalSIK”0”	FatalSIK”1” FatalSIK”1”
b) Assessment of selected model with the original sample/imbalanced data				
Target	**True Positives (TPs)**	**False Positives (FPs)**	**True Negatives (TNs)**	**False Negatives (FNs)**
Predicted Actual	FatalSIK”1” FatalSIK”1”	FatalSIK”1” FatalSIK”0”	FatalSIK”0” FatalSIK”0”	FatalSIK”0” FatalSIK”1”

The accuracy of the model measures the fraction of cases where the decision matches the actual target value. The accuracy rate (AR) in the training model is equivalent to the percentage of the cases predicted right by the model within the training sample. Equation 3.8 shows the calculation of “Accuracy Rate” as:

$$Accuracy\ Rate_{Training\ sample} = \frac{(TP + TN)}{(FN + TN + FP + TP)} \qquad \text{Equation 3.8}$$

On the other hand, the misclassification measures the fraction of cases where the decision does not match the actual target value. Equation 3.9 shows the misclassification rate:

$$Misclassification\ = \frac{(FN + FP)}{(FN + TN + FP + TP)} \qquad \text{Equation 3.9}$$

The validation process of a model is an important step to confirm that the developed model is likely to perform as expected in the field. The standard strategy in predictive modeling is the data splitting. Thus, a proportion would be used for fitting the model, which is the training data. The

remaining data would be used for empirical validation. However, with small or moderate data sets, data splitting is inefficient; the reduced sample size can severely degrade the fit of the model [43, 89]. Kononen et al. stated that "splitting-sample validation results in the validation of the model fit to a "training" dataset, but does not validate the model fit to the complete dataset, the objective of a predictive model" [43]. Computer-intensive methods, such as cross-validation and the bootstrap methodologies can be used for both fitting and honest assessment [83, 84, 89, 103].

Validation process relies on model assessment to predict new cases. However, for the selected model during the training process, the model was scored based on the training sample (with a stratified distribution of severe *vs.* non-severe cases). Firstly, the final selected model was score not only for the stratified sample (balanced proportion of severe and non-severe crashes), Table 3.6a). Secondly, it was scored using the original crash data (with original distribution of severe *vs.* non-severe cases). Table 3.6b) shows the assessment measure for the selected model score with the original sample. To predict new cases using the original imbalanced sample, the classification measures are as follows: True Positive (TPs), False Positives (FPs), True Negatives (TNs) and False Negatives (FNs).

Similarly to Equation 3.8, the performed accuracy for the final model was expressed as the percentage of the cases predicted right by the selected model when scoring the crash population. The accuracy rate within the entire crash dataset was calculated by Equation 3.10 and percentage of predicted right cases was derived from the accuracy rate*100% and the accuracy rate was in this case determined as:

$$Accuracy\ Rate_{\ Original\ sample} = \frac{(TPs + TNs)}{(TPs + FPs + TNs + FNs)} \qquad \text{Equation 3.10}$$

The selected model (final model) was evaluated for the prediction accuracy performance. The procedure developed to ensure a valid and reliable validation of the selected models is based on the k-fold cross validation from Crone and Finlay and Xie et al. [83, 93].

For the purpose of this study, and to ensure valid and reliable estimates of the experimental results despite small sample sizes, a resampling K random cross-validation was employed for the selected, essentially replicating each random sample k = 10 times (i.e., resampling). The resampling k random cross validation is to some extent different from k-fold cross validation by Crone and Finlay. For the Portuguese crash data analysis, a stratified random sampling was applied, with the events "1" (severe crashes) and events "0" (non-severe crashes) sampled with equal proportion. The crash dataset was segmented into k sections of equal size, with an equal proportion of severe crashes and non-severe crashes, within each fold. For the two-vehicle collisions (N=874), 10 stratified random samples (N=64) were developed including all severe crashes (32 events "1") and equal number of non-severe crashes (32 events "0"). For the validation

of the best selected models for the single dataset, 10 stratified random samples (N=76) were developed including all the severe crashes (38 events “1”) and equal number of non-severe crashes (38 events “0”).

The resampling k random cross validation is presented next, through step 1 to step 3.

Step 1: ***Stratified random samples cross validation***

To create each stratified sampled subset, observations were randomly excluded from the majority class (the non-severe crashes) until they equal the observations number of the minority class (severe crashes). Hence, 10 samples with balance classes were generated from the full dataset. The 10 stratified random balancing samples were chosen taking into account: time consuming, computing requirements and the need to obtain a reasonable number of samples, under the constrain of the available observations.

Step 2: ***Model Score with stratified random samples***

In Crone and Finlay k-fold cross validation approach, the stratified samples were used to construct k models for each cumulative percentage of the population [83]. Then for each model, all the N/k observations in the validation section were used to evaluate the model performance. In this research, the performance of accuracy prediction of the final model was evaluated by comparing the model score rates for the original crash dataset with the model score for each of those 10 stratified random samples. Hence, the final model was evaluated 10 times by score the final model with each of those 10 stratified samples subsets. Then, the model accuracy prediction rates for each of those subsets were recorded and the average of those 10 accuracy rates was estimated.

Step 3: ***Final Model Accuracy Rate Assessment Performance***

To conclude, the accuracy rate derived from the model application to the original crashes sample was compared with the model accuracy rate derived from the model application within the 10 stratified samples. Each accuracy rate obtained for each subset was subtracted from the accuracy rate of the final model (with the full sample). This procedure allowed evaluating the stability of accurate prediction rate of the final model through the 10 subsets (10 stratified random samples).

The experimental approach designed to evaluate the goodness-of-fits with: training sample, original sample (OS) and finally, validation with 10 stratified random samples (SS), is outlined in Figure 3.4. Figure 3.4 illustrates the resampling K random cross-validation developed in this study for the assessment of models performance and validation for the two-vehicle collisions.

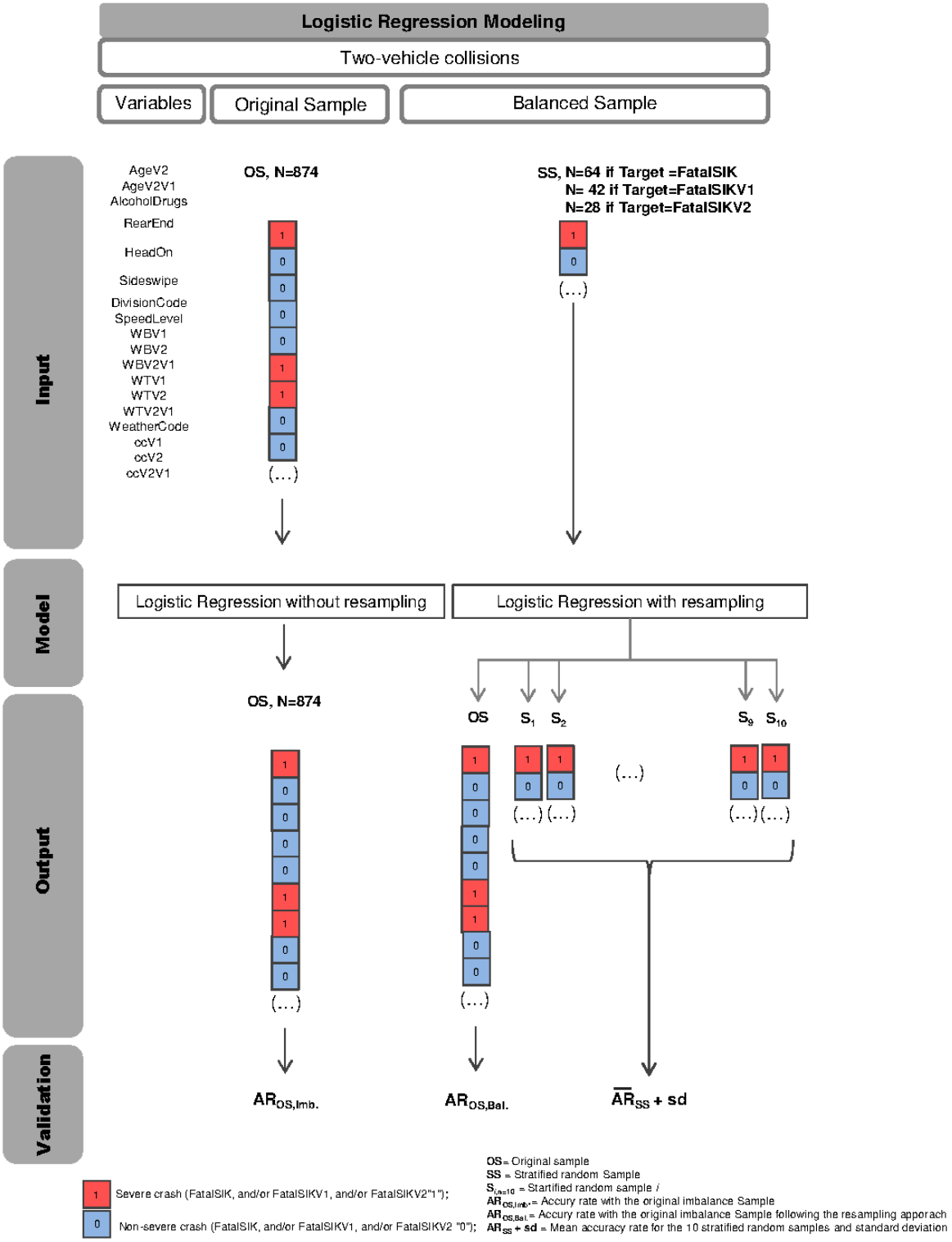


Figure 3.4 - Crash severity modeling using logistic with resampling strategy: training models assessment and validation for the two-vehicle collisions.

On the left top right of Figure 3.4, all the input variables are shown, as well as the original sample size. The “1”s squares in red illustrate the severe crashes which were less frequent than the non-severe-crashes, blue squares, in the original sample (OS) (N=874). Without the resampling approach, preliminary model training with the OS showed a poor fitting due to the high disproportion between target “1” and target “0”, bottom left side of Figure 3.4. Thus, a resampling approach, yield to training samples of equal proportion of sever crashes vs. non severe crashes (same proportion of red and blue square, on the top right of Figure 3.4. Subsequently, the model prediction accuracy was evaluated with the OS and then, validation was performed with the 10 stratified random samples S_{i+1} (N=42 and N=28 for FatalSIKV1 and fatalSIKV2 models assessment, respectively), on the bottom right of Figure 3.3. For example, the prediction accuracy rate for a selected model (developed with a balanced training ample) and then scored with the original sample is represented by “$AR_{,OS,Bal}$”, shown at the bottom of Figure 3.4. For FatalSIKV2 the structure would be the same, with category FatalSIK replaced by FatalSIKV1 or FatalSIKV2, depending on the target of interest. For the single-vehicle crashes the process is similar.

3.8 Concluding Remarks

The main constraint of the Portuguese crash sample was the limited number of observations (small sample size). In addition, a particular challenge was found when handling the imbalanced classes in the crash dataset, as result of the minority class of severe crashes in the sample. Due to the small dataset, data splitting would be inefficient, since the reduced sample size could reduce the fit of the model training and validation. The modest number of severe events (which were the target with interest for the modeling) generated an opportunity for a new modeling strategy: resampling and 10-fold cross validation procedure.

The safety analysis methodology presented in this chapter pursues the research goals as follows.

- Individual vehicle analysis to compare crash sample severity ratio with overall severity index for the national fleet.
- CART modeling to identify which variables are important to predict injury severity.
- Logistic regression modeling to evaluate the effect of vehicles attributes (risk factors) in injury severity prediction.

CHAPTER 4
CRASH DATA DESCRIPTIVE STATISTICS AND SEVERITY INDEX WITHIN THE PORTUGUESE FLEET

In this Chapter, initially descriptive statistics are presented for the crash sample with main focus on vehicles technical characteristics. Secondly, risk of exposure in the sample is presented based on injury severity and vehicle's engine size category. Thirdly, vehicle's individual brand analysis is discussed taking into account its involvement in crash severity outcomes. Then, brand's severity risk is compared with the overall severity within the Portuguese fleet. Main remarks are summarized.

4.1 Crash Data Descriptive Statistics

This section presents descriptive statistics for the crash sample comprising a total of 1,374 observations involving single-vehicle crashes and two-vehicle collisions. Whereas for single-vehicle crashes, the vehicles were defined as vehicle V1 only, for collisions, vehicles were register as vehicle V1 and vehicle V2, following the police record information (as explained in section 3.3). Descriptive statistics for All, Single and Two datasets are next.

4.1.1 General statistics

The crash sample revealed 22 crashes involving a drunk and/or drugged driver. Crash frequency distribution by road speed limit was as follows: 67.6%, 4.7%, 27.1% and 0.5, for 120 $km.h^{-1}$, 100 $km.h^{-1}$, 90 $km.h^{-1}$and 50 $km.h^{-1}$, respectively. Crashes registered at roads where the legal speed limit is the lowest, 50 $km.h^{-1}$, did not result in any severe case. On the other hand, crashes register in motorways, 120 $km.h^{-1}$, showed the highest percentage of severe observations, 3.4% (48/1374). This finding is consistent with previous studies that identified road speed as a key factor for crash severity risk [42, 43, 49, 91, 98].

The crash sample covers a total of 2,248 vehicles. The most frequent vehicle category was light passenger vehicles, which represented 74.3% of the vehicles, whereas light duty vehicles account for 25.7%. Diesel engines were the most common, corresponding to 58.9% of the analysed vehicles, following by the gasoline engines representing 40.7%. At a significant lower frequency: LPG (“GPL” at the Portuguese designation) and hybrid vehicles accounting only for 0.3% and 0.1%, respectively.

Regarding to vehicle technical characteristics, the mean values and its standard deviation (S.D.) for all the vehicles in the sample, vehicles’ weight, engine power, wheelbase and age were: 1238.1kg (S.D. 347.2), 1665.2 cm^3 (S.D. 504.4), 2591.9 mm (S.D. 270.2) and 8.5 yr (S.D. 5.1).

Relating to individual vehicles analysis, as V1 and as V2, descriptive statistics of continuous design variables with focus on vehicles characteristics is presented in Table 4.1. The oldest vehicle involved had 38 years, whereas the newest cars had one year, corresponding to 1972 and 2010 vehicle model year, respectively. Also, the heaviest vehicle in the crash dataset weighted six times more than the lighter passenger car, a 3500/584 weight ratio. Also, the largest vehicle’s wheelbase was almost three times larger than the smallest one, a 4325/1625 wheelbase ratio. Thus, results in Table 4.1 reflect a wide range of vehicles’ dimensions (weight, engine size and wheelbase) and vehicle model year (associated to vehicle’ age), as well. Therefore it is fundamental to take into account vehicle individual information for road safety analysis, since real crashes occur without any control among the vehicles categories and/or segments involved in the collision. In this Chapter, the statistics motivate the designed methodology to account for vehicle individual analysis, rather than the standard information, mainly restricted to vehicle type and vehicle model year, [43, 59-61,

85, 87, 93, 98]. Previous studies attempted to model overall crash severity without taking into account the effect of the opponent vehicle [43, 85, 86, 91, 93, 98, 102, 104]. Nonetheless, in multi-vehicle collisions, injury severity outcomes depend not only on the risk of the other vehicle involved, and also on the protective ability of the subject vehicle.

Table 4.1 - Descriptive statistics for vehicles selected variables in the crash dataset.

Symbol	N	Mean	S.D.	Minimum	Maximum
WTV1[1](Kg)	1374	1222.34	334.98	640	3200
ccV1[2](cm^3)	1374	1662.65	491.67	599	4104
WBV1[3](mm)	1374	2581.02	256.47	1625	4325
AgeV1[4](yr)	1374	8.48	5.06	1	25
WTV2[5](kg)	874	1262.85	364.46	584	3500
ccV2[6] (cm^3)	874	1700.94	522.18	698	4104
WBV2[7] (mm)	874	2609.00	289.88	1812	4100
AgeV2[8](yr)	874	8.54	5.26	1	38
WTV2V1[9](mm)	874	28.65	519.87	-2165	2860
ccV2V1[10](cm^3)	874	34.98	719.72	-2905	2909
WBV2V1[11](mm)	874	10.84	396.80	-2213	1918
AgeV2V1[12](yr)	874	<1	7.42	-20	28

1 Weight of Vehicle V1; **2** Engine size of Vehicle V1; **3** Wheelbase of Vehicle V1; **4** Age of vehicle V1; **5** Weight of Vehicle V2; **6** Engine size of Vehicle V2; **7** Wheelbase of Vehicle V2;**8** Age of vehicle V2; **9** Weight Differential between V2-V1, in two-vehicle collisions; **10** Engine size differential between V2-V1, in two-vehicle collisions; **11** Wheelbase Differential between V2-V1, in two-vehicle collisions; **12** Age Differential between V2-V1, in two-vehicle collisions.

At the crash reports, crash outcomes are classified in three injury levels: light injury (LI), serious injury (SI) and killed (K). Table 4.2 shows injury level distribution by number of vehicles involved and by vehicle recorded as V1 or V2, in the crash. Table 4.3 shows the frequency of severe observations expressed by the sum of serious injured and killed (SIK) by crash event.

Table 4.2 – Injury level distribution by vehicle position in the crash.

Datasets	Vehicle V1				Vehicle V2			Total			
	N[1]	LI	SI	K	LI	SI	K	LI[2]	SI[3]	K[4]	SIK[5]
Single	500	590	31	16	-	-	-	590	31	16	47
Two	874	643	14	9	732	16	2	1375	30	11	41
All	1374	1233	45	25	732	16	2	1965	61	27	88

1 Number of crashes observations; **2** Sum of light injuries; **3** Sum of serious injuries; **4** Sum of killed; **5** Sum of serious injured and/or killed.

Table 4.3 - Frequency of severe observations by number of severe injuries and/or killed and by vehicles involved**.**

Dataset	N[1]	SIK "1"[2]	SIK "2"[3]	SIK "3"[4]	Total SIK[5]
Single	500	31	5	2	38
Two	874	25	5	2	32
All	1374	56	10	4	70

1 Number of crashes observations; **2** Number of crashes having 1 occupant serious injured and/or killed; **3** Number of crashes having 2 occupants serious injured and/or killed; **4** Number of crashes having 3 occupants serious injured and/or killed; **5** Total Number of observation having a severe crash (either SI>0 and/or K>0)

Relating to crash severity risk of exposure in the sample, severe cases are presented based on vehicle involvement in single-vehicle crashes and vehicle involvement in two-vehicle collisions as V1 or as V2. Severe cases are related to an event that has resulted at least in a serious injured and/or killed among the occupants of the vehicle. For example in a severe collision, a severe injury can happen at one of the vehicle involved, or it can happen in both vehicles involved simultaneously. Table 4.4 shows the risk of exposure based on severe cases by the number of vehicles involvement and vehicle's age and engine size categories. Vehicle´s age was grouped by 5 categories: 1≤Age<5yr, 5≤Age<10yr, 10≤Age<15yr, 15≤Age<20yr and Age≥20yr. Engine size was grouped into three categories: c.c.<1400 cm^3, 1400≤c.c.<2000cm^3 and c.c.≥2000 cm^3. For single-vehicle crashes, the majority of vehicles fell in the engine size category c.c.<1400 cm^3, followed by the category 1400≤c.c.<2000 cm^3, with 219 vehicles involved in 13 severe crashes and 218 vehicles involved in 18 severe crashes, respectively, as shown in Table 4.4. Although the most frequent category was the vehicles in the small engine size category, it was in the middle engine size category that severe crashes were higher.

For two-vehicle collisions, vehicles V1 in the engine size category 1400≤c.c.<2000 cm^3, were the most frequent, with 390 vehicles involved in collisions that have resulted in 14 severe cases for the occupants of vehicle V1, Table 4.4. For vehicle V2, the most frequent engine size category was also 1400≤c.c.<2000 cm^3, 379 vehicles with three serve crash outcomes. However, for V2, the higher ratio of severe crashes was found for vehicles in the smaller engine size category, with 334 vehicles involved in collisions that had resulted in eight severe cases for its occupants. Appendix 7 provides information on Pearson correlation coefficients for all the variables in the crash dataset.

Table 4.4 – Crashes severe cases by: vehicles involvement in single-vehicle crashes or two-vehicle collisions, engine size and age categories.

Vehicle categories		Single-vehicle crashes		Two-vehicle collisions					
				As V1		As V2		V1+V2	
Engine size category	Age category	N	Severe Cases	N	Severe Cases	N	Severe Cases	N	Severe Cases
c.c.<1400 cm^3	1≤Age<5yr	63		60		76		199	
	5≤Age<10yr	77		121		99		297	
	10≤Age<15yr	48		93		98		239	
	15≤Age<20yr	27		67		50		144	
	Age≥20yr	4		5		11		20	
	Total	219	13	346	5	334	8	680	13
1400 cm^3≤c.c.<2000 cm^3	1≤Age<5yr	68		100		114		282	
	5≤Age<10yr	81		138		133		352	
	10≤Age<15yr	46		98		94		238	
	15≤Age<20yr	18		42		31		91	
	Age≥20yr	5		12		7		24	
	Total	218	18	390	14	379	3	769	17
c.c.≥ 2000 cm^3	1≤Age<5yr	12		40		37		89	
	5≤Age<10yr	32		52		66		150	
	10≤Age<15yr	16		33		35		84	
	15≤Age<20yr	3		10		17		30	
	Age≥20yr	0		3		6		9	
	Total	63	7	138	2	161	3	299	5
N		500	38	874	21	874	14	1,648	35

The overall crash severity was 5.1% (70/1374), reveling an unequal distribution of severe crashes compared to non-severe crashes, which were the most common events in the crash sample, showing a frequency of 94.9%. For the two-vehicle collisions, the ratio of the common event (non-severe crash) to the rare event (severe crash) was 26 (842/32). Thus, the non-severe crashes happened 26 times more frequently than the severe ones, yielding to an over represented of crashes with minor injuries. Therefore this crash data qualifies for imbalanced data. Whereas the percentage of severe crashes in two-vehicle collisions was 3.7% (32/874), for single-vehicle crashes the severity was 7.6% (38/500). Apart from unequal distribution of severe non-sever crashes, it is interesting to note that the overall severity was twice as higher for single-vehicle crashes than for the two vehicles crashes.

This disproportion between non-severe crashes and severe crashes imposed a challenge during the crash severity prediction. Next Chapter presents the approach designed in this research to handle imbalanced data.

4.1.2 Single-vehicle crashes descriptive statistics

In the Single dataset, the percentage of crashes involving drunk and/or intoxicated drivers was 2.0% (10/500). From those, three crashes that involved drunk and/or intoxicated drivers resulted in severe outcomes. As far as crashes distribution by road class speed limit, the frequency was: 375, 29, 95, and 1, for 120 $km.h^{-1}$, 100 $km.h^{-1}$, 90 $km.h^{-1}$and 50 $km.h^{-1}$, respectively. The roads that appeared more often were motorways: A4, A28, A3 and A29, with the frequency: 96, 86, 55, 55, respectively. A map with the identification of these roads was previously highlighted in Figure 3.2b). Crashes type distribution was as follows: 333 ran off road and 67 rollovers. The mean values for vehicle V1 technical characteristics were as follows. 1201.6Kg (S.D. 292.1), wheelbase of vehicle V1 was 2551.1mm (S.D. 205.0), for weight, engine size, and wheelbase, respectively. The mean vehicle's age was 7.8yr (S.D. 4.9).

Histograms are presented in Figure 4.1 to illustrated vehicles technical characteristics (independent variables) frequency distribution with crash severity (dependent variable). As shown by the histogram a), in Figure 4.1, the category 5≤AgeV1<10 is the most frequent and severe crashes were also more frequent for this category. For the majority of the vehicle's involved in single-vehicle crashes, had engine sizes in the categories ccV1<1400 cm^3 and 1400≤ccV1<2000 cm^3, which were also linked to more severe outcomes, histogram b), in Figure 4.1. Vehicles in the weight category, 1000≤WTV1≤1499 kg, were clearly the most frequent and also showed higher number of severe crashes, histogram c), in Figure 4.1. The two most frequent categories for vehicle's wheelbase were: 2000≤WBV1≤2499 and 2500≤WBV1≤2999 and with a higher number of crashes resulting in severe outcomes as well, histogram d), in Figure 4.1.

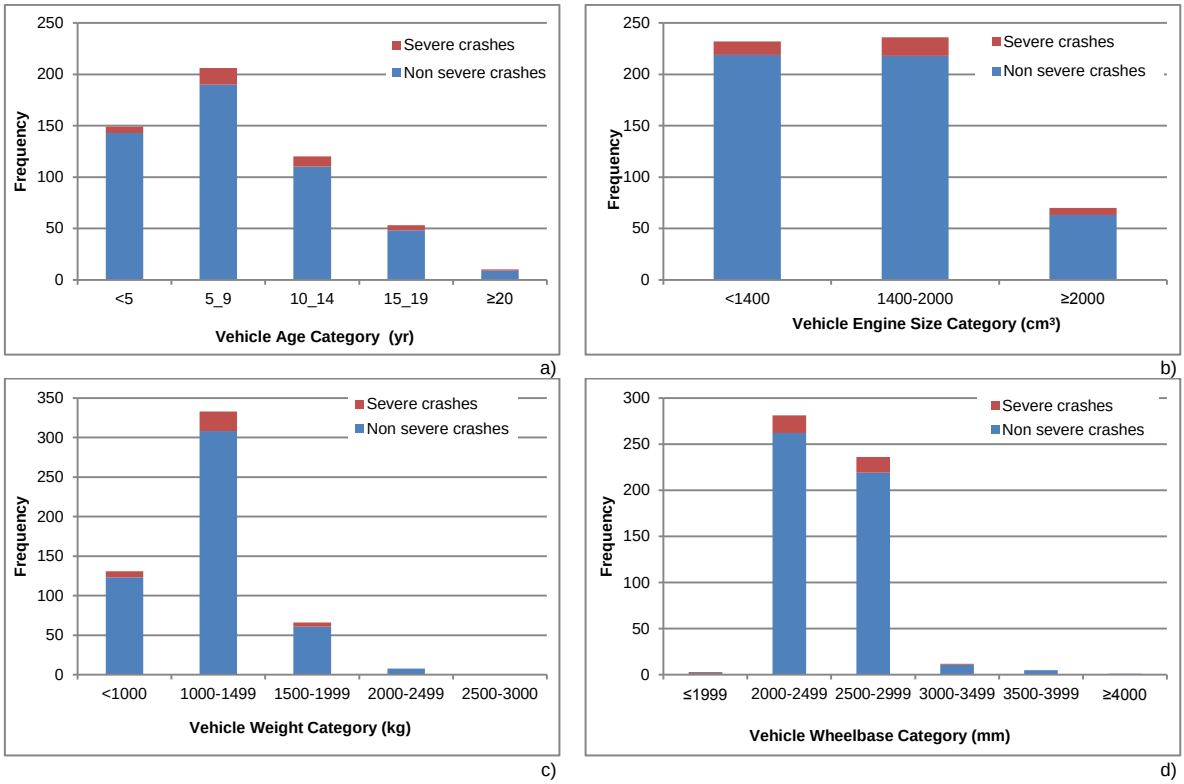


Figure 4.1 - Frequency distribution of vehicles' characteristics with crash severity, in single-vehicle crashes: a) AgeV1 category; b) ccV1 category; c) WTV1 category; d) WBV1 category.

4.1.2 Two-vehicle collisions descriptive statistics

In the Two dataset, the percentage of crashes involving drunk and/or intoxicated drivers was 1.4% (12/874), and three of them have resulted in severe collisions. The roads with higher frequency of collisions involving any type of injuries where: A4, A28, A3 and EN15, with 121, 112, 88 and 88 counts, respectively. A map with the identification of these roads was previously shown in Figure 3.2b). Regarding to the frequency of collisions by road class speed limits, the distribution was: 6, 278, 36 and 554, for 50km.h^{-1}, 90 km.h^{-1}, 100 km.h^{-1} and 120 km.h^{-1}, respectively. Crashes distribution by collision type was as follows: 311, 89, 67, and 407, for rear end, sideswipe, head on and others, respectively. The mean values for vehicles V1 and V2 weight were as follows: 1234.2 Kg (S.D. 356.8) and 1262.9 Kg (S.D. 364.5), respectively. The mean engine size for vehicle V1 and V2 was: 1665.0 cm^3 (S.D. 510.0) and 1700.9 cm^3 (S.D. 522.2), respectively. The mean wheelbase for vehicle V1 and V2, was: 2598.2 mm (S.D. 280.4) and 2609.0 mm (S.D. 289.9), respectively. The mean vehicle V1's age was 8.9 yr (S.D. 5.1), whereas, the mean vehicle V2's age was 8.5 yr (S.D. 5.3).

Comparison between Single and Two datasets, with 500 vehicles and 1,784 vehicles, respectively, is summarized next. The mean vehicles weight was 1248.5 Kg (S.D. 360.0) and 1201.6 Kg (S.D. 292.1), for Two and Single datasets, respectively. The mean engine size was 1683.5 cm^3 (S.D.

516.3) and 1601.6 cm^3 (S.D. 455.5), for Two and Single, respectively. The mean wheelbase was 2603.6 mm (S.D. 285.1) and 2551.1 mm (S.D. 205. 0.4), for Two and Single, respectively. The mean vehicles' age was 8.7 yr (S.D. 5.1) and 7.8 yr (S.D. 4.9). Despite of the difference in the number of observations for those datasets, it was noticeable that in average, vehicles involved in single-vehicles crashes were slightly lighter, with smaller engine size and smaller wheelbase, and almost one year younger than the vehicles involved in collisions.

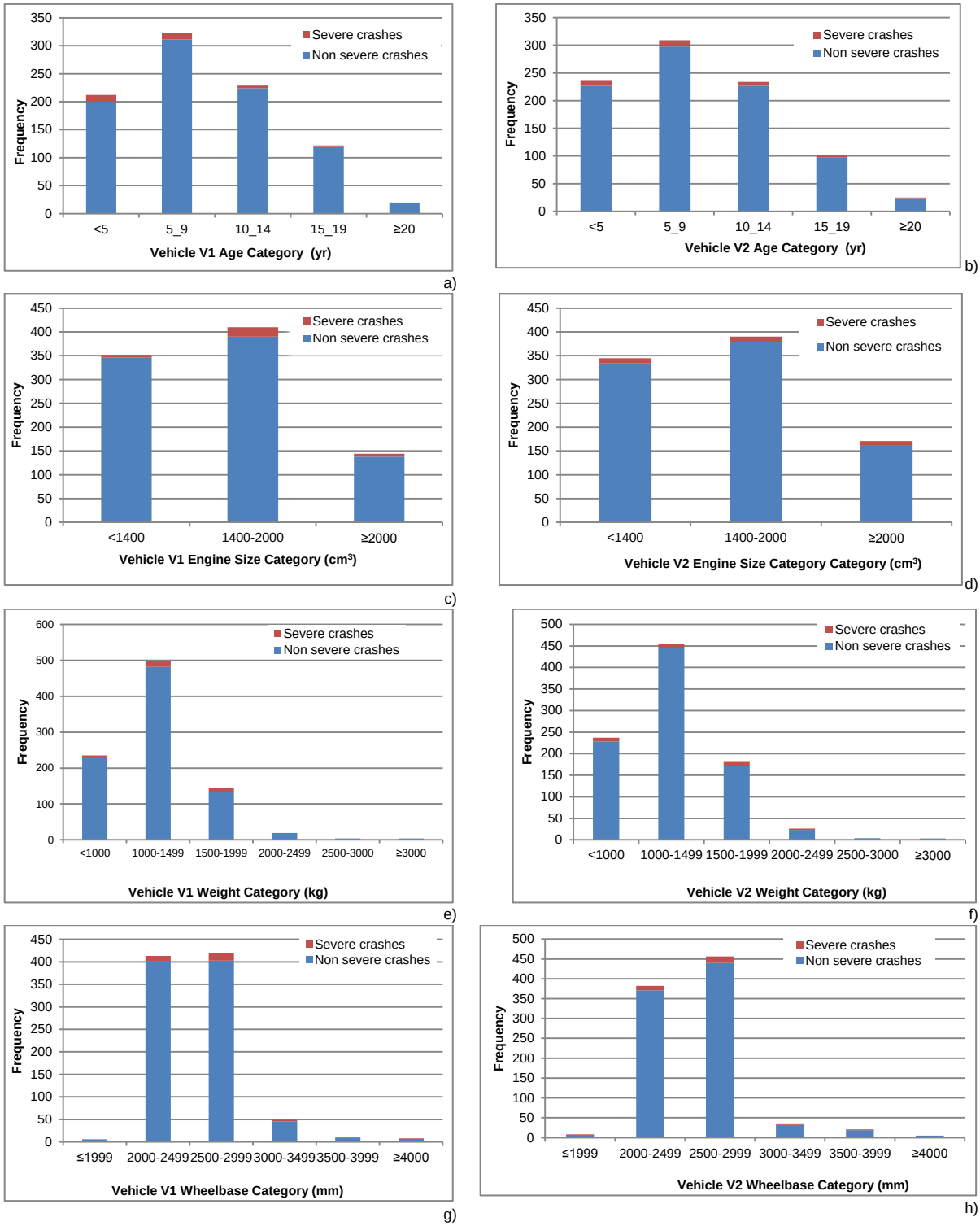


Figure 4.2 - Frequency distribution of vehicles' characteristics with crash severity, in two-vehicle collisions: a) AgeV1 category; b) AgeV2 category; c) ccV1 category; c) ccV2 category; d) WTV1 category; d) WTV1 category; e) WTV2 category; f) WBV1 category and f) WBV2 category.

4.2 Inference of Auto Brands in the Sample with the Portuguese Fleet

The vehicle's make individual analysis gives attention to the vehicle's auto brand distribution in the sample and the severity index for the crashes involving that specific brand being analyzed. First, it compares auto brand severity index with the sample severity index. Second, it compares the auto brand severity index with the overall severity at national level.

4.2.1 Vehicles brand severity ratio analysis in single and two-vehicle collisions and within the Portuguese fleet

The single-vehicle crashes included 500 vehicles representing 35 auto brands. Using the crash outcomes in Table 4.2 the crash severity index was 7.4% (47/639). The national level, road safety data for the single-vehicle crashes involving light vehicles only and during the period 2006 to 2008 showed the following injury distribution: 9,451 light injured, and 889 serious injured and killed, leading to an overall severity index of 8.6% (889/10341).

Table 4.5 shows the auto brands with the highest frequency at the crash sample. The brands involved in the single-vehicle crashes with higher frequency were: Renault (15.8%), Opel (9.2%), and Fiat (8.8%). Crashes involving a Renault had resulted in an increment of almost 1% in the severity ratio when compared to the overall severity at the sample: BSR for Renault was 8.3%, whereas the overall severity index at the crash sample was and 7.4%. However, when comparing this vehicle brand severity ratio with the overall severity index, it was slightly lower, 8.3% and 8.6%, for Renaults' BSR and Portuguese fleet, respectively. Based on the crash sample, Renault vehicles could be linked to lower lower protectiveness to its occupants since the severity index was 0.9% higher compared to the sample index. However, when Renaults' BSR is compared with OSI, it was 0.6% lower, thus suggesting that this brand provides better protection to its vehicle's occupant's than the average brand involved in the same crash type at national level.

Table 4.5 – Vehicle's brand severity ratio analysis across the crash sample for two-vehicle collisions and single-vehicle crashes

Vehicle Analysis by Brands	Frequency	LI[1]	SI[2]	K[3]	BSR[4]
Two-vehicle collisions					
Renault	14.7%	218	8	3	4.8%
Opel	10.8%	160	2	3	3.0%
Volkswagen	7.3%	105	0	0	0%
Single-vehicle crashes					
Renault	15.8%	99	6	3	8.3%
Opel	9.2%	53	1	2	5.36%
Fiat	8.8%	61	0	0	0%

[1] Number of light injured at vehicle's auto brand; [2] Number of serious injured at vehicle's auto brand; [3] killed at vehicle's auto brand; [4] brand severity ratio for the vehicle's brand.

The two-vehicle collisions sample in this study included 1,748 vehicles representing 41 auto brands. The crash outcomes for those collisions by injury level were as follows: 1,375 light injured, and 41 serious injured and killed, leading to a crash severity index of 2.9%, (41/1416). On the other hand, the overall severity index for the two-vehicle collisions involving light vehicles in the Portuguese fleet was 4.8%.

Table 4.5 shows the brands with the highest frequency for the two-vehicle collisions sample: Renault (14.7%), Opel (10.8%), and Volkswagen (7.3%). The two-vehicle collisions involving a Renault had resulted almost twice in the severity ratio for the overall crash sample, 4.8%, and 2.9% respectively. However this finding could not be used o drive a conclusion that the Renault brand showed a poor crashworthiness performance in general at the Portuguese roads. In fact, Renault's BSR when compared with the OSI (for the same type of crashes) showed the same severity ratio, 4.8%.

4.2.2 Expanding brand severity ratio analysis within the Portuguese fleet

Expanding the analysis of vehicles brand severity ratio with Portuguese overall severity index required an evaluation of those brands representativeness across the Portuguese fleet. For instance, if a brand has BSR higher than the OSI and their vehicles sales are low in the Portuguese fleet, it would suggest that probably the brands models would offer a poor crashworthiness. On the other hand, if a brand had a lower BSR than the OSI, and its vehicles sales are high in the nation; this brand could reflect good crashworthiness across the fleet. Based on the brands annual sales, each top brand identified earlier were normalized by the total number of light passenger vehicles and light duty vehicles registered at the annual calendar year, using data was provided by ACAP [134, 135].

In the case of Renault, it was the most common brand in the sample, this vehicles' brand were also the most exposure in the sample, hence increasing the risk of crash involvement. Therefore, it was also important to consider the share of Renault vehicles in the Portuguese fleet. This brand is in the top sales in Portugal, and, across the Portuguese fleet it would be expected more vehicles register under the Renault brand as in fact it is, as illustrated in Figure 4.3. BSR information does not support the statement about Renault vehicles crashworthiness because it is the most sale carmaker in Portugal; Renault vehicles have a higher probability to be involved in a crash because they are also more frequent at the fleet. In addition, the analysis presented in this study is limited to an analysis of average brand severity ratio, and different models of the same brand may perform differently.

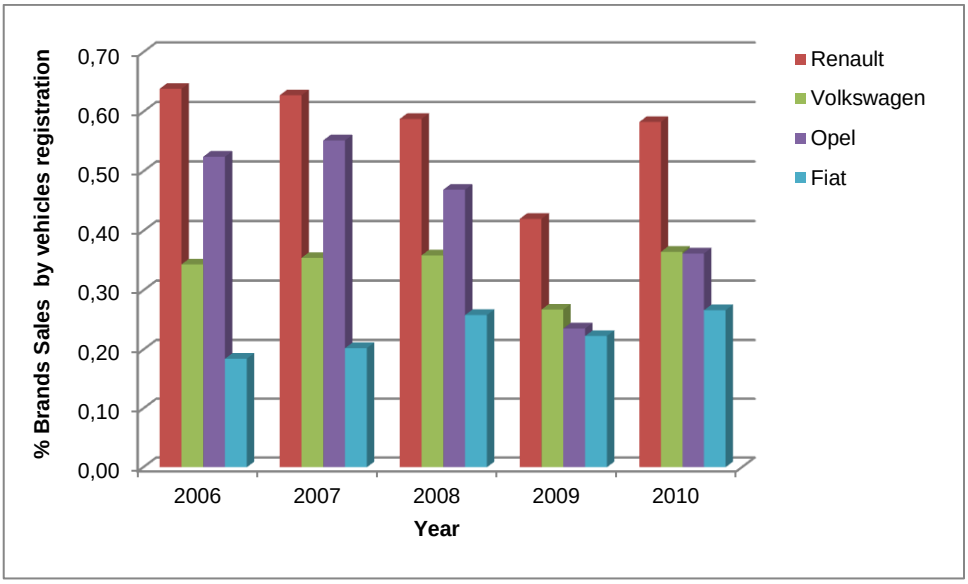


Figure 4.3 – Vehicle brand sales by the total vehicle, within the period 2006 to 2010.

However, it was interesting to notice that Volkswagen and Fiat vehicles, even though were found among the most popular brands in the Two and Single datasets, these brands crash involvement did not result in any severe consequences, since number of serious injured and/or killed was zero. The inference of these brands with the national fleet, also revealed that they are between the most representative in the vehicle fleet, in Figure 4.3. Despite of Volkswagen and Fiat high frequency in the fleet, its crash risk was smaller than for Renault. Based on the crash sample used in this study, the number of observations involving those vehicles was small to established further conclusions. Nevertheless, these differences in brand severity ratio among the most common brands are consistent with other study, which found Ford and Toyota as the most popular brands in Florida [143]. Even though the risk of exposure was the same for both brands, Ford showed better self-protective ability than Toyota [143].

4.3 Concluding Remarks

The results presented in this Chapter showed descriptive statistics of continuous design variables and vehicles characteristics, for the vehicles involved in two-vehicle collisions and single-vehicle crashes. The average weight and size of vehicles involved in the two-vehicle collisions was slightly larger than the average for the vehicles involved in single-vehicle crash. Regarding to crash outcomes, the overall crash severity was 5.1%, with high disproportion between non-severe crashes and few observations of severe crashes. Thus the crash sample qualifies for imbalanced data.

In this chapter, the most remarkable finding was related to the crash sample severity (either a serious injured and/or killed) distribution for the single-vehicle crashes and two-vehicle collisions: 7.6% and 3.7%, respectively. These findings are consistent with previous work which had stated that in crashes involving one car, the vehicle crashworthiness may be offset by the driver behavior that could be speeding, and thus increasing the risk of serious crash outcomes [49, 61, 99]. In addition, inference of sample severity index with the Portuguese overall severity index (serious injured and killed by the total number of injuries and killed) shows consistent values. At national level, for crashes involving one vehicle, the severity index was 7.4% and 8.6%, for the single-vehicle crash sample and population, respectively. For the crashes involving two vehicles, the severity index was 2.9% and 4.8%, for the two-vehicle collisions sample and population, respectively.

Regarding vehicles brand analysis, the most frequent brands were: Renault, Opel, Volkswagen and Fiat, with Renault showing the highest severity ratio. On the other hand, Volkswagen and Fiat, although among the most frequent brands, did not show any involvement in severe crashes. However, the inference of this brand with the Portuguese fleet showed that Renault's severity ratio was similar to the National overall severity index. For single-vehicle crashes, the brand severity ratio was 8.3% and the national crash severity ratio was 8.6. Furthermore, Renault brand has been in the top sales during the time period covered in this analysis, thus increasing the risk of exposure. It must be pointed out that severity risk reported in this vehicles' brand severity ratio analysis does not account for the total number of occupants in the vehicle, neither for the differences in annual kilometers driven, nor driver age or gender. In conclusion, the brands severity ratio inference analysis must be approached with care and always attending to the brands representativeness within the national fleet.

CHAPTER 5

DECISION CLASSIFICATION TREES ANALYSIS FOR CRASH SEVERITY PREDICTION

One benefit of decision tree compared to other modeling techniques is that these models provide decisions by making the answer “if-then” questions efficiently [104]. Researcher and traffic engineering can easily predict the injury likelihood of an accident simply by determining the value of splitters and tracing a path down the tree to a terminal node. The trees not only give the variables of importance, but also help to better interpret the results. The targets being predicted by CART models were: FatalSIK, FatalSIKV1 and FatalSIKV2, all of them having a categorical measurement level, “1” or “0” and therefore, the type of the prediction is a decision: severe or non-severe crash.

This Chapter is organized as follows. First, CART models are presented for all crashes, two-vehicle collisions and single-vehicle crashes, based on the original sample. Second, following a resampling procedure, CART models are presented for all crashes, two-vehicle collisions and single-vehicle based on balanced datasets. Third, CART models targeting individual vehicle injury severity classification are shown for the original sample distribution of two-vehicle collisions. Remarkable findings of crash severity analysis with CART methodology are also highlighted.

5.1 CART Analysis for FatalSIK with the Original Crash Sample-Imbalanced Datasets

In the beginning of this section, the decision tree is presented in a way that it will help to interpret the following trees in this Chapter. The trees' grow reflects a hierarchical group of relationships. Each branch is split further using the classes or categories of the other predictor variables. This process, known as recursive partitioning, continues until a stopping rule is satisfied, such as the minimum number of cases in the terminal leaf (5 counts). It must be noted that the root node split for each tree structure shows a branch that is highlighted bold, which shows the split with the larger number of cases. One of the two connecting lines showing the predictor split also displays the term "*missing*" for one of the categories. However it must be clear that this term appears by default at the CART diagrams. Fortunately in this study, there was no missing data, since all the predictor variable values were available for all the observations in the crash database. Also, the leaves' Node ID do not show an organized order. However, Node ID do reflect a decreasing order from the root node (which is always identified as Node ID:1). Each leaf/node contains information about the number of cases in the particular leaf, denoted by "count" term in the node. CART methodology was applied using SAS®v9.2 and SAS®Enterprise Miner™7.1 (EM7.1) software.

Following, the decision trees are discussed as prediction models for the crash severity target with interest for each dataset: All, Two and Single (as defined in section 3.3.). Figure 5.1 to Figure 5.8 show the decision trees models for the binary classification for crash severity.
In this section, section 5.1, CART results are discussed for the original sample distribution, which means that the proportion of the severe crashes (FatalSIK"1") vs. the non-severe crashes (FatalSIK"0") was kept the same as the original sample.

.

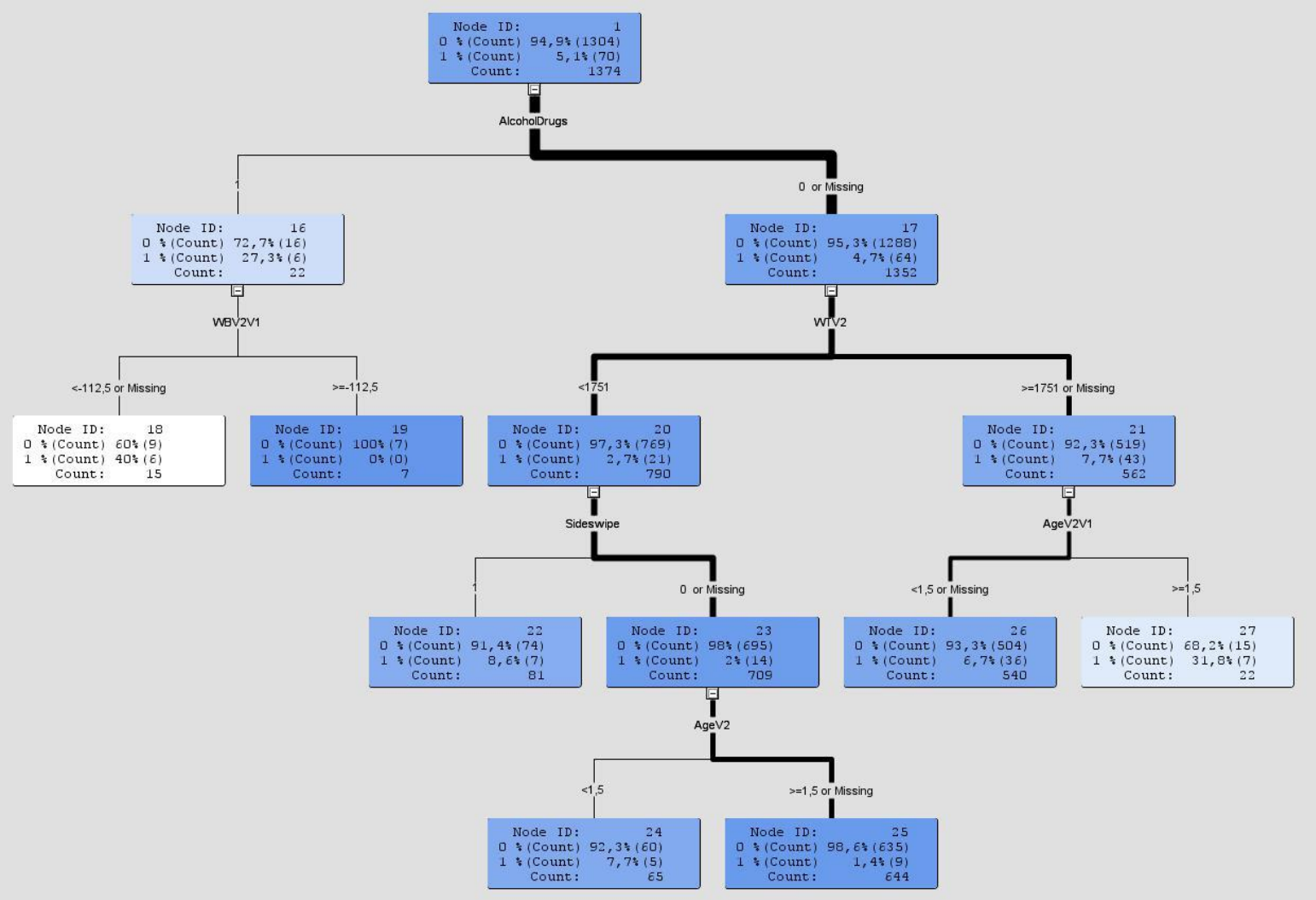


Figure 5.1 – Classification tree model for FatalSIK with all crashes using the original imbalanced sample.

5.1.1 CART for FatalSIK with all crashes- Imbalanced dataset

The original distribution of severe crashes in the crash sample was as follows: 5.1% of severe cases (corresponding to target FatalSIK“1”) and 94.9% of non-severe cases (corresponding to target FatalSIK”0”). Figure 5.1 shows the output of the CART prediction for FatalSIK using all crashes, including single-vehicle crashes and two-vehicle collisions. Twenty three independent variables (predictors) and one dependent variable (target) were defined for CART modeling, and these variables used as inputs and the target being modeled were identified in Table 3.5, Chapter 3.

The first selected variable for the decision tree split was alcohol and/or drugs, with the category for illegal drivers (alcohol or drugs use) associated with higher percentage of severe crashes, 27.3% of FatalSIK”1”. This node, node ID 16, was split by the differential of wheelbase between the vehicles involved in the collision, WBV2V1. As previously mentioned, crashes involving illegal driving (AlcoholDrugs “1”), and involving vehicles in the category WBV2V1 < -112.5 mm, were associated to the highest percentage (40%) of FatalSIK “1”. [87, 92, 93]. This decision tree model shows that alcohol and or drugs use plays a major role in increasing severity risk of crashes, despite of vehicle crashworthiness or collision type, and is in agreement previous research [92].

On the right branch of the tree, the category of crashes involving legal drivers (no alcohol or drugs use) was split by the weight of vehicle V2, WTV2. Then, crashes for legal drivers, with heavier vehicle V2 category, WTV2≥1751 kg, and involving a lower age differential (AgeV2V1<1.5 yr) showed the highest count (36) of crashes involving severe injuries or killed, node ID 26. This node showed 6.7% of severe crashes, which was higher than the overall rate at the crash sample, 5.1%. However, for the category AgeV2V1≥1.5yr the percentage of FatalSIK”1” was higher, 31.8%. The category of crashes involving a lighter vehicle V2, WTV2<1751kg, crash type other than sideswipe, were split by the age of vehicle V2. The category of newer vehicles V2, AgeV2<1.5yr presented higher percentage of FatalSIK”1”, node ID 24. A lower percentage of severity for cases involving newer vehicles models would be expected, but it must be noticed that the severe injured and killed were more frequent among occupants of vehicle V1; SIK distribution was as follows: 79.5% (70/88) and 20.5% (13/88), vehicle V1 and V2 respectively. Mendez et al. claimed that newer vehicle models have increased “agressivity”. Thus, it is possible that newer vehicle V2 models imposed more risk for occupants of V1. Whereas in two-vehicle collisions involving older V2 models, the impact on the compartment area of V1 could be less intrusive, leading to lower risk of severe injured. Thus, the risk imposed by newer vs. older V2 models could be a possible explanation for the differential concentration of severe crashes at the terminal nodes: node ID 24, and node ID 25, 7.7% and 5.5, respectively.

To assess the classification decision tree model for FatalSIK with all crashes, the Fisher’s exact test was conducted once some categories had less than five counts (node ID 19 showed zero

cases of target level “1” as observed in Figure 5.1). The p-value<$1.267E^{-12}$ denotes the significance level at which the terminal nodes affect the binary target being predicted, crash severity expressed by FatalSIK.

In addition to the graphical display, CART technique also provides information on the variable importance for all the variables in the decision tree model. The variables importance score indicates whether the presence or absence of a variable in the model (decision tree) will improve or degrade the efficiency of the model. For the FatalSIK decision tree model with all the crashes from the original sample, the variables relative importance score is as follows: AgeV1V2 (1), AlcoholDrugs (0.91), WTV2 (0.78), WBV2V1 (0.76), Sideswipe (0.49) and finally, AgeV2 (0.42). The most scored effects were the age difference between the two vehicles involved (AgeV1V2) and the effect of alcohol and/or drugs. The effect of AgeV1V2 can be explained when the vehicles involved in the collision differ by model year, it means that the vehicles structure may be different, and the safety equipment will also differ as well. It would be expected that newer vehicles models would be equipped with better safety equipment’s, hence providing a better protection to its occupants. These findings are coherent with Das, whose work found the use of alcohol and/or drugs use as the most important variable [87].

5.1.2 CART for FatalSIK with two-vehicle collisions- Imbalanced dataset

Figure 5.2 shows CART output for FatalSIK prediction using two-vehicle collisions. The original distribution of the Two dataset was as follows: 3.7% of severe cases (FatalSIK“1”) and 96.3% of non-severe cases (FatalSIK”0”). Twenty independent variables (predictors) and one dependent variable (FatalSIK) were defined for CART modeling. These variables used as inputs and the target being modeled are identified in Table 3.5, Chapter 3.

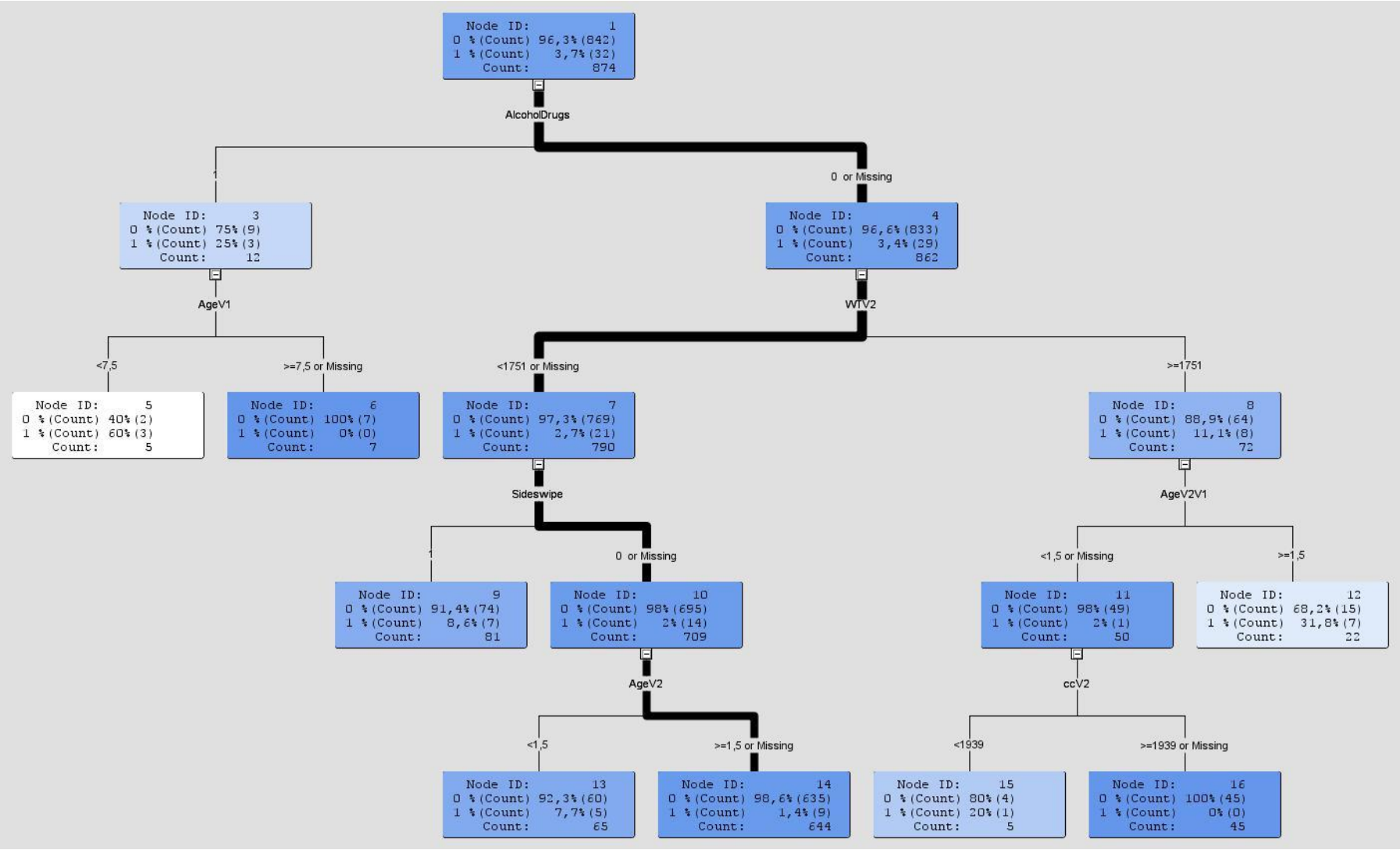


Figure 5.2 - Classification tree model for FatalSIK with two-vehicle crashes using the original imbalanced sample.

Figure 5.2 shows that the effect of alcohol and/or drugs was the first explanatory variable selected by CART methodology to split the 874 two-vehicle collisions. As shown in Figure 5.1, alcohol and/or drugs was also selected for the split of the original dataset containing all the crashes.

Crashes in which drivers were sober and involving a heavier category of vehicle V2, WTV2≥ 1751 kg, combined with the category of higher age differential, AgeV2V1≥1.5 yr, had a high concentration of severe crashes, 31.8% of FatalSIK"1". For the category AgevV2V1< 1.5 yr, severe cases were much less frequent, 2%. This finding suggests that the collision that involves vehicles of different ages, vehicles' crashworthiness and "agressitivy" performance also will be different. Newer vehicles models are better equipped with safety features, offering better protection to its occupants, but on the other hand, they may also imposed a higher risk for the towards the occupants of the other vehicle involved. This finding is consistent with [49, 61] that found increasing risk imposed by newer models. For collisions involving a lighter category of vehicle V2, WTV2< 1751 kg, the percent of severe crashes was lower (2.7%) than when V2 belonged to a heavier category (11.1%), as observed at nodes ID 7 and 8, respectively. This fining is consistent with previous research that found for collisions involving two cars of different masses, the fatality risk ratio of the heavier to the lighter car increases as a power function of mass ratio [53, 64]. Following, the node ID 7 was split by sideswipe crash type. The sideswipe collisions resulted in a higher concentration of severe crashes than a non-sideswipe collision, 8.6% and 2%, in nodes ID 9 and 10, respectively. This finding is consistent with other research, that found sideswipe impacts as the most serious crashes and substantially more likely to result in serious injury [43, 48, 144]. For the non-sideswipe crashes, the tree split by the age of vehicle V2, leading to a higher concentration of severe crashes (7.7%) when AgeV2< 1.5 yr, compared to 1.4% of severe collisions when category AgeV2≥1.5 yr was involved. This finding is consistent with Bédard et al. results that indicated an increased risk of fatalities [91]. Others, claimed that recent models are safer [57]. Newer vehicle models definably they offer better protection to its occupants, and when the other vehicle involved is an older model, probably its occupants face a greater risk. Thus, caution must be present in the interpretation of this finding because discrepancies between previous studies are likely explained by adopted methodology, variables use and samples.

Turning to the right side of the tree, for the collisions involving driving under the influence of alcohol and/or drugs, and a vehicle V1 newer than 7.5 yr, the risk of severe crash outcome was the highest, 60% for FatalSIK"1". On the other hand, collisions involving vehicle V1 with more than 7 yr only showed non-severe crashes, 100% for FatalSIK"0", node ID 6. As previously explained for the effect of AgeV2, this finding could seems counterintuitive since it would be expected that in general newer vehicles models show better safety performance than older models. One possible explanation would be that younger drivers usually underestimate the risk associated with alcohol and/or drugs use and driving faster. Kockelman and Kweon have stated that "young drivers involved in single-vehicle crashes are driving much more recklessly than middle-age drivers, leading to sufficiently more severe crashes that benefits of youth are outweighed by crash severity"

[92]. If this statement would be proven, it could be extended to two-vehicle collisions because younger driver keeps in a centrally way the same driving profile. Also, Kuhnert et al. classification tree model have identified drivers younger than 27 yr as the age group associated with higher concentration of severe crashes.

Since some of the categories had less than five counts, Fisher's exact test was performed for the eight terminal leafs showing a p-value $<6.516E^{-10}$. At the 5% significance level, the target FatalSIK and the above categories related to the tree terminal leafs cannot be considered independent.

For the classification tree model for FatalSIK with two vehicle-collisions with the original imbalanced dataset, the variables that have a major importance in predicting this target are as follows: AgeV2V1 (1), AgeV1 (0.87), AlcoholDrugs (0.64), WTV2 (0.59), Sideswipe (0.48), AgeV2 (0.41) and last, ccV2 (0.36). Similarly to the classification tree model for FatalSIK with all crashes, vehicles age differential, AgeV2V1, was the most important variable for the model. These results are consistent with Kockelman and Kweon that found vehicle´s age significant to predict crash severity for two-vehicle collisions [92]. As mentioned in the previous subsection, alcohol and drugs use have been identified as important factors related to increasing severity by several authors [87, 92, 137, 144]. Also, vehicles weight it is known as significant factor not only to address risk to occupants of vehicle, but also it affects the risk to the occupants of the opponent vehicle [53, 59, 64, 87].

5.1.3 CART for FatalSIK with single-vehicle crashes- Imbalanced dataset

This CART model to predict the target FatalSIK for single-vehicle crashes is discussed in this section. The original distribution of the Single dataset was as follows: 7.6% of severe cases (FatalSIK"1") and 92.4% of non-severe cases (FatalSIK"0"). CART output for this model is presented in Figure 5.3. Ten independent variables (predictors) and one dependent variable (target) were defined for CART modeling, and these variables used as inputs and target are identified in Table 3.5, Chapter 3.

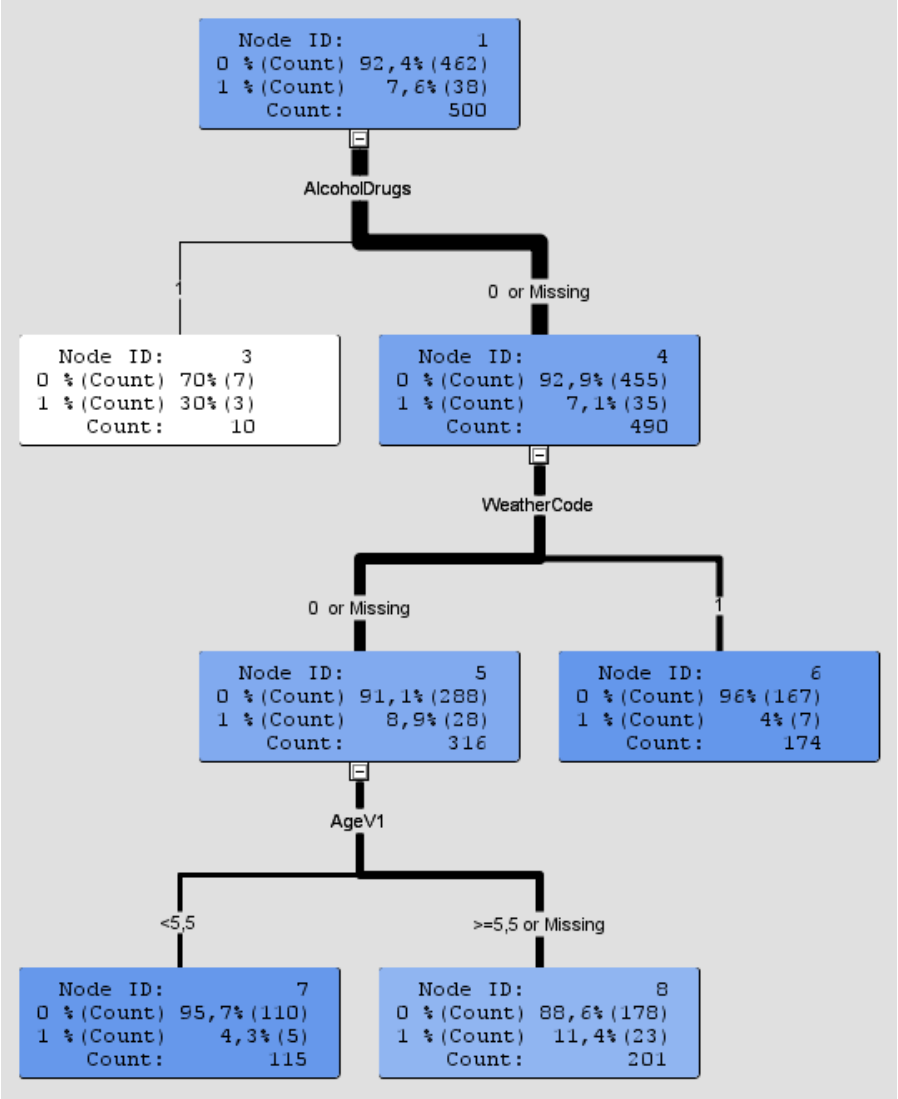


Figure 5.3 – Classification tree model for FatalSIK within single-vehicle crashes using an imbalanced sample.

Similar to what was found for the previous CART modes presented in sections 5.1.1 and 5.1.2, the initial split of node ID 1 was based on the alcohol and/or drugs use, and consistent with previous work [87, 137]. Subsequently, crashes involving illegal drivers, resulted in the highest percentage for severe crashes, 30% for FatalSIK"1", in node ID 3. These three severe crashes have already been analyzed in Figure 5.3. CART output for FatalSIK with single-vehicle crashes revealed that the presence of alcohol and/or drugs itself was linked to a higher crash severity, despite of vehicle characteristics. However only three severe cases were observed in node ID 3, hence caution must be presented in the previous statement. Whereas for the crashes where the effect of alcohol and/or drugs was not involved, the percentage of severe cases was lower, 7.1% in node ID 4. Subsequently, this node was split by the weather code, and the trees branch taking the value of 1 (meaning "bad" weather conditions) lead to a terminal node with lower percentage of severe

crashes (4%) compared to the good weather conditions (8.9%), nodes ID 6 and 5, respectively. This could seem counterintuitive since under bad weather conditions (due to rain, smog, and ice) crashes frequency is expected to increase because vehicles require longer distances to break. However, the higher proportion of severe crashes for good weather conditions is consistent with previous classification models [87, 104]. “Drivers could be less attentive when driving in good weather and road conditions” [87]. Then, node ID 5 was split by the age effect of the vehicle, predominantly recent models (AgeV1<5.5 yr) and older models (AgeV1≥5.5 yr.). It is interesting to notice that sober drivers, under good weather conditions and driving an older vehicle, (with 5.5 yr or more), showed the highest number of severe crashes (23 counts), node ID 8. On the other hand, keep the same conditions constant (no alcohol and/or drugs and good weather), when driving a vehicle model newer than 5.5 yr, the number of crashes resulted in severe consequences was smaller, 5 cases, terminal node ID7.

To test the association between the four terminal categories of the tree model discussed above and the target FatalSIK, Fisher’s exact test was used showing p-value<0.002. Thus, the null hypothesis is rejected and FatalSIK and its association with the presented categories of the tree terminal leafs cannot be considered independent.

Regarding to the variables importance for the classification tree model for severity prediction in single-vehicle crashes with the original dataset, the variables that have a major importance in predicting this target FatalSIK are as follows: AlcoholDrugs (1), AgeV1 (0.85) and WeatherCode (0.72). These findings are consistent with other researchers. The importance of alcohol and/or drugs in increasing severity is consistent with other studies [87, 92, 137, 144]. On the other hand, the importance of vehicle´s age and weather conditions has been also indicated by other research [57, 87].

5.2 CART Analysis for the FatalSIK with Resampling Approach

This section presents the CART analysis results for crash severity prediction using a balancing approach, leading to equal distribution between target levels. The resampling approach was applied to CART modeling more as an academic interest. Each crash dataset (All, Two and Single), had been balanced in order to include equal proportion of severe crashes and non-severe crashes. As previously explained in Chapter 3, the bias introduced by the resampling approach was correct by adjusting the prior probabilities within the crash subsets. As is going to be noticed in the graphical representation through the decision trees discussed in this section, the initial root node will reflect the original crash sample distribution, where the severe crashes were found at much lower proportion than the severe crashes. The predictor variables used in these models are the same used when modeling FatalSIK with the original crash sample, and those inputs are identified in Table 3.5, Chapter 3.

5.2.1 CART for FatalSIK with all crashes- Balanced dataset

The classification tree model for crash severity for all crashes presented in this section is presented in Figure 5.4. During the modeling phase, a resampling procedure was applied to the original crash sample, leading to a balanced dataset with equal proportion of target level “1” (70 counts for severe crashes) and target level “0” (70 counts for severe crashes), resulting in a total of 140 observations, as observed in the root node of the tree, node ID 1. As mentioned above, the decisions predicted with this tree model were corrected for the original sample distribution. Thus, seven counts in the root node, denoted the original 5.1% of severe cases (FatalSIK “1”) in the original crash sample, and the remaining 133 represented the original 94.9% for non-severe cases (FatalSIK”0”).

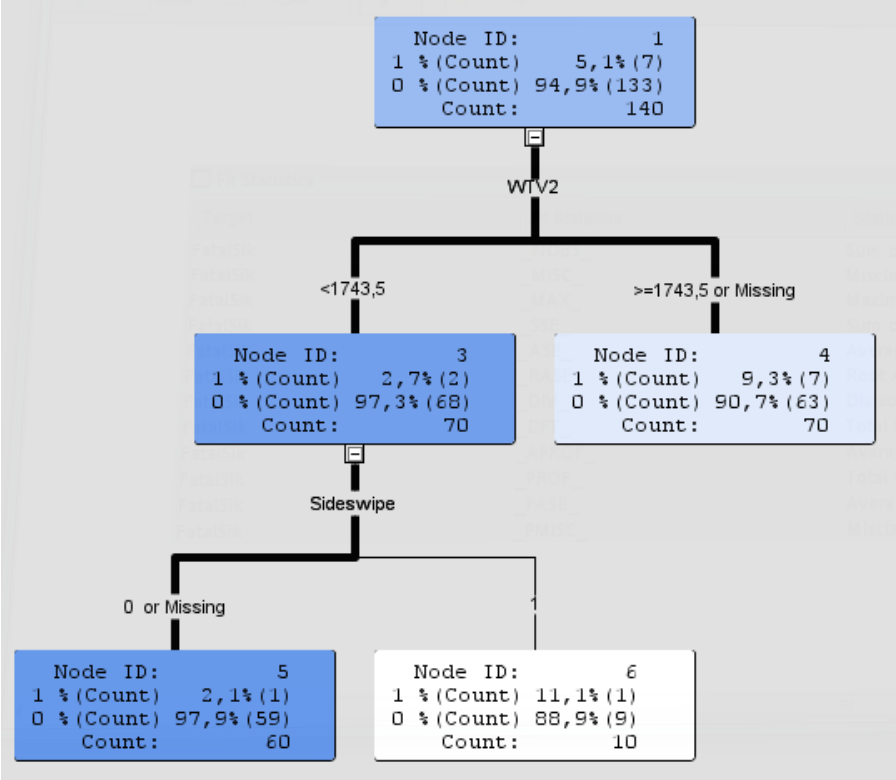


Figure 5.4 – Classification tree model for FatalSIK with all crashes for balanced sample.

CART output displayed in Figure 5.4 shows that the weight of vehicle V2 (WTV2) was the first variable used to split the observations at the root node. Collisions involving heavier vehicles for V2, WTV2≥1743.5 Kg, were associated with a higher percentage (9.3%) of severe crashes, node ID 4. On the other hand, crashes involving collisions with a vehicle V2 which follows into the lighter category, WTV2<1743.5 Kg, show a lower percentage of severe crash, 2.7%, node ID 3. This finding is consistent with previous research that indicates increasing risk of severity when the

weight of opponent vehicle increases [53, 56]. In addition, the category presented by node ID 4, can denote the effect of incompatibility between vehicles. The higher severity found for collision involving opponent vehicles having a weight ≥ 1743.5 kg can represent a collision involving a passenger car with a pick-up truck, thus higher severity may be expected for car occupants. This finding is consistent with Fredette at al. research that stated “drivers colliding a pickup truck rather than a car are 2.72 times more likely to die” [48]. Node ID 3 was further split by vehicle crash type, leading to sideswipe collisions to a higher percentage of severe crashes (11.1%) and non-sideswipe collisions with smaller percentage of severe crashes (2.1%), nodes ID 16 and 15, respectively. As previously mentioned for the tree model discussed in section 5.1.1, sideswipe collisions are known to increase the risk of severity. However, only one severe case is observed at each terminal node, nodes ID 15 and ID 16, and caution is needed in the interpretation of results relaying in few observations. The strength of association between the predicted target FatalSIK and the categories denoted by three terminal nodes was evaluated by Chi-sq test. Since two cells had expected counts less than 5, (1 observation for target level”1” in nodes ID 3 and ID 4), Fisher’s exact test was selected. Fisher’s exact p-value< 0.0164 and it implies that the FatalSIK cannot be considered independent from the weight of vehicle V2 and the collision type. The classification tree model indicates that the weight of vehicle V2 (WTV2) and crash type, were selected to classify a crash as severe FatalSIK”1”, or non-severe, FatalSIK “0”.

CART information for variable importance for the predictors included in decision tree model with a balanced dataset was as follows: WTV2 (1) and Sideswipe (0.66). As already explained, these predictors were also found important for modeling crash severity by other studies [48, 144].

5.2.2 CART for FatalSIK with two-vehicle collisions- Balanced dataset

The predictive decision tree model for two-vehicle collisions using a balanced dataset is presented in Figure 5.5. The resampling procedure lead to a balanced dataset with 0.5 ratio between the target level “1” (32 counts for severe crashes) and target being level “0” (32 counts for non-severe crashes), resulting in a total of 64 observations. To correct the bias from the over-representation of the target level “1”, prior probabilities were adjusted for the original dataset distribution, as observed in the root node o Figure 5.5. Thus, 2 counts represent the 3.7% of FatalSIK “1”, and 62 counts denoted the 96.3% of FatalSIK”0” for the original dataset distribution.

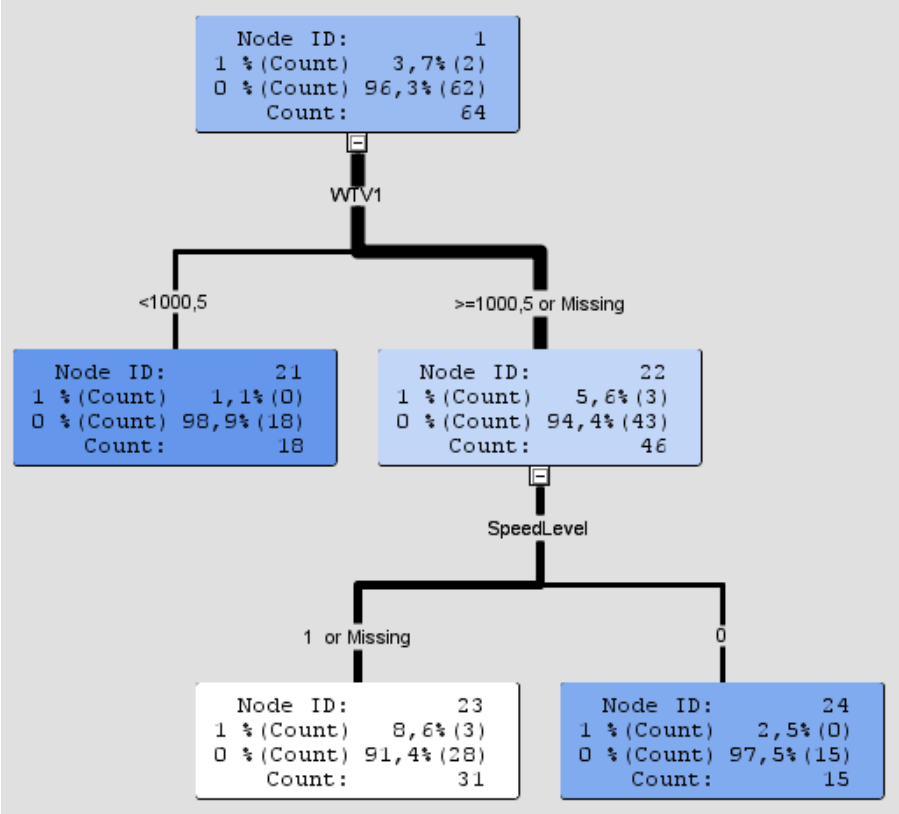


Figure 5.5 – Classification Tree for FatalSIK for two-vehicle crashes with a balanced sample.

The above classification tree shows that when the collision involved a lighter V1, WTV1<1000.5kg, 98.9% of the crashes were estimated non severe, node ID 21. On the other hand, collisions involving heavier vehicles V1, WTV1≥1000.5kg, were associated with a higher percentage of severe crashes (5.6%), node ID 22. Then, this node containing more severe crashes was split by speed level, showing that higher speeds (left branch with number 1") are associated with a higher proportion of severe crashes, leading to 8.6% for FatalSIK "1". On the other hand, collisions registered at roads with lower speed limits (right branch with number "0") showed a lower proportion of severe crashes, 2.5% for FatalSIK"0". These results are consistent with previous research that had identified the dominant effect of weight in increasing crash risk when a collision involves two cars of different weights [43, 53, 63, 64, 87]. Regarding to speed effects, the result is consistent with other research that had identified speeding as increasing risk of injury level [42, 43, 91, 93, 98]. This classification tree model has predicted the highest probability of 8.6% for severe crashes resulting from collisions involving heavier vehicle class and driving at higher speed level. Tracing the path down the tree to this terminal node, it can be noticed that the graphical representation of this model supports the Newtonian mechanism explored by Evans to evaluate injury risk based on mass ratio and changes in the velocity for the two vehicle involved [53, 145]. For this classification model, the strength of association between crash severity and the categories illustrated by the terminal nodes is proven by Fisher's exact test. The p-value< $1.7E^{-12}$ indicates

that the FatalSIK cannot be considered independent from the weight of the vehicles involved in the crash neither from the speed level.

CART output for variable importance for the classification model discussed above was as follows: WTV1 (1) and Speed Level (0.71). As already explained, these predictors were also found important for modeling crash severity by other studies mentioned earlier [42, 43, 53, 64, 91, 93, 98, 145].

5.2.3 CART for FatalSIK with single-vehicle crashes- Balanced dataset

The predictive decision tree model for single-vehicle crashes using a balanced dataset is presented in Figure 5.6. The resampling procedure was applied to obtain a balanced dataset with equal proportion of target level “1” (38 counts for severe crashes) and target being level “0” (38 counts for non-severe crashes). Hence a total of 76 observations were used as training sample for the decision tree development. To correct the bias from the over-representation of target level “1” (FatalSIK “1”), prior probabilities were adjusted for the original dataset distribution, as observed in the root node of Figure 5.6. Thus, 6 counts represent the 7.6% of FatalSIK “1”, and 70 counts denoted the 92.4% of FatalSIK”0” for the original dataset distribution.

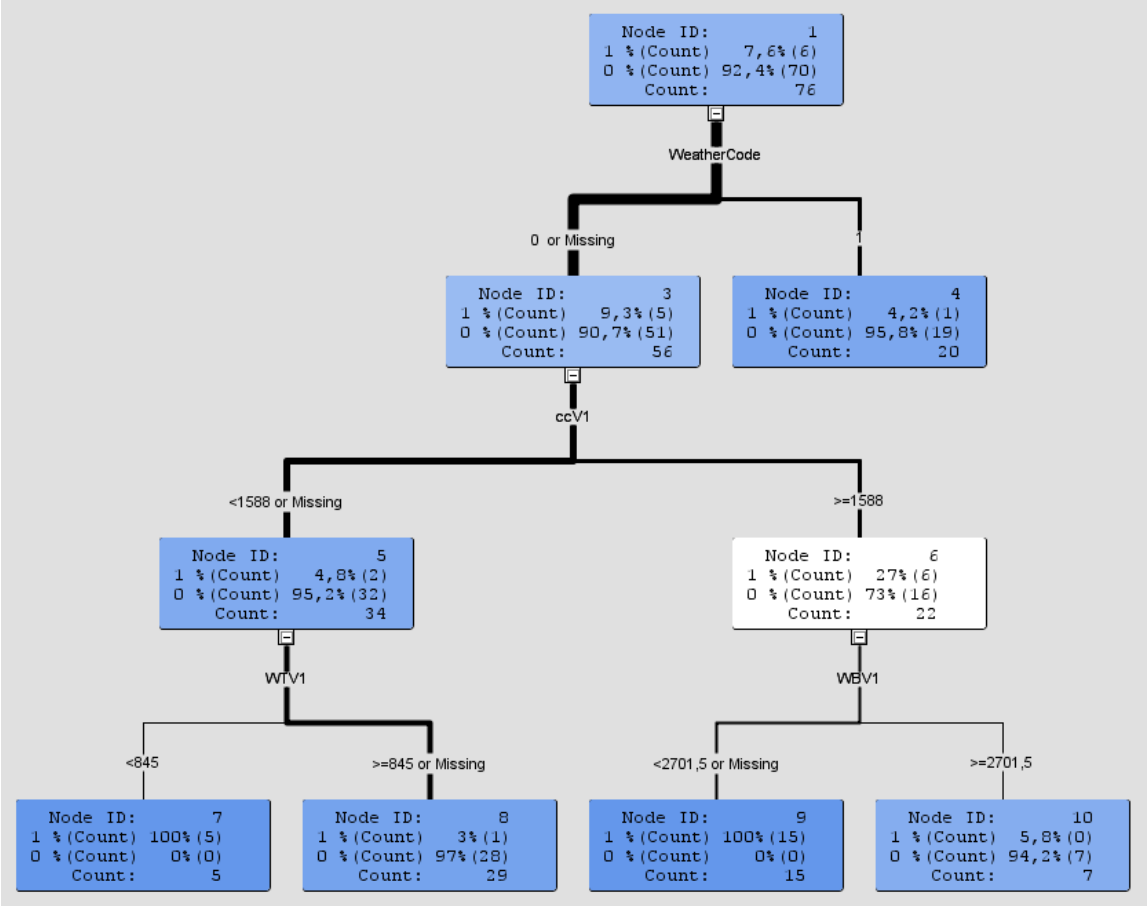


Figure 5.6 – Classification tree model for FatalSIK for single-vehicle crashes with a balanced sample.

The initial split at node ID 1 is based on the variable of weather conditions: crashes that happen under rain and/or bad weather conditions (variable taking up the value “1” at the right tree branch) showed a lower (4.2%) proportion of severe cases, node ID 4. On the other hand, crashes occurring under good weather conditions (variable taking up the value “0” at the left tree branch) showed a higher proportion of severe cases (9.3%), node ID 3. This node was split by vehicle´s engine size. Crashes involving lower vehicles engine size, ccV1<1588 cm^3, showed a lower crash severity, 4.8%. On the other hand, when vehicle with larger engine was involved, ccV1≥1588 cm^3, displayed a higher proportion of severe crashes, 27%, node ID 6. Following, node ID 5 in the left breach was split by vehicle’s weight into two terminal nodes. Crashes involving heavier vehicles, WTV1≥845 kg, were linked to smaller proportions of severe injuries (3%) than crashes involving lighter vehicles, WTV1< 845 kg, which was associated with 100% proportion of severe crashes based on the balanced dataset for single vehicle-crashes. Following down the path from node ID 6 (in the right), wheelbase of vehicle was used to split, leading two additional terminal categories as follows. Crashes including vehicles with larger wheelbase, WB≥ 2701.5 mm, revealed lower proportion of severe injuries, 5.8% (node ID 10). While crashes involving vehicles in the smaller

category of wheelbase, WB< 2701.5 mm, were predicted to result in severe injuries, 100%, node ID 9.

Fisher's exact test revealed a p-value $6.15E^{-18}$, showing that the FatalSIK cannot be consider independent from those four terminal categories. Comparison of this decision tree model with earlier studies, shows that good weather conditions have been linked to a higher incidence of severe crashes, as previously mentioned [87, 92, 144]. A possible explanation is that sunny days may result in higher speeds and more driver confidence [87]. For crashes involving lighter vehicles (WTV1< 845 kg) the probability of a severe crash was significantly higher than for the heavier vehicles. This finding supports the argument that any crash involving a vehicle with low mass will mostly be severe [53, 64, 145]. Very important to notice that, though vehicles with larger engines (ccV1≥1588 cm^3) suggests a higher probability of involvement in severe crashes, if those vehicles follow into the category of larger wheelbase distances (WBV1≥2701.5 mm), the injury risk could be reduced. This finding is consistent with Bédard et al. that suggested "25 cm increase in wheelbase translates into 10% reduction in the odds of a fatality" [91]. This model supports the protective value of larger vehicles independent of their drivers.

CART information for variables importance was as follows: ccV1 (1), WTV1 (0.93), WBV1 (0.78), and Weather (0.49). For the classification model discussed in this section, it is interesting to notice that vehicle technical characteristics were found significantly more important for FatalSIK prediction rather than crash information, denoted by the selection of only one variable (weather conditions) and its importance is less relevant than the variables linked to vehicles' technical data.

5.3 CART for FatalSIKV1 and FatalSIKV2 for Two-Vehicle Collisions- Original Sample

This section presents CART results for the innovative modeling strategy targeting the severity risk prediction for the occupants of each individual vehicle, in a two-vehicle collision. The original crash sample included a limit number of severe cases for FatalSIKV1 (21 observations) and FatalSIKV2 (14 observations). Therefore, the resampling strategy, as followed in section 5.2 for FatalSIK prediction, was not applied for FatalSIKV1 and FatalSIK2 modeling, since it would produce small balanced datasets: 42 and 28 observations, respectively. For these targets modeling, the original sample for two-vehicle collisions was used and results are presented next. For both models, the inputs were the same (20 independent variables), those variables, and targets are identified in Table 3.5, Chapter 3.

5.3.1 CART for FatalSIKV1 in two-vehicle collisions- Imbalanced dataset

This section presents CART results for crash severity prediction in the subject vehicle, (vehicle V1), by addressing the effect that the characteristics of opponent vehicle V2 might impose to the occupants of V1, and by taking into account the subject vehicle capability to protect its occupants (crashworthiness). The probability of serious injuries and/or fatalities within the occupants of vehicle V1 is expressed by FatalSIKV1. Classification tree model for FatalSIKV1 is shown in Figure 5.7.

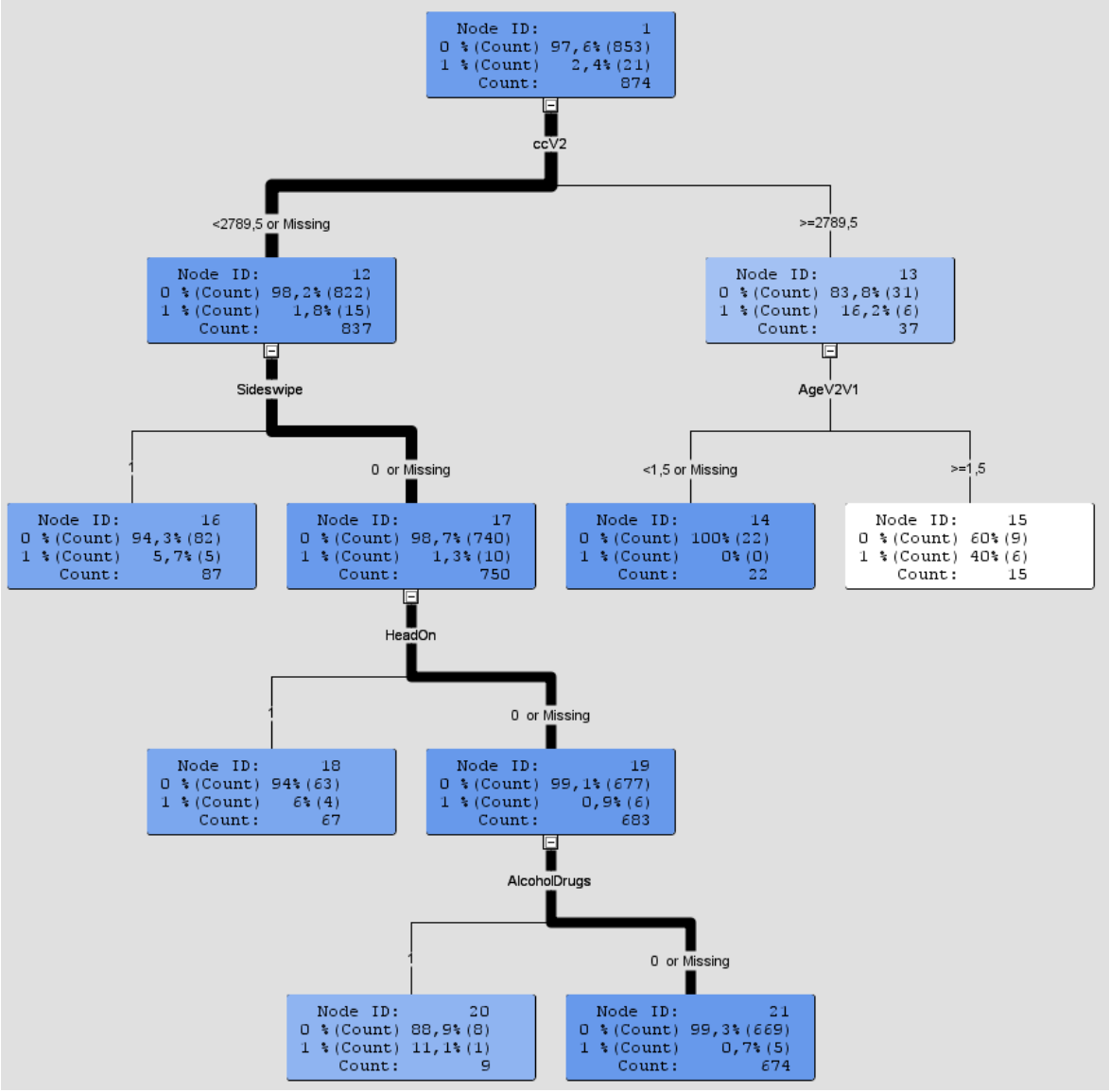


Figure 5.7 – Classification tree model for FatalSIKV1 in two-vehicle collisions with the original sample.

The engine size of vehicle V2, ccV2, was the first explanatory variable selected to split the original sample of 874 crashes. Collisions involving vehicle V2 with smaller engine size, ccV2< 2789.5 cm^3, showed a lower proportion of severity for occupants of vehicle V1 (1.8%), than when V2 had a larger engine size, ccV2≥ 2789.5 cm^3 (16.2%). Following, the type of crash and then by the presence of drivers tested for alcohol and/or drugs were used for the tree split. The terminal nodes at the left side of the tree clearly show that collisions involving a sideswipe collision or a head on have higher risk of severity for occupants of vehicle V1, 5.7% and 6%, in nodes ID 16 and 18, respectively. This finding is consistent with previous work that had identified these crash types as the most severe [43, 48, 61, 144]. Also, the effect of alcohol and/or drugs use is consistent with a large number of studies [87, 92, 99].

Following the right branch of the above tree, collisions where engines size of V2≥ 2789.5 cm^3 and AgeV2V1< 1.5 yr resulted in non-severe crashes (100% for FatalSIK"0" as observed in node ID 14). On the other hand, collisions involving AgeV2V1≥1.5 yr, were linked to the highest proportion of a severe outcome in the subject vehicle was the highest, 40% (in node ID 15). This analysis suggests that the characteristics of the opponent vehicle (vehicle V2) have an effect on the increased risk of serious and/or killed injuries in the subject vehicle V1. The association between the above categories and severe outcomes in vehicle V1 is confirmed by Fisher's exact test, which p-value <$1.96E^{-09}$ suggested that the FatalSIKV1 and the above selected categories cannot be considered independent at the 5% significance level.

In addition to the graphical display for the classification tree model for FatalSIKV1, CART also provides helpful information on the variables importance. For this model, variables importance was as follows: AgeV2V1 (1), ccV2 (0.72), HeadOn (0.33), Sideswipe (0.32) and AlcoholDrugs (0.26). Very interesting to notice that when predicting the probability of a severity for occupants of vehicle V1 involved in a collision with the counterpart vehicle V2, vehicles' characteristics play a more important role than variables relaying in crash type and presence of alcohol and/or drugs.

5.3.2 CART for FatalSIKV2 in two-vehicle collisions- Imbalanced dataset

This section presents CART results for crash severity prediction in the opponent vehicle, (vehicle V2), by addressing the effect that the characteristics of subject vehicle V1 might impose to the occupants of vehicle V2, and taking into V2 capability to protect its occupants. The probability of serious injuries and/or fatalities within the occupants of vehicle V2 is expressed by FatalSIKV2. Classification tree model for FatalSIKV2 is shown in Figure 5.8.

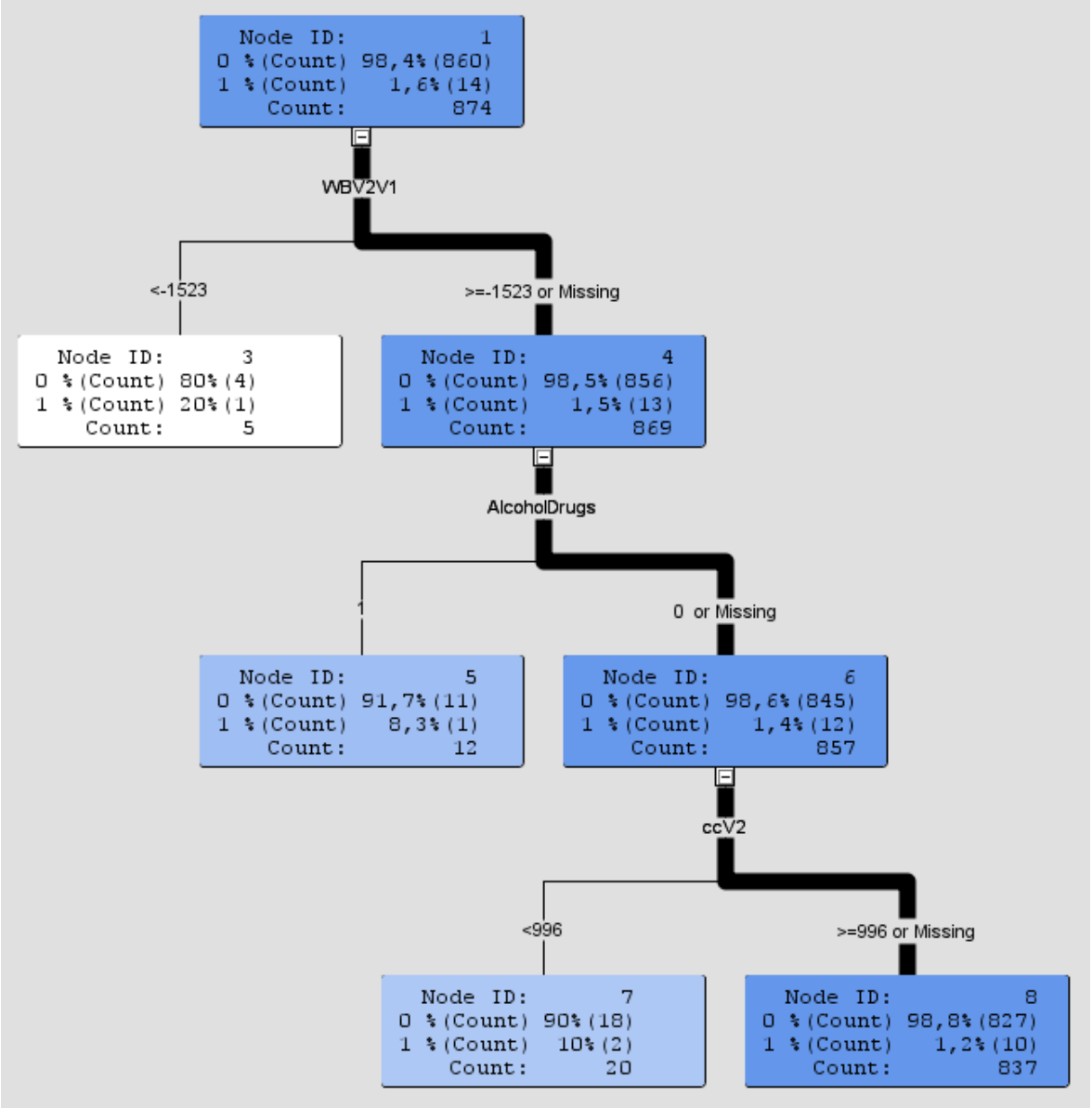


Figure 5.8 – Classification tree model for FatalSIKV2 in two-vehicle collisions with the original imbalanced sample.

The differential of wheelbase distance between the two vehicles, WBV2V1, was the first variable selected to split the crash. For collisions where the wheelbase of vehicle V2 was 1523 mm shorter than the wheelbase of the other vehicle involved, WBV2V1< -1523 mm, had resulted in higher proportion of severity for occupants of vehicle V2, 20% (in node ID 3). On the other hand, for collisions involving vehicles were WBV2V1≥ -1523 mm, the proportion of severe cases among vehicle V2 was smaller, 1.5% (node ID 4). Subsequently, the variable alcohol and/or drugs splits this node, and collisions involving this effect lead to a higher proportion of severity in vehicle V2, 8.3% (in node ID 5). For collisions where WBV2V1≥ -1523mm and involving sober drivers, the

proportion of severe cases in vehicle V2 was smaller, 1.4% (node ID 6). Following, this node was split by the engine size of vehicle V2, leading to two terminal nodes. Collisions where vehicle V2 follows in the category ccV2< 996 cm^3, were associated with higher proportion of severe cases than for vehicles in the category ccV2≥ 996 cm^3, (10% and 1.2%, respectively). Vehicles with larger engine size often are heavier; hence it would be possible to offer a better protection to its occupants. Even though the vehicles weight was not select for the tree development, it is possible that effect of vehicle´s weight could be in a certain way reflected in vehicle´s engine size categories. The results presented by this model are consistent with previous research that had associated vehicle crashworthiness with its size and mass [53, 60, 87, 145]. In the interpretation of this model, it must be aware that the training sample only had 14 cases for the target level with interest, FatalSIKV2"1". However, Fisher's exact test showed a p-value <0.0016, denoting that FatalSIKV2 and the differential of wheelbase, engine size and presence of alcohol and/or drugs cannot be considered independent, at the 0.05 significance.

As far as variables importance for the above model, it follows as: WBV2V1 (1), ccV2 (0.94), and AlcoholDrugs (0.57). Similarly to the previous model, for crash severity prediction, vehicle's characteristics for both involved in the collision were found more important predictors than variables relaying in crash information.

5.3.3 Comparison of FatalSIKV1 and FatalSIKV2 decision tree models

For both decision trees models For FatalSIKV1 and FatalSIKV2 (sections 5.3.1 and 5.3.2) vehicles' characteristics suggest to be more relevant for the injury severity prediction than variables related to crash information. It is noticed than for both models, vehicles differential for "specific" technical characteristic was found the most important predictor, denoting that it is important not only to consider vehicle's individual characteristics but also, its differential between the vehicle involved in the collision. The engine size of vehicle V2 was important for both targets prediction: severity among occupants of vehicle V1 and V2. A possible explanation why ccV2 was selected for both classification tree models could be related to the fact that mean values for engine size of vehicle V2 was larger than for vehicle V1, 1700.94 cm^3 (S.D. 522.18) and 1665.96 cm^3 (S.D. 509.98), respectively. In addition, it is conceivable that this variable contains the effect of vehicle weight; as a matter of fact, descriptive statistics seems to support this statement because the weight of vehicle V2 was also slight larger than the weight of vehicle V1, 1262.85 kg (S.D. 364.46) and 1234.20 kg (S.D. 356.82), respectively. In two-vehicle collisions, vehicle V2 due to its larger size and weight would raise the risk for occupants of the opponent vehicle, therefore larger vehicles categories of ccV2 would increase the severity risk for occupants of vehicle V1, as it was suggested with highest proportion for FatalSIKV1"1" (in node ID 15 inFigure 5.7). On the other hand, vehicle V2 would probably offer a larger compartment area to absorb the impact of the collisions, and they would decelerate more slowly following the impact, decreasing the risk of

injuries. Accordingly, collisions involving larger categories for ccV2 would decrease the severity risk for occupants of vehicle V2, (as observed in node ID 8 in Figure 5.8). For both models, the effect of alcohol and/or drugs use was linked to a higher proportion for severe cases. Information on crash type, even though those predictors were scored as important inputs for FatalSIKV1 prediction, they were irrelevant for FatalSIKV2 predication. An acceptable explanation for this difference is due to the fact that only 14 severe events cover target FatalSIKV2, while there were 21 severe events for target FatalSIKV2 modeling. Last, variables importance within the classification tree models for FatalSIKV1 and FatalSIKV2 prediction suggest that vehicles' characteristics play a more relevant role comparatively to other crash information.

5.4 Concluding Remarks

This Chapter presented CART results for crash severity prediction using two approaches: original imbalanced sample and balanced datasets. For the imbalanced approach (based on the original sample), the presence of alcohol and/or drugs was a common risk factor identified across all the classification decision tree models. These models showed evidence that alcohol and/or drug use play a major role in increasing crash severity risk, despite of vehicle crashworthiness and/or collision type. For single-vehicle crashes, this variable was found the most important for crash severity prediction, suggesting that crashes outcomes could be more influenced by drivers' behavior than vehicles' characteristics. For the balanced approach (following a resampling strategy), CART output revealed that the input alcohol and/or drugs was not present in any of the classification tree models. Severity prediction decision tree for two-vehicle collisions identified the effect of vehicle's weight as the most important predictor, suggesting an increasing proportion of severe crashes when one of the vehicles involved is heavier. For single-vehicle crashes, engine size was the most important factor for FatalSIK prediction.

The comparison of the two approaches, decision trees developed with the original sample and with the balanced sample, revealed that in general the application of the decision trees with the imbalanced sample resulted in larger trees, due to the larger number of observations used for the tree development. Thus, this approach resulted in trees with more splits, yielding to the identification of more risk factors for the classification of a crash event as sere or non-severe. On the other hand, the decision models developed with the balanced approach had resulted smaller, because fewer observations were used. Very interesting it was the finding that alcohol and/or drugs were identified as common risk factor across all crashes, two-vehicle collisions and single-vehicle crashes. Also, the age of the vehicles involved in the crash was identified as an important risk factor for all the decision trees models. However, when modeling the balanced sample, this risk factor was not selected by the decision tree models. Following the resampling approach, the weight of the vehicle was identified as an important risk factor across all the decision trees models, for: all crashes, two-vehicle collisions and single-vehicle crashes.

Regarding the individual vehicle injury severity analysis, classification tree models for FatalSIKV1 and FatalSIKV2 were developed using the original sample (imbalanced data). Owing the limited number of severe events in each vehicle involved (28 severe events in V1 and 14 events in V2), the resampling method was not applied. Decision trees also identified the effect of alcohol and/or drugs, although here the effect of alcohol and/or drugs was the less important variable for crash severity modeling. For FatalSIKV1, the most important risk factor was the age differential for the two vehicle involved in the collision. On the other hand, for FatalSIKV2, the most important risk factor was the wheelbase differential between the two vehicle involved. These findings suggest that for crashworthiness evaluation, it is important not only to consider vehicle's individual characteristics but also, its differential characteristics between the vehicles involved in the collision.

CHAPTER 6
LOGISTIC MODELS FOR SEVERITY PREDICTION IN SINGLE-VEHICLE CRASHES

This chapter discusses the injury severity risk sustained by the occupants of a vehicle involved in a single motor vehicle crash. For logistic regression models analysis, while the coefficients estimates provide a good interpretation for continuous independent variables, the odds ratio will be used for the interpretation of the categorical variables in the model.

Chapter 6 is organized as follows. Firstly, a model developed based on the original crash sample (imbalanced data) is presented. Secondly, the best models for FatalSIK prediction based on balanced approach are presented. Selected models are examined for its fit statistics and evaluated for prediction accuracy with the training sample and original sample, 10 stratified random sample used for validation. Finally, a recommended model for FatalSIK prediction is presented. Models presented in this chapter were developed with SAS® v9.2 and SAS®Enterprise Miner™7.2 software [84, 89, 117, 136].

6.1 Logistic Regression Analysis for FatalSIK with the Original Single Crash Sample- Imbalanced Data

This beginning section aims to exhibit the problem of prediction accuracy linked to logistic regression models using imbalanced data, rather than discussing the model itself. As previously explained in section 2.5, modeling rare events, such as sever crashes, imposes a challenge because the logit model would predicted right the most common event (non-sever crash) and will miss the prediction for the rare event (severe crash). As presented in Chapter 4, for single-vehicle crashes overall severity was 7.6%, thus yielding to 92.4% of non-severe crashes in the sample. The logistic model developed to predict FatalSIK using the original crash sample is presented in Table 6.1. This model exhibits the problem of prediction accuracy when dealing with imbalanced classes' distribution at the Portuguese crash sample. Thus it supports the need to perform logistic regression modeling for the Portuguese crash data, based on a balanced training sets.

Table 6.1 - Imbalanced-Model-S results for FatalSIK prediction with logistic regression performed for the original single-vehicle crashes sample.

Imbalanced- MODEL-S								
Fit Statistics								
Test for Global Null Hypothesis		**Analysis of Maximum Likelihood Estimates and Odds Ratio**						
DF	Pr<ChSq	**Parameter**	DF	Estimate	SE[1]	Pr>ChiSq	OR[2]	95% CI[3]
4	0.0015	Intercept	1	2.0201	2.6143	0.4397		(-3.1039_ 7.1441)
		AlcoholDrugs (0)	1	-0.8263	0.3665	0.0242	0.192	(-1.5446_ -0.1080)
		WBV1	1	-0.00233	0.0011	0.0415	0.998	(-0.0046_ -0.0001)
		WeatherCode (0)	1	0.4269	0.2175	0.0496	2.349	(0.0007-0.8532)
		ccV1	1	0.0012	0.0004	0.0032	1.001	(0.0004_ 0.0020)
Obs.	500							
ASE	0.07							
MISC	0.07							
Accuracy Performance								
Accuracy Rate with Training Sample (N=500)								
FN[4]		TN[5]		FP[6]		TP[7]		
37		462		0		1		

1 Standard Error; 2 Odds Ratio Estimate; 3 95% Confidence Interval; 4 False Negative; 5 True Negative; 6 False Positive; 7 True Positive.

As observed in Table 6.1, the model predicted right all the cases of non-severe crashes, (TN=462). However, only one severe crash was correctly predicted, whereas the remaining severe ones were incorrectly predicted as non-severe (FN=37). Thus, without a resampling strategy, model training prediction accuracy for the severe crashes would be unsatisfactory, 2.7% (1/37). Next, logistic regression models results for crash severity prediction based on the resampling strategy are presented.

6.2 Logistic Regression Analysis for FatalSIK with Resampling Approach

This section presents the logistic regression modeling results for the probability of a serious injury and/or fatality given a single-vehicle crash. Several candidate models were developed based on a balanced training sample and the best candidate models were selected for further accuracy performance evaluation. During the modeling stage, four models were selected for FatalSIK prediction in single-vehicle crashes: Model-IA-S, Model-IB-S, Model-IC-S and Model-ID-S. For single-vehicle crashes there is only one target, FatalSIK, denoted by “I”, and the alphabetic terms “A, B, C and D” are used to indicate the best four candidate models, and “S” stands for single-vehicle crashes. Model-IA-S and Model-IB-S were selected to be presented and discussed in this section. Model-IC-S and Model-ID-S are provided in Appendix 8.

Independent variables used as models inputs are identified in Table 6.2. Models results for single-vehicle crashes are discussed based on hypothesis testing for the selected variables (model parameters estimates). The parameters (predictors) that are statistically significant at 0.05 level are shown with an “*”. Lower and upper bound of 95% confidence interval of estimates are shown in brackets. The ASE and MISC are of most interest in model fit statistics. The ASE measures the difference between the prediction estimate and the observed FatalSIK value. Also, misclassification measures the fraction of cases where the decision does not match the actual target value, as defined in Equation 3.9 and Equation 3.10. For the selected best models candidates, accuracy performance was evaluated as follows. Firstly, each selected model was evaluated based on its prediction accuracy with the original sample (500 observations). Secondly, each of the selected models was evaluated using 10 stratified random samples (76 observations), based on the K-fold cross validation explained in section 3.7.4.

Table 6.2 – Description of design variables (inputs) and targets when modeling crash severity for single-vehicle crashes with logistic regression.

Variable	Description	Abbreviation	Model Identification
Independent Variables Used as Inputs			
Age of Vehicle 1	$AgeV_1$ (yr) was calculated based on the year of the crash event minus the year of the first vehicle registration.	AgeV1	Model-IA-S, Model-IB-S, Model-IC-S and Model-ID-S
Alcohol and/or Drugs	The Driver´s test for alcohol and or drugs is presented as: Code=0, legal; Code=1, illegal	AlcoholDrugs	Model-IA-S, Model-IB-S, Model-IC-S and Model-ID-S
Crash type	Ran off road Rollover	RanOff Rollover	Model-IA-S, Model-IB-S, Model-IC-S and Model-ID-S -
Divided/undivided	Existence or absence of physical median: Code=0, undivided Code=1, divided	DivisionCode	Model-IA-S, Model-IB-S, Model-IC-S and Model-ID-S
Speed Level	The speed level was coded as follow: If Speed limit<=90 $km.hr^{-1}$, then code=0 If Speed limit>90 $km.hr^{-1}$, then code=1	SpeedLevel	Model-IA-S, Model-IB-S, Model-IC-S and Model-ID-S
Wheelbase of Vehicle 1	Wheelbase of vehicle (V1) (mm)	WBV1	Model-IA-S, Model-IB-S, Model-IC-S and Model-ID-S
Weight of Vehicle 1	Weight of vehicle 1 (V1) (kg)	WTV1	Model-IA-S, Model-IB-S, Model-IC-S and Model-ID-S
Weather Conditions	Weather conditions at the moment of the crash: Code=0, Clear and/or dry pavement Code=1, rain and/or wet pavement	WeatherCode	Model-IA-S, Model-IB-S, Model-IC-S and Model-ID-S
Engine Size of Vehicle 1	Engine size of vehicle (V1) (cm^3)	ccV1	Model-IA-S, Model-IB-S, Model-IC-S and Model-ID-S
Dependent Variable used as Target			
Serious and/or Killed	FatalSIK is a categorical response for a crash outcome used to predict either a serious injury, or fatality in a crash event. FatalSIK=1, if SI>0 and/or K>0, else, FatalSIK=0	FatalSIK	Model-IA-S, Model-IB-S, Model-IC-S and Model-ID-S

6.2.1 Model-IA-S Analysis

Model-IA-S was developed using logistic regression for FatalSIK prediction in crashes involving one single vehicle, with forward selection for the inputs signalized at Table 6.2. As already mentioned forward selection method begins with no candidate inputs in the model and adds inputs until the entry significance level is met. For this model design, the entry level was set 0.1, similarly to Li modeling research [100], the p-values less than or equal to the 0.1 level of significance are considered.

Table 6.3 summarizes Model-IA-S fitting results and performance evaluation. The test for the global null hypothesis shows that at least one of the predictor´s regression coefficient is not equal to zero in the model, p-value<0.0004. From a total of nine inputs (in Table 6.2), the final model has four predictors: AgeV1, WBV1, ccV1 and WeatherCode. All these predictors are statistically significant at 0.1 level. Model intercept was not found statistically significant at 0.1 level, p-value<0.30. “Too much focus on statistical significance can lead to the false conclusion that a variable is “important” explaining “Y”, even though its estimated effect is modest” [146]. In addition, for smaller size, some authors are willing to use larger significance levels, reflecting the fact that it is harder to find significance with smaller sample sizes (the estimators are less precise) [146]. For instance, at one of the models developed by Li to predict crash severity in work zones, a larger criterion of 0.3 was set [100]. On the other hand, focus only in the predictors at the model, with exception for the

wheelbase of the vehicle, (p-value<0.0593), all the selected predictors in the model were found to be statistically significant at the 0.05 level (p-value<0.0144, p-value<0.0418, and p-value<0.0031, for AgeV1, WeatherCode(0) and ccV1, respectively. The model fit statistics yield an ASE of 0.187 and MISC of 0.237.

Table 6.3 - Model-IA-S results for FatalSIK prediction with logistic regression performed for a balanced dataset of single-vehicle crashes.

MODEL-IA-S								
Fit Statistics								
Test for Global Null Hypothesis		**Analysis of Maximum Likelihood Estimates and Odds Ratio**						
DF	Pr<ChSq	Parameter	DF	Estimate	SE[1]	Pr>ChiSq	OR[2]	95% CI[3]
4	0.0004	Intercept	1	5.1730	5.0151	0.3023		(-4.6565_ 15.00)
		AgeV1	1	0.1519	0.0621	0.0144*	1.164	(0.0302_ 0.2736)
		WBV1	1	-0.0045	0.0024	0.0593	0.996	(-0.0092-0.0002)
		WeatherCode (0)	1	0.6879	0.3380	0.0418*	3.958	(0.0255-1.3504)
		ccV1	1	0.00297	0.0010	0.0031*	1.003	(0.0010_ 0.0049)
Obs.	76							
ASE	0.187							
MISC	0.237							

Prediction Accuracy Performance											
Accuracy Rate with Training Sample (N=76)					Accuracy Rate with Original Sample (N=500)					Prediction Accuracy for 10 Stratified Random Samples	
FN[4]	TN[5]	FP[6]	TP[7]	AR[8]%	TPs[9]	FPs[10]	TNs[11]	FNs[12]	AR[13]%	Mean%[14]	S.D.[15]
10	30	8	28	76.3	17	97	365	21	76.4	62.0	2.3

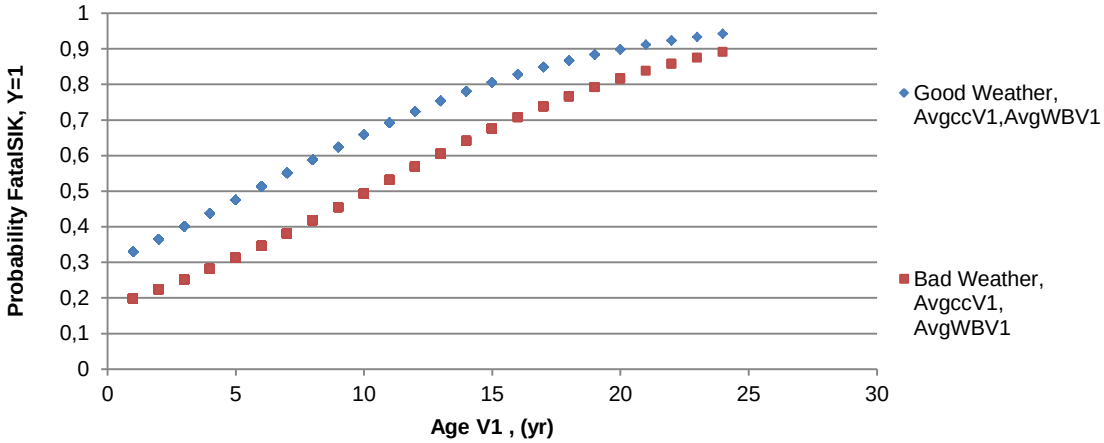
1 Standard Error; 2 Odds Ratio Estimate; 3 95% Confidence Interval; 4 False Negative; 5 True Negative; 6 False Positive; 7 True Positive; 8 Percentage of Accuracy Rate; 9 True Positives; 10 False Positives; 11 True Negatives; 12 False Negatives; 13 Percentage of Accuracy Rate; 14 Mean of Prediction Accuracy for the 10 stratified random samples; 15 Standard Deviation for the Prediction Accuracy of the 10 stratified random samples; *Statistically significant at 5% level.

The logistic regression equation developed for Model-IA-S is presented below.

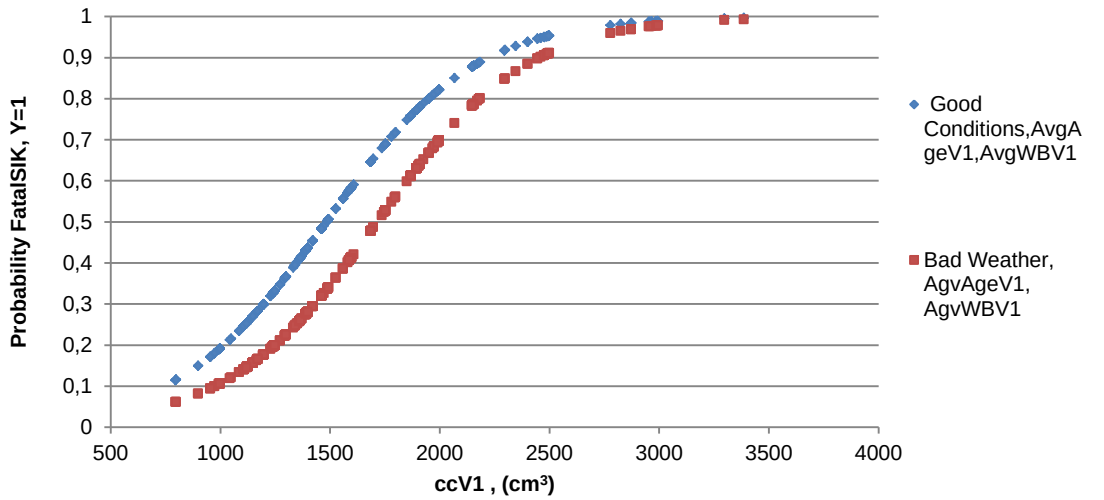
$$P\,(FatalSIK = 1) = \frac{\exp(5.1730 + 0.1519 * AgeV1 - 0.0045 * WBV1 + 0.6879 * WeatherCode(=0) + 0.00297 * ccV1)}{1 + \exp(5.1730 + 0.1519 * AgeV1 - 0.0045 * WBV1 + 0.6879 * WeatherCode(=0) + 0.00297 * ccV1)} \quad \text{Equation 6.1}$$

The interpretation of the Model-IA-S shows a positive relationship between vehicle engine size and age and good weather conditions with the probability of severe crashes, FatalSIK"1". Therefore, the model parameters: AgeV1, WeatherCode and ccV1 show positive sign at the above equation, Equation 6.1. On the other hand, as the vehicle wheelbase increases there is a decrease in the probability of a FatalSIK"1". Thus the parameter WBV1 shows a negative sign on Equation 6.1. Crashes occurring under good weather condition are associated with a significant increased risk of crash severity, as shown by the odds ratio. In Table 6.3, odds ratio of a severe crash increases in good weather condition almost by four compared to the bad weather conditions. Graphical representation for this model FatalSIK prediction equation, Equation 6.1, is illustrated in Figure 6.1.

a)



b)



c)

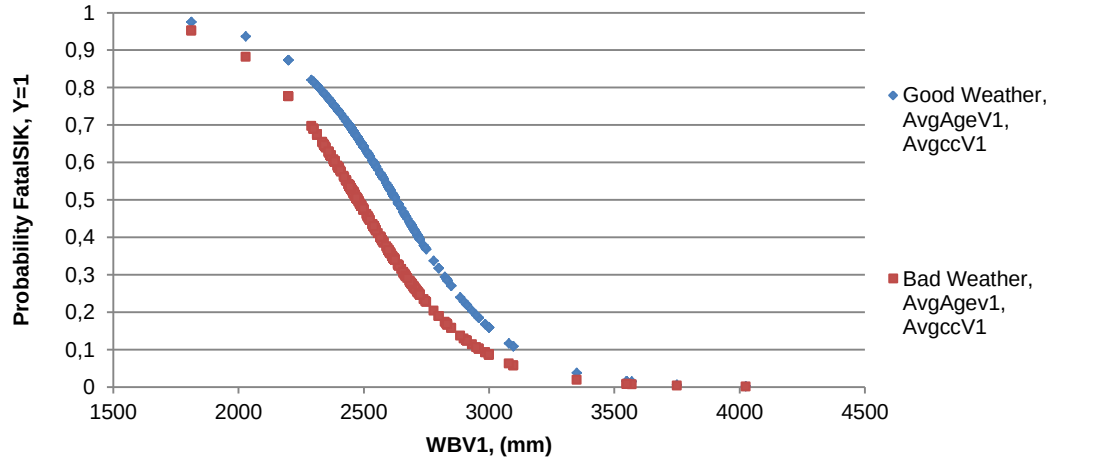


Figure 6.1 – Probability of a serious injury and/or killed by Model-IA-S for single-vehicle crashes with: a) age of the vehicle; b) engine size of vehicle; and c) wheelbase of the vehicle.

In Figure 6.1a), FatalSIK probability is predicted a function of the age of the vehicle, controlling for vehicle wheelbase and the engine size (at 2551 mm and 1602 cm^3, average wheelbase (AvgWBV1) and engine size (AvgccV1), respectively). A similar approach was used for Figure 6.1b) and c). Figure 6.1a) shows that as the age of the vehicle increases, the probability of a FatalSIK also increases. This model finding supports previous conclusions that recent cars protect their drivers better than older cars [49, 59, 60, 85]. Figure 6.1b) shows that as the engine size of the vehicle increases, the probability of a FatalSIK also increases. The effect of the engine size may be interacting with travel speed, since drivers of more powerful cars tends to accelerate more. This finding also supports previous studies that argued "higher engine performance and power could be associated with greater speeds and greater injury risk" [64]. Figure 6.1c) shows that as the wheelbase size of the vehicle decreases, the probability of a FatalSIK also increases. The size of vehicle's wheelbase in the decreasing risk of a serious and/or fatal crash may be interpreted by the fact that one of the vehicles attribute most related to the injury severity level is vehicle size [53, 91]. A larger vehicle, offers a greater area for the energy dissipation following the crash impact force, hence reducing the energy change to which occupants in the compartment area may be exposed, thus reducing the risk. This finding is consistent with previous research which suggested that "25 cm increase in wheelbase translates into 10% reduction in the odds of a fatality" [91]. For the risk factors explained above, crashes occurring under the good weather conditions are worsen, the probability of FatalSIK is higher than for bad weather conditions, as observed by logit curves blue and red, respectively. Comparison with earlier crash severity prediction models, good weather conditions have been linked to a higher incidence of severe crashes, as previously mentioned [87, 92, 144].

The assessment of the Model-IA-S shows a good performance, as observed in Table 6.3. The accuracy rate when running the model with the training sample, which was stratified in 38 severe crashes and 38 non-severe crashes, correctly predicted 76.3% of the cases. In the training sample, the model correctly predicted 28 severe crashes (TP) and 30 non-severe crashes (TN). When compared with the previous model in Table 6.1, Imbalanced-Model-S, it is clear the improvement in model accuracy prediction. The model developed with the original imbalanced sample predicted 37 severe crashes as non-severe, leading to unsatisfactory results for TP, (TP=1). On the other hand, Model-IA-S performed with the balanced approach, was able to predict right 28 severe crashes (out of 38). When assessing the performance of this model with the original crash sample, the prediction accuracy, was also good, 76.4%. From a total of 500 crashes events, Model IA-S correctly predicted 17 severe crashes out of 38. In addition, the model correctly predicted 365 of the non-severe events out of 462 non-severe events at the entire sample. The model predicted right more severe crashes in the training sample, than in the original dataset, 28, and 17, respectively. However, it is noted that the model overall accuracy within the original sample was slightly higher than in the training sample, 76.4% and 76.3%, respectively. The evaluation of the

model overall performance with 10 stratified random samples was very satisfying; 62% (S.D. 2.3) prediction accuracy rate. EM output for Model-IA-S is provided in Appendix 8.

6.2.2 Model-IB-S Analysis

Model-IB-S is an alternative to FatalSIK prediction for single-vehicle crashes. This model was developed using logistic regression for FatalSIK prediction in crashes with backward selection for the inputs signalized at Table 6.2. As explained in Chapter 3, backward selection begins with all candidate effects (inputs) in the model and removes effects until the stay significance level is met. For this model design, the entry level was set 0.05.

Table 6.4 summarizes Model-IB-S fitting results and accuracy performance evaluation. The test for the global null hypothesis shows that at least one of the predictor´s is not equal to zero in the model, p-value<0.0013. From a total of nine independent variables entered as inputs, only two were selected by the model: AgeV1 and ccV1. These predictors are statistically significant at 0.05 level: p-value<0.0079 and p-value<0.0229, Age and ccV1, respectively. The model fit statistics shows ASE of 0.206 and MISC of 0.276, respectively.

Table 6.4 - Model-IB-S results for FatalSIK prediction with logistic regression performed for a balanced dataset of single-vehicle crashes.

MODEL-IB-S									
Fit Statistics									
Test for Global Null Hypothesis		**Analysis of Maximum Likelihood Estimates and Odds Ratio**							
DF	Pr<ChSq	Parameter	DF	Estimate	SE[1]	Pr>ChiSq	OR[2]	95% CI[3]	
2	0.0013	Intercept	1	-3.4443	1.1651	0.0031		(-5.7279_ -1.1607)	
		AgeV1	1	0.1572	0.5922	0.0079*	1.164	(0.0411_ 0.2732)	
		ccV1	1	0.00139	0.0006	0.0229*	1.003	(0.0002_ 0.0026)	
Obs.	76								
ASE	0.206								
MISC	0.276								
Prediction Accuracy Performance									
Accuracy Rate with Training Sample (N=76)					Accuracy Rate with Original Sample (N=500)				Prediction Accuracy for 10 Stratified Random Samples	
FN[4]	TN[5]	FP[6]	TP[7]	AR[8]%	TPs[9]	FPs[10]	TNs[11]	FNs[12]	AR[13]%	Mean%[14]	S.D.[15]
10	27	11	28	72.4	14	96	366	24	76.0	58.0	3.1

1 Standard Error; 2 Odds Ratio Estimate; 3 95% Confidence Interval; 4 False Negative; 5 True Negative; 6 False Positive; 7 True Positive; 8 Percentage of Accuracy Rate; 9 True Positives; 10 False Positives; 11 True Negatives; 12 False Negatives; 13 Percentage of Accuracy Rate; 14 Mean of Prediction Accuracy for the 10 stratified random samples; 15 Standard Deviation for the Prediction Accuracy of the 10 stratified random samples; *Statistically significant at 5% level.

The logistic regression equation developed for Model-IB-S is presented below.

$$P\,(FatalSIK = 1) = \frac{\exp(-3.4443 + 0.1572 * AgeV1 + 0.00139 * ccV1)}{1 + \exp(-3.4443 + 0.1572 * AgeV1 + 0.00139 * ccV1)} \qquad \text{Equation 6.2}$$

The positive regression estimates for AgeV1 and ccV1 shows a positive effect of vehicle engine size and vehicle age on crash severity risk, FatalSIK"1". Graphical representation of the logit curve for FatalSIK prediction with Equation 6.2, is illustrated in Figure 6.2. Figure 6.2 shows the probability of a serious injury and/or fatality predicted by Model-IB-S for single-vehicle crashes with engine size of the vehicle and taking into account the effect of vehicle's age consecutively.

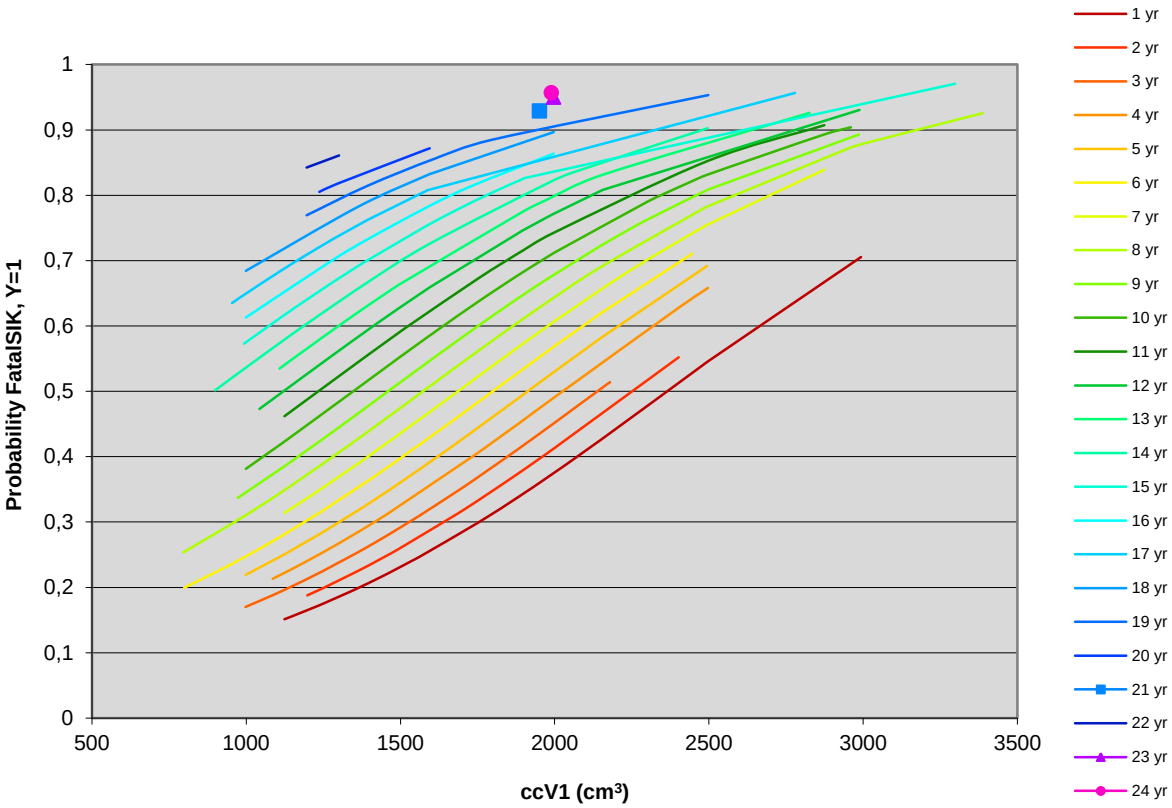


Figure 6.2 – Probability of a serious injury and/or fatality predicted by Model-IB-S with engine size of the vehicle and age of the vehicle, for single-vehicle crashes.

Figure 6.2 clearly shows that the probability of a severe crash increases as the engine size increases. As explained in Chapter 4, the engine size for the 500 vehicles in the Single dataset has minimum of 796 cm^3 and a maximum of 3387 cm^3. On the other hand, the newest vehicles in the crash data had 1yr old, while the oldest vehicle model was 24 years old. The color lines at the chart not only illustrate the age, but also the frequency of vehicles at that age level. Following this explanation, it is easy to follow the effect of vehicles age, as the engine size increases, resulting in a higher probability of a severe crash outcome.

Model-IB-S results, which have identified vehicles age and engine size as significant predictor of crash severity, are consistent with other research. The effect of vehicle´s age have been widely discussed for single and two-vehicle collisions analysis that claimed that recent cars protect their

drivers better than older cars models [49, 59, 60, 85]. Comparing to vehicle age, vehicle´s engine effect on crash severity have not been so widely explored. However, previous research have mention that larger engine size (as a proxy of vehicle power) could be associated with greater speeds and thus, severity risk [58, 64]. It is possible that the injury severity risk associated to engine powerful cars would reflect the way that vehicles are driven, rather than to inherent characteristics of vehicles engine themselves.

The assessment of the Model-IB-S confirms a good performance, as observed in Table 6.4. When using the training sample, the model correctly predicted 72.4% of the cases. In the training sample, the model correctly predicted 28 severe crashes (TP) and 27 non-severe crashes (TN). When assessing the performance of this model with the original crash sample, the prediction accuracy, was even better than for the training sample, 76%. From a total of 500 crashes, Model IB-S correctly predicted: 14 severe crashes (out of 38) and 365 non-severe events (out of 462). The evaluation of the model overall performance accuracy rate with 10 stratified random samples was also satisfactory; 58.0% (S.D. 3.1). EM output for Model-IB-S is provided in Appendix 8.

The comparison of selected Models for FatalSIK analysis in single-vehicle crashes is presented next. Both models, Model-IA-S and Model-IB-S had identified the effect of vehicle´s age and engine size in crash severity analysis. Models prediction accuracy for the original sample was almost the same for models, 76.4% and 76.0%, for Model-IA-S and Model-IB-S, respectively. However, when evaluating prediction accuracy with 10 stratified random samples, Model-IA-S was slight better than Model-IB-S, 62.0% (S.D. 2.3) and 58.0% (S.D. 3.1), respectively. Regarding to the other two additional models developed for FatalSIK prediction, a brief comparison is presented as follows. Considering, model complexity and comprehensive interpretation, accuracies rates and average estimated values, Models-IA-S and Model-IB-S were better than Model-IC-S and Model-ID-S (Appendix 8). Model-IC-S and Model-ID-S, accuracy rate evaluation with the original sample was also very satisfactory, (76.4% and 79.2%) respectively. Models performance assessment for the 10 stratified random sample was also good: 65.3% (S.D. 2.6) and 56.6% (S.D. 1.9) for Model-IC-S and Model-ID-S, respectively. Despite of these two alternative models had achieved good performance accuracy, they are more complex and hence they would be more complex to apply for real world crash scenarios prediction.

6.3 Concluding Remarks

The presented models for FatalSIK prediction in single-vehicle, Model-IA-S and Model-IB-S, crashes had identified vehicle's characteristics associated with crash severity risk. Model-IA-S with four degrees of freedom and p-value< 0.0004 selected the effect of vehicle´s inherent characteristics (age, engine size and wheelbase) and also crash circumstances (linked to weather conditions) for crash severity prediction. All these selected predictors were statistically significant at 0.05 level, with exception for wheelbase (p-value<0.0593) and the intercept (p-value<0.3023). On the other hand, Model-IB-S with two degrees of freedom and p-value<0.0013 selected the effect of vehicle´s age and engine size for crash severity prediction, with both predictors being statistically significant at 0.05 level. Model-IA-S showed lower MISC rate than Model-IB-S, (0.237 and 0.276). Models prediction accuracy for the original sample was almost the same for Model-IA-S and Model-IB-S, (76.4% and 76.0%, respectively). However, when evaluating prediction accuracy with 10 stratified random samples, Model-IA-S was better than Model-IB-S, 62.0% (S.D. 2.3) and 58.0% (S.D. 3.), respectively. For single-vehicle crash analysis, Model-IB-S is recommended for severity prediction, owing to the three main reasons presented next.

1. This model, the association between the selected predictors (AgeV1 and ccV1) and intercept is stronger than for Model-IA-S parameters, since for the first model all parameters were statistically significant at 0.05 level.
2. Model overall prediction accuracy rate was slightly better for Model-IA-S, Model-IB-S showed better prediction accuracy for the original sample, compared to the training sample, (76.0% and 72.4%, respectively).
3. Model-IB-S is simpler to apply and easy to interpreter.

Comparison of crash severity prediction models using CART and logistic regression is summarized next. Although the input parameters were the same for both techniques, CART model (Figure 5.6 in pp 101) showed the contribution of vehicle characteristics and weather conditions in risk. Small engine size with low weight vehicles and larger engine size in vehicles with smaller wheelbase increased the likelihood of a severe crash. On the other hand, the logistic Model-IB-S (pp 116) identified the age and the engine size of the vehicle as important factors for crash severity prediction. Similarly to CART, larger engine size vehicles were linked to an increased risk.

Often, the selection of statistical models is recommended based on models purpose objective, hence this model provides a good way to judge the practical (as opposed to the statistical) importance of the model in the target of interest prediction [89]. Both models support the conclusion that, for single vehicle crash severity analysis vehicle engine size and age are statistically significant for crash severity prediction. Models results clearly show that recent vehicles protect their occupants better than older vehicles models in the event of a crash. In addition, both models showed good overall prediction accuracy for the original imbalanced data, despite of crash sample limitations.

CHAPTER 7

LOGISTIC MODELS FOR CRASH SEVERITY PREDICTION IN TWO-VEHICLE COLLISIONS

Logistic regression modeling results for the probability of serious injuries and/or fatalities in a crash involving two vehicles is discussed next. Important inputs are ascertained by parameters estimates and odds ratio. The best model to predict the overall crash severity (conveyed as FatalSIK) in two-vehicle collisions was identified as Model I-T. Model II-T was designed to estimate the probability of a serious injured and/or killed in vehicle V1 (expressed by FatalSIKV1). On the other hand, Model III-T was developed to estimate the probability of a serious injured and/or killed in vehicle V2 (defined as FatalSIKV2). This modeling strategy for two-vehicle collisions differentiates from previous modeling approaches mainly for two reasons. Firstly, it integrates new design variables to express the differential of technical characteristics for the two vehicles involved. Secondly, these models were able to model simultaneously and independently the contributing effect of each individual vehicle in the risk of severity sustained by the occupants of the vehicle being analyzed. Models were developed with SAS® v9.2 and SAS®Enterprise Miner™7.2 software [84, 89, 117].

Chapter 7 is organized as follows. First, presentation of best models to estimate the probability of a serious injured and/or killed in the event of two-vehicle collisions. Second, models prediction accuracy and performance assessment, and validation are discussed. Finally, main remarks are summarized.

7.1 Logistic Regression Analysis for FatalSIK with the Original Crash Sample- Imbalanced Data

To prove the problem of prediction accuracy of logistic regression models using the imbalanced sample for two-vehicle collisions, the Model-T is shown in Table 7.1. This model predicted right all the cases of non-severe crashes, (TN=842). However, all the severe crashes were incorrectly predicted as non-severe crashes (FN=32).

Table 7.1 - Imbalanced-Model-T results for FatalSIK prediction with logistic regression performed for the original sample of two-vehicle collisions.

Imbalanced- MODEL-T								
Fit Statistics								
Test for Global Null Hypothesis		**Analysis of Maximum Likelihood Estimates and Odds Ratio**						
DF	Pr<ChSq	**Parameter**	DF	Estimate	SE[1]	Pr>ChiSq	OR[2]	95% CI[3]
3	0.0013	Intercept	1	-4.7726	1.3192	0.0003		(-7.358_ -2.187)
		AlcoholDrugs (0)	1	-1.1648	0.3507	0.0009	0.097	(-1.852_ -0.477)
		Sideswipe (0)	1	-0.5223	0.2258	0.0207	0.352	(-0.965_ -0.08)
		WBV1	1	0.0011	0.0005	0.0198	1.001	(0.0001_0.002)
Obs.	874							
ASE	0.03							
MISC	0.4							
Accuracy Performance								
Accuracy Rate with Training Sample (N=874)								
FN[4]		TN[5]			FP[6]		TP[7]	
32		842			0		0	

1 Standard Error; 2 Odds Ratio Estimate; 3 95% Confidence Interval; 4 False Negative; 5 True Negative; 6 False Positive ; 7 True Positive.

Owing to constrain of the Portuguese crash sample nature and size, the resampling strategy described earlier was applied for the two-vehicle collisions crash severity prediction modeling.

7.2 Logistic Regression Analysis for FatalSIK with Resampling Approach

This section presents the logistic regression modeling results for the probability of a serious injury and/or fatality given any level of injuries in a vehicle crash involving two vehicles. As previously explained in sections 3.7.2 and 3.7.3 of the Safety Analysis Methodology Chapter, several candidate models were developed based on a balanced training sample and three of the best candidate models were selected for further accuracy performance evaluation: first with the original sample (874 observations) and then, using 10 stratified random samples (64 observations). The best three models to predict the overall crash severity in two-vehicle collisions is labeled as: Model IA-T, Model IB-T, and Model IC-T. This model labels are explained as: “I”, designs the model

number for the target of interest, “A, B, and C” indicates the three best model candidates for the target being predicted, and “T” stands for two-vehicle collisions. Among the tree best candidate model to predict FatalSIK, only the recommended model is presented in this section, Model IA-T. The other two best models for FatalSIK prediction, Model-IB-T and Model-IC-T are shown in Appendix 9. Independent variables used as inputs and models’ targets are identified in Table 7.2.

Table 7.2 – Description of design variables (inputs) and targets when modeling crash severity for two-vehicle collisions with logistic regression.

Variable	Description	Abbreviation	Model Identification
Independent Variables Used as Inputs			
Age of Vehicle 1	AgeV1 (yr) was calculated based on the year of the crash event minus the year of the first vehicle registration.	AgeV1	Model-IA-T, Model-IB-T, Model-IC-T, Model-II and Model-III-T
Age of Vehicle 2	AgeV2 (yr) was calculated based on the year of the crash event minus the year of the first vehicle registration.	AgeV2	Model-IA-T, Model-IB-T, Model-IC-T, Model-II and Model-III-T
Age Difference between vehicles (V_2) and (V_1)	AgeV2V1 (yr) stands for age of vehicle V_2 minus the age of vehicle V_1, crash observation.	AgeV2V1	Model-IA-T, Model-IB-T, Model-IC-T, Model-II and Model-III-T
Alcohol and/or Drugs	The Driver´s test for alcohol and or drugs is presented as: Code=0, legal; Code=1, illegal	AlcoholDrugs	Model-IA-T, Model-IB-T and Model-IC-T
Crash type for collisions	Rear End, Head-On, Sideswipe or Other	RearEnd HeadOn Sideswipe Other	Model-IA-T, Model-IB-T and Model-IC-T Model-IA-T, Model-IB-T and Model-IC-T Model-IA-T, Model-IB-T and Model-IC-T -
Divided/undivided	Existence or absence of physical median: Code=0, undivided Code=1, divided	DivisionCode	Model-IA-T, Model-IB-T and Model-IC-T
Speed Level	The speed level was coded as follow: If Speed limit<=90 km.h^{-1}, then code=0 If Speed limit>90 km.h^{-1}, then code=1	SpeedLevel	Model-IA-T, Model-IB-T, Model-IC-T, Model-II and Model-III-T
Wheelbase of Vehicle 1	Wheelbase of vehicle (V_1) (mm)	WBV_1	Model-IA-T, Model-IB-T, Model-IC-T, Model-II and Model-III-T
Wheelbase of Vehicle 2	Wheelbase of vehicle (V_2) (mm)	WBV_2	Model-IA-T, Model-IB-T, Model-IC-T, Model-II and Model-III-T
Wheelbase Difference between vehicles (V2) and (V1)	WBV2V1 stands for wheelbase of vehicle V2 minus the wheelbase of vehicle V1, at crash observation, (mm).	WBV2V1	Model-IA-T, Model-IB-T, Model-IC-T, Model-II and Model-III-T
Weight of Vehicle 1	Weight of vehicle 1 (V1) (kg).	WTV1	Model-IA-T, Model-IB-T, Model-IC-T, Model-II and Model-III-T
Weight of Vehicle 2	Weight of vehicle 2 (V2) (kg).	WTV2	Model-IA-T, Model-IB-T, Model-IC-T, Model-II and Model-III-T
Weight Difference between vehicles (V2) and (V1)	WTV2V1 stands for weight of vehicle V2 minus the engine size of vehicle V1, at crash observation (kg).	WTV2V1	Model-IA-T, Model-IB-T, Model-IC-T, Model-II and Model-III-T
Weather Conditions	Weather conditions at the moment of the crash: Code=0, Clear and/or dry pavement Code=1, rain and/or wet pavement	WeatherCode	Model-IA-T, Model-IB-T and Model-IC-T
Engine Size of Vehicle 1	Engine size of vehicle (V1) (cm^3).	ccV1	Model-IA-T, Model-IB-T, Model-IC-T, Model-II and Model-III-T
Engine Size of Vehicle 2	Engine size of vehicle (V2) (cm^3).	ccV2	Model-IA-T, Model-IB-T, Model-IC-T, Model-II and Model-III-T
Engine Size Difference between vehicles (V2) and (V1)	ccV2V1 stands for engine size of vehicle V2 minus the engine size of vehicle V1, at crash observation, (cm^3).	ccV2V1	Model-IA-T, Model-IB-T, Model-IC-T, Model-II and Model-III-T
Dependent Variables used as Targets			
Serious injured and/or killed	FatalSIK is a categorical response for a crash outcome used to predict either a serious injury, or fatality in a crash event. FatalSIK=1, if SI>0 and/or K>0, else, FatalSIK=0	FatalSIK	Model-IA-T, Model-IB-T, Model-IC-T, Model-II and Model-III-T
Serious injured and/or killed for vehicle 1 (V1) occupants	FatalSIKV1 is a categorical response for a crash outcome used to predict either a serious injury, or fatality or both for occupants in vehicle 1 in a crash event. FatalSIKV1=1, if SI>0 and/or K>0, else, FatalSIKV1=0	FatalSIKV1	Model-II-T
Serious injured and/or killed for vehicle 2 (V2) occupants	FatalSIKV2 is a categorical response for crash outcome for a crash outcome used to predict either a serious injury, or fatality or to both for occupants in vehicle 2 in a crash event. FatalSIKV2=1, if SI>0 and/or K>0, else, FatalSIKV2=0	FatalSIKV2	Model-III-T

Regarding to Model-IA-T design, it was developed using logistic regression for FatalSIK prediction in crashes involving two-vehicle collision, with forward selection for the inputs signalized at Table 7.2. During the forward selection, the modeling begins with no candidate inputs in the model and adds inputs until the entry significance level is met. Table 7.3 summarizes Model-IA-T fitting results and performance evaluation. The test for the global null hypothesis shows that at least one of the predictor´s regression coefficient is not equal to zero in the model, p-value <0.0054. From a total of 19 potentially explanatory variables exanimated with forward selection method, only two of them are found to be statistically significant at 0.05 level. Age of vehicle V1 (AgeV1) and non-head-on collisions are significant factors to estimate the crash severity; p-values 0.0084 and 0.0346, respectively. In this model, as the age of vehicle V1 increases, the risk of a severe crash outcome is lower. Also, crashes rather than head-on collisions were associated with a decrease in crash severity. In the sample, just 7.6% of the crashes were as head-on collisions (67/874). The remaining 808 observations were distributed as follows: 311 were rear end collisions, 89 were sideswipe collisions and 408 were reported as other. Those head-on collisions have resulted 12% (4/32) of severe events in the crash dataset. The model fit statistics shows the following values of 0.211 and 0.328, for the ASE and MISC, respectively.

Table 7.3 - Model-IA-T results for FatalSIK prediction with logistic regression performed for a balanced dataset of two-vehicle collisions.

MODEL-IA-T								
Fit Statistics								
Test for Global Null Hypothesis		**Analysis of Maximum Likelihood Estimates and Odds Ratio**						
DF	Pr<ChSq	Parameter	DF	Estimate	SE[1]	Pr>ChiSq	OR[2]	95% CI[3]
2	0.0054	Intercept	1	2.6230	0.9736	0.0071		(0.7147_ 4.5312)
		AgeV1	1	-0.1769	0.0671	0.0084*	0.838	(-0.3084_ -0.0454)
		HeadOn (0)	1	-1.3964	0.6610	0.0346*	0.061	(-2.6920_ -0.1008)
Obs.	64							
ASE	0.211							
MISC	0.328							

Prediction Accuracy Performance											
Accuracy Rate with Training Sample (N=64)					Accuracy Rate with Original Sample (N=874)					Prediction Accuracy for 10 Stratified Random Samples	
FN[4]	TN[5]	FP[6]	TP[7]	AR[8]%	TPs[9]	FPs[10]	TNs[11]	FNs[12]	AR[13]%	Mean%[14]	S.D.[15]
10	21	11	22	67.2	8	148	694	24	80.3	54.4	1.7

NOTA: [1] Standard Error; [2] Odds Ratio Estimate; [3] 95% Confidence Interval; 4 False Negative; 5 True Negative; 6 False Positive; 7 True Positive; 8 Percentage of Accuracy Rate; 9 True Positives; 10 False Positives; 11 True Negatives; 12 False Negatives; 13 Percentage of Accuracy Rate; 14 Mean of Prediction Accuracy for the 10 stratified random samples; 15 Standard Deviation for the Prediction Accuracy of the 10 stratified random samples; *Statistically significant at 5% level.

The logistic regression equation developed for Model-IA-T is presented next.

$$P\,(FatalSIK = 1) = \frac{\exp(2.623 - 0.1769 * AgeV1 - 1.3964 * HeadOn(= 0))}{1 + \exp(2.623 - 0.1769 * AgeV1 - 1.3964 * HeadOn(= 0))} \qquad \text{Equation 7.1}$$

The interpretation of the Model-IA-T with the odds ratio, in Table 7.3, shows that the odds of a FatalSIK crash in a non-head-on collision is 0.061 the odds in a head-on collision. In other words,

the odd of a severe crash increases by 16 times for head-on collisions. Also, the odds for the continuous variable AgeV1, 0.838, shows that an increased risk of a FatalSIK is associated with the decrease for the age of vehicle V1. Figure 7.1 shows a graphical representation of crash severity prediction estimates logit curve using Model-IA-T equation. As observed, the logit curve for the estimated probability of FatalSIK for head-on collisions is higher than the estimated target values for all the others crash types, such as, rear-end and sideswipe.

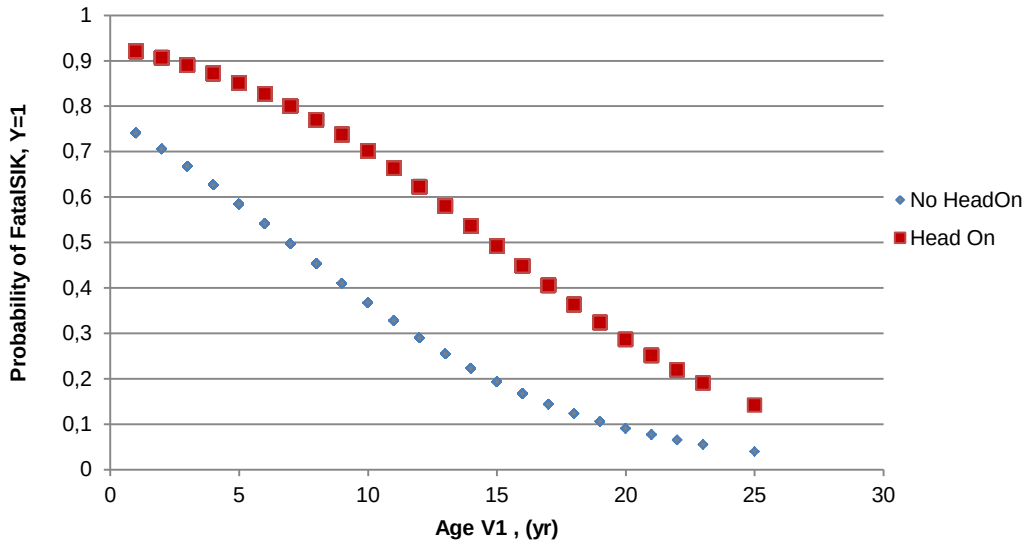


Figure 7.1 – Probability of a serious injury and/or fatality with age of vehicle V1, in two-vehicle collisions, using Model-IA-T.

Model-IA-T shows a positive effect of head-on collision in crash severity risk (or a negative effect of non-head on collisions), which is supported by several other works which found head-on collision contribution to more severe injuries levels [48, 92, 144]. The most severe crash configuration is front-to-side impact, which imposes a higher risk of being killed in the side-impacted vehicle [49]. On the other hand, as the age of the vehicle V1 increases, the overall crash severity tends to decrease. Some studies have related newer vehicle models with an increased risk for the accounts of the other vehicle involved [49, 59, 60, 85]. Thus, as age of V1 increases, it would be possible that the vehicle would be less “aggressive” during the event of a collision. Hence, the occupants of the other vehicle involved would face a lower risk, and it could contribute to a decrease in the overall crash severity.

The assessment of the Model-IA-T shows a good performance, as observed in Table 7.3. The accuracy rate when running the model with the training sample, which was stratified in 32 severe crashes and 32 non severe crashes, correctly predicted 67.2% of the cases. In the training sample, the model correctly predicted 22 severe crashes (TP) and 21 non severe crashes (TN). When

compared with the previous model in Table 7.1, Imbalanced-Model-T, it is straightforwardly to notice the improvement in model accuracy prediction. The model developed with the original imbalanced sample predicted all the severe crashes as non-severe, leading to unsatisfactory results for TP, which were none. On the other hand, Model-IA-T performed with the balanced approach, was able to predict right 22 severe crashes, out of 32. When assessing the performance of this model within the original crash sample, it shows high prediction accuracy, 80.3%. From a total of 874 crashes observations, Model IA-T correctly predicted 8 severe crashes out of the 32 severe collisions. In addition, the model correctly predicted 694 of the non-severe out of the observed 842 non-severe events in the entire dataset. The model predicted right more severe crashes in the training sample, than in the original dataset, (22 and 8, respectively). However, it is noted that the model overall accuracy within the original population was higher than the model accuracy within the training sample, (80.3% and 67.2%, respectively). The evaluation of the model performance with 10 stratified random samples was also satisfactory; the mean prediction accuracy rate was 54.4% (S.D. 1.7). EM output for Model-IA-T is provided in Appendix 9.

Model-IA-T, when compared with the FatalSIK prediction candidates, Model-IB-T and Model-IC-T (Appendix 9) showed slight lower prediction accuracy, 80.3%, 82.6% and 82.8%, respectively. Model-IA-T was selected because its prediction accuracy was good and since it is easier to interpret, its application in real world crash scenarios could be more helpful.

7.3 Logistic Regression Analysis for FatalSIKV1 and FatalSIKV2 with Resampling Approach

The original crash sample included a limit number of severe cases for FatalSIKV1 (21 cases) and FatalSIKV2 (14 cases). Whereas the resampling strategy was not applied for CART modeling of those targets due to this technique sensitivity to sample size (as explained in Chapter section 5.3), the resampling approach was applied for the logistic modeling.

To model FatalSIKV1 and FatalSIKV2, the design variables focus in each individual vehicle characteristics, and differential of vehicle characteristics. In addition, the variable SpeedLevel was also used as input during the modeling stage, since speed is known as increasing risk of injury level [42, 43, 91, 93, 98]. It must be mentioned that only the best models for FatalSIKV1 and FatalSIKV2 are presented. During the modeling stage several candidate models were developed using the same design variables, a total of 19 predictors, as they were used for FatalSIK modeling, such as AlcoholDrugs, DivisonCode, WeatherCode and variables related to crash type. However those models showed a poor performance and only the best models for each target are discussed in section 7.3. For FatalSIKV1 and FatalSIKV2 models, the inputs were the same, 13 independent variables, and models' targets, are identified in Table 7.2. Interpretation of FatalSIKV1 and FatalSIKV2 logistic models: Model-II-T and Model-III-T, respectively, is next.

7.3.1 Model-II-T Analysis

Model-II-T was developed using logistic regression for FatalSIKV1 prediction models with a balanced training sample, which was stratified in 21 severe crashes and 21 non severe crashes, for two-vehicle collisions. Forward method was used for selection of the inputs in Table 7.2. Forward method was used for selection of the inputs in Table 7.2. Since the model development was based on sample training containing a limited number of observations (42 crashes), the 5% level was not applied, but 10% level. Therefore, model entry level was set to 10% (p-value<0.1). Statistical support is provided bellow.

Some researchers argued to use smaller significance levels as the sample size increases, partly to offset the fact that standard errors are getting smaller. Some authors feel comfortable using 5% level when is a few hundred, thus they might use 1% level when n is a few thousand [146]. Additional information for Model-II-T is provided in Appendix 9. As previously mentioned in the previous chapter (section 6.2.2), for samples with smaller size, some authors are willing to use larger significance levels, reflecting the fact that it is harder to find significance with smaller sample sizes (the estimators are less precise). For small sample sizes, it can be use a larger p-value, as 0.2, but there is no hard rules [84, 146].

Table 7.4 summarizes Model-II-T fitting results and performance evaluation. As explained above, due to the small training sample size (N=42), a larger p-value is used, p-value<0.1. The test for the global null hypothesis shows that at least one of the predictor´s regression coefficient is not equal to zero in the model, p-value <0.0594. From a total of 13 variables entered as inputs during the modeling stage, only the engine size of the opponent vehicle is statistically significant at 0.10 level, p-value<0.0762.

Table 7.4 - Model-II-T results for FatalSIKV1 prediction with logistic regression performed for a balanced dataset of two-vehicle collisions.

MODEL-II-T								
Fit Statistics								
Test for Global Null Hypothesis		**Analysis of Maximum Likelihood Estimates and Odds Ratio**						
DF	Pr<ChSq	Parameter	DF	Estimate	SE[1]	Pr>ChiSq	OR[2]	95% CI[3]
1	0.0594	Intercept	1	-2.0657	1.1961	0.0842		(-4.4101_0.2786)
		ccV2	1	0.00108	0.0006	0.0762	1.001	(-0.0001_0.0023)
Obs.	42							
ASE	0.239							
MISC	0.357							

Prediction Accuracy Performance											
Accuracy Rate with Training Sample (N=42)					Accuracy Rate with Original Sample (N=874)					Prediction Accuracy for 10 Stratified Random Samples	
FN[4]	TN[5]	FP[6]	TP[7]	AR[8]%	TPs[9]	FPs[10]	TNs[11]	FNs[12]	AR[13]%	Mean%[14]	S.D.[15]
10	16	5	11	64.3	6	41	812	15	93.6	61.2	2.4

1 Standard Error; 2 Odds Ratio Estimate; 3 95% Confidence Interval; 4 False Negative; 5 True Negative; 6 False Positive; 7 True Positive; 8 Percentage of Accuracy Rate; 9 True Positives; 10 False Positives; 11 True Negatives; 12 False Negatives; 13 Percentage of Accuracy Rate; 14 Mean of Prediction Accuracy for the 10 stratified random samples; 15 Standard Deviation for the Prediction Accuracy of the 10 stratified random samples; *Statistically significant at 10% level.

The logistic regression equation developed for Model-II-T is presented next.

$$P\,(FatalSIKV1 = 1) = \frac{\exp(-2.0657 + 0.00108 * ccV2)}{1 + \exp(-2.0657 + 0.00108 * ccV2)} \qquad \text{Equation 7.2}$$

The signs of coefficient estimates are directly related to their influence on probability of the target being modeling. As can be observed in Table 7.2, the estimate for ccV2 has a positive sign (0.00108). Graphical representation Equation 7.2 for Model-II-T is illustrated in Figure 7.2, showing that as the engine size of vehicle V2 increases, the probability of severe injury sustained by the occupants of vehicle V1 also increases.

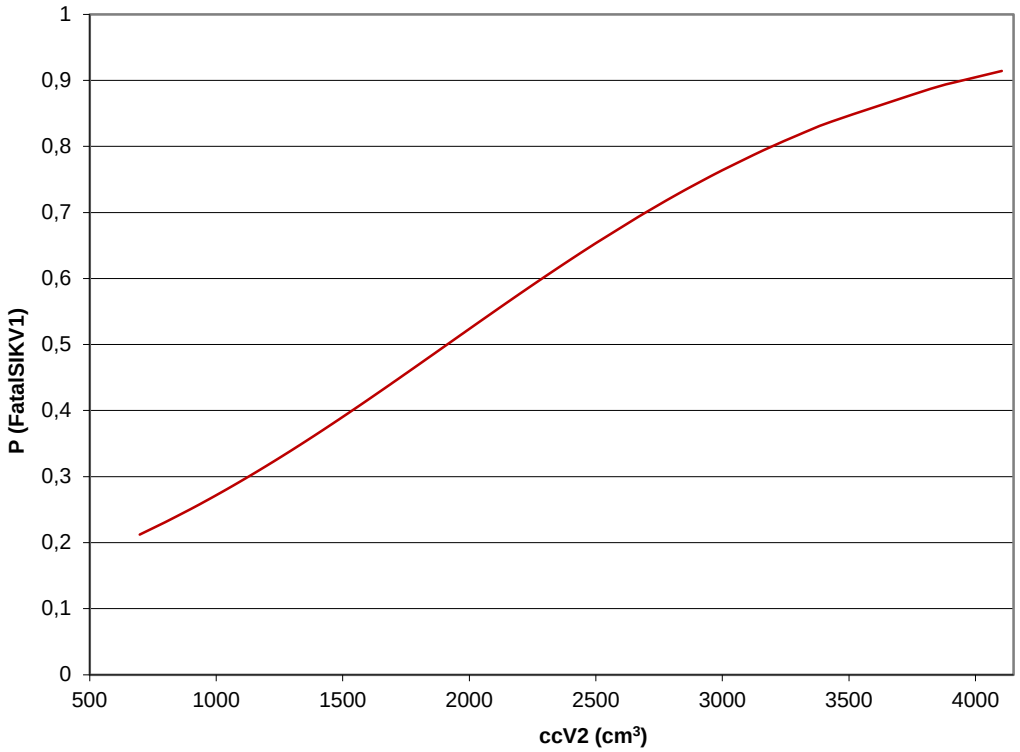


Figure 7.2 –Estimated probability of a serious injury and/or killed among the occupants of vehicle V1 with the engine size of the opponent vehicle, ccV2, in two-vehicle collisions, using Model-II-T.

In the interpretation of Model-II-T revealed that when analyzing the risk that occupants of the subject vehicle V1 are exposed, the model did not select any variable related to this vehicle crashworthiness, rather a variable that seems to be related to the “risk agressivitive” imposed by the other vehicle involved in the collision. The engine size of the opponent vehicle increases the probability of major injuries and/or fatalities among the occupants of the subject vehicle. It is possible that effect of mass of the opponent vehicle could be reflected in vehicle´s engine size. Often, vehicles with larger engine size are heavier; hence it would be expected higher risk following the collision. Model-II-T results are supported by previous work which agree that in a two-vehicle collisions severity risk rises with size and mass of the other vehicle involved [53, 60, 87, 145].

The assessment of the Model-II-T shows a good performance, as observed in Table 7.3. The accuracy rate when running the model with the training sample, correctly predicted 64.3% of the cases. In the training sample, the model correctly predicted 11 severe crashes (TP) and 16 non severe crashes (TN). When assessing the performance of this model within the original crash sample, it shows great prediction accuracy, 93.6%. From a total of 874 crashes observations, Model II-T correctly predicted 6 severe crashes out of the 21 severe collisions. In addition, the model correctly predicted 694 of the non-severe out of the observed 842 non-severe events in the

entire dataset. As expected, the model predicted right more severe crashes in the training sample, than in the original dataset, 11, and 6, respectively. However, the model overall accuracy within the original population was higher than the model accuracy within the training sample, 93.6% and 64.3%, respectively. It must be pointed out that this model was able to predict six out of the 21 severe cases in vehicle V1 for the entire sample containing 853 non-severe cases and only 21 severe cases. As a matter of fact, at the original sample, the non-severe outcomes for individual vehicle V1 were almost 41 times more frequent than severe outcomes (853/21). As a validation approach for the model discussed in this section, the evaluation of the Model-II-T performance with 10 stratified random samples is also satisfactory; showing a mean prediction accuracy rate of 61.2% (S.D. 2.4).

7.3.2 Model-III-T Analysis

Model-III-T was developed using logistic regression for FatalSIKV2 prediction models with a balanced training sample stratified in 14 severe crashes and 14 non severe crashes, for two-vehicle collisions. Backward method was used for selection of the inputs in Table 7.2. Additional information for Model-III-T is provided in Appendix 9. Table 7.5 summarizes Model-III-T fitting results and performance evaluation. The test for the global null hypothesis shows that at least one of the predictor´s regression coefficient is not equal to zero in the model, p-value <0.0201. From a total of 13 variables entered as inputs during the modeling stage, only the engine size of the opponent vehicle is found to be statistically significant at 0.05 level, p-value<0.0387.

Table 7.5 - Model-III-T results for FatalSIKV2 prediction with logistic regression performed for a balanced dataset of two-vehicle collisions.

MODEL-III-T											
Fit Statistics											
Test for Global Null Hypothesis		**Analysis of Maximum Likelihood Estimates and Odds Ratio**									
DF	Pr<ChSq	Parameter	DF	Estimate	SE[1]		Pr>ChiSq	OR[2]	95% CI[3]		
1	0.0201	Intercept	1	-3.5969	1.78		0.0433		(-7.0856_-0.1082)		
		ccV1	1	0.00205	0.0010		0.0387	1.002	(-0.0001_0.0040)		
Obs.	28										
ASE	0.231										
MISC	0.286										
Prediction Accuracy Performance											
Accuracy Rate with Training Sample (N=28)					Accuracy Rate with Original Sample (N=874)					Prediction Accuracy for 10 Stratified Random Samples	
FN[4]	TN[5]	FP[6]	TP[7]	AR[8]%	TPs[9]	FPs[10]	TNs[11]	FNs[12]	AR[13]%	Mean%[14]	S.D.[15]
4	10	4	10	71.4	5	133	727	9	83.8	40.5	2.1

1 Standard Error; 2 Odds Ratio Estimate; 3 95% Confidence Interval; 4 False Negative; 5 True Negative; 6 False Positive; 7 True Positive; 8 Percentage of Accuracy Rate; 9 True Positives; 10 False Positives; 11 True Negatives; 12 False Negatives; 13 Percentage of Accuracy Rate; 14 Mean of Prediction Accuracy for the 10 stratified random samples; 15 Standard Deviation for the Prediction Accuracy of the 10 stratified random samples; *Statistically significant at 10% level.

The logistic regression equation developed for Model-III-T is presented next.

$$P\ (FatalSIKV2 = 1) = \frac{\exp(-3.5969 + 0.00205 * ccV1)}{1 + \exp(-3.5969 + 0.00205 * ccV1)} \qquad \text{Equation 7.3}$$

The signs of coefficient estimates are directly related to their influence on probability of the target being modeling. As can be observed in Table 7.5, the sign of ccV1 estimate has a positive sign, showing that its effect is associated with an increase probability for FatalSIKV2. Graphical representation for this Model-III-T equation (Equation 7.3) is illustrated in Figure 7.3, showing that as the engine size of vehicle V1 increases, it raises the probability of severe injuries in the occupants of vehicle V2 also increases.

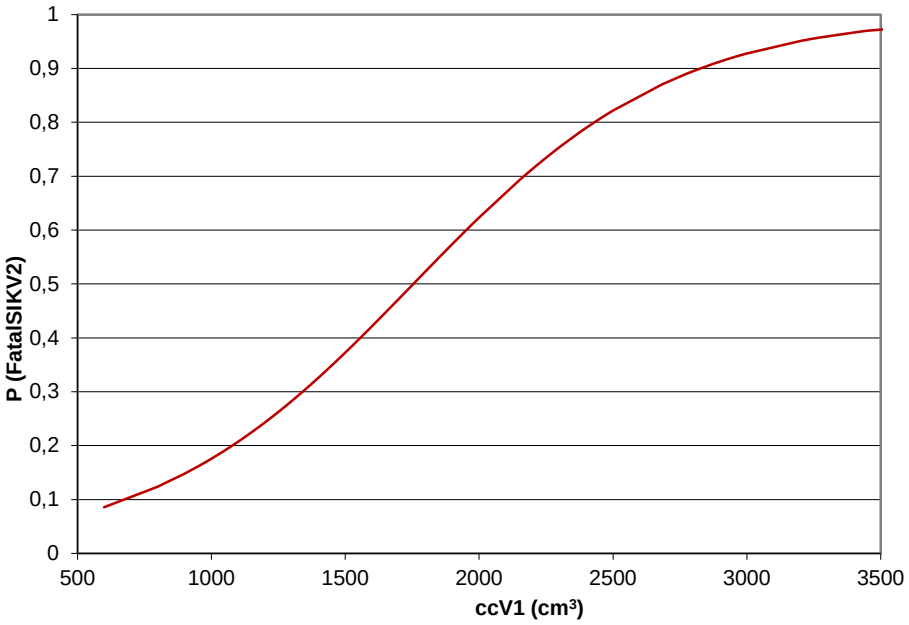


Figure 7.3 – Estimated probability of a serious injury and/or fatality among the occupants of vehicle V2 with the engine size of the opponent vehicle, ccV1, in two-vehicle collisions, using Model-III-T.

Similarly to Model-II-T, Model-III-T shows the effect of a predictor that seems to express the risk imposed by the other vehicle involved in the collision, ccV1. As previously mentioned, Model-III-T fining supports other research that had identified the size of the opponent vehicle (which encompasses vehicle mass, engine size and length) as a risk factor for serious injuries and/or fatalities among the occupants of the other vehicle involved in the collision [53, 60, 87, 145].

The assessment of the Model-III-T shows a great performance, as observed in Table 7.5. The accuracy rate with the training sample, correctly predicted 71.4% of the cases. In the training sample, the model correctly predicted 10 severe crashes (TP) and 10 non severe crashes (TN). When assessing the performance of this model within the original crash sample, it shows good prediction accuracy, 83.8%. Based on the 874 collisions observations, it is important to notice that

Model-III-T was able to predict 5 out of the 14 severe cases of the original sample containing only 1.6% cases for severe crash outcomes in vehicle V2. In addition, the model correctly predicted 727 of the non-severe cases out of the observed 842 non-severe cases in the entire sample. As expected, the model predicted right more severe crashes in the training sample, than in the original dataset, since the first was a balanced dataset; the second was the original sample that was highly imbalanced. As a validation approach for Model-III-T, the evaluation performance for the 10 stratified random samples was also suitable; showing a mean prediction accuracy rate of 40.5% (S.D. 2.1).

Following the discussion of Model-II-T for FatalSIKV1 prediction and Model-III-T for FatalSIKV2, the consistency of both models is analyzed. As previous explained each of these models targets to predict the probability of a serious injured and/or killed in the subject vehicle, by modeling this vehicle crashworthiness simultaneously with the opponent vehicle risk. Both models found the engine size of the opponent vehicle as a significant factor contributing towards an increased risk of severity injuries sustained by the occupants of the vehicle being analyzed. Figure 7.4 integrates the effect of engine size in crash severity risk for each vehicle involved in a two-vehicle collision.

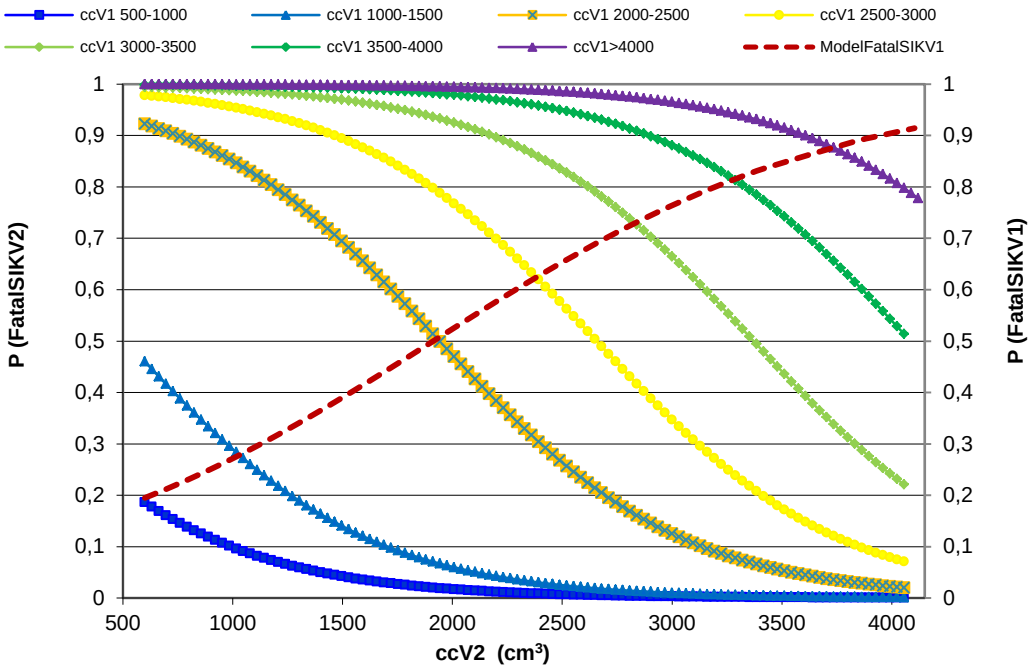


Figure 7.4 – Effect of engine size of the opponent vehicle in the probability of a serious injury and/or fatality among the occupants of vehicle being analyzed, in two-vehicle collisions.

In Figure 7.4, the logit curve for Model-II-T is presented in red and denotes an increasing in crash severity risk for V1 as the engine capacity of the other vehicle involved increases. Whereas, for

Model-III-T several curves illustrates how crash severity risk for V2 varies with several categories of opponent vehicles engine size, ccV1 in series. It must be noticed that, since this methodology strategy design takes into account not only own vehicle protection, but simultaneously, the risk caused by the opponent vehicle, it cannot be “directly” compared with previous research because in the literature few studies have considered the effect of the other vehicle involved and those did not integrate simultaneously each individual vehicle contribution. However, the findings for these models support previous work, that in a two-vehicle collisions severity risk rises with size and mass of the other vehicle involved [53, 60, 87, 145].

Regarding to Model-II-T and Model-III-T fit statistics to the crash sample, it was notice that the first model the parameters were significant at 10% level, even though it had a larger training sample size (N=42). On the other hand, for Model-III-T, the parameters were found significant at 5% significant level, even though the training sample was very small, only 28 observations. Despite of the smaller sample size (with only 14 severe events for the target being predicted), Model-III-T showed a lower misclassification rate than Model-II-T, (0.286 and 0.357, respectively). Even though both models showed a good prediction performance, care must be present in the interpretation of these models because logist regression modeling was performance with very small samples.

7.4 Concluding Remarks

This Chapter presented logistic regression results to examine the probability of a serious injured and/or killed as an outcome of two-vehicle collisions. Logistic regression performed with the original imbalanced sample, Imbalanced-Model-T, revealed poor accuracy performance (with none true positive (TP) for crash severity prediction). On the other hand, re-sampling procedure adopted for the logistic modeling (by randomly removing the majority class of non-severe cases to a balanced proportion of severe cases) has resulted in improvements in TP without increasing significantly the FP. Some training information is lost, but it is counterbalanced by the improvements in the minority class accuracy, i.e, crash severity prediction.

Following logistic models presentation, the Model-IA-T is recommended to predict the overall crash severity following a collision. Regarding to Model-IA-T findings, it shows that when a collision involves an older vehicle, the risk of a severe crash outcome decreases. Newer vehicles models are known to show improved crashworthiness, however they also have been linked to impose higher “agressitivity” towards the occupants of the other vehicle involved in the collision. As expected, head-on collision contribute to more severe injuries levels. When analyzing the risk of severity to each individual vehicle involved in two-vehicle collisions, Models-II-T and Models-III-T are recommended. Model-II-T targets the prediction of a serious injuries and/or fatality for occupants of vehicle V1. Model-III-T targets the prediction of a serious injuries and/or fatality for occupants of vehicle V2. Both models are consistent and both reinforce the finding that the engine size of the opponent vehicle involved in the collision is a significant variable in explaining crash severity suffered by the occupants of the vehicle being analyzed.

CHAPTER 8

VEHICLE EMISSIONS MODELING

This Chapter aims at the analysis of the vehicle's emissions and fuel consumption. Based on this analysis, the vehicle environmental performance will be developed for further application in vehicle integrated analysis presented in Chapter 9.

Firstly, it briefly summarizes the methodology applied for pollutant emissions and fuel consumption estimation. Secondly, it centers on pollutant vehicle's emissions modeling. It begins by explaining the design methodology to fit emissions estimations results to linear regression models. Then, it highlights the most relevant trends for pollutant vehicles' emissions and fuel use for the vehicles included in the crash database. Thirdly, it presents vehicle's emissions models for selected pollutants. Main remarks present key findings for vehicle's emissions models developed based on the sample explored in this Thesis.

8.1 Methodology

This section summarizes the methodology applied to estimate pollutant vehicle emissions and fuel consumption for the vehicles included in the crash dataset. For the purpose of this research, the CORINAIR methodology was applied [147], which is based on the European emission standards that are related to the acceptable limits for emissions of new vehicles sold in EU member states Since the crashes in the sample occurred in main roads and motorways, it was assumed that engines were in stabilized operation. Thus, CORINAIR methodology was used to estimate "hot" emissions, which better reflect the driving conditions for the vehicles registered in the crash dataset since the majority of sample represents highway or motorway driving. For the environmental performance analysis of the vehicles included in the crash database, carbon monoxide (CO), nitrogen oxide (NO_x) and particle matter (PM) were selected. In addition to the above pollutants, CORINAIR methodology was also applied to estimate the fuel consumptions, based on the CO_2 emissions. The PM emissions factors refer to $PM_{2.5}$ coarse fraction. This choice is justified because this fraction travel deeper in the lungs and are more toxic, so these can have worse health effects.

8.1.1 Vehicles classification

As explained earlier in Chapter 3, this research focuses exclusively on the analysis of crash reports involving light vehicles. For the estimations of emissions and fuel consumption for light passenger vehicles (LPV) and light duty vehicles (LDV) based on CORINAIR methodology, the following inputs were used:

- Vehicle category;
- Fuel type;
- Engine size category;
- Technology level (Emission standard);
- Average speed;
- And driving share.

The above inputs are explained as follows.

Henceforth, CORINAIR methodology was applied to the following two vehicles categories in the dataset: LPV and LDV which weight is lower than 3.5 tons. For these vehicles fuel type are subdivided into: gasoline (G), diesel (D), liquefied petroleum gas (LPG) and hybrid (H). The LPV with gasoline are then subdivided by the engine size (c.c.) into three categories, whereas LPV with diesel are subdivided into two categories.

Following CORINAIR methodology, road vehicles are usually classified according to their level of emission control technology, which is actually defined in terms of the pollutant emission legislation with which they are compliant. Table 8.1 summarizes the vehicle technology (emissions standards) based on CORINAIR methodology [147] used in this work. The nomination "ECE" and "Euro"

reflect the legislative regulation, and “Improved conventional” or “conventional” refer to applied technology. In 1992, “Euro” standards became mandatory in all Members States.

Table 8.1 – Vehicles legislation technology adopted by CORINAIR [147].

Vehicle Category	Fuel type	Engine Size	Legislation/Technology	
Light Passenger Vehicle	Gasoline	<1.4L 1.4-2.0L >2.0L	ECE 15/00-01	
			ECE 15/02	
			ECE 15/03	
			ECE 15/04	
			Improved conventional	
			Euro 1	91/441/EEC
			Euro 2	94/12/EC
			Euro 3	98/69/EC Stage 2000
			Euro 4	98/69/EC Stage 2005
			Euro 5	EC 715/2007
	Diesel	<2.0L >2.0L	Conventional	
			Euro 1	91/441/EEC
			Euro 2	94/12/EC
			Euro 3	98/69/EC Stage 2000
			Euro 4	98/69/EC Stage 2005
			Euro 5	EC 715/2007
	LPG	-	Euro 1	91/441/EEC
			Euro 2	94/12/EC
			Euro 3	98/69/EC Stage 2000
			Euro 4	98/69/EC Stage 2005
			Euro 5	EC 715/2007
	Hybrid	<1.6L	Euro 4	98/69/EC Stage 2005
Light-duty vehicles	Gasoline	<3.5t	Conventional	
			Euro 1	93/59/EEC
			Euro 2	96/69/EC
			Euro 3	98/69/EC Stage 2000
			Euro 4	98/69/EC Stage 2005
			Euro 5	EC 715/2007
	Diesel	<3.5t	Euro 1	93/59/EEC
			Euro 2	96/69/EC
			Euro 3	98/69/EC Stage 2000
			Euro 4	98/69/EC Stage 2005
			Euro 5	EC 715/2007

Euro 1 was officially introduced by Directive 91/441/EEC in July 1992. In the subsequent years, new legislatives steps leads to Euro 2 to Euro 5 and Euro 6, with more restrictions in emissions levels and succeeding pollutants reductions. Euro 5 emissions standards came into effective in September 2009, leading to further 25% reduction NO_x, compared to Euro 4. Euro 6 was not represented in the above table since the vehicle in the crash database were previous to the introduction of this European emission standard.

Emissions control-technology for LDV follows the technology for LPV with a delay of one or two years. For LPG category, vehicles were grouped as conventional for those vehicles prior to 91/441/EEC. Otherwise, the same Euro norms were applied as those relating to gasoline and diesel cars. The legislation classes for hybrid vehicles comply with the Euro 4 and Euro 5 European Emissions Standards [147].

Another, required input in the CORINAIR methodology is the average speed, however, since the vehicle real speed is unknown, the legal speed limit was used as a proxy of vehicle speed (see more details on Chapter 4). The speed profile was obtained since the police reports provide information for the road name and road type. Input speed values were: 50 km.hr^{-1}, 90 km.hr^{-1}, 100 km.hr^{-1} and 120 km.hr^{-1} for vehicles involved in crashes at: urban roads, main roads, complementary roads and motorways.

8.1.2 Emission and fuel consumption estimation

Vehicle emissions are strongly dependent on the engine operation conditions. Emissions depend on several factors, such as: distance that the vehicle travels, its speed, road type, vehicle's age, vehicles engine size and weight. Vehicle speed has a major influence on exhaust emissions as well as in the fuel consumption. Equation 8.1 represents the formula for estimating hot emissions (g.km^{-1}) for a generic pollutant [147].

$$Emission\,(g) = EmissionFactor\left(\frac{g}{km}\right) * number\ of\ vehicles(vehicles) * mileage\ per\ vehicle\left(\frac{km}{vehicle}\right) \quad \text{Equation 8.1}$$

The CO_2 was obtained directly from the fuel consumption. Further detail for each selected pollutant and its emissions factors based on the vehicle category covered in this study are found in the CORINAIR methodology [147].

8.1.3 Modelling vehicle's environmental performance

Subsequently, to the application of CORINAIR for emissions estimation, the obtained data was fit in order to develop models for further application in vehicle's environmental performance, as part of the vehicle's integrated analysis, presented in Chapter 9.

Although CORINAIR methodology is valuable, it requires specific iterations and are time consuming. Therefore, vehicle's environmental performance evaluation would benefit from having access to straight forward mathematical equation models for emissions estimation. As a starting point, emissions data was obtained for the 2,248 vehicles included in the dataset. Following, for each vehicle category and fuel type (section 8.1.1) a methodology approach was used to develop estimation models for the selected targets: CO_2, CO, NO_x and PM and fuel consumption, using as inputs vehicle's engine size category, speed and Norm, among others. Since those targets pollutants are continuous variables, linear regression was selected for modelling [146].

A linear regression model is described by the following equation:

$$Y = \beta 0 + \beta 1 * \mathrm{x1} + \ldots + \beta \mathrm{k} * \mathrm{xk} \quad \text{Equation 8.2}$$

Where Y is the response variable (target being modelling), β0 is the intercept, β1 is the estimate for the parameter x1, and so one. The linear regression is broadly used for estimations modelling of

continuous targets. Since all these targets (selected air pollutants) were continuous variables, linear regression modelling was selected [84, 89, 146]. The estimation modelling with linear regression approach is explained next, through step 1 to step 4. Models were developed using SAS® v9.2 and SAS®Enterprise Miner™7.2 software [84, 89, 117].

Step 1: ***Setting Emissions Training Database***

At the original emissions estimation database covering a total 2,248 vehicles, some vehicles' categories were represented by few observations and where be removed in order to avoid bias [146]. Hence, models training were performed based on a database which covered the emissions estimation values for 2,236 vehicles.

Step 2: ***Reorganizing Training Database by Vehicle Category and Fuel Type***

Training data was organized by vehicle category and fuel type. Hence the training database from step 1 (N=2,236 vehicles) was split yielding to the following groups:

- Light Passenger Gasoline Vehicles (LPGV): N=889 vehicles;
- Light Passenger Diesel Vehicles (LPDV): N=769 vehicles;
- Light Duty Diesel Vehicles (LDDV): N=556 vehicles;
- and Light Duty Gasoline Vehicles (LDGV): N=22 vehicles.

The dataset referring to LDGV has only 22 vehicles and was not used due to the insufficient limited number of observations. Following, each dataset was addressed for model the most relevant pollutants associated with vehicles category. Despite of improvements due to catalytic converters, gasoline engines have been associated with higher CO emissions. On the other hand, diesel engines have been associated with significant emissions rates for NO_x and PM than gasoline engines [125, 148]. Diesel engines generally produce larger amounts of NO_x than gasoline engines due to higher combustion temperatures. Also, they emit greater amounts of PM. Since CO results from the incomplete combustion of vehicle fuels, gasoline engines emit a lighter proportion of CO than diesel engines, due to the lower combustion temperatures. Thus, the LPGV dataset was used to model CO emissions, whereas, LPDV and LDDV datasets were used to model NO_x and PM emissions. On the other hand, for CO_2 emissions modeling, the three datasets (LPGV, LPDV and LDDV) were used in order to address fuel consumption for those categories.

Step 3: ***Linear Regression Modeling***

The response variables (targets) with interest for this study were: CO_2, CO, NO_x, and PM. The explanatory/predictor variables (inputs) used during the modeling stage were: engine size category

(cc), wheelbase (WBV1), weight (WTV1), technological level (Norm) and speed limit (SpeedLimit, was used as a proxy of traveling speed). For each pollutant, several candidates' models were developed and the best models were selected using the goodness-of-fit measures to the three datasets mentioned in step 2.

Step 4: ***Assessment of explanatory variables belonging to the model***

For the final model assessment, adjusted R-Square (Adj R-Sq) parameter was used for evaluation of goodness-of-fit and the analysis of the maximum likelihood estimates (AMLE) for evaluation of parameters and to test its statistics significance in the model [89]. As an example, the Enterprise Miner output for CO_2 modeling based on LPGV dataset is illustrated in Figure 8.1. The analysis of variance and effects showed p-value<0.0001. Model fit statistics revealed Adj R-Sq explained 94.2% of the variation in the CO_2 estimations. However, the AMLE displays a non-statistically significant value for the parameter Euro 4 (Euro IV in the Figure 8.1), p-value <0.3799. Thus, any variable/parameter that is not statistically significant must be analyzed individually in order to keep that parameter in the model or removed it from the model. The factor that should determine whether an explanatory variable belongs in a model is whether the explanatory variable has a nonzero partial effect on "Y" in the population, which means, its population coefficient is zero [146]. During the modeling phase some variables were dropped from the model, those that were not adding useful information to the variability of the response variable [85]. Following Wooldrige, Tolouei and Al-Ghamdi, the observations related to LPGV with Norm 4 were dropped from LPVG dataset because they were not statistically significant [64, 85, 146]. The new model is displayed in Figure 8.2. Even though the size of the training sample was reduced (N=817), the model revealed good performance, with all the parameters in the model being statistically significant at 5% level, as observed in AMLE in Figure 8.2. In addition, Adj R-Sq (used for evaluation of goodness-of-fit) shows a very satisfying value, 0.9473. In fact, Adj R-Sq slightly improved after dropping Norm 4 observations, 0.9426 and 0.9473, for model goodness-of-fit with and without Norm 4, respectively.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	10	282844	28284	1458.02	<.0001
Error	878	17033	19.399235		
Corrected Total	888	299877			

Model Fit Statistics

R-Square	0.9432	Adj R-Sq	0.9426
AIC	2647.0241	BIC	2649.3425
SBC	2699.7151	C(p)	9.3013

Type 3 Analysis of Effects

Effect	DF	Sum of Squares	F Value	Pr > F
Norm	5	19234.6574	198.30	<.0001
SpeedLimit	3	180808.505	3106.80	<.0001
cc	2	75401.5491	1943.42	<.0001

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept		1	172.6	0.7183	240.24	<.0001	171.2	174.0
Norm	ECE15-00/04	1	8.6837	0.5869	14.80	<.0001	7.5334	9.8341
Norm	Euro I	1	3.6542	0.5169	7.07	<.0001	2.6411	4.6672
Norm	Euro II	1	-5.5133	0.4993	-11.04	<.0001	-6.4919	-4.5347
Norm	Euro III	1	-3.0808	0.5020	-6.14	<.0001	-4.0646	-2.0970
Norm	Euro IV	1	-0.5431	0.6182	-0.88	0.3799	-1.7549	0.6686
SpeedLimit	100	1	-2.6833	0.6511	-4.12	<.0001	-3.9595	-1.4072
SpeedLimit	120	1	21.2175	0.4548	46.66	<.0001	20.3262	22.1088
SpeedLimit	50	1	-8.1688	1.1888	-6.87	<.0001	-10.4989	-5.8387
cc	1.4-2.	1	-8.0139	0.4459	-17.97	<.0001	-8.8878	-7.1401
cc	<1.41	1	-24.0533	0.3968	-60.63	<.0001	-24.8309	-23.2756

Figure 8.1 – Linear regression output for CO_2 modeling with LPGV dataset using SAS®Enterprise Miner™7.2 software.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	275323	30591	1629.81	<.0001
Error	807	15147	18.769931		
Corrected Total	816	290470			

Model Fit Statistics

R-Square	0.9479	Adj R-Sq	0.9473
AIC	2405.5916	BIC	2407.8740
SBC	2452.6480	C(p)	8.6095

Type 3 Analysis of Effects

Effect	DF	Sum of Squares	F Value	Pr > F
Norm	4	19432.8927	258.83	<.0001
SpeedLimit	3	178939.681	3177.77	<.0001
cc	2	71582.6248	1906.84	<.0001

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept		1	172.2	0.7574	227.40	<.0001	170.7	173.7
Norm	ECE15-00/04	1	8.7158	0.6271	13.90	<.0001	7.4866	9.9450
Norm	Euro I	1	3.6156	0.5710	6.33	<.0001	2.4964	4.7348
Norm	Euro II	1	-5.5574	0.5576	-9.97	<.0001	-6.6502	-4.4646
Norm	Euro III	1	-3.2152	0.5598	-5.74	<.0001	-4.3124	-2.1179
SpeedLimit	100	1	-3.2140	0.6610	-4.86	<.0001	-4.5096	-1.9183
SpeedLimit	120	1	21.6372	0.4525	47.82	<.0001	20.7503	22.5240
SpeedLimit	50	1	-7.9800	1.1705	-6.82	<.0001	-10.2741	-5.6859
cc	1.4-2.	1	-8.3384	0.4431	-18.82	<.0001	-9.2069	-7.4700
cc	<1.4l	1	-23.7741	0.3923	-60.60	<.0001	-24.5430	-23.0053

Figure 8.2 – Linear regression output for CO_2 modeling following the removal of Euro 4 observations from the LPGV dataset, obtained with SAS®Enterprise Miner™7.2 software.

For CO, NO_x and PM modeling, the procedure was similar. Following the optimization of the estimation models with removal of not statistically significant parameters, the training size datasets are shown in Table 8.2. The model for NO_x estimation for LPDV revealed all the parameters being significant at 5% level, hence there was not need to remodel and therefore the training sample was kept at the original size for the LPDV dataset, 769 with "*" in Table 8.2.

Table 8.2 – Training sample size by vehicle category for selected pollutants modeling

Model	Dataset	Training Sample Size (N)	
		Previous	Following Optimization
Modeling CO_2	LDGV	889	817
	LPDV	769	344
	LDDV	556	335
Modeling CO	LDGV	889	847
Modeling NO_x	LPDV	769	769*
	LDDV	556	535
Modeling PM	LPDV	769	731
	LDDV	556	533

* All the parameters were found statistically significant without need to remove any parameter from the training set.

After dropping the variables that were not useful for the models, Adj R-Sq was very satisfactory for final models and all the parameters in the models were statistically significant. Final models are presented next.

8.2 Results

This section presents the most significant trends for emissions estimation of the selected pollutants and fuel consumption for the crash dataset (see section 8.2.1). Then, it presents the results for fitting the emissions database into linear regressions models for CO_2 and local pollutants emissions estimation, as basis for vehicle's environmental performance analysis.

8.2.1 Emissions and fuel consumption trends

Based on the crash sample explored in this study with 2,248 vehicles, trends on the emissions for the selected pollutants using CORINAIR methodology can be summarized as follows. The mean CO emissions were higher for gasoline than diesel engines: 2.07 $g.km^{-1}$ and 0.40 $g.km^{-1}$, for a sample with 914 gasoline vehicles (LPGV and LDGV) and 1,325 diesel vehicles (LPDV and LDDV), respectively. On the other hand, PM emissions were significantly higher for diesel than gasoline engines, 0.109 $g.km^{-1}$ and 0.002 $g.km^{-1}$, respectively. For NO_x emissions estimation, diesel engines also revealed a mean value higher than for gasoline engines, 1.04 $g.km^{-1}$ and 0.53 $g.km^{-1}$, respectively. Regarding to CO_2 emissions, it appears to be higher for the diesel engines than for gasoline engines in the crash sample, 241 $g.km^{-1}$ and 164 $g.km^{-1}$, respectively. The effect of engine size is relevant for the interpretation of these results in the crash sample, since diesel vehicles has a higher engine size. The mean engine size was: 1912 cm^3 (S.D. 471) and 1309 cm^3 (S.D. 295) for the 1325 diesel vehicles and 914 gasoline vehicles in the sample. While the majority of gasoline vehicles in the sample fells in the category c.c.<1400 cm^3, diesel vehicles are very closer to 2000 cm^3, and the disproportion of engine size may affect CO_2 emissions results for the sample used in this study.

8.2.2 Environmental performance analysis

This section presents the results for fitting the emissions database into linear regressions models. Models equations are presented for each selected pollutant based on vehicle category. Emissions models are identified as “Model-E-*i+1*” were “E” stands for emissions and “*i+1*” identifies the model number. Results for models fit statistics and analysis of maximum likelihood estimates are summarized in Table 8.3. Though, engine size categories appear in L units in the CORINAIR methodology, in this study be consistent with previous sections, engine size categories were converted to cm^3 for model equation presentation.

Table 8.3 – Emissions estimations models results for selected pollutants using a linear regression approach.

Model	Target	Vehicles Category	Model Fit Statistics			Analysis of Maximum Likelihood Estimates
			Pr>F	Adj R-Sq	ASE	
MODEL-E-1	CO_2	LPGV	<0.0001	0.9473	18.54	
MODEL-E-2	CO_2	LPDV	<0.0001	0.8643	145.34	
MODEL-E-3	CO_2	LDDV	<0.0001	0.9877	68.75	
MODEL-E-4	CO	LPGV	<0.0001	0.9762	0.03	

MODEL-E-1

Parameter		DF	Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept		1	172.2	0.7574	227.40	<.0001	170.7	173.7
Norm	ECE15-00/04	1	8.7158	0.6271	13.90	<.0001	7.4866	9.9450
Norm	Euro I	1	3.6156	0.5710	6.33	<.0001	2.4964	4.7348
Norm	Euro II	1	-5.5574	0.5576	-9.97	<.0001	-6.6502	-4.4646
Norm	Euro III	1	-3.2152	0.5598	-5.74	<.0001	-4.3124	-2.1179
SpeedLimit	100	1	-3.2140	0.6610	-4.86	<.0001	-4.5096	-1.9183
SpeedLimit	120	1	21.6372	0.4525	47.82	<.0001	20.7503	22.5240
SpeedLimit	50	1	-7.9800	1.1705	-6.82	<.0001	-10.2741	-5.6859
cc	1.4-2.	1	-8.3384	0.4431	-18.82	<.0001	-9.2069	-7.4700
cc	<1.41	1	-23.7741	0.3923	-60.60	<.0001	-24.5430	-23.0053

MODEL-E-2

Parameter		DF	Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept		1	187.5	1.3421	139.67	<.0001	184.8	190.1
Norm	Conventional	1	17.3330	1.6170	10.72	<.0001	14.1637	20.5023
Norm	Euro III	1	-7.9800	1.2766	-6.25	<.0001	-10.4821	-5.4780
SpeedLimit	120	1	18.5736	0.7643	24.30	<.0001	17.0755	20.0716
cc	<2.01	1	-29.9821	0.7704	-38.92	<.0001	-31.4920	-28.4722

MODEL-E-3

Parameter		DF	Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept		1	269.1	2.0747	129.70	<.0001	265.0	273.2
Norm	Conventional	1	31.1490	1.3046	23.88	<.0001	28.5919	33.7060
Norm	Euro I	1	5.0301	1.0357	4.86	<.0001	3.0002	7.0601
Norm	Euro II	1	-2.7867	0.8793	-3.17	0.0016	-4.5101	-1.0633
Norm	Euro III	1	-2.6600	0.8352	-3.18	0.0015	-4.2970	-1.0230
Norm	Euro IV	1	-2.7850	1.0089	-2.76	0.0060	-4.7624	-0.8077
SpeedLimit	120	1	117.2	2.0141	58.18	<.0001	113.2	121.1
SpeedLimit	50	1	-76.8725	3.9636	-19.39	<.0001	-84.6411	-69.1040
cc	<2.01	1	-1.4847	0.3987	-3.72	0.0002	-2.2662	-0.7032

MODEL-E-4

Parameter		DF	Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept		1	1.5339	0.0281	54.53	<.0001	1.4788	1.5890
Norm	ECE15-00/04	1	2.8646	0.0243	117.69	<.0001	2.8169	2.9123
Norm	Euro I	1	0.5828	0.0213	27.34	<.0001	0.5410	0.6246
Norm	Euro II	1	-0.8589	0.0206	-41.64	<.0001	-0.8994	-0.8185
Norm	Euro III	1	-0.2333	0.0207	-11.25	<.0001	-0.2739	-0.1926
Norm	Euro IV	1	-1.1593	0.0258	-45.00	<.0001	-1.2098	-1.1088
SpeedLimit	120	1	0.7568	0.0224	33.81	<.0001	0.7130	0.8007
SpeedLimit	50	1	-0.4766	0.0431	-11.07	<.0001	-0.5610	-0.3922

Model	Target	Vehicles Category	Model Fit Statistics			Analysis of Maximum Likelihood Estimates
			Pr>F	Adj R-Sq	ASE	
MODEL-E-5	NO_x	LPDV	<0.0001	0.7941	0.01	

Parameter		DF	Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept		1	0.6569	0.0125	52.68	<.0001	0.6325	0.6813
Norm	Conventional	1	0.2069	0.0104	19.81	<.0001	0.1864	0.2273
Norm	Euro I	1	-0.0293	0.00900	-3.25	0.0012	-0.0469	-0.0116
Norm	Euro II	1	0.0197	0.00678	2.90	0.0038	0.00641	0.0330
Norm	Euro III	1	0.1205	0.00578	20.86	<.0001	0.1092	0.1319
Norm	Euro IV	1	-0.0398	0.00646	-6.15	<.0001	-0.0524	-0.0271
SpeedLimit	100	1	-0.0363	0.0152	-2.39	0.0169	-0.0660	-0.00659
SpeedLimit	120	1	0.2160	0.0122	17.76	<.0001	0.1922	0.2399
SpeedLimit	50	1	-0.1087	0.0342	-3.18	0.0015	-0.1757	-0.0417

MODEL-E-6	NO_x	LDDV	<0.0001	0.7609	0.04	

Parameter		DF	Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept		1	1.1220	0.0520	21.57	<.0001	1.0200	1.2240
Norm	Conventional	1	0.4662	0.0325	14.32	<.0001	0.4024	0.5300
Norm	Euro I	1	0.3659	0.0260	14.08	<.0001	0.3150	0.4169
Norm	Euro II	1	0.1399	0.0221	6.34	<.0001	0.0966	0.1831
Norm	Euro III	1	-0.0992	0.0209	-4.74	<.0001	-0.1402	-0.0582
Norm	Euro IV	1	-0.3994	0.0252	-15.82	<.0001	-0.4488	-0.3499
SpeedLimit	120	1	0.4470	0.0505	8.85	<.0001	0.3480	0.5460
SpeedLimit	50	1	-0.2474	0.0994	-2.49	0.0132	-0.4423	-0.0525

MODEL-E-7	PM	LPDV	<0.0001	0.8910	0.0001	

Parameter		DF	Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept		1	0.0772	0.00101	76.70	<.0001	0.0752	0.0792
Norm	Conventional	1	0.1286	0.00226	56.89	<.0001	0.1242	0.1331
Norm	Euro I	1	0.0594	0.00195	30.48	<.0001	0.0555	0.0632
Norm	Euro II	1	-0.0206	0.00147	-13.96	<.0001	-0.0235	-0.0177
Norm	Euro III	1	-0.0304	0.00124	-24.43	<.0001	-0.0328	-0.0280
Norm	Euro IV	1	-0.0500	0.00140	-35.75	<.0001	-0.0527	-0.0473
SpeedLimit	120	1	0.0154	0.000714	21.52	<.0001	0.0140	0.0168

MODEL-E-8	PM	LDDV	<0.0001	0.8639	0.0001	

Parameter		DF	Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept		1	0.1717	0.00296	57.94	<.0001	0.1659	0.1775
Norm	Conventional	1	0.2006	0.00567	35.38	<.0001	0.1895	0.2117
Norm	Euro I	1	0.0915	0.00454	20.16	<.0001	0.0826	0.1004
Norm	Euro II	1	0.0198	0.00385	5.13	<.0001	0.0122	0.0273
Norm	Euro III	1	-0.0479	0.00364	-13.13	<.0001	-0.0550	-0.0407
Norm	Euro IV	1	-0.1253	0.00440	-28.50	<.0001	-0.1339	-0.1167
SpeedLimit	120	1	0.0491	0.00168	29.26	<.0001	0.0458	0.0524

8.2.2.1 Models for CO_2 Emissions Estimation

Model-E-1 estimates the emissions for CO_2 from LPGV. As observed in Table 8.3, the AMLE shows that for LPGV category older vehicles models (ECE15-00/04) and/or driving at higher speeds (120 km.h^{-1}) significantly contribute to increase CO_2 emissions, estimates 8.7 and 21.6, respectively. On the other hand, for the same vehicle category, when models have a small engine size (c.c.<1400 cm^3, labelled as category cc<1.4L in Table 8.3), CO_2 emissions decrease considerably, due to the parameter value of 23.8. Also, all these predictor variables in the model were statistically significant, p-value<0.0001. The linear regression equation developed for Model-E-1 is presented below.

$$
\begin{aligned}
&CO_2\,(g.km^{-1}) \\
&= \ 172.2000 + 8.7158 * (if Norm\text{ECE15-00/04"}) + 3.6156 \\
&* (if"Norm\ \text{Euro1"})\text{-}5.5574\text{*(if"NormEuro2")-}3.2152\text{*(if"NormEuro3")} \\
&- 7.9800 * (if"SpeedLimit50") - 3.2140 * (if"SpeedLimit100") + 21.6372 \\
&* (if"SpeedLimit120") - 23.7741 * (if"c.c.< 1400") - 8.384 \\
&* (if"c.c.1400 - 2000")
\end{aligned}
\qquad \text{Equation 8.3}
$$

Where "if" implies a condition to be satisfied by the categorical variable (model parameter), otherwise the term in the equation will be zero. Since the equations were developed for individual vehicles, only one category for each component (Norm, engine size and speed) can be satisfied. Additional information for Model-E-1 is provided in Appendix 10. The following example is demonstrated. Considering a light passenger gasoline vehicle with 1300 cm^3, complying with Norm 2 and driving at 120 km.h^{-1}, Equation 8.4 will be simplified as:

$$CO_2\,(g.km^{-1}) = \ 172.2 + 8.7 * 0 + 3.6 * 0 - 5.6 * 1 - 3.2 * 0 - 8.0 * 0 - 3.2 * 0 + 21.6 * 1 - 23.8 * 1 - 8.3 * 0$$

Thus for the vehicle in this example, CO_2 emissions would be estimated of 188.2 g.km^{-1}.

Model-E-2 estimates CO_2 emissions for LPDV. As shown in Table 8.3, for LPDV category earlier technological legislation (conventional) and/or driving at higher speeds (120 km.h^{-1}) expressively increase CO_2 emissions. On the other hand, these vehicles models when in the smaller engine size category (c.c.<2000 cm^3), CO_2 emissions decrease significantly. Also, all model predictor variables were statistically significant, p-value<0.0001. The linear regression equation developed for Model-E-2 is presented underneath.

$$
\begin{aligned}
CO_2\,(g.km^{-1}) = \ & 187.5000 + 17.3330 * (if\ "NormConventional)" - 7.9800 \\
& * (if"NormEuro3") + 18.5736 * (if"SpeedLimit120") \\
& - 39.9821 * (if"c.c.< 2000")
\end{aligned}
\qquad \text{Equation 8.4}
$$

Model-E-3 estimates CO_2 emissions for LDDV. As shown in Table 8.3 for LDDV category earlier technological level (conventional) and/or driving at 120 km.hr-1 expressively increase CO_2 emissions. On the other hand, for the same category, vehicle with smaller engine size (c.c.<2000 cm^3) and/or driving slow (50 $km.h^{-1}$) significantly contribute towards CO_2 emissions reduction. Model-E-3 equation is as follows.

$$CO_2\,(g.km^{-1}) = 269.1000 + 31.1490 * (if\ "Norm\text{Conventional}") + 5.0301 * (if"Norm\ Euro1") - 2.7867 * (if"NormEuro2") - 2.6600 * (if"NormEuro3") - 2.7850 * (if"NormEuro4") - 76.8725 * (if"SpeedLimit50") + 117.2000 * (if"SpeedLimit120") - 1.4847 * (if"c.c.< 2000") \quad \text{Equation 8.5}$$

Although in this Chapter, emissions estimation models are presented for LPVG, LPDV and LDDV, in the following Chapter, vehicle's safety, fuel efficiency and green integrated analysis is presented for LPGV and LPDV. Thus, Model-E-1 and Model-E-2 were applied for the evaluation of these vehicles' fuel efficiency.

8.2.2.2 Models for local pollutants emissions estimation

This section presents the results for fitting the local pollutants emissions database into linear regressions models. Model-E-4 estimates CO emissions for LPGV. Table 8.3 shows that for LPGV category earlier technological level (ECE15-00/04) yields to an increase in CO emissions. All these predictor variables in the model were statistically significant, p-value<0.0001. Model-E-4 equation is presented below.

$$CO\,(g.km^{-1}) = 1.5339 + 2.864 * (if\ "Norm\text{ECE15-00/04}") + 0.5828 * (if"NormEuro1") - 0.8589 * (if"NormEuro2") - 0.2333 * (if"NormEuro3") - 1.1593 * (if"NormEuro4") + 0.7568 * (if"SpeedLimit120") - 0.4766 * (if"SpeedLimit50") + 0.8 * (if"SpeedLimit120") \quad \text{Equation 8.6}$$

Model-E-5 estimates NO_x emissions for LPDV. Table 8.3 shows that for LPDV category earlier technological level (Conventional) and/or driving at 120 $km.h^{-1}$ increase NO_x emissions. As expected, when driving at lower speed (50 $km.h^{-1}$) the emissions estimations for NO_x decrease. All predictor variables in the model were statistically significant, p-value<0.0001. Model-E-5 equation is next.

$$NO_x\,(g.km^{-1}) = \; 0.6569 + 0.2069 * Norm(if\ \text{"Conventional"}) - 0.0293 * (if\text{"}NormEuro1\text{"}) + 0.0197 * (if\text{"}NormEuro2\text{"}) + 0.1205 * (if\text{"}NormEuro3\text{"}) - 0.0398 * (if\text{"}NormEuro4\text{"})0.1087 * (if\text{"}SpeedLimit50\text{"}) - 0.0363 * (if\text{"}SpeedLimit100\text{"}) + 0.2160 * (if\text{"}SpeedLimit120\text{"})$$

Equation 8.7

Model-E-6 estimates NO_x emissions for LDDV. The analysis of maximum likelihood estimates in Table 8.3 confirms that for LDDV category, earlier technological level (conventional and Euro 1) and/or driving at 120 km.h^{-1} contribute towards an increase of NO_x emissions. On the other hand, for those LDVD, newer models with Euro 4 and/or when driving slow (50 km.h^{-1}), the emissions estimations for NO_x decrease. All predictor variables in the model were statistically significant, p-value<0.0001. Model-E-6 equation is below.

$$NO_x\,(g.km^{-1}) = \; 1.1220 + 0.4662 * (if\ \text{"}Norm\text{Conventional"}) + 0.3659 * (if\text{"}Norm\ Euro1\text{"}) + 0.1399 * (if\text{"}NormEuro2\text{"}) - 0.0992 * (if\text{"}NormEuro3\text{"}) - 0.3994 * (if\text{"}\ NormEuro4\text{"}) - 0.2474 * (if\text{"}SpeedLimit50\text{"}) + 0.4470 * (if\text{"}SpeedLimit120\text{"})$$

Equation 8.8

Model-E-7 estimates PM emissions for LPDV. The analysis of maximum likelihood estimates shows that convectional vehicles had the biggest impact in PM emissions. This result was expected since vehicles with earlier technological level were not equipped with particle filters. On the other hand, after Euro 2, there were refinements of fuel injection and LPDV started to be equipped with particle filters, thus contributing to reductions in PM, as observed in Table 8.3. Also, all these predictor variables in the model were statistically significant, p-value<0.0001. Model-E-7 equation is below.

$$PM\,(g.km^{-1}) = \; 0.0772 + 0.1286 * (if\ \text{"}Norm\text{Conventional"}) - 0.0594 * (if\text{"}Norm\ Euro1\text{"}) - 0{,}0206 * (if\text{"}NormEuro2\text{"}) - 0{,}0304 * (if\text{"}NormEuro3\text{"}) - 0.0500 * (if\text{"}Norm\ Euro4\text{"}) + 0.0154 * (if\text{"}SpeedLimit120\text{"})$$

Equation 8.9

Model-E-8 estimates PM emissions for LDDV, as shown in Table 8.3. Similarly to LPDV, for LDDV models, earlier technological level (conventional) increases PM emissions. In addition, driving at 120 km.h^{-1} also shows a positive effect in PM emissions. On the other hand, newer LDDV models with Euro 4 contribute towards to PM reductions. Predictor variables in the model were statistically significant, p-value<0.0001. Model-E-8 equation is below.

$$PM\,(g.km^{-1}) = \ 0.1717 + 0.2006 * (if\ "Norm\text{Conventional}") + 0.0915 * N(if"\ NormEuro1") + 0.0198 * (if"NormEuro2") - 0.0479 * (if"NormEuro3") - 0.1253 * N(if"Norm\ Euro4") + 0.0491 * (if"SpeedLimit120")$$

Equation 8.10

8.2.2.3 Assessment of vehicle's emissions estimation models

As previously mentioned (in section 8.2.2.1), since in the vehicle's safety, fuel efficiency and green emissions analysis presented in Chapter 9, only light passenger vehicles are discussed, the assessment of models goodness-of-fit is presented for LPGV and LPDV. For LPGV, CO_2 and CO emissions estimation based on Model-E-1 and Model-E-4 explained 94.7% and 97.6% of data variability, respectively as shown in Table 8.3. For LPDV, Model-E-2 showed a good fit to the CO_2 emissions estimation data, with Adj R-Sq explaining 86.4% of the data, Table 8.3. Also for those, NOx and PM emissions estimation based on Model-E-5 and Model-E-7 explained 79.4% and 89.1% of data, respectively. All these models revealed very satisfying results for goodness-of-fit, and will be further apply for vehicle's environmental performance evaluation. Although goodness-of-fit models results are very promising, they are based on the crash sample explored in this study with CORINAIR methodology. If a different sample was used, or if more vehicles information would be added to the crash database, the emissions estimation models may change. More information for the emissions models is found in Appendix 10.

8.3 Concluding Remarks

In this Chapter, CORINAIR methodology was used to develop an emissions estimation database for the vehicles included in the crash sample explored in this study. Then, the emissions data were fit into linear regression models. The models were developed to estimate the most relevant selected pollutants for gasoline and diesel vehicles. Emissions estimation models revealed very satisfactory results for goodness-of-fit, as summarized next. For light passenger gasoline vehicles, CO_2 and CO emissions estimation models, showed an adjusted R-square explaining 94.7% and 97.6% of the data emission, respectively. For light passenger diesel vehicles, CO_2, NO_x and PM emissions estimation models, showed adjusted R-square values explaining 98.8%, 79.4% and 89.9% of the data, respectively. Thus, the developed models are helpful for further application on the vehicle's environmental performance evaluation, which is part of the vehicle's integrated analysis in Chapter 9.

Based on the developed emissions estimation models, its predictor variables and its estimate values, the following statements can be drawn, focusing the effect of model predictor variables (sign and magnitude of the predictor estimate) has on the model response, air pollutant. For all the selected pollutants, CO_2, CO, NO_x and PM emissions models for gasoline and diesel engines there is an increase of these pollutants emissions for earlier technological levels (as shown by the positive sign associated to the former emissions regulation). Driving at higher speeds (120 $km.h^{-1}$) contributes to a general increase for all the above air emission pollutants and fuel consumption. NO_x emissions models for diesel engines showed that earlier technology level (Conventional) contribute to higher emissions, because vehicles were not equipped with emissions control systems, such as exhaust gas recirculation and diesel oxidation catalyst. PM emissions models had identified vehicles complying with earlier technological levels (Conventional and Euro 1) as contributing to a significant increase on particulate matters, because few vehicles were equipped with particle filters.

CHAPTER 9

INTEGRATED ANALYSIS OF VEHICLE'S SAFETY, EFFICIENCY AND GREEN PERFORMANCE

The main goal of this Chapter is to present a methodology which combines the vehicle's safety and environmental evaluation into an integrated analysis in order to provide a rate classification. SEG (for **S**afety, Fuel **E**fficiency and **G**reen) is the integrated indicator that was developed. This chapter combines the results from Chapters 3, 6, 7 and 8 and is organized as explained next. First, the methodology to develop the integrated analysis methodology is explained. Second, the results for a scenario base analysis are presented. Third, final combined score, SEG itself, is discussed for several vehicles categories. Finally, the most relevant findings are highlighted.

9.1 Methodology

SEG integrated analysis examines the trade-off between a vehicle's safety and its environmental performance. As will be explained in this section the SEG methodology was designed to explore the conflict that apparently seems to exist between larger and heavier cars with smaller and lighter cars' safety and environmental performances. Bigger and heavier cars are considered safer but they use more fuel and emit more CO_2 among other air pollutants. On the other hand, smaller, lighter cars are more affordable, they use less fuel, and thus, they earn higher environmental performance, but they could do a relative poor job of protecting their occupants. To examine this potential conflict, SEG rates the vehicle performance for each domain: safety, fuel efficiency and emissions. Figure 9.1 illustrates the basic steps of SEG methodology overview for each of those three domains.

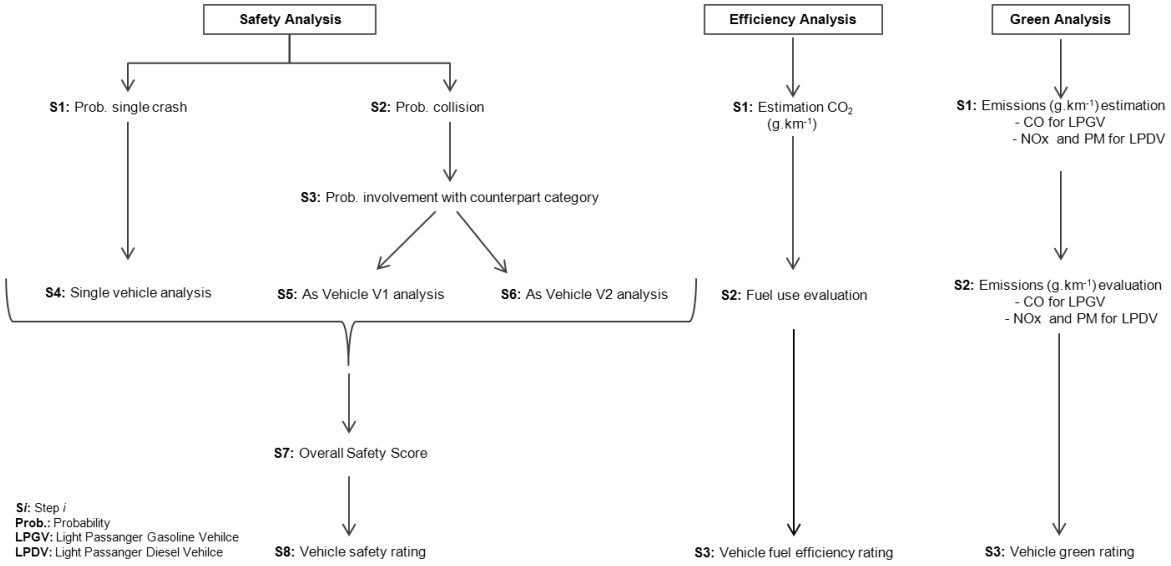


Figure 9.1 – SEG methodology overview.

As shown in Figure 9.1, safety analysis follows eight basic steps which comprise the probability of the vehicle being involved in a single-vehicle crash and in a collision, step 1 "S1" and step 2 "S2", respectively. For two vehicle collisions, the probability of the vehicle being analyzed to be involved with a counterpart vehicle with an engine size category is calculated, "S3". Following, the risk of severe crash outcome is evaluated for the crashes were only one vehicle is involved in the crash, "S4" and for the collisions, where the vehicle is considered to be vehicle V1 and as V2, steps "S5" and "S6", respectively. Following, in step 7, "S7", the vehicle overall safety score is calculated based on each safety component derived from the steps "S4-S6". Vehicles' fuel efficiency analysis covers three basic steps: mainly estimation of CO_2 emissions ($g.km^{-1}$), evaluation and rating, represented by the steps "S1", "S2" and "S3", respectively.

Finally, vehicle's green emissions analysis start with emissions estimation, based on vehicle fuel type, as illustrated in step 1, "S1". Following, emissions estimations are evaluated at step 2 "S2" and vehicle is rated for its emissions "S3". Concluding, vehicles' safety, fuel efficient and green ratings are combined into a single score.

9.1.1 Methodology for a vehicle safety rating

Safety rating measures vehicle's crashworthiness (capability to protect vehicle's occupants) on a qualitative scale. In this research, the procedure evaluates vehicle crashworthiness both in single-vehicle and two-vehicle collisions. The overall safety score (OSS) evaluates the vehicle on both risk of exposure and the probability that a crash would result in a severe outcome. Therefore OSS is the product of the probability that certain vehicle categories would be involved in a crash and the probability of crash injury severity itself.

Previous to the OSS methodology, crash severity distribution as presented earlier in Chapter 3 must be recalled for a better comprehension of the risk of exposure in the crash sample (previously presented in section 4.1.1). Table 4.4, presented earlier, showed the distribution of vehicles by number of vehicles involved in the crash, engine size category and crash severity outcomes for each vehicle involved. For single-vehicle crashes, 43.8% of the vehicles fell in the engine size category c.c.<1400 cm^3, and those vehicles were involved in 34.2% of the severe crashes. Also, 43.6% of the vehicles fell in the engine size category 1400≤c.c.<2000 cm^3 those were involved in 47.4% of the severe crashes in the single-vehicle crashes, as shown in Table 4.4.

For two-vehicle collisions, risk of exposure is also shown in Table 4.4. The example for vehicles in the intermediate engine size category is presented. In Table 4.4, 44.6% of vehicles V1 fell in the intermediate engine size category, 1400≤c.c.<2000 cm^3, and those were involved in 43.8% of the severe collisions. Also, for vehicle V2, the intermediate engine size category was the most frequent, covering 43.4% of vehicle V2, but it accounted for a smaller proportion of involvement in severe crashes, 9.4%.

Therefore, it must be pointed out that the risk of exposure is derived from the crash sample used in this study. If a different crash sample were used, the distribution of vehicle involvement by crash type, engine size and age would vary, and hence, the risk of exposure would be affected as well. The method to determine vehicle safety rating can be updated as more crashes are added to the current sample.

The OSS methodology is presented as follows. The OSS has mainly three components derived from: the risk of an event involving the vehicle in a single-vehicle crash, the risk of a crash event in a collision where the vehicle being analyzed is assumed to be as vehicle V1 and the risk associated when the vehicle being analyzed is assumed to be as vehicle V2. OSS is performed following step 1 through 7, as previously summarized in Figure 9.1. An example will be provided

through these steps yielding the overall safety score. A vehicle one year old and with an engine size capacity of 1300 cm^3 is evaluated.

Step 1: ***Probability of involvement in a single-vehicle crash***

The first step of the OSS is the estimation of the probability of exposure as a single vehicle involved in the crash. Considering the vehicle mentioned above, 1 year old and 1300 cm^3 engine, and based on Table 4.4, the probability of this vehicle being involved in a single-vehicle crash is 0.028 (63/2248).

Step 2: ***Probability of involvement in two-vehicle collisions***

Step 2 is based on the calculation of the probability of vehicle involvement in a collision and also the probability of involvement with a counterpart engine size category. For the two-vehicle crashes scenario, the probability that this vehicle is involved in a collision as vehicle V1 is 0.027, whereas the probability to be involved as vehicle V2 is 0.034.

The probability that the vehicle is involved in a crash event with certain counterpart vehicle category is determined based on the engine size category into which falls the other vehicle involved in the collision. In this probability of event calculation, the engine size category for a counterpart vehicle being V2 is considered, as well as the engine size category for a counterpart vehicle being V1. The following example illustrates better the step 2 calculations using Table 4.4. In the scenario when the vehicle being analyzed is assumed to be vehicle V1, the probability that is going to be involved in a collision with a counterpart vehicle, vehicle V2, in the category of engine size <1400 cm^3 is 0.146 (334)/2248. The probability that the vehicle being analyzed is involved with a V2 in the engine size category 1400-2000 cm^3 is 0.169 (379)/2248, and so forth. Similarly, the vehicle being analyzed could be considered as vehicle V2, and thus the counterpart vehicle would be V1. In this scenario, the probability that V2 is involved with V1 for each engine size category: <1400 cm^3, would be (346)/2248=0.1540, and so forth.

Step 3: ***Probability of exposure-vehicle involvement in a crash with opponent category***

The probability of exposure is the product of vehicle involvement in a collision and the probability of involvement with a counterpart engine size category, both calculated in step 2.

Following the example, the probability of exposure for a vehicle 1 year old with 1300 cm^3 engine is calculated, as explained next. The probability that this vehicle would be involved with a counterpart vehicle, V2, with c.c.<1400 cm^3 is 0.0039 (0.027x0.146). The probability that the vehicle would be involved with V2 with c.c.1400-2000 cm^3 and V2 with c.c.≥2000 cm^3, are 0.0046 and 0.0019,

respectively. The same procedure would be followed to cover the scenario where the vehicle being analyzed would be vehicle V2 and the counterpart vehicle would be V1.

Step 4: ***Component from a single-vehicle crash event***

For the single crash scenario, as discussed in Chapter 6, the probability of a serious injury and/or fatality (FatalSIK"1") is given by Model-IB-S, presented in Equation 6.2. Component for vehicle overall safety score from a single-vehicle event is the product of the probability estimated in step 1 and the probability of FatalSIK.

Following the example, the probability of the vehicle being analyzed being involved in a single crash is 0.028, as explained in step 1. The probability of FatalSIK using Model-IB-S for the vehicle being analyzed, that is 1 yr old and with 1300 cm^3, is 0.1854. Thus, the component from a single-vehicle crash event towards vehicle overall safety score is 0.520% (0.028x0.1854x100).

Step 5: ***Component from vehicle involvement, as vehicle V1, in collision with opponent V2***

For a two-vehicle collision scenario, the subject vehicle can be either V1 or V2. Step 5 assumes the subject vehicle is V1, and opponent vehicle as V2. As explained in section 7.3, Chapter 7 for severe crashes prediction in two-vehicle collisions, for a subject vehicle V1 the probability of a serious injury and/or fatality (FatalSIKV1"1") is given by Model-II-T, presented in Equation 7.2. Since V2 can fall in one of three engine size categories, this component integrates the probability of a crash event for: ccV2<1400 cm^3, 1400≤ccV2<2000 cm^3 and ccV2≥2000 cm^3.

Following the example, the probability for FatalSIKV1 when V2 with engine of 1300 cm^3 is involved is 0.340. As the engine size of the other vehicle involved in the collision increases, the probability FatalSIKV1 also increases since Model-II-T depends on ccV2 engine size only. For counterpart vehicles with engine sizes of 1700 cm^3 and 2500 cm^3, the probability of FatalSIKV1 would be 0.4428 and 0.6535, respectively. The probability of exposure was already determined in step 3. The contribution from this collision event is the product of the probability of exposure and the probability of a severe crash outcome in V1. In this case, it is 0.0039x0.340, yielding a value of 0.135%. Similarly, the contribution from the collision event involving a counterpart vehicle with 1700 cm^3 is 0.199% (0.00449x0.4428x100). Finally, the contribution from a collision event involving a counterpart vehicle in the largest engine size category, ccV2=2500 cm^3, is 0.125% (0.019x0.653x100).

Step 6: ***Component from vehicle involvement, as vehicle V2, in collision with opponet V1***

Step 6 focuses on the subject vehicle as V2, whose safety score takes into account the probability to be involved with an opponent vehicle, V1. When the vehicle being analyzed is V2, the probability

of a serious injury and/or fatality (FatalSIKV2"1") is given by Model-III-T, presented in Equation 7.3. Similarly to step 5, this component integrates the probability of a crash event involving the vehicle being analyzed with an opponent vehicle for each engine size category: ccV1<1400 cm^3, 1400≤ccV1<2000 cm^3 and ccV1≥2000 cm^3.

Following the example, the probability for FatalSIKV2 the opponent vehicle has engine size of 1300 cm^3 is 0.2825. For opponent vehicles with engine sizes of 1700 cm^3 and 2500 cm^3, the probability of FatalSIKV2 would be 0.4720 and 0.8217, respectively. The probability of exposure was already determined in step 3. The contribution from this collision event is the product of the probability of exposure and the probability of a severe crash outcome in V2. The collision with a V1 in the category of c.c.< 1400 cm^3, is 0.147% (0.0052x0.2825x100). The collision with a V1 in the category of 1.4≤c.c.< 2000 cm^3 is 0.277% (0.0058x0.4720x100). The collision with a V1 in the category of c.c.≥ 2000 cm^3 is 0.171% (0.0021x0.8217x100).

Step 7: ***Overall safety score***

The overall safety score is the result of steps 1 through 6. OSS includes three components scores:

- risk associated with the vehicle being involved in a single-vehicle crash, estimated in step 4;
- risk associated with vehicle being V1 and involved with the tree categories of engine size of V2, estimated in step 5;
- and risk associated with vehicle being V2 and involved with the tree categories of engine size of V1, estimated in step 6.

Using the same example, mentioned in the above steps, the overall safety score for a vehicle that is with 1 yr old and with an engine size 1300 cm^3capacity would be: 0.520% + 0.135% + 0.199% + 0.125% + 0.147% + 0.277% + 0.175%. Thus, the vehicle will achieve a score of 1.573%.

Step 8: ***SEG vehicle safety rating***

Two approaches were established for SEG safety rating: one is based on the overall safety score the other alternative is based on the vehicle severity risk score (SRS). The evaluation of preliminary results reveled that safety rating was very dependent on the risk of exposure, which is affected by the vehicle category distribution in the crash sample. SRS is part of OSS, however does not take into account the risk of exposure, but focuses exclusively on vehicle crashworthiness. SRS is calculated as the mean value for the probability of risk of severity for each target component: FatalSIK, FatalSIKV1 and FatalSIKV2, as the subject vehicle is considered in single-vehicle crash event, as vehicle V1 in a collision and as vehicle V2 in a collision, respectively.

SEG rating based on OSS was defined as: good, if OSS is lower than 1.99%, moderate if OSS is lower than 2.75% and poor if OSS is higher or equal to 2.75%. The criteria to establish the limit values to differentiate between good and moderate and moderate and poor safety ratings were established based on the maximum and minimum values of OSS using a training data scenario, [0.887%; 3.915%]. The lowest value, 0.887%, is associated to the vehicle with best safety performance, on the other hand, the highest value, 3.915%, is associated with the poorest safety performance for the vehicles tested with the scenario based analysis. Hence, based on the OSS range scale, the value of 1.99% was selected as cut off point for vehicle differentiation between good and moderate safety ratings. The value of 2.75% was selected as cut point for vehicle differentiation between moderate and poor safety ratings. As result, SEG safety rating based on OSS was defined as good, if OSS is lower than 1.99%, moderate if lower than 2.75% and poor if equal or higher than 2.75%.

A similar criteria set was established for SRS evaluation based in its training data scenario range [0.457; 0.559]. SEG safety rating based on SRS was defined as: good, if SRS is lower than 0.503%, moderate if OSS is lower than 0.521% and poor if OSS is higher or equal to 0.521%. The criteria to establish those values were based on the maximum and minimum values of SRS using a training data scenario, [0.457%; 0.559%].

9.1.2 Vehicle's fuel efficiency rating

SEG designed methodology for vehicle fuel efficiency evaluation is based on CO_2 emissions, since they are a direct function of vehicles fuel use [32, 147]. Vehicle fuel efficiency evaluation was performed following step 1 through 3.

Step 1: ***CO_2 estimation based on vehicle category***

For each vehicle category, CO_2 emissions ($g.km^{-1}$) were calculated using Model-E-1 and Model-E-2, which were developed in section 8.2.2.1. For LPGV CO_2 emissions were estimated using equation 8.3, whereas for LPDV, CO_2 emissions were estimated using equation 8.4.

Step 2: ***CO_2 criteria for vehicle fuel efficiency rating***

The criteria to assess vehicles CO_2 emissions were developed based on a recent study from Kok [72], in which the author published CO_2 emissions by vehicle class (from mini cars to executive and SUVs) and fuel type from 2000 to 2010 [72]. This data was further combined to develop Table 9.1.

Table 9.1 – Criteria for CO_2 ($g.km^{-1}$) evaluation in the SEG vehicle efficiency rating.

Vehicle	CO_2 ($g.km^{-1}$) by year										
Category	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
LPGV	179	179	176	176	174	173	167	165	156	146	138
LPDV	159	158	161	163	161	161	163	163	158	152	128

For gasoline vehicles, the emissions values estimated from 2000 to 2004 were used to estimate an average CO_2 emission value, 177 $g.km^{-1}$, as shown in Table 9.1. As expected, older vehicle models emitted more CO_2. Thus, the value of 177 $g.km^{-1}$ was used to set the criteria for the lowest and middle scores for fuel efficiency differentiation. On the other hand, advanced efficiency technology in newer vehicles models is known to reduce CO_2 emissions and fuel use. Thus, the emissions values from 2005 to 2010 were used to estimate the average CO_2 emissions, 158 $g.km^{-1}$, as shown in Table 9.1.

For diesel vehicles the procedure was quite similar. CO_2 emissions values from 2000 to 2005 were used to estimate the average CO_2 emission value, 161 $g.km^{-1}$, as observed in Table 9.1. On the other hand, the emissions values from 2006 to 2000 were combined into the average value of 153 $g.km^{-1}$, to differentiate a vehicle from being fuel efficient or not.

Step 3: ***SEG vehicle fuel efficiency rating***

Vehicle fuel efficiency rating, based on estimated CO_2 average values (from Step 2 of section 9.1.2) were further explored to establish rating criteria as follows. For example, for gasoline engines, a vehicle will reach a good rating for fuel efficiency if the CO_2 emissions are lower than 158 $g.km^{-1}$. A vehicle with CO_2 emissions equal or above 158 $g.km^{-1}$and lower than 177 $g.km^{-1}$, will reach the moderate rating. On the other hand, a vehicle with emissions equal or above 177 $gCO_2.km^{-1}$ will be scored as poor for fuel efficiency. A similar procedure was developed for diesel vehicles fuel efficiency rating, based on the estimated CO_2 average values for a diesel vehicles fleet. Diesel engines with CO_2 emissions lower than $g.km^{-1}$ will raise a good efficiency rating, CO_2 emissions equal or above 153 $g.km^{-1}$ and lower than 161 $g.km^{-1}$ will reach a moderate rating and CO_2 emissions equal or above 161 $g.km^{-1}$ will reach a poor rating.

9.1.3 Vehicle's Green Emissions Rating

SEG design methodology for vehicle green rating is based on the CO emissions for gasoline vehicles and NO_x and PM for diesel vehicles. Vehicle green evaluation is described through step 1 to step 3.

Step 1: ***Selected pollutants estimation for each vehicle category***

For each vehicle category, air emissions (g.km^{-1}) were calculated using Model-E-4, Model-E-5 and Model-E-7 developed in section 8.2.2.2. For LPGV, CO emissions were estimated using Equation 8.6. For LPDV, NO_x and PM emissions, estimations were obtained using Equations 8.7 and 8.9, respectively.

Step 2: ***Criteria for green evaluation***

The SEG rating for green evaluation is designed using emission factors for passengers cars and light duty vehicles, extracted from CORINAR [147]. For the green rating criteria the emissions limits are established taking as reference Euro 4 and Euro 2. Euro 4 vehicles benefit from advanced engine technology and improvements in the after treatment monitoring (for NO_x reduction and PM oxidation) and control [147]. Thus, Euro 4 emission factors were chosen to differentiate between good and moderate score. On the other hand, Euro 2 vehicles were equipped with three-way catalyst but they were not equipped with particle filters [147]. Thus, Euro 2 emission factors were chosen to differentiate between moderate and poor score.

For gasoline vehicles, the green evaluation focuses CO emissions. As an example of green rating for a gasoline vehicle, let is imagining that the engine size is 1300 cm^3. Thus, the vehicle will fell in the engine size corresponding to c.c.<1400 cm^3. For this engine size category and LPGV, if the vehicle emits lower than 0.710g gCO.km^{-1}, the vehicle is scored with good. For the same vehicle category, if the emissions are between 0.710 g.km^{-1}≤CO<2.39 g.km^{-1}, than the attributed score is moderate. If CO ≥2.39 g.km^{-1}, than the attributed score is poor.

For diesel vehicles, the green evaluation focuses on NO_x and PM emissions, and for LPDV the criteria evaluation values are independent of engine size, based on CORINAR [147]. Regarding to NO_x analysis, if emissions are lower than 0.601 g.km^{-1}, the vehicle is scored with good. If emissions are between 0.601 g.km^{-1}≤NO_x<0.726 g.km^{-1}, than the attributed score is moderate. If NO_x ≥0.726 g.km^{-1}, than the attributed score is poor. Regarding to PM analysis, if emissions are lower than 0.0324 g.km^{-1}, the vehicle is scored with good. If PM emissions are between 0.0342 g.km^{-1} ≤PM<0.0549 g.km^{-1}, than the attributed score is moderate. If PM ≥0.0549 g.km^{-1}, than the attributed score is poor.

Step 3: ***SEG vehicle green emissions rating***

For gasoline engines, the following ranges: <0.710 gCO.km^{-1}, 0.710g.km^{-1}≤CO<2.39 g.km^{-1}, and ≥2.39 gCO.km^{-1}, will lead to the attribution of good, moderate or poor green ratings, respectively. However, for diesel engines, the vehicle green final rating is the combination of NO_x and PM emissions scores. Table 9.2 illustrates the final green rating for diesel vehicles evaluation, taking into account the score attributed based on NO_x and PM emissions values. If a vehicle has good for

NO_x emissions evaluation and good for PM emissions evolution, the final green score will be good. However, in order to render the SEG rating more demanding, the following rule was established: a "lower" score is dominant when combined with a "higher" score.

Table 9.2 – Light passenger gasoline and diesel vehicles final green emissions rating.

Vehicle type	CO	NO_x	PM	Final green rating
Gasoline Vehicle	Good	-	-	Good
	Moderate	-	-	Moderate
	Poor	-	-	Poor
Diesel vehicle	-	Good	Good	Good
	-	Moderate	Good	Moderate
	-	Poor	Good	Moderate
	-	Good	Moderate	Moderate
	-	Moderate	Moderate	Moderate
	-	Poor	Moderate	Poor
	-	Good	Poor	Moderate
	-	Moderate	Poor	Poor
	-	Poor	Poor	Poor

9.1.4 SEG integrated rating

The criteria and rating score for SEG integrated analysis are summarized in Table 9.3. The best rating corresponds to the brightest yellow (since yellow is the standard color for crash testing) and is associated with the "Good" rating. Thus, the brightest yellow is adopted for all the three domains (safety, efficiency and green) reaching "Good". The medium rating is represented by middle yellowish, following by orange (which denotes awareness), for "Moderate" and "Poor" ratings, respectively. SEG rating leads to a qualitative classification of vehicle performance for each domain being analyzed: safety, efficiency and green. The final output of the SEG analysis, described on previous sections, is a combined score which transforms vehicle SEG rating into a quantitative score, designed as SEG. SEG final combined score assumes two principles:

1. On a descending order, the lowest number corresponds to a better vehicle performance, whereas the largest number relates to the poorest performance.
2. The combined score for a vehicle reaching the poorest rating for all the three domains will end up being one, assuming that the weighting factor attributed to each domain is the same.

Table 9.3 – Ranting criteria for SEG integrated analysis based on vehicle category.

Domain	SAFETY		EFFICIENCY		GREEN			
Vehicle category	LPV[1]		LPGV[2]	LPDV[3]	LPGV[2]		LPDV[3]	
Target	**OSS%[4]**	**SRS%[5]**	**CO_2[6] (g/km)**		**CO[7] (g/km)**		**NO_x (g/km)[8]**	**PM (g/km)[9]**
Evaluation criteria	OSS<1.99	SRS<0.503	CO_2<158	CO_2<153	c.c.< 1400 cm^3: 1400≤cc<2000 cm^3: c.c.> 2000 cm^3:	CO<0.710 CO<0.658 CO<0.549	NO_x<0.601	PM<0.0342
	1.99≤OSS<2.754	0.503≤SRS<0.521	158≤CO_2<177	153≤CO_2<161	c.c.< 1400 cm^3: 1400≤cc<2000 cm^3: c.c.> 2000 cm^3:	0.710≤CO<2.39 0.658≤CO<2.18 0.549≤CO<1.74	0.601≤NO_x<0.726	0.0342≤ PM<0.0594
	OSS≥2.754	SRS≥0.521	CO_2≥177	CO_2≥161	c.c.< 1400 cm^3: 1400≤cc<2000 cm^3: c.c.> 2000 cm^3:	CO≥2.39 CO≥2.18 CO≥1.74	NO_x≥0.726	PM≥0.0594

1 Light Passenger Vehicle; **2** Light Passenger Gasoline Vehicle; **3** Light Passenger Diesel Vehicle; **4** Overall Safety Score; **5** Severity Risk Score; **6** Carbon Dioxide emissions in g/km; **7** Carbon Monoxide emissions in g/km; **8** Nitrogen Oxides emissions in g/km; **9** Particulate Matter in g/km.

Table 9.4 shows the conversion of SEG quantitative rating into qualitative score. Similarly to OSS, as SEG increases, vehicle's performance decreases.

Table 9.4 – Converting SEG quantitative rating into a qualitative score.

SEG rating			
Qualitative	Good	Moderate	Poor
Quantitative	0.1	0.5	1

Equation 9.1 shows the generic calculations for SEG final combined score.

$$SEG = \frac{SR * WF_S + ER * WF_E + GR * WF_G}{WF_S + WF_E + WF_G} \quad \text{Equation 9.1}$$

Where: "SR" is the safety rating , "WF_S" is the weighting factor attributed to the safety rating, "ER" is the efficiency rating, "WF_E" is the weighting factor for efficiency rating and "GR" is the green rating, "WF_G" is the weighting factor for green rating. Since SEG aims to provide a flexible classification tool for vehicle performance evaluation, the weighting factor attributed to each domain can be changed, as illustrated in Table 1.1. In scenario 1, Sc.1, is assumed for a neutral user/consumer which would tend to equate each evaluation domain with the same weight, 0.333. In scenario 2, Sc.2, for a user more interested in vehicle safety evaluation, SEG combined score could be calculated given a weighting factor of 75% to the safety rating and 12.5% to efficiency rating and 12.5% to green rating, and so forth, as explained in Table 9.5.

Table 9.5 – Weighting factors for SEG final combined score applying different users profiles

	User profile	SEG final combined score		
		WFs	WF_E	WF_G
Scenarios	**Sc1. Neutral user**	0.333	0.333	0.333
	Sc2. Safety-Conscious user	**0.750**	0.125	0.125
	Sc3. Efficiency-Conscious user	0.125	**0.750**	0.125
	Sc4. Eco-Conscious user	0.125	0.125	**0.750**

The final combined score, SEG, ranges from 0.1 to 1, or 10-100%. Similarly to OSS, as SEG increases, vehicle's performance decreases. A vehicle achieving a SEG rating of "Good", for safety, efficiency and green performance, respectively, will lead to the quantitative scores: 0.1, 0.1, and 0.1, yielding a SEG of 10%, assuming that weighting factor is 0.333. On the other hand, a vehicle raising Poor performance for all the three domains will be scored with a SEG of 100%.

9.2 Results

SEG rating results are integrated in to a scenario base analysis covering several situations. This scenario base was carried out to allow conducting SEG evaluation for the vehicles categories LPGV and LPDV, covering the three engine size categories and emissions standards from Euro 1 through Euro 5. Based on the levels of service standards (LOS) A and F from the Highway Capacity Manual, two speed profiles were assumed for the Portuguese motorways: 120 $km.hr^{-1}$ and 60 $km.hr^{-1}$, [149]. Whereas 120 $km.hr^{-1}$ would represent free flow conditions, and 60 $km.hr^{-1}$ would represent unusual traffic conditions on motorway and/or when the driver is taking the ramp for exit, and the maximum allowed speed is 60 $km.hr^{-1}$.

The scenario base is presented as a matrix, where the variables/information added was as follows: vehicle category, vehicle's engine size and age, vehicle's Euro norm, road speed, emissions for selected pollutants, and vehicle's safety analysis for vehicle involvement in a crash as a single vehicle and vehicle involvement in two-vehicle collisions. Since engine size affects not only vehicle emissions but its safety, scenarios were created to cover all engine size categories. Regarding the safety analysis, crash severity was estimated for the situations were only the vehicle being analyzed was involved in the crash event, and thus the effects of the vehicle belonging to smaller, medium and larger engine sizes categories are model. In addition, for a scenario where the vehicle was involved in a collision with another vehicle, crash severity was estimated taking into account the possible combinations of engine size category for the counterpart. The resulting scenario base matrix has 73linesx74 columns. Only selected scenarios are discussed in this section. First, SEG results for vehicle safety analysis are discussed. Second, SEG results for Euro 1 and Euro 5 vehicles safety and environmental performances driving at 120 $km.hr^{-1}$ vs 60 $km.hr^{-1}$ are presented. Third, results for final combined score, SEG, are presented covering selected scenarios for different users and/or consumers profiles.

9.2.1 Safety analysis

Vehicle safety score is presented for both alternative measures: SRS and OSS, (as explained in section 9.1.1, step 8). SRS is the mean value for severity risk calculated for each target component, FatalSIKV1 and FatalSIKV2. On the other hand, OSS is based on a conditional probability that takes into account the risk of a serious and/or fatality in the vehicle being analysed and also the risk of exposure to the crash event for the vehicles categories involved. For the final combined score, vehicle safety rating is provided by OSS measure. Results for vehicle safety score is presented in Table 9.6. Results are presented for vehicles at each engine size category and then by decreasing order of vehicle age.

Vehicle safety rating would be the same for both driving scenarios, 120 $km.h^{-1}$and 60 $km.h^{-1}$, hence results in Table 9.6 could be apply to those speed levels. It could be expected that vehicle safety

performance would be poor at 120 km.h^{-1}, and it could, if vehicles were truly driving at 120 km.h^{-1} at the time of the crash. No doubt collision speed is a very important variable for crash severity analysis; however speed data is not available from police crash records. For vehicle's injury severity risk modelling, legal speed limit was used, but this variable was not selected by the crash severity prediction models. Even though speed would be selected by the models, that variable would be a categorical variable informing on the legal speed limit only and not on vehicle's driving speed at the moment of the crash.

Table 9.6 – SEG results for vehicle safety.

Subject		As Single		As V1						As V2						Safety score		Safety rating	
				ccV2<1400cm^3:		1400≤cc<2000cm^3		ccV2≥2000cm^3		ccV2<1400cm^3:		1400≤cc<2000cm^3		ccV2≥2000cm^3					
c.c (cm^3)[1]	Age (yr)[2]	FatalSIK[3]	CS%[4]	FatalSIKV1[5]	CS%[6]	FatalSIKV1[5]	CS%[7]	FatalSIKV1[5]	CS%[8]	FatalSIKV2[9]	CS%[10]	FatalSIKV2[9]	CS%[11]	FatalSIKV2[9]	CS%[12]	SRS[13]	OSS[14]%	SRS[3]	OSS[4]
	14	0.637	1.361	0.340	0.209	0.443	0.309	0.653	0.194	0.283	0.190	0.472	0.357	0.822	0.220	0.521	2.839	Poor	Poor
	10	0.484	1.033	0.340	0.209	0.443	0.309	0.653	0.194	0.283	0.190	0.472	0.357	0.822	0.220	0.500	2.511	Good	Moderate
1.3	7	0.369	1.264	0.340	0.272	0.443	0.402	0.653	0.252	0.283	0.192	0.472	0.361	0.822	0.222	0.483	2.964	Good	Poor
	4	0.267	0.749	0.340	0.135	0.443	0.199	0.653	0.125	0.283	0.147	0.472	0.277	0.822	0.171	0.469	1.803	Good	Good
	1	0.185	0.520	0.340	0.135	0.443	0.199	0.653	0.125	0.283	0.147	0.472	0.277	0.822	0.171	0.457	1.573	Good	Good
	14	0.754	1.543	0.340	0.220	0.443	0.325	0.653	0.204	0.283	0.182	0.472	0.342	0.822	0.211	0.538	3.028	Poor	Poor
	10	0.620	1.269	0.340	0.220	0.443	0.325	0.653	0.204	0.283	0.182	0.472	0.342	0.822	0.211	0.519	2.754	Moderate	Poor
1.7	7	0.505	1.819	0.340	0.310	0.443	0.458	0.653	0.287	0.283	0.257	0.472	0.485	0.822	0.298	0.503	3.915	Moderate	Poor
	4	0.389	1.176	0.340	0.225	0.443	0.332	0.653	0.208	0.283	0.221	0.472	0.415	0.822	0.256	0.486	2.833	Good	Poor
	1	0.284	0.859	0.340	0.225	0.443	0.332	0.653	0.208	0.283	0.221	0.472	0.415	0.822	0.256	0.471	2.516	Good	Moderate
	14	0.903	0.643	0.340	0.074	0.443	0.110	0.653	0.069	0.283	0.068	0.472	0.128	0.822	0.079	0.559	1.169	Poor	Good
	10	0.832	0.592	0.340	0.074	0.443	0.110	0.653	0.069	0.283	0.068	0.472	0.128	0.822	0.079	0.549	1.119	Poor	Good
2.5	7	0.756	1.076	0.340	0.117	0.443	0.173	0.653	0.108	0.283	0.128	0.472	0.240	0.822	0.148	0.538	1.990	Poor	Moderate
	4	0.659	0.352	0.340	0.090	0.443	0.133	0.653	0.083	0.283	0.072	0.472	0.135	0.822	0.083	0.525	0.947	Poor	Good
	1	0.547	0.292	0.340	0.090	0.443	0.133	0.653	0.083	0.283	0.072	0.472	0.135	0.822	0.083	0.509	0.887	Moderate	Good

1 Engine size of vehicle being analyzed; **2** Age of vehicle being analyzed; **3** Probability of a serious injured and/or killed in the crash involving only the subject vehicle; **4** Component associated with the risk of a single crash event; **5** Probability of a serious injured and/or killed in vehicle V1; **6** Component associated with the risk of V1 involvement with a counterpart with ccV2<1400 cm^3; **7** Component associated with the risk of V1 involvement with a opponent with 14≤ccV2<2000 cm^3; **8** Component associated with the risk of V1 involvement with a counterpart with ccV2≥2000 cm^3; **9** Probability of a serious injured and/or killed in vehicle V2; **10** Component associated with the risk of V2 involvement with a counterpart with ccV1<1400 cm^3; **11** Component associated with the risk of V2 involvement with a counterpart with 1400≤ccV1<2000 cm^3; **12** Component associated with the risk of V2 involvement with a counterpart with ccV1≥2000 cm^3; **13** Severity Risk Score; **14** Overall Safety Score, (see ratings in Table 9.3).

The highest OSS was 3.915, associated with 7 yr vehicles and 1700 cm^3 engines, Table 9.6. On the other hand, the best safety score, lowest crash severity and risk of exposure was at 0.887, for the newest and larger engine size vehicle category, 1 yr vehicles with 2500 cm^3 engine. OSS calculations are dependent on vehicles characteristics, but also in vehicles category distribution in the sample. For OSS analysis, the worst safety performance was estimated for vehicles in the categories: 1400≤c.c.<2000 cm^3 and 5≤Age<10 yr. The best safety performance was predicted for vehicles in the categories: c.c.≥2000 cm^3 and 1≤Age<5 yr. For vehicles categories, 1400≤c.c.<2000 cm^3 and 5≤Age<10 yr, the probability FatalSIK was 0.505. However, these vehicles categories represent the highest fraction at the sample, 15.7%. On the other hand, for the categories, c.c.≥2000 cm^3 and 1≤Age<5 yr, the probability FatalSIK was higher, 0.547. Nonetheless, these vehicles categories represent the lowest fraction at the sample, 4.0%. Based on SRS analysis, the highest severity risk, 0.559, was associated to the oldest vehicles in the largest engine size category: 14 yr and 2500 cm^3, Table 9.6. The lowest severity risk, 0.457, was attributed to the newest vehicle models in the smallest engine size category: 1 yr and 1300 cm^3. The SRS results clearly show that a better crashworthiness (lowest risk) is associated to the newest vehicle models, showing that auto-industry have achieved significant improvements during the last years. These results are consistent with previous work that claimed that recent cars protect their drivers better than older cars [49, 59-61]. During the last years the auto industry has significantly improved vehicles' crashworthiness (secondary safety) but also, active safety. These technological developments involve the structure of the vehicles, with progressive crumple zones and a more rigid survival cell, restrain systems (as pretensioning seat belts) and impact absorption systems (as airbags) [49].

Table 9.6 further illustrates the differences in SEG safety rating based on the SRS or OSS measures. For instance, a vehicle with 2500 cm^3 and 14 yr old yield a poor safety rating using SRS, but a good safety performance using OSS measure. As already discussed, older vehicles have poorest crashworthiness. In addition, larger engine size vehicles are associated to more powerful vehicles and they have been linked to potentiate speeding [64]. Hence, crashes involving that vehicle category can increases the probability that its occupants would sustain severe injuries and/or fatalities. In addition, they impose more risk the other vehicle involved in the collision. However, SEG using OSS has rating that vehicle category, c.c.≥2000 cm^3 and 1≤Age<5 yr, with good safety performance, mainly because the probability that as crash is going to involved that category is low. Thus, the risk of exposure is reduced, and hence OSS takes benefit of that, as explained above. It is important to mention that, as more crashes would be added to the crash sample, the probability of crash severity and risk of exposure would become more stable and SRS and OSS would be more accurate.

9.2.2 Environmental performance

Vehicle's emissions models developed in Chapter 8 were applied to estimate emissions for selected pollutants based on vehicles categories and driving scenarios. Following, SEG methodology (see sections 9.1.2 and 9.1.3) was used in the environmental performance evaluation of those vehicles categories. In addition to the analysis of vehicles by engine size and age category, Euro Norms were added as a complement of vehicle's age. Although environmental performance results were obtained for all the vehicles categories discussed in section 9.2.1, for the environmental analysis vehicle's 14 yr old and 1 yr old categories are presented in order to allow the discussion for the earlier and most recent Euro Norms. In Table 9.7, the results for vehicle environmental performance are presented for selected vehicles categories complying with Euro 1 and Euro 5 emission standards, assuming free flow (120 $km.h^{-1}$) and congested scenarios on motorway (60 $km.h^{-1}$).

Regarding to fuel consumption and emissions, vehicles are clearly affected by the driving speed scenarios using the emissions estimation models. For Euro 1 vehicles driving in free flow conditions, 120 $km.h^{-1}$, fuel consumption was significantly higher compared to congestion (60 $km.h^{-1}$). For example, fuel consumption, expressed in terms of CO_2 emissions, for a gasoline vehicle with 1.7 L and 14 yr old, was 189.11 $g.km^{-1}$ and 167.48 $g.km^{-1}$, at 120 $km.h^{-1}$ and 60 $km.h^{-1}$, respectively, Table 9.7. On the other hand, for 1 yr old car complying with Euro 5, in the categories mentioned above, fuel consumption was 159.25 $g.km^{-1}$ and 127.65 $g.km^{-1}$, at 120 $km.h^{-1}$ and 60 $km.h^{-1}$, respectively, Table 9.7.

Concerning emissions, for LPGV, a Euro 1 vehicle with 1.3L engine, driving at 60 $km.h^{-1}$, CO emissions were lower than at 120 $km.h^{-1}$, 2.117 $g.km^{-1}$ and 2.874 $g.km^{-1}$, respectively. The same trend was found for Euro 5 vehicles under the same driving scenarios. Comparing CO emissions for LPGV Euro 5 and Euro 1, there were a noteworthy reduction of this pollutant, 2.783 $g.km^{-1}$ and 0.885 $g.km^{-1}$, as shown in Table 9.7 at 120 $km.h^{-1}$. Similarly to LPGV, for LPDV emissions reductions were also detected from the 120 $km.h^{-1}$ to 60 $km.h^{-1}$ driving scenarios and when comparing older vehicle models with newer ones. Assuming that Euro 5 and Euro 1 vehicles were driving under the same conditions, a Euro 5 vehicle would emit less 0.244 NO_x g per kilometres driven than a Euro 1 vehicle, 0.600 $g.km^{-1}$ and 0.844 $g.km^{-1}$, respectively

Table 9.7 – Selected results for a scenario using Euro 1 and Euro 5 vehicles analysis in SEG methodology.

Subject Vehicle Characteristics					Emissions (g.km^{-1}) for 120 (km.h^{-1}):				Emissions (g.km^{-1}) for 60 (km.h^{-1})				Safety	
Norm[1]	Vehicle category		Subject vehicle		Efficiency	Green			Efficiency	Green				
	LPGV[2]	LPDV[3]	c.c (cm^3)[4]	Age (yr)[5]	**CO_2**[6]	**CO**[7]	**NO_x**[8]	**PM**[9]	**CO_2**[6]	**CO**[7]	**NO_x**[8]	**PM**[9]	SRS[10]	OSS[11]
Euro 1	1	0	1.3	14	173.679	2.874	NA	NA	152.042	2.117	NA	NA	0.521	2.839
	1	0	1.7	14	189.114	2.874	NA	NA	167.477	2.117	NA	NA	0.538	3.028
	1	0	2.5	14	197.453	2.874	NA	NA	175.816	2.117	NA	NA	0.559	1.169
Euro 5	1	0	1.3	1	146.254	0.885	NA	NA	127.646	0.559	NA	NA	0.457	1.573
	1	0	1.7	1	159.529	0.885	NA	NA	140.921	0.559	NA	NA	0.471	2.516
	1	0	2.5	1	166.700	0.885	NA	NA	148.092	0.559	NA	NA	0.509	0.887
Euro 1	0	1	1.3	14	176.092	NA	0.844	0.152	157.518	NA	0.628	0.137	0.521	2.839
	0	1	1.7	14	176.092	NA	0.844	0.152	157.518	NA	0.628	0.137	0.538	3.028
	0	1	2.5	14	206.074	NA	0.844	0.152	187.500	NA	0.628	0.137	0.559	1.169
Euro 5	0	1	1.3	1	137.792	NA	0.600	0.005	120.265	NA	0.444	0.003	0.457	1.573
	0	1	1.7	1	150.295	NA	0.600	0.005	132.768	NA	0.444	0.003	0.471	2.516
	0	1	2.5	1	157.049	NA	0.600	0.005	139.523	NA	0.444	0.003	0.509	0.887

1 Emission standard norm; **2** Light Passenger Gasoline Vehicle; **3** Light Passenger Diesel Vehicle; **4** Engine size of vehicle being analyzed; **5** Age of vehicle being analyzed; **6** Carbon Dioxide emissions in g/km; **7** Carbon Monoxide emissions in g/km; **8** Nitrogen Oxides emissions in g/km; **9** Particulate Matter in g/km; **10** Severity Risk Score; **11** Overall Safety Score; **NA** Means that the pollutant was not applicable to the vehicle category.

9.2.3 SEG integrated ratings

In this section, SEG rating results are first presented as a qualitative evaluation of vehicles' performance. Additionally, SEG final combined score, as quantitative score are also presented. Since SEG aims to provide a flexible classification tool for vehicle performance evaluation based on the user and/or consumer profile, vehicles' performance is discussed based on different users profiles and domain interests.

9.2.3.1 SEG rating

SEG ratings results are shown in Table 9.8. When comparing Euro 5 and Euro 1 vehicles, significant differences in safety and environmental performances were found, as explained next.

First results are discussed for gasoline vehicles at 120 $km.h^{-1}$ driving scenario. In Table 9.8, for vehicles in the older category (complying with Euro 1) with the smaller engine size category, c.c.<1400 cm^3, SEG rating was: poor, moderate and poor. On the other hand, Euro 1 vehicles with c.c.≥2000 cm^3, SEG rating was: good, poor and poor. While for the smaller engine size category, vehicles reached moderate fuel efficiency performance, for the larger engine size, vehicles revealed poor efficiency performance, since fuel consumption was larger. For vehicles complying with Euro 5 with c.c.<1400 cm^3, SEG rating was: good, good and moderate. For Euro 5 vehicles with c.c.≥2000 cm^3, SEG rating was: good, moderate and moderate. For newer vehicles, safety improvements as well as environmental performance are evident. For the larger engine size category, vehicle's use more fuel for driving in the same conditions, as vehicles with the smaller engine size. SEG results showed that newer models are safer, suggesting protecting its occupants in ran off road or rollover crash, but also when involved in collision with other vehicle. SEG safety findings supports other research that concluded that drivers of recent cars are better protected than drivers of older vehicles [49, 59, 61, 99]. The improved vehicle efficiency when comparing the earlier Euro 1 models with the recent Euro 5 models could be explained due to the fact that newer vehicles when introduced in the market benefit from advanced engine technology and optimize fuel injection leading to a better fuel efficiency. Improvements in the after treatment monitoring and control yield to emissions reductions in general, such as on CO emissions, contributing to improvements in environmental performance. These findings are consistent with previous research, showing that during the last years, improvements in vehicles design have contributed to improve green performance allowing significant reductions in exhaust emissions [37, 62].

Second, for diesel vehicles at 120 $km.h^{-1}$ driving scenario, no significance differences were found between diesel and gasoline vehicles in the older vehicle category (Euro 1), with the exception that for the smaller engine size category, c.c.<1400 cm^3, SEG rated gasoline vehicles as more efficient than diesel vehicles, moderate and poor, respectively. For vehicles complying with Euro 5, environmental performance was better than for gasoline vehicles. As mentioned, safety

performance is the same as for gasoline vehicles, since safety was not influenced by fuel type. For instance, while for gasoline vehicles, green performance was moderate for all the three engine size categories, for diesel ones were good for the smaller and medium size categories. SEG finding is consistent with previous work, revealing inherent efficiencies of diesel engines and higher energy content of diesel [150]. For green performance, gasoline vehicles achieved a moderate rate for all the three engine size categories, whereas diesel vehicles raised good performance for all categories. Whereas none of the gasoline vehicles reached a good rating simultaneously for all the three domains, diesel vehicles raised good rating for safety, efficiency and green, for the category c.c.<1400 cm^3.

Third, considering 60 $km.h^{-1}$ driving scenario, vehicles safety rating was the same as for the 120 $km.h^{-1}$, as previously explained in section 9.2.1. However for vehicles environmental performance, technological improvements were significant. For Euro 1 vehicles with c.c.<1400 cm^3, SEG rating was: poor, good and moderate, whereas for c.c.≥2000 $cm^{3,}$ SEG rating was: good, moderate and poor, Table 9.8. For vehicles complying with Euro 5 with c.c.<1400 cm^3, SEG rating was: good, good and good. For Euro 5 vehicles with c.c.≥2000 cm^3, SEG rating was: good, good and moderate. Newer vehicles models with the smaller engine size category driving at lower speed yield to a good rating for all the three domains.

For diesel vehicles, in general SEG ratings improved and more vehicle categories raised good rating for safety, efficiency and green performance simultaneously. For example, vehicles complying with Euro 5 and c.c.<1400 cm^3, SEG rating was: good, good, good, for both driving scenarios: 120 $km.h^{-1}$ and 60 $km.h^{-1}$, as observed in Table 9.8. As far as safety performance, results were not affected neither by the scenario speed, neither by vehicle's fuel type, since crash severity prediction were not function of speed neither of fuel type.

Table 9.8 – SEG rating results for Euro 1 and Euro 5 vehicle's safety, fuel efficiency and green performances.

Vehicle					Rating results for 120 (km.h^{-1})				Rating results for 60 (km.h^{-1})			
Norm[1]	LPGV[2]	LPDV[3]	c.c (cm^3)[4]	Age (yr)[5]	**SAFETY SRS[6]**	**SAFETY OSS[7]**	**EFFICIENCY[8]**	**GREEN[9]**	**SAFETY SRS[6]**	**SAFETY OSS[7]**	**EFFICIENCY[8]**	**GREEN[9]**
Euro 1	1	0	1300		Poor	Poor	Moderate	Poor	Poor	Poor	Good	Moderate
	1	0	1700	14	Poor	Poor	Poor	Poor	Poor	Poor	Moderate	Moderate
	1	0	2500		Poor	Good	Poor	Poor	Poor	Good	Moderate	Poor
Euro 5	1	0	1300		Good	Good	Good	Moderate	Good	Good	Good	Good
	1	0	1700	1	Good	Moderate	Moderate	Moderate	Good	Moderate	Good	Good
	1	0	2500		Moderate	Good	Moderate	Moderate	Moderate	Good	Good	Moderate
Euro 1	0	1	1300		Poor	Poor	Poor	Poor	Poor	Poor	Moderate	Poor
	0	1	1700	14	Poor	Poor	Poor	Poor	Poor	Poor	Moderate	Poor
	0	1	2500		Poor	Good	Poor	Poor	Poor	Good	Poor	Poor
Euro 5	0	1	1300		Good	Good	Good	Good	Good	Good	Good	Good
	0	1	1700	1	Good	Moderate	Good	Good	Good	Moderate	Good	Good
	0	1	2500		Moderate	Good	Moderate	Good	Moderate	Good	Good	Good

1 Emission standard norm; **2** Light Passenger Gasoline Vehicle; **3** Light Passenger Diesel Vehicle; **4** Engine size of vehicle being analyzed; **5** Age of vehicle being analyzed;**6** Safety rating using Severity Risk Score; **7** Safety rating using Overall Safety Score.

9.2.3.2 SEG final combined score

When evaluating vehicle's performance with SEG combined score, results are similar to SEG rating, although combined into a single score. Final combined score for SEG results are presented in Table 9.9, using similar selected scenarios to the previous ones used for SEG rating. In addition, based on Table 9.5 four profiles were added to differentiate vehicle performance evaluation according to which the user/consumer favors or not: neutral, safety, efficiency or ecology. For simplicity, results are shown for free flow conditions, 120km.hr^{-1}, considering normal traffic conditions for Portuguese motorways. Although, driving scenarios in roads with different speed limits are possible to be considered.

For gasoline vehicles, (LPGV), none vehicle had reached the best combined performance score, 0.100, in Table 9.9. For newer vehicles in the smaller engine size category, c.c.<1400 cm^3, SEG was very good, 0.150, either from the perspective of a safer profile, either from the perspective of an efficient profile (whom may be concerning with vehicles fuel consumption). For an ecologist user, the vehicles in this category could not be desirable since SEG was 0.400. However, for a neutral user, SEG of 0.234 could be accepted as a sufficient vehicle performance.

For diesel engines, (LPDV), the newer vehicles complying with Euro 5 and in the small engine size category reached the best SEG performance, achieving 0.100 for all the profiles, in Table 9.9. As explained in section 9.1.4, the best score, maximum vehicle performance, is attributed to vehicle's reaching 0.100. Thus, vehicles in this category could result very appealing for any user style: the safety-conscious, or eco-conscious user, and even for the neutral user. For a user to whom vehicle safety performance would be the most important, either the above category, either Euro 5 vehicles with the larger engine size category would be preferable, SEG combined scores of 0.100 and 0.150, respectively. For example, a safety-conscious consumer interested in a larger car for work proposed or family comfort, and who seeks for safety as a priority, new diesel vehicles in the engine size category c.c.≥2000 cm^3 would be recommend, since the SEG combined score for this category was 0.150. However, for a user more interested in fuel consumption, efficiency-conscious user, this category would not be so appealing, SEG combined score of 0.400. For the efficient user, whom saving fuel is the most important, to would be recommended to shows between the following categories: On the other hand, for an environmental-friendly user, Euro 5 vehicles in the intermediate engine size category, 1400 cm^3 ≤c.c.<2000 cm^3, would result very appealing, due to SEG of 0.150 for both efficiency-conscious user and eco-conscious user. Even though this vehicle category would save fuel and emissions, a user and/or consumers in favor of safety, could not consider this category so tempting due to 0.4 SEG score.

As presented above, vehicle's performance evaluation using SEG combined score offers an easier approach for faster user compression since vehicle evaluation is summarized into a single score. On a different approach, SEG rating exhibits an individualized and separated evaluation of safety, fuel efficiency and green performance.

Table 9.9 – Selected combined score results for a scenario using vehicles Euro 1 and Euro 5.

Vehicle					Analysis for 120 (km.h^{-1})							Analysis for 60 (km.h^{-1})						
					SEG quantitative score			SEG combined score				SEG quantitative score			SEG combined score			
								User profile							User profile			
Norm[1]	LPGV[2]	LPDV[3]	c.c (cm^3)[4]	Age (yr)[5]	Safety	Efficiency	Green	Sc1[6]	Sc2[7]	Sc3[8]	Sc4[9]	Safety	Efficiency	Green	Sc1[6]	Sc2[7]	Sc3[8]	Sc4[9]
Euro 1	1	0	1300		1	0.5	1	0.833	0.938	0.625	0.938	1	0.1	0.5	0.533	0.825	0.263	0.513
	1	0	1700	14	1	1	1	1	1	1	1	1	0.5	0.5	0.666	0.875	0.563	0.563
	1	0	2500		0.1	1	1	0.700	0.325	0.888	0.888	0.1	0.5	1	0.533	0.263	0.5125	0.825
Euro 5	1	0	1300		0.1	0.1	0.5	0.234	0.150	0.150	0.400	0.1	0.1	0.1	0.100	0.100	0.100	0.100
	1	0	1700	1	0.5	0.5	0.5	0.500	0.500	0.500	0.500	0.5	0.1	0.1	0.233	0.400	0.150	0.150
	1	0	2500		0.1	0.5	0.5	0.366	0.200	0.450	0.450	0.1	0.1	0.5	0.233	0.150	0.150	0.400
Euro 1	0	1	1300		1	1	1	1	1	1	1	1	0.5	1	0.832	0.938	0.625	0.938
	0	1	1700	14	1	1	1	1	1	1	1	1	0.5	1	0.832	0.938	0.625	0.938
	0	1	2500		0.1	1	1	0.700	0.325	0.888	0.888	0.1	1	1	0.699	0.325	0.888	0.888
Euro 5	0	1	1300		0.1	0.1	0.1	0.100	0.100	0.100	0.100	0.1	0.1	0.1	0.100	0.100	0.100	0.100
	0	1	1700	1	0.5	0.1	0.1	0.233	0.400	0.150	0.150	0.5	0.1	0.1	0.233	0.400	0.150	0.150
	0	1	2500		0.1	0.5	0.1	0.233	0.150	0.400	0.150	0.1	0.1	0.1	0.100	0.100	0.100	0.100

1 Emission standard norm; **2** Light Passenger Gasoline Vehicle; **3** Light Passenger Diesel Vehicle; **4** Engine size of vehicle being analyzed; **5** Age of vehicle being analyzed; **6** Neutral user; **7** Safety-Conscious user; **8** Efficiency-Conscious user; **9** Eco-Conscious user.

9.3 Concluding Remarks

Based on the crash sample, SEG major findings for a scenario base analysis are summarized as follows.

As SEG rating, gasoline vehicles in an older category (complying with Euro 1) with the smaller engine size category, c.c.<1400 cm^3, achieved: poor, moderate and poor. Euro 1 vehicle with c.c.≥2000 cm^3, achieved: good, poor and poor. Smaller engine size use less fuel then larger engines. For these vehicles, although in the same age category, vehicles with larger engine size revealed good safety performance, whereas vehicles in the smaller engine size showed poor performance. Thus, larger vehicles (probably with more weight and extra length) seemed to offer better protection to its occupants. Recent vehicles (complying with Euro 5) with c.c.<1400 cm^3 achieved a SEG rating as: good, good and moderate. Also Euro 5 vehicles, but with c.c.≥2000 cm^3, SEG rating was: good, moderate and moderate. Thus, for newer vehicles, safety performance seemed not to be affected by engine size category, but it affects fuel consumption. When comparing the earlier vehicles with more recent vehicles in the crash sample, improvements in vehicle design, and fuel injection moved vehicles towards performance optimization. For diesel vehicles, SEG rating revealed better performance than the same age and engine size categories in gasoline vehicles. Several categories reached good rating for all the three domains, whereas for gasoline vehicles only Euro 5 vehicles with c.c.<1400 cm^3 raise good ratings for safety, efficiency and green. As SEG final combined score, results vary between 1 to 0.100, for the worst and the best vehicle performance, respectively. Recent gasoline vehicles with c.c.<1400 cm^3, achieved 0.150, either from the viewpoint of either a safety-conscious user or a efficiency-conscious user. However, this vehicle evaluation under an eco-conscious user, yield to a final combined score of 0.400, and thus, this vehicle category would not be recommended. Newer diesel vehicles complying with Euro 5 and in the smaller engine size category reached the best performance, 0.100 for all the users profiles. As a conclusion, main advantages of SEG are highlighted.

1. Is designed to be easy-to-use tool to assist consumers in vehicle's selection based on users profile style: neutral, safety-conscious, efficient-conscious or eco-conscious.
2. Allows the evaluation of vehicle's safety performance for single-vehicle crashes and for two vehicle collisions, as well as the comparison between vehicles above a 113 kg weight range.
3. Allows the evaluation of vehicle's efficiency and green performance ratings in a flexible scale for different scenarios and taking into account vehicles' engine size and age, category.
4. Overall safety rating is for the first time provided for the analysis of single-vehicle crash but also for the situation where the vehicle is involved in collision. It takes into account the effect of vehicle characteristics in crashworthiness. In addition, it includes risk of exposure for the vehicle category being analysed.

CHAPTER 10

CONCLUSIONS AND FUTURE WORK

Chapter 10 presents the key concluding remarks of the present research. It is organized as follows. First, conclusions are presented based on the stated research objectives at the beginning of the dissertation. Next, the major findings are highlighted, followed by the scope and limitations of the methodology and findings. Finally, recommendations for future work are offered.

10.1 Conclusions

The following main conclusions can be drawn in terms of meeting the research objectives, most of which were fully accomplished.

The **1st objective** was to determine if vehicles characteristics affect crash outcomes, and to identify which factors are more significant in predicting crash injury severity.

This objective was fully achieved and research findings have shown the impact of vehicles characteristics on crash injury risk. Mainly, vehicle's age, engine size, weight, and wheelbase have been identified as important predictors of crash severity. These findings were further explored, as the subsequent objectives were accomplished, and are explained next.

The **2nd objective** was to develop decision models to predict the probability of a serious injury and/or fatality in single-vehicle and two-vehicle collisions, based on the technical characteristics of the vehicles involved and crash information.

Based on the original crash sample, in single-vehicle crashes the presence of alcohol and/or drugs was linked to a higher crash severity. A classification and regression tree analysis revealed that the presence of alcohol and/or drugs was the most important risk factor, followed by the age of the vehicle and weather conditions, yielding values of (1), (0.85) and (0.72), for variable importance, respectively.

For two-vehicle collisions, the decision model for classifying overall crash severity prediction, expressed by the binary target FatalSIK, identified the age differential between the two vehicles involved as the most important factor in predicting crash severity, followed by the age of the vehicle V1, alcohol and/or drugs and weight of vehicle V2, yielding values of (1), (0.87), (0.64) and (0.59), for variable importance, respectively. When focusing on the crash severity sustained by the occupants of the subject vehicle V1, expressed by the binary target FatalSIKV1, the decision tree model also identified age differential between the two vehicles involved as the most important factor for crash severity prediction, followed by the engine size of the opponent vehicle, yielding (1) and (0.72) for variable importance, respectively. When analyzing the risk of severe injuries in the opponent vehicle, expressed by binary FatalSIKV2, the most important risk factors were: (a) wheelbase differences between the two vehicles involved, (b) engine size of vehicle V2 and (c) presence of alcohol and/or drugs, yielding values of (1), (0.94), (0.57), for variables importance, respectively. These findings confirm that it is important not only to consider vehicle's individual characteristics but also its differential between the vehicles involved in the collision. Also, the variables importance within the classification tree models for FatalSIKV1 and FatalSIKV2 prediction suggest that vehicles' characteristics play a more relevant role comparatively to other crash related variables. Indeed, it is interesting to note that the engine size of vehicle V2 was important

for both targets prediction, thought the effect of this variable is the opposite for each target. For FatalSIKV1, larger engine sizes of the opponent vehicle increased the probability of FatalSIKV1. This suggests that occupants of vehicle V1 are at higher risk when the opponent vehicle has a larger engine size. On the other hand, when predicting the probability of FatalSIKV2, the involvement of a larger engine size of the subject vehicle, V2 in this case, in the collision were associated with lower probability of a serious injury and/or fatality among its occupants. This finding suggest that vehicles with larger engine size offer its occupants a better protection, however they impose higher risk for the occupants of the other vehicle involved, occupants of vehicle V1.

The **3rd objective** was to develop advanced logistic regression models to predict the probability of a serious injury and/or fatality in single-vehicle crashes and in two-vehicle collisions, based on the technical characteristics of the vehicles involved.

Regarding crash severity prediction (expressed by FatalSIK) for single-vehicle crashes, Model-IB-S (pp 116) helps to explain the effect of vehicle's characteristics on crash outcomes. This model showed that the age of the vehicle and engine size were associated with an increase probability of FatalSIK. Model predictors such as vehicle's age and engine size were statistically significant, with p-values<0.0079 and <0.0229, respectively. The auto industry has improved not only vehicles' crashworthiness (secondary safety), but also active safety, thus occupants in a newer vehicle are better protected than in an older vehicle. Model accuracy rate was estimated at 58.0% (S.D. 3.1)

For two-vehicle collisions, models were developed to predict injury severity risk for each individual vehicle occupants taking into account not only the vehicle's own capability to protect its occupants, but also the risk posed by the opponent vehicle. When predicting crash severity in vehicle V1, FatalSIKV1, Model-II-T (pp 128) suggests that the engine size of the opponent vehicle, vehicle V2, increases the probability of major injuries and/or fatalities among the occupants of the subject vehicle, vehicle V1. The engine size of the opponent vehicle was found to be significant at 10% significance level. Model-II-T yielded good performance with a mean prediction accuracy rate of 61.2% (S.D. 2.4). When analyzing crash severity for occupants in the other vehicle involved, vehicle V2, Model-III-T (pp 131) predicted that the engine size of the opponent vehicle (vehicle V1) heightened the probability of severe injury sustained by the occupants of vehicle V2. The engine size of the opponent vehicle was a significant predictor, with a p-value<0.0387. Model-III-T shows good performance, with mean prediction accuracy rate of 40.5% (S.D. 2.1). It is clear that the consistency between Model-II-T and Model-III-T magnifies the effect of engine size of the opponent vehicle as a significant risk factor when predicting the injury severity suffered by the occupants of the subject vehicle. As vehicle mass is highly correlated with engine size the same conclusion between a collision involving a vehicle of heavier mass and crash severity to occupants of a lighter vehicle can be drawn.

Notwithstanding the constraint of using balanced training samples, final models performance was evaluated using the original sample, where the imbalanced severity was distributed as: 92.4% of non-severe crashes and 7.6% of severe crashes and 96.3% of non-severe crashes and 3.7% of severe crashes, for single vehicle crashes and two-vehicle collisions, respectively. Prediction accuracy for Model-IB-S, Model-II-T and Model-III-T using the original crash sample was as follows: 76.0%, 93.6% and 83.8%, respectively. Next, each model was validated using 10 stratified random samples, and the mean prediction accuracy for Model-IB-S, Model-II-T and Model-III-T was satisfactory, (58.0%, 61.2% and 40.5%, respectively).

In summary, the proposed models' mean prediction accuracy rates were good, simple to apply, provide additional understanding about vehicles' characteristics which contribute to crash severity and they tend to support previous research results in the literature. Some studies seem to be more concerned with the predictive accuracy and the traditional validation (using new data) but fail to reflect other objectives such as interpretability and resource efficiency (in both time and costs), which also determine the empirical adequacy of different algorithms in practice. Beyond balanced approach, the interpretability of models presented in this research is often of even greater importance. Still, further analysis with larger samples size is highly recommended to confirm the validity of the models.

The **4th objective** was to attempt to contrast vehicle brands insofar as their severity involvement in the crash sample occurred as well as within the larger Portuguese fleet.

This analysis resulted in the identification of vehicles from Renault as the most frequent auto brand (14.7%) involved in collisions, among the 1,748 vehicles in the crash sample. The two-vehicle collisions involving a Renault vehicle resulted in almost twice the severity ratio of the overall crash sample, 4.8%, vs. 2.9%, respectively. At the national level, for the same time period (2006-2010) the overall severity ratio for two vehicle collisions was 4.8%. Thus, the above findings could not lead to the conclusion that the Renault brand has a poor crashworthiness performance. Instead, Renault's severity ratio is exactly the same as for the Portuguese two-vehicle collisions fleet. In the case of single-vehicle crashes, Renault was also the most frequent, accounting for 15.8% of the 500 vehicles in the crash sample. Renault's severity ratio was slightly above the severity ratio of the crash sample, 8.3% and 7.4% respectively, which may not be statistically meaningful. However, this brand inference with the Portuguese entire fleet was slightly lower, 8.6% and 8.3%, respectively, but again probably within the margin of error. Because of the above comparisons, brands severity ratio inference analysis must be viewed with extreme caution, and always contrasted in terms of representativeness within the national fleet. In fact, different models of the same brand may perform differently within the fleet.

The **5th and last objective** was to develop a safety, efficiency and environmental performance combined score (herein termed the SEG score) to estimate the impact of vehicle characteristics from the perspectives of crash severity, fuel consumption and pollutants emissions, respectively.

The accomplishment of the last objective allowed a full successful integration of all the domains covered by this research: vehicle's injury severity risk prediction and vehicle's safety performance, emissions estimation and vehicle's fuel efficiency and green performance.

The most relevant differences in vehicle's SEG rating were found between older vehicle models and newer ones and between newer vehicle models using gasoline and diesel fuel, under 120 $km.h^{-1}$ driving scenario. For the smaller engine size category, c.c.<1400 cm^3, SEG rating was: poor, moderate and poor, and good, good, and moderate (for safety, efficiency and green performance), for Euro 1 and Euro 5 gasoline vehicles, respectively. For the larger engine size category, c.c.≥2000 cm^3, SEG rating was: poor, poor and poor and good, moderate, and moderate (for safety, efficiency and green performance), for Euro 1 and Euro 5 gasoline vehicles, respectively. When comparing vehicles in those categories, the age differential between those vehicles models was around 13 years. Thus, improvements in vehicle's stiffness structure, passive safety and active safety features explained the good rating for vehicles safety performance for newer vehicles. Similarly, vehicles' fuel injection improvements have contributed to fuel efficiency and hence, CO_2 emissions have been decreasing within same vehicle category. On the other hand, SEG rated older vehicles category (Euro 1) as poor in terms of green performance for all the three engine size categories. Euro 5 vehicles yielded moderate green performance for all the three engine size categories. When comparing SEG ratings for diesel with gasoline vehicles in the older vehicles category, those performances were similar. Diesel vehicles safety performance was the same as gasoline vehicles since fuel type did not affect injury risk. However, major differences were found between newer vehicles models using gasoline and using diesel. Considering a 120 $km.h^{-1}$ driving scenario, among all tested categories, only Euro 5 diesel vehicles with c.c.<1400 cm^3 raised a good SEG rating for all the three domains: safety, efficiency and green. Euro 5 diesel vehicles achieved good efficiency performance for c.c.<1400 cm^3 and 1400≤c.c.<2000 cm^3. Also, green performance was good for all the Euro 5 diesel engine size categories, whereas for Euro 5 gasoline engine size categories, green performance was moderate based on SEG raking. The improved vehicle efficiency is the result of advanced engine technology and optimized fuel injection leading to a better fuel use. While vehicles of earlier model year were equipped with initial catalyst, manufactures have installed in recent models: after treatment of exhaust emissions (such as NO_x reduction and PM oxidation), particle filters (in diesel vehicles) and more efficiency catalytic converters (in gasoline vehicles). In a 60 $km.h^{-1}$ driving scenario, vehicles' performance was better, and more vehicle categories achieved good rating for safety, fuel efficiency and green performance simultaneously. Interesting to notice that reducing driving speed, newer vehicle models achieved good efficiency performance despite of engine size category and fuel type. Thus, vehicle design

matters, but the way vehicle is driven also plays an important role, in particular, in fuel consumption.

SEG has the potential to provide an important selection base of information for consumers, the general public, road transportation technicians and automotive engineers.

Concluding remarks based in the crash sample explored in this research are summarized below.

1. Crash severity for single-vehicle crashes was twice as higher as the crash severity for two-vehicle collisions. This finding may suggest that for crashes involving one car, vehicle crashworthiness may be offset by the driver speeding behavior yielding an increased risk of a severe crash outcome.
2. Engine size of the vehicle was identified as a significant predictor for crash severity across all crash severity prediction models. The effect of this risk factor depends on the number of vehicles involved in the crash. For two-vehicle collisions, as the engine size of the other vehicle involved increases, the probability of severity injury increases for the subject vehicle. Engine size seemed to suggest a protective effect for vehicle's occupants and at same time imposes an increased risk towards the occupants of the opponent vehicle. On the other hand, for single-vehicle crashes, engine size may mask the effects of driver behaviour. Larger engine size (as a proxy of vehicle power) could be associated with greater speeds and thus, yielding an increased severity risk. This is especially true in the case of luxury and sport cars. For two-vehicle collisions there is evidence that engine size reflects the effect of vehicle characteristics on crash severity risk. For single-vehicle crashes, although there is no factual evidence based on the crash sample, it could be possible that engine size may emphasize driver aggressiveness.
3. Vehicle safety performance was dependent on the vehicle' technical characteristics but also on risk of exposure based on vehicle's category frequency in the crash sample. Additionally, the composition of the car fleet also will affect vehicle safety crashworthiness in two-vehicle collisions.
4. SEG findings clearly confirm the progress achieved by the auto industry in vehicle design, as well as the positive effects of law enforcement and emissions regulations for road vehicles. Thus, the SEG results allow us to answer the question: "Is there a trade-off between vehicle's safety, efficiency and green performances?". The simple answer is "No". The results presented in this research showed that newer vehicles are safer, use less fuel and hence, fewer emissions, when compared with older vehicle models in the same weight range. Mainly, advanced technology and improved vehicle design are very much reflected in SEG ratings, and it is evident that newer vehicles achieve good performance on all three criteria. Newer vehicles models, however, should not be downsized, but rather, take advantage of new technologies of mass reduction and materials, such as aluminum and

high-strength steel, to be lighter and resistant, not smaller. Vehicle size matters in protecting vehicle occupants; but this should not impose a conflict with the goal of improved fuel efficiency and emissions control technologies. What is required is decision making and setting agreements to make advanced technologies accessible to auto brands in order to improve the performance of car fleet. Safety goals and environmental goals drive together and save lives.

10.2 Research Limitations

1. Police accident reports are used worldwide for crash analysis and road safety. However several authors have claimed the misclassification of injury severity among road casualties in police reports. Studies have claimed that police reports overestimate injury severity significantly [95]. Whereas fatal casualties are quite clearly defined and well reported, non-fatal casualties could be biased. In this research, injury level was recorded as stated in crash reports. However the author is aware that injuries classification could possibly be biased namely because the injury condition may change after the victims' entrance in the hospital. In addition, only in 2010 did Portugal start recording road victims on the 30 days basis. Thus, crash outcomes collected from police records underestimate any fatality that may have happened following 24 hours after the crash.
2. This research would be improved if crash report records would provide information on the number of vehicle's occupants, whether injured or not, vehicle kilometers driven, and the speed of the vehicle at the moment of the crash. While speed has been identified as the most important factor to affect crash severity outcomes, this key variable is not available on Portuguese crash records. Hence for the crash severity prediction, the legal speed limit has been used as a proxy of vehicle's speed. Also for the emissions estimation, the vehicle's travelling speed was assumed to be the legal speed limit for the road where the vehicle was traveling at the time that was involved in the crash. Incorporating additional variables will improve models accuracies.
3. In Portugal, crash data are not available in digital files to download, which are easily accessible across the globe. Instead, the author was required to manually collect data from police crash records at the Police Office in Oporto. In addition, crash, vehicle and road safety data are not centralized, depending on the type of information requested; at least three key players are needed: Police Forces, IMT (former IMTT) and ANSR. Hence, complementary data needed for the crash database development involved another institution, IMT, that manages a database on vehicle technical features to match vehicle registration plate (extracted from crash reports). In an earlier phase of this research, vehicle technical features for two hundred of vehicles were obtained from the IMTT Oporto in a voluntary act. Due to this nature, this assistance is much appreciated but was very

time consuming and dependent on the goodwill of participants. The full crash database development requested the access to model technical data for 2,248 vehicles in total. By the end of 2011 cooperation with IMTT Lisbon made it possible to access to vehicles technical details, and hence, the crash database was developed.

4. Finally, this research faced several challenges due to the sample size limitations. It would be beneficial to have had access to larger samples, and having access to the population of Portuguese collisions involving any level of injuries would be desirable. To be able to work with a population of crashes the time needed to collect data would be infeasible under the study program that the corresponding author is accomplishing. The final crash database has been completed for 1,374 crashes and included a large number of vehicles, 2,248. However the collected crash sample showed a very low proportion of severe relative to non-severe events. It must be remembered that for the entire sample, only 70 observations were related to severe events. Thus the targets with interest for crash severity modeling were distributed as follows: 38 severe crashes for single vehicle analysis, 32 overall severe crashes for two-vehicle collisions and among those, 21 resulted in severe outcomes for vehicle V1 and 14 for vehicle V2. As result, modeling all the designed targets in this research would benefit from a larger sample size which would provide more targets for crash severity modeling.

10.3 Future Work

As a final remark of the conclusions section, driver's behavior was suggested as a factor potentially influencing the risk of exposure to a crash and also, vehicle's fuel consumption and emissions. This finding motivates the following future work needs.

1. Collecting a larger crash sample to improve the development of crash severity training models. To meet international regulations, in 2010 Portugal started to record road deaths on the basis of 30 days-definitions. Collecting a larger sample of crashes will significant improve prediction models robustness. On the other hand, it will allow the portioning of the data for training, testing and validation. In addition, collecting a new sample with crashes after 2010 will reduce the bias associated with the possible misclassification of injury level by police forces.
2. Analyzing the Portuguese drivers' heterogeneity using, a driving simulator. This lab experiment could support data collection on driver's performance and behavioral factors affecting crash involvement and crash severity.

3. Obtaining vehicle technical features from IMTT, in a similar manner to this Doctoral Research. In addition, adding new variables will allow a better comprehension of vehicle's technical dimensions, safety equipment and maintenance conditions. For example, including vehicles kilometers traveled, will improve risk analysis as well as emissions estimation.
4. The severity of occupant injuries is subject to the restraint devices and impact absorption by airbags (if available), structure of the vehicle, position of the occupants in the vehicle and their individual ability to withstand the impact. Future work would benefit using through the use of simulated crash scenarios on high-performance computers to ensure accurate and robust models for crash severity prediction.
5. CORINAIR is one of the most popular tools to estimate vehicles emissions but has some limitations. It does not consider, for instance accelerations that increase energy use and emissions. SEG integrated analysis for vehicles' safety, efficiency and green performance should be developed using AI methods to conduct a multi-objective analysis. This will enable the analyst to assess how vehicles' technical characteristics can be optimized in order to promote vehicles performance for the three domains: safety, fuel efficiency and green emissions reduction.

REFERENCES

1. WHO. Global plan for the decade of action for road safety 2011-2020. 2010.
2. WHO. Global status report on road safety 2013. Supporting a decade for action 2013 ISBN 978 92 4 156456 4.
3. WHO. Air quality and health. 2011; Available from: http://www.who.int/mediacentre/factsheets/fs313/en/.
4. Zervas E. Analysis of the CO2 emissions and of the other characteristics of the European market of new passenger cars. 1. Analysis of general data and analysis per country. Energy Policy. 2010;38(10):5413-5425.
5. EC. Tackilng Climate Change. Summaries of EU Legislation. . 2013 [20 February 2013]; Available from: http://europa.eu/legislation_summaries/environment/tackling_climate_change/l24007_en.htm.
6. EC. White Paper - European transport policy for 2010: time to decide Office for Official Publications of the European Communities 2001. Available from: http://www.central2013.eu/fileadmin/user_upload/Downloads/Document_Centre/OP_Resources/EU-transportpolicy2010_en.pdf.
7. IRTAD. Road Safety. Annual Report 2011. OECD/ITF 2012.
8. ERSO. Traffic Safety Basic Factos 2011. Car occupants. 2012.
9. ETSC. A Challenging Start towards the EU 2020 Road Safety Target . 6th Road Safety PIN Report. 2012.
10. Donário A, Santos R. Economical and Social Cost of Road Accidents in Portugal. . EDIUAL, editor2012.
11. WHO. World report on road traffic injury prevention. 2004. Available from: http://whqlibdoc.who.int/publications/2004/9241562609.pdf.
12. Haddon W. Advances in the Epidemiology of Injuries as a Basis for Public Policy Landmarks in American Epidemiology 1980;95(5).
13. EC. Horizon 2020 Transport Challenge Work Programme 2014-2015 2013.
14. Observatory ERS. Annual Statistics Report 2011. 2011; Available from: http://ec.europa.eu/transport/road_safety/pdf/statistics/dacota/dacota-3.5-asr-2011.pdf.
15. Lopes CV. Estratégia Nacional de Segurança Rodoviária. Simpósio de Segurança Viária. Porto: Faculdade de Engenharia do Porto; 2011.
16. EC. Mobility and Transport. Road Safety. Statistics Accident Data. 2013 [19 February 2013]; Available from: http://ec.europa.eu/transport/road_safety/specialist/statistics/index_en.htm.
17. EU road fatalities [database on the Internet]. 2013 [cited 5th April 2013]. Available from: http://ec.europa.eu/transport/road_safety/pdf/observatory/trends_figures.pdf.
18. ETSC. 2010 Road Safety Target Outcome:100,000 fewer deaths since 2001. 5th Road Safety PIN Report. 2011.
19. Road Safety Evolution EU [database on the Internet]. 2012. Available from: http://ec.europa.eu/transport/road_safety/pdf/observatory/historical_evol.pdf.
20. ANSR. National Road Safety Startegy 2008-2015 2009.
21. Diário da República, 2.ª Serie, N. 252. Despacho n.º 27808/2009. 2009.
22. ANSR. Annual Report for the Year 2010. Road Fatalities in the basis of 30 days. 2011.
23. ANSR. Annual Report for the Year 2011. Road Fatalities in the basis of 30 days. . 2013.
24. ANSR and ISCTE Institute of University of Lisbon. National Road Safety Strategy. Supporting Document Mid-term 2012-2015. 2012.
25. EC. EU transport in figures. Statistical Pocketbook 2012: Publications Office of the European Union, 2012; 2012. Available from: http://ec.europa.eu/transport/publications/statistics/doc/2012/pocketbook2012.pdf.
26. ANSR and ISCTE Institute of University of Lisbon. Term review document to support the National Road Safety Startegy 2012-2015. 2012.
27. EC. Keep Europe moving - Sustainable mobility for our continent Mid-term review of the European Commission's 2001 Transport White Paper. Communication from the Comission to the Council and the European Parliament 2006.
28. Commission EE, cartographer Energy, transport and environment indicators. Eurostat Pocketbooks. : Publications Office of the European Union; 2012.
29. EEA. The contribution of transport to air quality - TERM 2012: Transport indicators tracking progress towards environmental targets in Europe. 2012 Contract No.: No 10/2012.

30. EPA. Most car manufacturers on track to meet 2012 CO2 targets 2012 [23 April 2013]; Available from: http://www.eea.europa.eu/highlights/most-car-manufacturers-on-track/.
31. EC. Regulation (EC) No 443/2009. Setting emission performance standards for new passenger cars as part of the Community's integrated approach to reduce CO2 emissions from light-duty vehicles. 2009.
32. EC. Climate Action. Reducing CO2 emissions from passenger cars. 2012 [23 April 2013]; Available from: http://ec.europa.eu/clima/policies/transport/vehicles/cars/index_en.htm.
33. EC. White Paper- Roadmap to a Single European Transport Area – Towards a competitive and resource efficient transport system. Brussels,
: 2011 28.3.2011. Report No.: Contract No.: COM(2011) 144 final.
34. Hermans E, Brijs T, Wets G, Vanhoof K. Benchmarking road safety: Lessons to learn from a data envelopment analysis. Accident Analysis & Prevention. 2009;41(1):174-182.
35. Koppel S, Charlton J, Fildes B, Fitzharris M. How important is vehicle safety in the new vehicle purchase process? Accident Analysis and Prevention. 2008;40(3):994-1004.
36. Euro NCAP. Frontal Impact 2011 [26 Oct., 2011]; Available from: http://www.euroncap.com/tests/frontimpact.aspx.
37. Chen C, Ren Y. Exploring the relationship between vehicle safety and fuel efficiency in automotive design. Transportation Research Part D: Transport and Environment. 2010;15(2):112-116.
38. NCAP Programmes. 2012 [11 December 2012]; Available from: http://www.globalncap.org/NCAPProgrammes/Pages/IIHS.aspx.
39. IIHS. Safety research and communications 2012 [11th December 2012]; Available from: http://www.iihs.org/.
40. Coelho MC, Andrade J, Soares D, Frey HC, Rouphail NM. A Vehicle Energy Use and Safety Information Support System. 89th Annual Meeting of the Transportation Research Board. Washington, D.C. 2010.
41. Euro NCAP. Comparable Cars. 2011 [26 Oct., 2011]; Available from: http://www.euroncap.com/Content-Web-Page/0f3bec79-828b-4e0c-8030-9fa8314ff342/comparable-cars.aspx.
42. Abdel-Aty M. Analysis of driver injury severity levels at multiple locations using ordered probit models. Journal of Safety Research. 2003;34(5):597-603.
43. Kononen DW, Flannagan CAC, Wang SC. Identification and validation of a logistic regression model for predicting serious injuries associated with motor vehicle crashes. Accident Analysis and Prevention. 2011;43(1):112-122.
44. NHTSA. Traffic Safety Facts. 2010 Data [database on the Internet]. 2012. Available from: http://www-nrd.nhtsa.dot.gov/Pubs/811630.pdf.
45. Delen D, Sharda R, Bessonov M. Identifying significant predictors of injury severity in traffic accidents using a series of artificial neural networks. Accident Analysis & Prevention. 2006;38(3):434-444.
46. Pakgohar A, Tabrizi RS, Khalili M, Esmaeili A. The role of human factor in incidence and severity of road crashes based on the CART and LR regression: a data mining approach. Procedia Computer Science. 2011;3(0):764-769.
47. Awadzi KD, Classen S, Hall A, Duncan RP, Garvan CW. Predictors of injury among younger and older adults in fatal motor vehicle crashes. Accident Analysis & Prevention. 2008;40(6):1804-1810.
48. Fredette M, Mambu LS, Chouinard A, Bellavance F. Safety impacts due to the incompatibility of SUVs, minivans, and pickup trucks in two-vehicle collisions. Accident Analysis and Prevention. 2008;40(6):1987-1995.
49. Martin JL, Lenguerrand E. A population based estimation of the driver protection provided by passenger cars: France 1996-2005. Accident Analysis and Prevention. 2008;40(6):1811-1821.
50. Tefft BC. Risks older drivers pose to themselves and to other road users. Journal of Safety Research. 2008;39(6):577-582.
51. Pompili M, Serafini G, Innamorati M, Montebovi F, Palermo M, Campi S, et al. Car accidents as a method of suicide: A comprehensive overview. Forensic Science International. 2012;223(1–3):1-9.
52. Vrkljan BH, Anaby D. What vehicle features are considered important when buying an automobile? An examination of driver preferences by age and gender. Journal of Safety Research. 2011;42(1):61-65.
53. Evans L. How to Make a Car Lighter and Safer. Bloomfield, Michigan, USA: Science Serving Society 2004.
54. Lie A, Tingvall C. How Do Euro NCAP Results Correlate with Real-Life Injury Risks? A Paired Comparison Study of Car-to-Car Crashes. Traffic Injury Prevention. 2002;3:288-293.
55. Wood DP. Safety and the car size effect: A fundamental explanation. Accident Analysis and Prevention. 1997;29(2):139-151.

56. Wenzel TP, Ross M. The effects of vehicle model and driver behavior on risk. Accident Analysis and Prevention. 2005;37(3):479-494.
57. Robertson LS. Prevention of motor-vehicle deaths by changing vehicle factors. Injury Prevention. 2007;13(5):307-310.
58. IIHS. Car size and weight are crucial. Statut Report 2009;44(4).
59. Broughton J. Car driver casualty rates in Great Britain by type of car. Accident Analysis and Prevention. 2008;40(4):1543-1552.
60. Broughton J. The influence of car registration year on driver casualty rates in Great Britain. Accident Analysis and Prevention. 2012;45:438-445.
61. Méndez ÁG, Aparicio Izquierdo F, Ramírez BA. Evolution of the crashworthiness and aggressivity of the Spanish car fleet. Accident Analysis & Prevention. 2010;42(6):1621-1631.
62. Zachariadis T. The effect of improved safety on fuel economy of European cars. Transportation Research Part D: Transport and Environment. 2008;13(2):133-139.
63. Huang H, Siddiqui C, Abdel-Aty M. Indexing crash worthiness and crash aggressivity by vehicle type. Accident Analysis & Prevention. 2011;43(4):1364-1370.
64. Tolouei R, Titheridge H. Vehicle mass as a determinant of fuel consumption and secondary safety performance. Transportation Research Part D: Transport and Environment. 2009;14(6):385-399.
65. Tolouei R, Maher M, Titheridge H. Vehicle mass and injury risk in two-car crashes: A novel methodology. Accident Analysis & Prevention. 2013;50(0):155-166.
66. Keall MD, Newstead S. Are SUVs dangerous vehicles? Accident Analysis & Prevention. 2008;40(3):954-963.
67. Brozović N, Ando AW. Defensive purchasing, the safety (dis)advantage of light trucks, and motor-vehicle policy effectiveness. Transportation Research Part B: Methodological. 2009;43(5):477-493.
68. Baker BC, Nolan JM, O'Neill B, Genetos AP. Crash compatibility between cars and light trucks: Benefits of lowering front-end energy-absorbing structure in SUVs and pickups. Accident Analysis & Prevention. 2008;40(1):116-125.
69. Richter M, Pape H-C, Otte D, Krettek C. Improvements in passive car safety led to decreased injury severity – a comparison between the 1970s and 1990s. Injury. 2005;36(4):484-488.
70. Broughton J. The benefits of improved car secondary safety. Accident Analysis & Prevention. 2003;35(4):527-535.
71. Euro NCAP. Protocols for Crash Testing. 2013 [16th April 2013]; Available from: http://www.euroncap.com/Content-Web-Page/fb5e236e-b11b-4598-8e20-3eced15ce74e/protocols.aspx.
72. Kok R. New car preferences move away from greater size, weight and power: Impact of Dutch consumer choices on average CO2-emissions. Transportation Research Part D: Transport and Environment. 2013;21(0):53-61.
73. Daziano RA. Taking account of the role of safety on vehicle choice using a new generation of discrete choice models. Safety Science. 2012;50(1):103-112.
74. Kullgren A, Lie A, Tingvall C. Comparison Between Euro NCAP Test Results and Real-World Crash Data. Traffic Injury Prevention. 2010;11(6):587-593.
75. Newstead SV, Keall MD, Watson LM. Rating the overall secondary safety of vehicles from real world crash data: The Australian and New Zealand Total Secondary Safety Index. Accident Analysis & Prevention. 2011;43(3):637-645.
76. Hauer E. Speed and Safety. Transportation Research Record. 2009(2103):10-17.
77. IIHS. Top Safety Picks 2012 2012 [11th December 2012]; Available from: http://www.iihs.org/ratings/tsp_current.aspx.
78. Lund CMFaak. Trends over time in the risk of driver death: what if vehicle design had not improved? Traffic Injury Prevetion. 2006;7.
79. Farmer CM. Crash Avoidance Potential of Five Vehicle Technologies Insurance Institute for Higway Safety, 2006.
80. Jermakian JS. Crash avoidance potential of four passenger vehicle technologies. Accident Analysis & Prevention. 2011;43(3):732-740.
81. IIHS and HLDI. Estimated Time of Arrival. 2012;47.
82. IIHS and HLDI. Crash Avoidance. They are working Insurance claims data show which new technologies are preventing crashes. Status Report. 2012;47(5).
83. Crone SF, Finlay S. Instance sampling in credit scoring: An empirical study of sample size and balancing. Int J Forecasting. 2012;28(1):224-238.
84. SAS Institute Inc. S. Statistics I: Introduction to Anova, Regression, and Logistic Regression. Cary, NC, USA: SAS Institute Inc.; 2009.

85. Ali S A-G. Using logistic regression to estimate the influence of accident factors on accident severity. Accident Analysis & Prevention. 2002;34(6):729-741.
86. Chang LY, Wang HW. Analysis of traffic injury severity: An application of non-parametric classification tree techniques. Accident Analysis and Prevention. 2006;38(5):1019-1027.
87. Das A, Abdel-Aty M, Pande A. Using conditional inference forests to identify the factors affecting crash severity on arterial corridors. Journal of Safety Research. 2009;40(4):317-327.
88. Meng Q, Weng J. Classification and Regression Tree Approach for Predicting Drivers' Merging Behavior in Short-Term Work Zone Merging Areas. Journal of Transportation Engineering. 2012;138(8):1062-1070.
89. SAS Institute Inc. S. Applied Analytics Using SAS®Enterprise MinerTM 5. Instructor-based training. Cary, NC, USA: SAS Institute Inc.; 2007.
90. Savolainen PT, Mannering FL, Lord D, Quddus MA. The statistical analysis of highway crash-injury severities: A review and assessment of methodological alternatives. Accident Analysis and Prevention. 2011;43(5):1666-1676.
91. Bédard M, Guyatt GH, Stones MJ, Hirdes JP. The independent contribution of driver, crash, and vehicle characteristics to driver fatalities. Accident Analysis & Prevention. 2002;34(6):717-727.
92. Kockelman KM, Kweon Y-J. Driver injury severity: an application of ordered probit models. Accident Analysis & Prevention. 2002;34(3):313-321.
93. Xie YC, Zhao KG, Huynh N. Analysis of driver injury severity in rural single-vehicle crashes. Accident Analysis and Prevention. 2012;47:36-44.
94. Hauer E. The frequency-severity indeterminacy. Accident Analysis and Prevention. 2006;38(1):78-83.
95. Tsui KL, So FL, Sze NN, Wong SC, Leung TF. Misclassification of injury severity among road casualties in police reports. Accident Analysis & Prevention. 2009;41(1):84-89.
96. Amoros E, Martin J-L, Laumon B. Under-reporting of road crash casualties in France. Accident Analysis & Prevention. 2006;38(4):627-635.
97. Montella A, Andreassen D, Tarko AP, Turner S, Mauriello F, Imbriani LL, et al. Critical Review of the International Crash Databases and Proposals for Improvement of the Italian National Database. Procedia - Social and Behavioral Sciences. 2012;53(0):49-61.
98. Boufous S, Finch C, Hayen A, Williamson A. The impact of environmental, vehicle and driver characteristics on injury severity in older drivers hospitalized as a result of a traffic crash. Journal of Safety Research. 2008;39(1):65-72.
99. Chen D, Kockelman K, editors. The role of vehicle footprint, height, and weight in crash coutcomes: application of a heteroscedastic ordered proibit model. 91st Annual Meeting of the Transportation Research Board; 2012; Washington, DC 20001 USA.
100. Li Y, Bai Y. Development of crash-severity-index models for the measurement of work zone risk levels. Accident Analysis & Prevention. 2008;40(5):1724-1731.
101. Chang L-Y, Wang H-W. Analysis of traffic injury severity: An application of non-parametric classification tree techniques. Accident Analysis & Prevention. 2006;38(5):1019-1027.
102. Kuhnert PM, Do K-A, McClure R. Combining non-parametric models with logistic regression: an application to motor vehicle injury data. Computational Statistics & Data Analysis. 2000;34(3):371-386.
103. Harrell J, Frank E. Regression Modeling Strategies With Applications to the Linear Models, Logistic Regression, and Survival Analysis. 2001 0-387-95232-2.
104. Kashani AT, Shariat-Mohaymany A, Ranjbari A. A Data Mining Approach to Identify Key Factors of Traffic Injury Severity. Promet-Traffic & Transportation. 2011;23(1):11-17.
105. Sobhani A, Young W, Logan D, Bahrololoom S. A kinetic energy model of two-vehicle crash injury severity. Accident Analysis & Prevention. 2011;43(3):741-754.
106. He H, Garcia EA. Learning from Imbalanced Data. Knowledge and Data Engineering, IEEE Transactions on. 2009;21(9):1263-1284.
107. Kotsiantis S, Kanellopoulos K, Pintelas P. Handling imbalanced datasets: A review. GESTS International Transactions on Computer Science and Engineering. 2006;30.
108. Joshi MV, Kumar V, Agarwal RC. Evaluating boosting algorithms to classify rare cases: comparison and imporvements. First IEEE International Conference on Data Mining2001. p. 257-264.
109. King G, Zeng L. Logistic Regression in Rare Events Data. Political Analysis. 2001;9(2):137-163.
110. Japkowicz N, Stephen S. The class imbalance problem: A systematic study. Intell Data Anal. 2002;6(5):429-449.

111. Chawla NV. C4.5 and Imbalanced Data sets: Investigating the effect of sampling method, probabilistic estimate, and decision tree structure. Workshop on Learning from Imbalanced Datasets II, ICML Washington DC 2003.
112. Cramer JS. Predictive Performance of the Binary Logit Model in Unbalanced Samples. Journal of the Royal Statistical Society: Series D (The Statistician). 1999;48(1):85-94.
113. Maloof MA, editor. Learning when data sets are imbalanced and when costs are unequal and unknown. ICML'03 workshop on learning from imbalanced data sets; 2003; Washington DC.
114. Maalouf M, Trafalis TB. Robust weighted kernel logistic regression in imbalanced and rare events data. Computational Statistics & Data Analysis. 2011;55(1):168-183.
115. Chan PK, Stolfo SJ. Toward scale learning with non-uniform class cost distributions: a case study in credit card fraud detection. Fourth International
Conference on Knowledge Discovery and Data Mining 2001. p. 164-168.
116. Anderson R. The Credit Scoring Toolkit. Theory and Practice for Retail Credit Risk Management and Decision Automation In: Press OU, editor. 2007.
117. Sarma KS, Institute. S. Predictive modeling with SAS Enterprise Miner: practical solutions for business applications. Cary, NC: SAS Institute; 2007.
118. Nisbet R, Elder J, Miner G. Handbook of Statistical Analysis and Data Miming Applications: Elservier Inc. ; 2009.
119. Weiss GM. Mining with rarity: a unifying framework. ACM SIGKDD Explorations Newsletter. 2004;6(1):7-19.
120. Torrão G, Coelho M, Rouphail N. Modeling the impact of subject and opponet vehicle on crash severity in two-vehicle collisions Washington D.C., : 93rd Annual Meeting of the Transportation Research Board; 2014.
121. Torrão G, Coelho M, Rouphail N. Binary Classification and Logistic Regression Models Application to Crash Severity. World Conference on Transportation Research (WCTR). Rio de Janeiro, Brazil.2013.
122. Torrão G, Coelho M, Rouphail N. Effect of Vehicle Size and Weight on Crash Severity: Portugal Experience. World Conference on Transportation Research (WCTR 2010), Lisbon 2010.
123. Fontaras G, Dilara P. The evolution of European passenger car characteristics 2000–2010 and its effects on real-world CO2 emissions and CO2 reduction policy. Energy Policy. 2012;49(0):719-730.
124. Leduc G, Mongelli I, Uihlein A, Nemry F. How can our cars become less polluting? An assessment of the environmental improvement potential of cars. Transport Policy. 2010;17(6):409-419.
125. Franco V, Kousoulidou M, Muntean M, Ntziachristos L, Hausberger S, Dilara P. Road vehicle emission factors development: A review. Atmospheric Environment. 2013;70(0):84-97.
126. Bampatsou C, Zervas E. Critique of the regulatory limitations of exhaust CO2 emissions from passenger cars in European union. Energy Policy. 2011;39(12):7794-7802.
127. Lutsey N. Review of Technical Literature and Trends Related to Automobile Mass-Reduction Technology. 2010 May Report No.: UCD-ITS-RR-10-10.
128. EC. Regulation (EC) No 19/2011. Official Journal of the European Union. 11 January 2011 ed2011.
129. Wenzel T. Analysis of the Relationship Between Vehicle Weight/Size and Safety, and Implications for the Federal Fuel Economy Regulation. Final Report prepared for the Office of Energy Efficiency and Renewable Energy. In: Energy UDo, editor. US2010.
130. Park HS, Dang XP, Roderburg A, Nau B. Development of plastic front side panels for green cars. CIRP Journal of Manufacturing Science and Technology. 2013;6(1):44-52.
131. Coelho M, Torrão G, Emani N, Grácio J. Nanotechnology in automotive industry: research strategy and trends for the future-small objects, big impacts. Journal of Nanoscience and Nanotechnology. 2012;12(8): 6621-6630.
132. Coelho MC, Torrão, G., Grácio, J., editor. Research strategy for the use of nanomaterials in automotive applications – A risk analysis based approach. International Conference on Advanced Nano-Materials; 2010; Agadir, Marrocos.
133. ANSR. Annual Report for the Year 2011. Road Casulaties. 2012.
134. ACAP. Estatísticas do Sector Automóvel 2010 2011.
135. ACAP. Estatísticas do Sector Automóvel 2009. 2010.
136. Refaat M. Data preparation for data mining using SAS. San Francisco: Morgan Kaufmann Publishers; 2007. xxi, 399 p.
137. Chang L-Y, Chien J-T. Analysis of driver injury severity in truck-involved accidents using a non-parametric classification tree model. Safety Science. 2013;51(1):17-22.

138. Chang L-Y, Chen W-C. Data mining of tree-based models to analyze freeway accident frequency. Journal of Safety Research. 2005;36(4):365-375.
139. Cieslak DA, Chawla NV. Learning Decision Trees for Unbalanced Data. In: Daelemans W, Goethals B, Morik K, editors. Machine Learning and Knowledge Discovery in Databases, Part I, Proceedings. Berlin: Springer-Verlag Berlin; 2008. p. 241-256.
140. SAS Institute Inc. S. Applied Analytics Using SAS®Enterprise MinerTM 5. Instructor-based training. Cary, NC, USA: SAS Institute Inc.; 2007.
141. Batista G, Prati R, Monard M. A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data. SIGKDD Explorations. 2004;6(1).
142. Fawcett T. An introduction to ROC analysis. Pattern Recognition Letters. 2006;27(8):861-874.
143. Huang H, Hu S, Abdel-Aty M. Indexing crash worthiness and crash aggressivity by major car brands. Safety Science. 2014;62(0):339-347.
144. Zhang J, Lindsay J, Clarke K, Robbins G, Mao Y. Factors affecting the severity of motor vehicle traffic crashes involving elderly drivers in Ontario. Accident Analysis & Prevention. 2000;32(1):117-125.
145. Evans L. Driver injury and fatality risk in two-car crashes versus mass ratio inferred using Newtonian mechanics. Accident Analysis & Prevention. 1994;26(5):609-616.
146. Wooldridge JM. Introductory Econometrics A Modern Approach. 3rd ed: Thomson South-western; 2006.
147. CORINAIR. European Monitoring and Evaluation Programme (EMEP). Emission Inventory Guidebook - 2009: Exhaust emissions from road transport. European Environment Agency, 2012.
148. Kiely G. Environemental Engineering: McGFraw Hill; 1999.
149. Highway Capacity Manual 2010. edition t, editor: Transportation Research Board; 2010.
150. An F, Earley R, Green-Weiskel L. Global Overview on Fuel Efficiency and Motor Vehicle Emission Standards: Policy Options and Perspectives for International Cooperation. 2011 [updated May 2011]; Available from: https://cleanenergysolutions.org/content/global-overview-fuel-efficiency-and-motor-vehicle-emission-standards-policy-options-and-pers.
151. Safercar.gov. Safety Technology. 2012 [11 December 2012]; Available from: http://www.safercar.gov/staticfiles/safetytech/st_landing_ca.htm#st_tabs.

APPENDICES

Appendix 1: Approaches for risk factors linked to road traffic injuries

		FACTORS		
PHASE		HUMAN	VEHICLES AND EQUIPMENT	ENVIRONMENT
Pre-crash	Crash prevention	Information Attitudes Impairment Police enforcement	Roadworthiness Lighting Braking Handling Speed management	Road design and road layout Speed limits Pedestrian facilities
Crash	Injury prevention during the crash	Use of restraints Impairment	Occupant restraints Other safety devices Crash protective design	Crash-protective roadside objects
Post-crash	Life sustaining	First-aid skill Access to medics	Ease of access Fire risk	Rescue facilities Congestion

Figure 1 The Handdon matrix [11].

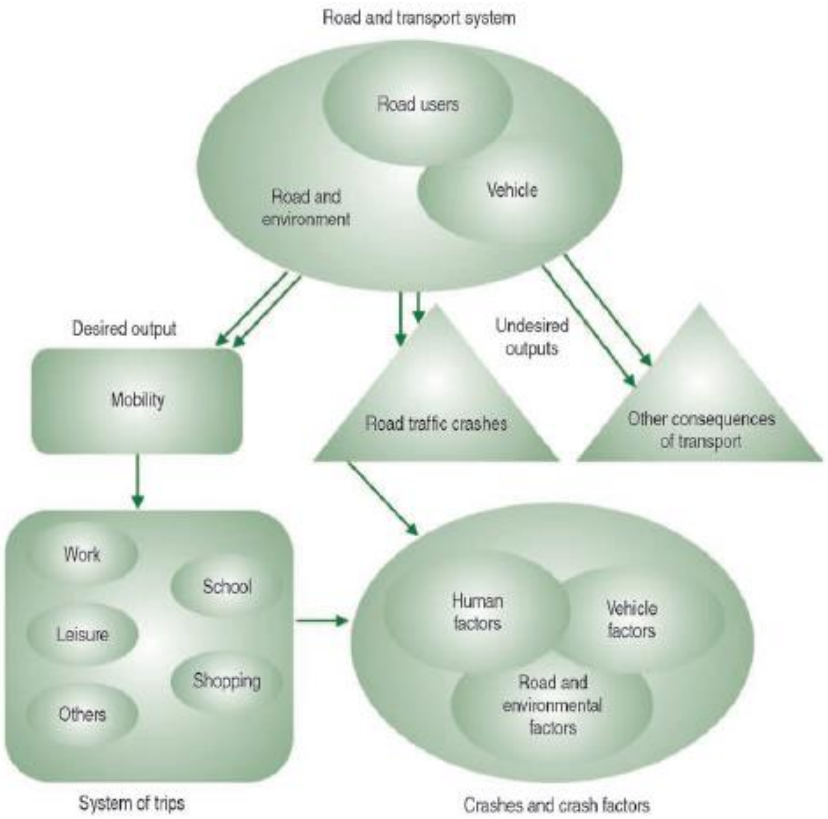


Figure 2 System approach to analysis the risk factors for road traffic injuries [11, 12]. .

Appendix 2: Advanced safety technologies

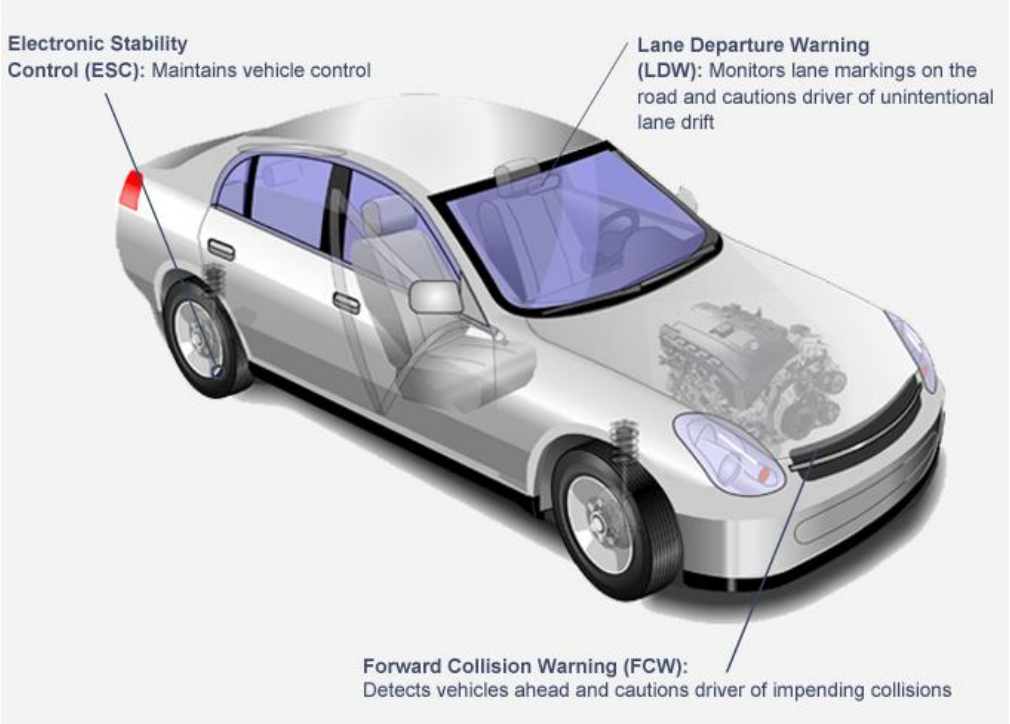


Figure 3 - Vehicle safety technology for crash avoidance, adaptation [151].

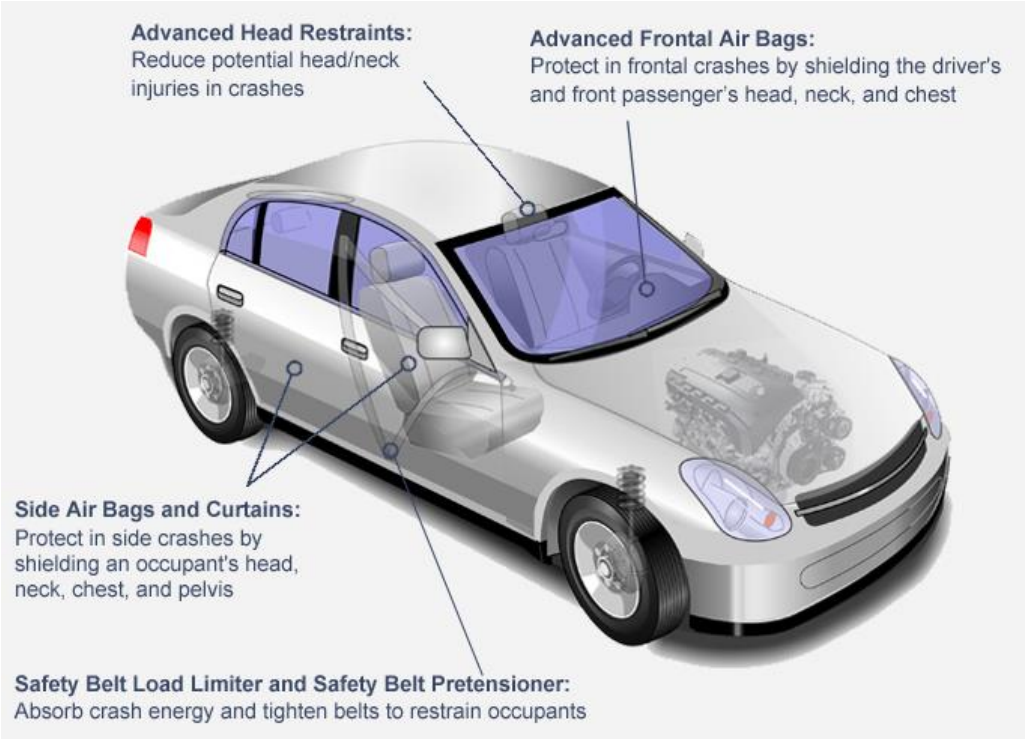
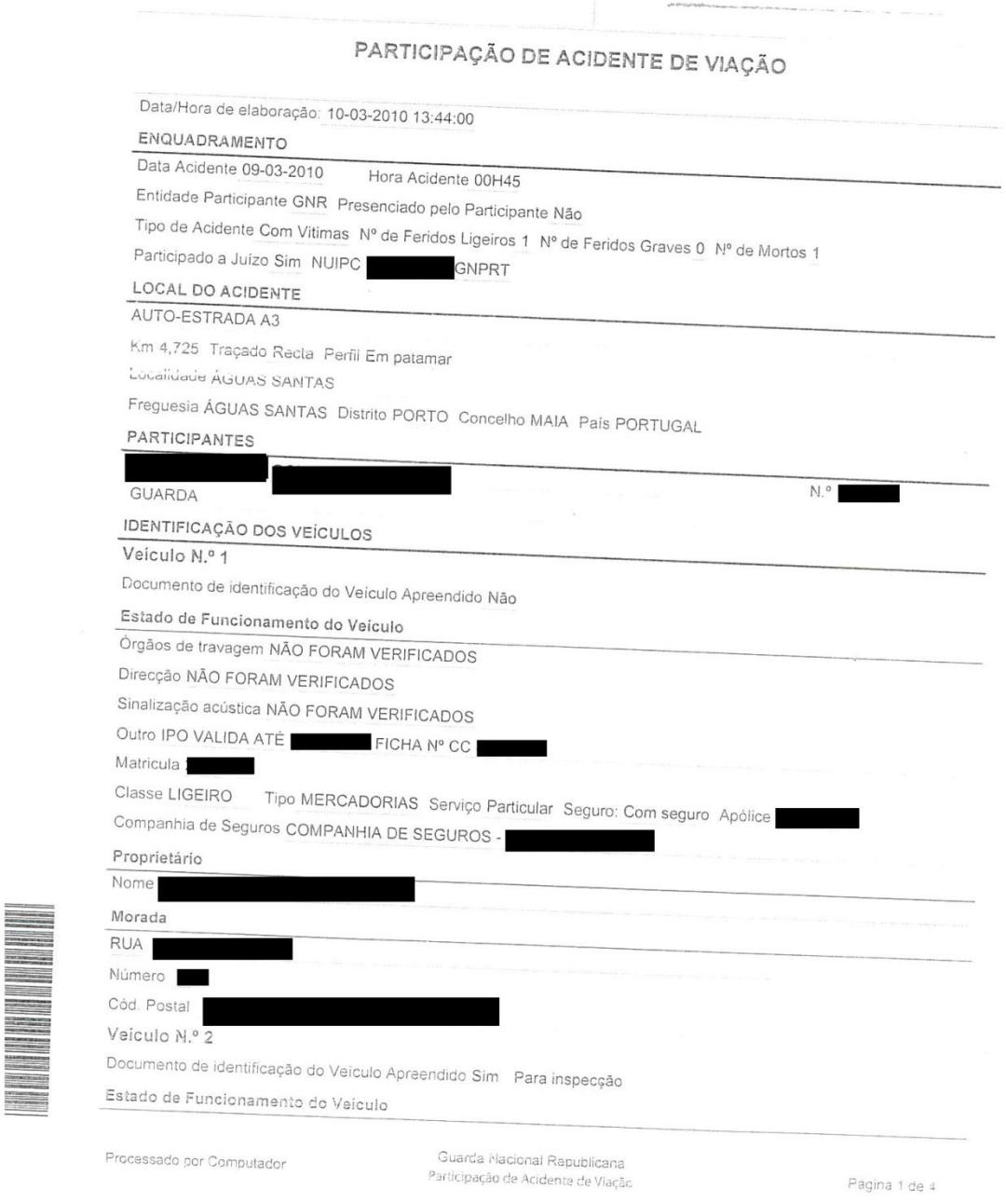


Figure 4- Vehicle safety technology for crash protection, adaptation [151].

Appendix 3: GNR crash report



PARTICIPAÇÃO DE ACIDENTE DE VIAÇÃO

Data/Hora de elaboração: 10-03-2010 13:44:00

ENQUADRAMENTO

Data Acidente 09-03-2010 Hora Acidente 00H45

Entidade Participante GNR Presenciado pelo Participante Não

Tipo de Acidente Com Vitimas Nº de Feridos Ligeiros 1 Nº de Feridos Graves 0 Nº de Mortos 1

Participado a Juízo Sim NUIPC [redacted] GNPRT

LOCAL DO ACIDENTE

AUTO-ESTRADA A3

Km 4,725 Traçado Recta Perfil Em patamar

Localidade ÁGUAS SANTAS

Freguesia ÁGUAS SANTAS Distrito PORTO Concelho MAIA País PORTUGAL

PARTICIPANTES

[redacted]

GUARDA N.º [redacted]

IDENTIFICAÇÃO DOS VEÍCULOS

Veículo N.º 1

Documento de identificação do Veículo Apreendido Não

Estado de Funcionamento do Veículo

Órgãos de travagem NÃO FORAM VERIFICADOS

Direcção NÃO FORAM VERIFICADOS

Sinalização acústica NÃO FORAM VERIFICADOS

Outro IPO VALIDA ATÉ [redacted] FICHA Nº CC [redacted]

Matricula [redacted]

Classe LIGEIRO Tipo MERCADORIAS Serviço Particular Seguro: Com seguro Apólice [redacted]

Companhia de Seguros COMPANHIA DE SEGUROS - [redacted]

Proprietário

Nome [redacted]

Morada

RUA [redacted]

Número [redacted]

Cód. Postal [redacted]

Veículo N.º 2

Documento de identificação do Veículo Apreendido Sim Para inspecção

Estado de Funcionamento do Veículo

Processado por Computador

Guarda Nacional Republicana
Participação de Acidente de Viação

Página 1 de 4

6829330

Órgãos de travagem NÃO FORAM VERIFICADOS

Direcção NÃO FORAM VERIFICADOS

Sinalização acústica NÃO FORAM VERIFICADOS

Outro IPO VALIDA ATÉ [redacted] FICHA Nº CD [redacted]

Matricula [redacted]

Classe LIGEIRO Tipo PASSAGEIROS Serviço Particular Seguro: Com seguro Apólice [redacted]

Companhia de Seguros Seguradora - [redacted]

Proprietário

Nome [redacted]

Morada

RUA [redacted]

Número [redacted]

Cód. Postal [redacted]

IDENTIFICAÇÃO DOS CONDUTORES

Condutor da Viatura [redacted] - CITROEN - XSARA

Lesões Morto Titulo de Condução Apreendido Não

Teste de Álcool Não submetido por lesão ou morte decorrente do acidente

Levantado Auto C.O. Não

Tem Habilitação de Condução Sim

Nome [redacted]

Filiação: Pai [redacted]

Mãe [redacted]

Nascido(a) [redacted]

Naturalidade

Freguesia [redacted] Distrito LISBOA Concelho LISBOA

País PORTUGAL

Nacionalidade PORTUGAL Estado Civil [redacted]

Documento de Identificação

Bilhete de Identidade Nº [redacted] Data de Emissão [redacted] Emitido por Arquivo Identificação

Local Emissão LISBOA

Carta de Condução Nº L-[redacted] Data de Emissão [redacted] Emitido por IMTT Local Emissão LISBOA

B [redacted]

Morada

RUA [redacted]

Número [redacted]

Cód. Postal [redacted]

Condutor da Viatura [redacted] - SUBARU - LEGACY

Lesões Ferido Leve Titulo de Condução Apreendido Sim Passada Guia Não

Teste de Álcool Outra

Teste de Substâncias Psicotrópicas Submetido

Levantado Auto C.O. Não

Tem Habilitação de Condução Sim

Nome [redacted]

Filiação: Pai [redacted]

Mãe [redacted]

Nascido(a) [redacted]

6829330

Naturalidade

Freguesia [redacted] Distrito PORTO Concelho PORTO País PORTUGAL

Nacionalidade PORTUGAL Estado Civil [redacted]

Documento de Identificação

Bilhete de Identidade Nº [redacted] Data de Emissão [redacted] Emitido por Arquivo Identificação

Local Emissão PORTO

Carta de Condução Nº P-[redacted] Data de Emissão [redacted] Emitido por IMTT Local Emissão PORTO

B [redacted]

Morada

RUA [redacted]

Número [redacted]

Cód. Postal [redacted]

AMBIENTE

Cond. Amb. Meteorológicas Bom Tempo

Cond. Amb. Luminosidade Noite - iluminação insuficiente

Cond. Amb. Visibilidade Boa

CARACTERISTICAS DO ACIDENTE

Tipo Acidente Colisão Tipo Colisão Lateral_Positiva Circulação Reduzido Situação do Acidente Em plena Via

DESCRIÇÃO DO ACIDENTE

Pelas declarações verbais e por escrito do condutor do veículo nº2, posicionamento de ambos os veículos e vestígios no local o acidente ter-se-á dado com passo a descrever:
O condutor do veículo nº2, declarou por escrito que: "Circulava na A3, no sentido Porto - Braga, na via central esquerda sensivelmente ao quilómetro 4,700, quando deparei com um veículo acidentado atravessado na faixa de rodagem onde eu circulava, não tendo sido possível evitar o choque, não vi qualquer pessoa na via".
Do acidente resultaram a morte do condutor do veículo nº1, e ferimentos leves no condutor do veículo nº2.

VESTÍGIOS NO LOCAL

DESTRÇOS DOS VEÍCULOS.

CAUSAS PROVAVEIS

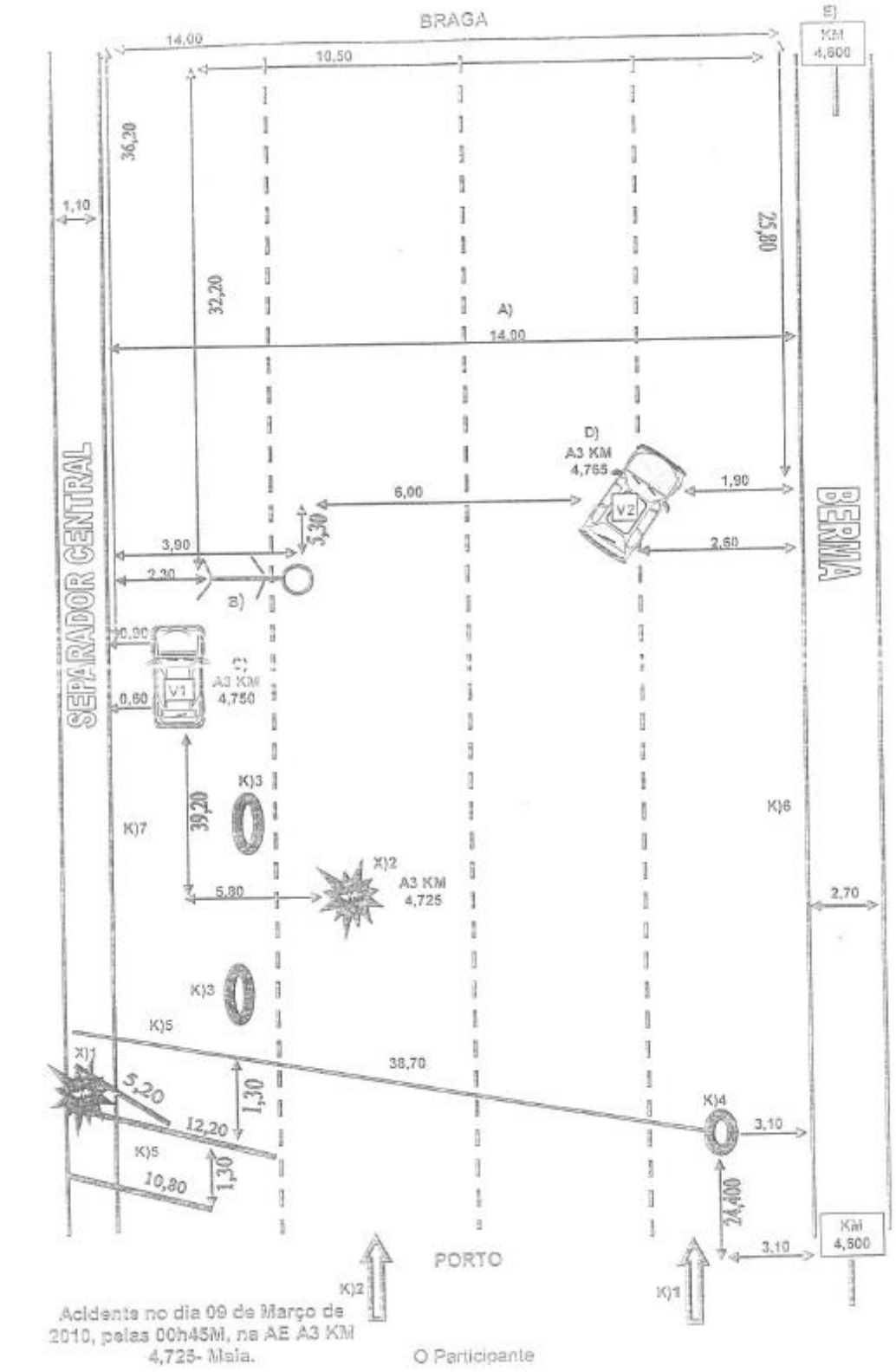
Desconhecidas pelo participante.

OUTROS DANOS

Danos materiais em ambos os veículos, para a Brisa 1 deliniadores, 2 extintores e rasgos no pavimento.

OUTRAS INFORMAÇÕES

Foi feita colheita sanguinia ao condutor do veículo nº2, para exame de confirmação de substâncias psicotrópicas e análise para quantificação da taxa de álcool no sangue, com o selo nº 006496.
No local estiveram o INEM, os Bombeiros Voluntários de Fafe e Bombeiros Voluntários de Ermesinde.
O cadaver foi transportado para o Instituto de Medicina Legal do Porto, pelos Bombeiros Voluntários de Ermesinde.
O veículo nº1 apresentava no capôt vestigios de ter capotado antes de ter acontecido o embate entre o veículo nº1 e veículo nº2.


BRAGA
14,00
10,50
36,20
1,10
32,20
25,80
A)
14,00
E)
KM
4,600
SEPARADOR CENTRAL
BERMA
D)
A3 KM
4,765
V2
6,00
3,30
1,90
2,60
3,90
2,30
B)
0,90
0,60
V1
C)
A3 KM
4,750
K)3
39,20
K)7
K)6
X)2
A3 KM
4,725
5,80
2,70
K)3
K)5
X)1
38,70
5,20
1,30
12,20
K)5
10,30
1,30
K)4
3,10
24,400
3,10
KM
4,500
PORTO
K)2
K)1
Acidente no dia 09 de Março de 2010, pelas 00h45M, na AE A3 KM 4,725- Maia.
O Participante

Appendix 4: Vehicle specific technical information

Vehicle Specific Technical Information extracted from “Ficha Homologação” which was provided by IMTT Porto.

Instituto da Mobilidade e dos Transportes Terrestres, I.P.

FOLHA DE APROVAÇÃO DE MODELO

CONSULTA

Nº DE HOMOLOGAÇÃO NACIONAL	200210000117	extensão 0064	SITUAÇÃO: REGULAR
Nº DE HOMOLOGAÇÃO CE	e11*2001/116*0181*06		DESPACHO EM: 2004-09-01

CARACTERÍSTICAS GERAIS			
0.1	MARCA	TOYOTA	780
0.2	MODELO	ZZ2T (COROLLA)	4084
	VARIANTE	NDE120 (N)	
	VERSÃO	NDE120L-DMNNTW (1G)	
0.2.1	DESIGN. COMERCIAL		
0.4	CATEGORIA	LIGEIRO	1
	TIPO	PASSAGEIROS	1
	CÓDIGO CE		
0.5	FABRICANTE	TOYOTA MOTOR EUROPE MARKETING & ENGINEERIN	10
0.6	CLASSE		
0.7	TIPO DE MÁQUINA		

CONSTITUIÇÃO GERAL DO VEÍCULO			
1	N.º de Eixos	F 1 R 1 T	
	N.º de Rodas	F 2 R 2 T	
2	N.º de Eixos Motores	F 1 R	
3	Distância entre Eixos	2600	
5	Largura de Via	1 2 3 4	
	Mínima		
	Máxima		
6.1	Comprimento		
7.1	Largura	1710	
8	Altura		
9.1	Dist. Eixo-Apoio		
9.2	Dist. Eixo-Frente/Frente		
11	Eixo Ret. à Retaguarda		
11.1	Dist. Eixos Consecutivos	1-2 2-3 3-4	
11.2	Avanço do Prato	Máx. Mín.	

PESOS			
12.1	Tara	F R T 1360	
14.1	Peso Bruto (TOTAL)	1695	
14.2	Distribuição do PB	Frente Retaguarda	
14.3	Máximo Admissível	1 #00930 2 #00905 3 4	
14.4	Peso Máximo no Engate		
17	Peso Bruto Rebocável	com Travão 1000 sem Travão	
18	Peso Bruto Conjunto		

MOTOR			
	Homologação	200110003830 Extensão 0090	
20	Marca	TOYOTA	780
21	Modelo	1ND	
23	N.º de Cilindros	4	
24	Cilindrada	1364	
25	Combustível / Energia	GASOLEO	2
26	Potência Máxima	66 Kw a 3800 rpm	
27	%Máx. Biocombustível		

TRANSMISSÃO - SUSPENSÃO - TRAVÕES			
28	Caixa de Velocidades	MANUAL	1
32	Pneumáticos Frente	195/60 R 15	
32.1	Retaguarda	195/60 R 15	
33	Suspensão Frente	MECANICA	3
33.1	Retaguarda	MECANICA	3
35	Travões Serviço	HIDRAULICO	2
36	Estacionamento	MECÂNICO	3

CAIXA			
37	Tipo	FECH.C/S TECTO ABRIR	47
37.1	Categoria Europeia		
37.2	Comprimento Exterior	Máx. Mín.	
37.3	Comprimento Interior	Máx. Mín.	
37.4		Total	
37.5	Altura Interior	Máx. Mín.	
37.6	Largura Exterior	Máx.	
37.7	Avanço Centro Gravidade	Máx. Mín.	
37.8	Avanço cg/calculo	Máx. Mín.	
41	N.º de Portas	4	
42.1	Lotação Total	5	
42.2	Sentada	F 2 M R 3	
42.3	Em Pé		

EMISSÕES POLUENTES E CONSUMO			
45	Nível Sonoro Estacionário	75 dB (A) a 2850 rpm	
46.1	Nível Sonoro em Movim.	72 dB (A)	
46.1	Emissões CO (Tipo I)	.17 g / Km	
	CO (Tipo II)	%	
	HC	g / Km	
	NOx	.17 g / Km	
	HC+NOx	.19 g / Km	
	Particulas	.021 g / Km	
46.2	CO_2	Combustível I Combustível II	
		GASOLEO	
	Urbano	153 g / Km g / Km	
	Extra Urbano	113 g / Km g / Km	
	Combinado	128 g / Km g / Km	
46.3	Cons. Combustível Urb.	5.8 l / 100 Km	
	Extra Urbano	4.3 l / 100 Km	
	Combinado	4.8 l / 100 Km	
46.4	Redução Emissão CO_2	g / Km	
46.5	Tecnologia Inovativa		
46.6	Autonomia baterias Valor	Km	
46.7	Fonte		

50 - Anotações
PNEUS 195/55R16

Appendix 5: Logit models development using Enterprise Miner

SAS Institute defines data mining as the process of Sampling, Exploring, Modifying, Modeling, and Assessing (SEMMA). At the Enterprise Miner software, a graphical user interface (GUI) provides an advance use for the SEMMA data mining process:

a) **S**ample the data by creating one or more data tables.
b) **E**xplore the data by searching for anticipated relationships, unanticipated trends, and anomalies in order to gain understanding and ideas.
c) **M**odify the data by creating new variables, selecting, and transforming the existent variables to be included in the model.
d) **M**odel search for a combination of the data that reliably predicts a desired outcome.
e) **A**ssess the data by evaluating the reliability of the findings from the modeling process.

The crash data mining analysis at the Enterprise Miner interface, started by importing each data sets, Two and Single, into the process flow workplace. Then, the data mining process was developed, including all the above SEMMA steps, and some were necessary to repeat one or more of the steps several times before a satisfied result were obtained. At the end of the assess phase of the SEMMA process, the best models were scored to new data.

The diagram process flow was developed by applying the following tools for Sample, Modify, Model and Assess phases of the SEMMA process. Since a previous correlation analysis amongst the response variable (FatalSIK) and the independent variable was performed withSAS9.2 PROC CORR procedure, the Explore phase of SEMMA process was applied previously to the models process flow diagrams to generate graph reports and summary association statistics for the training subsets. The logistic regression models presented in this dissertation were developed by the application of specific features of the EM software, as explained next though step 1 to step 10.

Step1: Input data source node

The data table generate at SAS 9.2 was launched to Enterprise MinerTM 6.2. Metadata was specified for the data set. For each variable used in the modeling process the role was set as input or target, and the measurement level was selected as: Interval for continuous variables, Nominal for category variables and binary for the target variable. Then the input data node was used as training data to estimate the parameters of the model. Two inputs data source were imported into each EM diagram process flow: Two input data source (containing the data set for two-vehicles collisions) and Single input data source (containing the data set for single vehicles crash).

Step 2: Sample node

A sample node, which is part of the Sample from the SEMMA data mining process, it was connected to the input data to create a stratified random training sample. The selected stratified criterion was “Level Base” and the sample proportion 50.0. As a result, the new subset used during the training included all the observations of minority class being predicted, FatalSIK”1”, and an equal proportion of the majority class, FatalSIK”0”, which was randomly selected. The stratified random sample for each data set was described Chapter 3 of the Thesis. The subsets samples had the following proportion. The training sample for Single included 38 observations of FatalSIK”1” and 38 observations of FatalSIK”0”. The training sample for Two included 32 observations of FatalSIK”1” and 32 observations of FatalSIK”0”.

Step 3: Drop node

Drop node, which is part of the Modify phase of SEMMA was used as an optional path for some models candidates and it was connected to the Two training sample to hide variables from the metadata. The effect of the differential characteristics

of the vehicles involved in the collision was explored by using the following inputs: AgeV2V1, ccV2V1, WTV2V1 and WBV2V1. On the

other hand, using the drop node allowed hiding the individual vehicle characteristics as follows: AgeV1, AgeV2, ccV1, ccV2, WTV1, WTV2, and WBV1 and WBV2.

Step 4: Transformation node

The transform variables node, which is also part of the Modify phase of SEMMA, it enables to create new variables and also enables to transform class variables and to create interaction variables. Transformations are useful when the researcher want to improve the fit of a model to the data, (SAS EM7.1 Reference Help, 2010). For example, transformations can be used to stabilize variances, remove nonlinearity, improve additivity, and correct non-normality in variables (SAS EM7.1 Reference Help, 2010). The transformation node was set to bucket. This option allows creating by dividing the data into evenly spaced intervals based on the difference between the maximum and minimum values. For the models path that including a transformation node, four bins were create for those variables mentioned above: AgeV1, AgeV2, ccV1, ccV2, WTV1, WTV2, WBV1, WBV2, AgeV2V1, ccV2V1, WTV2V1 and WBV2V1. If the path included a drop node, the bins were only created for the variables: AgeV2V1, ccV2V1, WTV2V1 and WBV2V1. For those interval inputs, the default transformation method, bucket, was applied.

Step 5: Regression node

The modeling phase of SEMMA was performed with the incorporation of regression nodes into the workspace. At each regression node properties the logistic regression type and logit link function were selected. The logit option specifies the inverse of the cumulative logistic distribution function. During the training, four selection methods were chosen, as follows:

- Backward- begins with all candidate effects (inputs) in the model and removes effects until the Stay Significance Level or the Stop Criterion is met. Inputs are sequentially removed from the model with the highest p-value. The sequence terminates when all the remaining inputs have a p-value in excess of the predetermined stay cutoff. It creates a sequence of models decreasing complexity, SAS Institute Inc., 2009.
- Forward- begins with no candidate inputs in the model and adds inputs until the Entry Significance Level or the Stop Criterion is met. In contrast with backward selection creates a sequence of models of increasing complexity, SAS Institute Inc., 2009. Improvement is quantified by the measurement of significance, p-value. A small p-value indicates a significant improvement. The forward selection procedure terminates when no p-value is below a predetermined entre cutoff, SAS Institute Inc., 2009.
- Stepwise- begins as in the forward model but may remove inputs already in the model. This procedure sequentially adds inputs with the smallest p-value below the entry cutoff. As each input is added, the algorithm re-evaluates the statistical significance of all included inputs in the model. If p-value of the selected inputs exceeds a stay cutoff, the input is removed from the model, SAS Institute Inc., 2009. This procedure terminates when all the inputs available for addition in the model have a p-value in excess of the entry cutoff, SAS Institute Inc., 2009.
- None- all inputs are used to fit the model.

During the models training, at the EM process flow diagrams, several regression nodes were used in the training and all the above four input selection criteria were explored. If one of these methods were chosen: forward, backward or stepwise, the selection criteria for the model comparison can be specified. Misclassification rate was used to select the model from the several candidate models being developed at the EM process flow. Hence the model comparison node selected the model with the smallest misclassification rate. Some regressions nodes for the selections methods described above were run with the default setting entry significance level, which is 0.05. Others regression nodes were training with the entry significance level of the regression node was specified for 0.1 to add variables in forward and stepwise regressions.

Step 6: Cutoff node

Cutoff node belongs to the Assess category in the SAS data mining process SEMMA. The node provides graphical information to determine the appropriate probability cutoff point for decision making with binary target models, (SAS EM7.1 Reference Help, 2010). The establishment of a cutoff decision point entails the risk of generating false positives and false negatives, but an appropriate use of the Cutoff node can help minimize those risks, (SAS EM7.1 Reference Help, 2010). During the models training, the optimal cutoff value was obtained for 0.69. This optimal cutoff value selected by taking into account which cutoff value would result in a higher overall classification rate and the prior probabilities for the severe crashes in the data set.

Step 7: Control point

Control point node was used to simplify the distribution of connections between process flow steps that have multiple interconnected nodes. The control running a process flow diagram from the Control Point node will run or update all preceding paths, and this tool was very helpful during the diagrams development.

Step 8: Model comparison node

The Model Comparison node belongs to the Assess category in the SAS SEMMA and enables to compare the performance of competing models using various criteria. For binary targets the Model Comparison node provides information about:

-Classification Measures, which include the Receiver Operating Characteristic (ROC) charts and corresponding area under the curve, and classification rates.

-Data Mining Measures, which include lift and gain measures and profit and loss measures.

-Statistical Measures, which include Bayesian Information Criterion (BIC), Akaike's Information Criterion (AIC), Gini statistics, and Kolmogorov-Smirnov statistics, among others.

Several measures can be used to choose the best model out of a group of several candidate models. The comparative measures types of analysis are: statistical, classification, and data mining. The selection of those three measures types depends on the preference of who evaluates the training modes. An illustration example is extracted from the SAS EM7.1 Reference Help, 2010: “while statisticians might be more familiar with stopping measures such as Mallows' Cq, analysts might be more comfortable using ROC chart analysis to choose the best model, and direct marketers might prefer using lift and gains tables to benchmark model performance”.

Step 9: Score Node

The Score node is part the end process of the Assess phase of SEMMA data mining process. This node creates predictions using the model deemed best by the Model Comparison node, described above. Alternatively, the score node into the diagram workspace at EM can be directly link to any desired model. To evaluate the performance of the selected model from the training procedure, new a data source must be dragged into to diagram workspace. Hence the the original data set, containing the original crash population, was dragged again into the diagram and connected to the score node was well. While for the training models development the data set’s role was set to “raw”, for the score stage, the data set was set to score role. This attribute allows the score node to use the data set to generate predicted values for a data set that might not contain a target.

Step 10: SAS Score Code Node

Finally, at the end of the models development path, a sas score code node was linked to the score node, (as explained in step 9). This tool was used to generate a new sas code into the process flow diagram to create a customized scoring data output. At the SAS code node's properties panel from the code editor, a sas code was written to generate report output for

the score node predicted results. The generated report output creates the scores results for the classification assessment as follows: True Positives (TPs), False Positives (FPs), True Negatives (TNs), and False Negatives (FNs). The specific sas code was written for this specific crash analysis in order to assist with the models evaluation. This sas code enable an easier comparison between the selected model classification measures, expressed by TN, FN, FP and TP, as explained previously. and the assessment of the model performance score results, expressed by TNs, FNs, FPs and TPs (as explained previously.

Selection of Best Models for Injury Severity Prediction

Following the development of several models alternatives, the best models to predict the target FatalSIK were selected amongst the candidate models based on the goodness of fit of the model to the crash data. For the models selection, the next analysis parameters were evaluated: model fit statists, test for the null hypothesis, type 3 analysis of effects and event classification output.

The **Model Fit Statistics** provides the following information:

a) Akaike Information Criteria (AIC), which can be used for the comparison of nonnested models on the sample.

b) Schwarz Criterion (SC), which penalizes for the number of predictors in the model, (UCLA; 2012).

c) -2 Log L is the negative two times the log-likelihood, which is used in the hypothesis tests for nested models, however its value is.

The **Test of the null hypothesis** (β=0) relies on three equivalent Chi-Square tests, and all them test against the null hypothesis that at least one of the predictors' coefficients is not equal to zero in the model. These three tests are presented next.

a) The Likelihood Ratio Chi-Square test that at least one of the predictor´s coefficients is not equal to zero in the model. The Likelihood Ratio Chi-Square statistic can be calculated by -2 Log L(model with intercept only) - 2 Log L(model with Intercept and Covariates).

b) The Score Chi-Square Test that at least one of the predictors' regression coefficient is not equal to zero in the model.

c) Wald Chi-Square Test tests that at least one of the predictors' regression coefficient is not equal to zero in the model.

The Chi-Square test statistics for those tests provides the degrees of freedom (DF) and associated p-value (Pr>ChiSq) corresponding to the specific test that all of the predictors are simultaneously equal to zero. The DF defines the distribution of the Chi-Square test statistics and is defined by the number of predictors in the model. The Pr>ChiSq can be understand as a specified alpha level, related to the acceptance of type I error, (usually 0.05 or 0.01). The small p-value from the all three tests would lead to conclude that at least one of the regression coefficients in the model is not equal to zero.

The **Type 3 Analysis of Effects** tests the statistical significance of adding a new input to the model that is being developed. The statistical significance measures range from <0.0001, which is associated to highly significant inputs, to 0,9997, which means that the input is dubious, (SAS Institute Inc., 2007). This analysis output provides information for each effect (input variable) in the model, its DF and the respective Pr>Chi-Square for the selected effect.

If decisions predictions are of interest, model fit can be evaluated by the misclassification. If estimates are of interest, model fit can be assessed by the average square error. A small Average Square error shows a better model.

The **Analysis of the Maximum Likelihood Estimates** (AMLE) output provides information for each parameter in the model, intercept and input variables, (including BIN groups for those variables, if there were Bin transformations performed). The AMLE also presents for each parameter its: estimates, DF, Standard error and Pr>Chi-Square.

a) The DF in this analysis define the Chi-Square distribution to test whether the individual regression coefficient is zero, given the others predictors in the model.

b) Estimates are the binary logit regression estimates for the Parameters in the model. The logistic regression model models the log odds of a positive response (for the target FatalSIK=1 in this research) as a linear combination of the predictor variables. This is written as

where p is the probability that FatalSIK is 1, thus the crash would be severe.

The parameter estimates can be understood as follows: for a one unit change in the predictor variable, the difference in log-odds for a positive outcome is expected to change by the respective coefficient, given the other variables in the model are held constant.

c) Standard Errors are related to the individual regression coefficients. They are used in both the 95% Wald Confidence Limits, and the Chi-Square test statistic.

d) The Chi-Square and Pr > ChiSq are the test statistics and p-values, respectively, testing the null hypothesis that an individual predictor's regression coefficient is zero, given the other predictor variables are in the model. The Chi-Square test statistic is the squared ratio of the Estimate to the Standard Error of the respective predictor, (UCLA, 2012). The Chi-Square value follows a central Chi-Square distribution with degrees of freedom given by DF, which is used to test against the alternative hypothesis that the Estimate is not equal to zero, (UCLA, 2012). The probability that a particular Chi-Square test statistic is as extreme as, or more so, than what has been observed under the null hypothesis is defined by Pr>ChiSq.

e) The Effect refers to the predictor variables that are interpreted in terms of odds ratios.

f) The Point Estimate underneath are the odds ratio corresponding to selected Effects in the model. The odds ratio is obtained by the estimate. The difference in the log of two odds is equal to the log of the ratio of these two odds. The log of the ratio of two odds is the log odds ratio. Hence, the interpretation of Estimate-the coefficient was interpreted as the difference in log-odds-could also be done in terms of log-odds ratio. When the Estimate, the log-odds ratio becomes the odds ratio. We can interpret the odds ratio as follows: for a one unit change in the predictor variable, the odds ratio for a positive outcome is expected to change by the respective coefficient, given the other variables in the model are held constant.

g) The 95% Wald Confidence Limits is the Wald Confidence Interval (CI) of an individual odds ratio, given the other predictors are in the model. For a given predictor variable with a level of 95% confidence, the interpretation is as follows: there is 95% confident that upon repeated trials, 95% of the CI's would include the "true" population odds ratio. The CI is equivalent to the Chi-Square test statistic: if the CI includes one, it would fail to reject the null hypothesis that a particular regression coefficient equals zero and the odds ratio equals one, given the other predictors are in the model. An advantage of a CI is that it is illustrative; it provides information on where the "true" parameter may lie and the precision of the point estimate for the odds ratio.Additionally, the Enterprise Miner logistic output provides a list with all the fit statistics labels used statistical analysis, such as the following examples:

- -AIC (explained previously)
- -ASE (Average Squared Error)
- -MSE (Mean Squared Error)
- -RMSE (Root Mean Squared Error)
- -SBC (explained previously)
- -SSE (Sum of Squared Error)
- -MISC (Misclassification Rate).

Appendix 6: SAS Code

SAS Code created for Crash Data Analysis

```
**ReadAllCrashes.sas CRASH DATASET;

libname  crash '.';

data crash.All;
   infile 'All.csv' dsd firstobs=2 missover lrecl=500;


     Input Record $   RoadName $ Location $ LocalSpeedLimit  RoadClass $ LanesSD LanesOD
     DivisionCode $   SpeedLimit  Speed50kmHr $  Speed90kmHr $   Speed100KmHr $  Speed120kmHr $
     VehiclesInvolved    WeatherCode $   AlcoholDrugs $
     RanOff $ Rollover $ RearEnd $ HeadOn $  Sideswipe $ Other  $
     (PlateV1 BrandV1 ModelV1)(:$16.) CategoryV1 :$1. WTV1 ccV1 WBV1 LengthV1  FuelV1 :$1. MileageV1 YrV1 AgeV1
     (PlateV2 BrandV2 ModelV2)(:$16.) CategoryV2 :$1. WTV2 ccV2 WBV2 LengthV2  FuelV2 :$1. MileageV2 YrV2 AgeV2
     (PlateV3 BrandV3 ModelV3)(:$16.) CategoryV3 :$1. WTV3 ccV3 WBV3 LengthV3  FuelV3 :$1. MileageV3 YrV3 AgeV3
     LightInjuryV1   SeriousInjuryV1 KilledV1
     LightInjuryV2   SeriousInjuryV2 KilledV2
     LightInjuryV3   SeriousInjuryV3 KilledV3
     SUMLI   SUMSI   SUMK  SEVERITY :$2. CostCrashSeverity CostInjuriesSeverity
     WTDiff  ccDiff WBDiff LengthDiff AgeDiff ;

run;
proc print data=crash.All ;

run;
*************************
***DEPENDENT VARIABLES***
*************************;
data crash.All; set crash.All;

if (SeriousInjuryV1 >0 or KilledV1 >0)  then FatalSikV1 = 1;
else FatalSikV1 = 0;

if (SeriousInjuryV2 >0  or KilledV2>0) then FatalSikV2 = 1;
else FatalSikV2 = 0;

if (SeriousInjuryV3 >0  or KilledV3 >0)  then FatalSikV3 = 1;


SIK = SUMSI + SUMK;
if (SUMSI > 0 or SUMK > 0) then FatalSik = 1;
else FatalSik = 0;
if SUMK > 0 then FatalK = 1;
else FatalK = 0;
SIKRatio = SIK/(SIK + SUMLI);
SocietyCost = CostInjuriesSeverity/3366388;
If SIKRatio = 1 then FatalSIKRatio = 1;
Else FatalSIKRatio = 0;

if SpeedLimit > 90 then SpeedLevel = 1;
else SpeedLevel = 0;

run;
```

Appendix 7: Variables Correlation

For single-vehicles crashes dataset

Pearson Correlation Coefficients							
Prob > |r| under H0: Rho=0							
Number of Observations							
	WTV1	**ccV1**	**WBV1**	**YrV1**	**AgeV1**	**SIK**	**FatalSIK**
WTV1	100.000	0.78814	0.74717	0.31989	-0.30815	0.02101	0.01743
		<.0001	<.0001	<.0001	<.0001	0.6393	0.6973
	500	500	500	500	500	500	500
ccV1	0.78814	100.000	0.64347	0.01273	-0.00418	0.08133	0.08153
	<.0001		<.0001	0.7765	0.9257	0.0692	0.0685
	500	500	500	500	500	500	500
WBV1	0.74717	0.64347	100.000	0.10844	-0.09973	-0.02459	-0.03141
	<.0001	<.0001		0.0153	0.0257	0.5833	0.4834
	500	500	500	500	500	500	500
YrV1	0.31989	0.01273	0.10844	100.000	-0.98898	-0.10393	-0.09870
	<.0001	0.7765	0.0153		<.0001	0.0201	0.0273
	500	500	500	500	500	500	500
AgeV1	-0.30815	-0.00418	-0.09973	-0.98898	100.000	0.09057	0.08044
	<.0001	0.9257	0.0257	<.0001		0.0429	0.0723
	500	500	500	500	500	500	500
SIK	0.02101	0.08133	-0.02459	-0.10393	0.09057	100.000	0.91198
	0.6393	0.0692	0.5833	0.0201	0.0429		<.0001
	500	500	500	500	500	500	500
FatalSIK	0.01743	0.08153	-0.03141	-0.09870	0.08044	0.91198	100.000
	0.6973	0.0685	0.4834	0.0273	0.0723	<.0001	
	500	500	500	500	500	500	500

Variables Correlation for two-vehicles collisions dataset

Variables	WTV1	ccV1	WBV1	AgeV1	WTV2	ccV2	WBV2	AgeV2	FatalSik	FatalSikV1	FatalSikV2	WTV2V1	ccV2V1	WBV2V1	AgeV2V1
WTV1	100.000	0.77864	0.79309	-0.38996	-0.03888	0.02414	-0.00736	0.04873	0.04169	-0.00156	0.05564	0.37721	0.23207	0.35636	-0.06405
		<.0001	<.0001	<.0001	0.2508	0.4760	0.8281	0.1500	0.2182	0.9634	0.1002	<.0001	<.0001	<.0001	0.0584
	874	874	874	874	874	874	874	874	874	874	874	874	874	874	874
ccV1	0.77864	100.000	0.68193	-0.08996	-0.02998	0.02770	-0.01008	0.05194	0.04502	0.00120	0.07448	0.26921	0.29787	0.28430	-0.03470
	<.0001		<.0001	0.0078	0.3760	0.4135	0.7660	0.1250	0.1836	0.9717	0.0277	<.0001	<.0001	<.0001	0.3055
	874	874	874	874	874	874	874	874	874	874	874	874	874	874	874
WBV1	0.79309	0.68193	100.000	-0.22840	-0.00619	0.03709	0.03193	0.01286	0.07679	0.01498	0.09135	0.31161	0.21265	0.47075	-0.03950
	<.0001	<.0001		<.0001	0.8550	0.2734	0.3458	0.7042	0.0232	0.6583	0.0069	<.0001	<.0001	<.0001	0.2434
	874	874	874	874	874	874	874	874	874	874	874	874	874	874	874
AgeV1	-0.38996	-0.08996	-0.22840	100.000	0.05290	0.02279	0.04196	-0.02169	-0.07285	-0.04212	-0.05146	-0.00007	0.00985	-0.05374	0.15078
	<.0001	0.0078	<.0001		0.1181	0.5011	0.2153	0.5219	0.0313	0.2135	0.1285	0.9983	0.7711	0.1124	<.0001
	874	874	874	874	874	874	874	874	874	874	874	874	874	874	874
WTV2	-0.03888	-0.02998	-0.00619	0.05290	100.000	0.79150	0.80250	-0.36028	0.05388	0.08872	-0.02544	0.39900	0.27249	0.34747	0.02160
	0.2508	0.3760	0.8550	0.1181		<.0001	<.0001	<.0001	0.1114	0.0087	0.4525	<.0001	<.0001	<.0001	0.5236
	874	874	874	874	874	874	874	874	874	874	874	874	874	874	874
ccV2	0.02414	0.02770	0.03709	0.02279	0.79150	100.000	0.65409	-0.07678	0.06629	0.11583	-0.04438	0.30482	0.35106	0.25770	0.05206
	0.4760	0.4135	0.2734	0.5011	<.0001		<.0001	0.0232	0.0501	0.0006	0.1900	<.0001	<.0001	<.0001	0.1241
	874	874	874	874	874	874	874	874	874	874	874	874	874	874	874
WBV2	-0.00736	-0.01008	0.03193	0.04196	0.80250	0.65409	100.000	-0.22826	0.05016	0.09713	-0.04350	0.32149	0.19470	0.43747	0.02081
	0.8281	0.7660	0.3458	0.2153	<.0001	<.0001		<.0001	0.1385	0.0041	0.1988	<.0001	<.0001	<.0001	0.5389
	874	874	874	874	874	874	874	874	874	874	874	874	874	874	874
AgeV2	0.04873	0.05194	0.01286	-0.02169	-0.36028	-0.07678	-0.22826	100.000	-0.01771	-0.03175	-0.02697	-0.06525	0.02591	-0.08716	0.17126
	0.1500	0.1250	0.7042	0.5219	<.0001	0.0232	<.0001		0.6012	0.3485	0.4258	0.0538	0.4442	0.0099	<.0001
	874	874	874	874	874	874	874	874	874	874	874	874	874	874	874
FatalSik	0.04169	0.04502	0.07679	-0.07285	0.05388	0.06629	0.05016	-0.01771	100.000	0.80485	0.65448	0.01746	0.04173	0.04089	0.01938
	0.2182	0.1836	0.0232	0.0313	0.1114	0.0501	0.1385	0.6012		<.0001	<.0001	0.6062	0.2178	0.2272	0.5671
	874	874	874	874	874	874	874	874	874	874	874	874	874	874	874
FatalSikV1	-0.00156	0.00120	0.01498	-0.04212	0.08872	0.11583	0.09713	-0.03175	0.80485	100.000	0.15852	0.02908	0.04665	0.00359	0.01796
	0.9634	0.9717	0.6583	0.2135	0.0087	0.0006	0.0041	0.3485	<.0001		<.0001	0.3905	0.1682	0.9156	0.5960
	874	874	874	874	874	874	874	874	874	874	874	874	874	874	874
FatalSikV2	0.05564	0.07448	0.09135	-0.05146	-0.02544	-0.04438	-0.04350	-0.02697	0.65448	0.15852	100.000	-0.02661	-0.01722	0.03951	0.02593
	0.1002	0.0277	0.0069	0.1285	0.4525	0.1900	0.1988	0.4258	<.0001	<.0001		0.4320	0.6112	0.2433	0.4439
	874	874	874	874	874	874	874	874	874	874	874	874	874	874	874
WTV2V1	0.37721	0.26921	0.31161	-0.00007	0.39900	0.30482	0.32149	-0.06525	0.01746	0.02908	-0.02661	100.000	0.60990	0.67570	0.13338
	<.0001	<.0001	<.0001	0.9983	<.0001	<.0001	<.0001	0.0538	0.6062	0.3905	0.4320		<.0001	<.0001	<.0001
	874	874	874	874	874	874	874	874	874	874	874	874	874	874	874
ccV2V1	0.23207	0.29787	0.21265	0.00985	0.27249	0.35106	0.19470	0.02591	0.04173	0.04665	-0.01722	0.60990	100.000	0.48223	0.01562
	<.0001	<.0001	<.0001	0.7711	<.0001	<.0001	<.0001	0.4442	0.2178	0.1682	0.6112	<.0001		<.0001	0.6447
	874	874	874	874	874	874	874	874	874	874	874	874	874	874	874
WBV2V1	0.35636	0.28430	0.47075	-0.05374	0.34747	0.25770	0.43747	-0.08716	0.04089	0.00359	0.03951	0.67570	0.48223	100.000	0.05789
	<.0001	<.0001	<.0001	0.1124	<.0001	<.0001	<.0001	0.0099	0.2272	0.9156	0.2433	<.0001	<.0001		0.0872
	874	874	874	874	874	874	874	874	874	874	874	874	874	874	874
AgeV2V1	-0.06405	-0.03470	-0.03950	0.15078	0.02160	0.05206	0.02081	0.17126	0.01938	0.01796	0.02593	0.13338	0.01562	0.05789	100.000
	0.0584	0.3055	0.2434	<.0001	0.5236	0.1241	0.5389	<.0001	0.5671	0.5960	0.4439	<.0001	0.6447	0.0872	
	874	874	874	874	874	874	874	874	874	874	874	874	874	874	874

Appendix 8: Models for Crash Severity Prediction - Single

Model-IA-S

```
*------------------------------------------------------------*
User:                  mina
Date:                  May 25, 2012
Time:                  09:35:43
*------------------------------------------------------------*
* Training Output
*------------------------------------------------------------*
```

Variable Summary

Role	Measurement Level	Frequency Count
INPUT	INTERVAL	4
INPUT	NOMINAL	6
REJECTED	INTERVAL	13
REJECTED	NOMINAL	8
TARGET	BINARY	1

Model Events

Target	Event	Measurement Level	Number of Levels	Order	Label
FatalSik	1	BINARY	2	Descending	

Predicted and decision variables

Type	Variable	Label
TARGET	FatalSik	
PREDICTED	P_FatalSik1	Predicted: FatalSik=1
RESIDUAL	R_FatalSik1	Residual: FatalSik=1
PREDICTED	P_FatalSik0	Predicted: FatalSik=0
RESIDUAL	R_FatalSik0	Residual: FatalSik=0
FROM	F_FatalSik	From: FatalSik
INTO	I_FatalSik	Into: FatalSik

The DMREG Procedure

Model Information

Training Data Set	EMWS20.SMPL2_DATA.DATA
DMDB Catalog	WORK.REG7_DMDB
Target Variable	FatalSik
Target Measurement Level	Ordinal
Number of Target Categories	2
Error	MBernoulli
Link Function	Logit
Number of Model Parameters	11
Number of Observations	76

Target Profile

Ordered Value	FatalSik	Total Frequency
1	1	38
2	0	38

Forward Selection Procedure

Step 0: Intercept entered.

The DMREG Procedure

Newton-Raphson Ridge Optimization

Without Parameter Scaling

Parameter Estimates 1

Optimization Start

Active Constraints	0	Objective Function	52.679185723
Max Abs Gradient Element	0		

The selected model is the model trained in the last step (Step 4). It con

Intercept AgeV1 WBV1 WeatherCode ccV1

Likelihood Ratio Test for Global Null Hypothesis: BETA=0

-2 Log Likelihood Intercept Only	-2 Log Likelihood Intercept & Covariates	Likelihood Ratio Chi-Square	DF	Pr > ChiSq
105.358	84.650	20.7082	4	0.0004

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
AgeV1	1	5.9839	0.0144
WBV1	1	3.5566	0.0593
WeatherCode	1	4.1424	0.0418
ccV1	1	8.7255	0.0031

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	5.1730	5.0151	1.06	0.3023
AgeV1		1	0.1519	0.0621	5.98	0.0144
WBV1		1	-0.00450	0.00239	3.56	0.0593
WeatherCode	0	1	0.6879	0.3380	4.14	0.0418
ccV1		1	0.00297	0.00100	8.73	0.0031

Odds Ratio Estimates

Effect	Point Estimate
AgeV1	1.164
WBV1	0.996
WeatherCode 0 vs 1	3.958
ccV1	1.003

Effect		Estimate
AgeV1		1.164
WBV1		0.996
WeatherCode	0 vs 1	3.958
ccV1		1.003

--
* Score Output
--

--
* Report Output
--

Fit Statistics

Target=FatalSik

Fit Statistics	Statistics Label	Train
AIC	Akaike's Information Criterion	94.650
ASE	Average Squared Error	0.187
AVERR	Average Error Function	0.557
DFE	Degrees of Freedom for Error	71.000
DFM	Model Degrees of Freedom	5.000
DFT	Total Degrees of Freedom	76.000
DIV	Divisor for ASE	152.000
ERR	Error Function	84.650
FPE	Final Prediction Error	0.213
MAX	Maximum Absolute Error	0.889
MSE	Mean Square Error	0.200
NOBS	Sum of Frequencies	76.000
_NW_	Number of Estimate Weights	5.000
RASE	Root Average Sum of Squares	0.432
RFPE	Root Final Prediction Error	0.462
RMSE	Root Mean Squared Error	0.447
SBC	Schwarz's Bayesian Criterion	106.304
SSE	Sum of Squared Errors	28.422
SUMW	Sum of Case Weights Times Freq	152.000
MISC	Misclassification Rate	0.237

Classification Table

Data Role=TRAIN Target Variable=FatalSik

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count
0	0	75.0000	78.9474	30
1	0	25.0000	26.3158	10
0	1	22.2222	21.0526	8
1	1	77.7778	73.6842	28

Event Classification Table

Data Role=TRAIN Target=FatalSik

False Negative	True Negative	False Positive	True Positive
10	30	8	28

Model-IB-S

```
 1 *------------------------------------------------------------
   --*
 2 User:                  MINA
 3 Date:                  02-12-2012
 4 Time:                  22H07m
 5 *------------------------------------------------------------
   --*
 6 * Training Output
 7 *------------------------------------------------------------
   --*
 8
 9
10
11
12 Variable Summary
13
14              Measurement    Frequency
15 Role             Level         Count
16
17 INPUT          INTERVAL            4
18 INPUT          NOMINAL             5
19 REJECTED       INTERVAL           13
20 REJECTED       NOMINAL             9
21 TARGET         BINARY              1
22
23
24
25
26 Model Events
27
28                                        Number
29                        Measurement       of
30  Target      Event         Level       Levels      Order
    Label
```

Predicted and decision variables

Type	Variable	Label
TARGET	FatalSik	
PREDICTED	P_FatalSik1	Predicted: FatalSik=1
RESIDUAL	R_FatalSik1	Residual: FatalSik=1
PREDICTED	P_FatalSik0	Predicted: FatalSik=0
RESIDUAL	R_FatalSik0	Residual: FatalSik=0
FROM	F_FatalSik	From: FatalSik
INTO	I_FatalSik	Into: FatalSik

The DMREG Procedure

Model Information

Training Data Set	WORK.EM_DMREG.VIEW
DMDB Catalog	WORK.REG2_DMDB
Target Variable	FatalSik
Target Measurement Level	Ordinal
Number of Target Categories	2
Error	MBernoulli
Link Function	Logit
Number of Model Parameters	10
Number of Observations	76

Target Profile

The selected model is the model trained in the last step (Step 7). It consists of the following effects:

Intercept AgeV1 ccV1

Likelihood Ratio Test for Global Null Hypothesis: BETA=0

-2 Log Likelihood		Likelihood		
Intercept Only	Intercept & Covariates	Ratio Chi-Square	DF	Pr > ChiSq
105.358	92.055	13.3037	2	0.0013

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
AgeV1	1	7.0485	0.0079
ccV1	1	5.1782	0.0229

Analysis of Maximum Likelihood Estimates

Fit Statistics

Target=FatalSik

Fit Statistics	Statistics Label	Train
AIC	Akaike's Information Criterion	98.055
ASE	Average Squared Error	0.206
AVERR	Average Error Function	0.606
DFE	Degrees of Freedom for Error	73.000
DFM	Model Degrees of Freedom	3.000
DFT	Total Degrees of Freedom	76.000
DIV	Divisor for ASE	152.000
ERR	Error Function	92.055
FPE	Final Prediction Error	0.223
MAX	Maximum Absolute Error	0.897
MSE	Mean Square Error	0.215
NOBS	Sum of Frequencies	76.000
_NW_	Number of Estimate Weights	3.000
RASE	Root Average Sum of Squares	0.454
RFPE	Root Final Prediction Error	0.473
RMSE	Root Mean Squared Error	0.463
SBC	Schwarz's Bayesian Criterion	105.047
SSE	Sum of Squared Errors	31.358
SUMW	Sum of Case Weights Times Freq	152.000
MISC	Misclassification Rate	0.276

Event Classification Table

Data Role=TRAIN Target=FatalSik

False Negative	True Negative	False Positive	True Positive
10	27	11	28

Model-IC-S

Table for Model-IC-S Characteristics for Single-Vehicle Crashes.

MODEL IC-S											
Fit Statistics											
Test for Global H_0		**Analysis of Maximum Likelihood Estimates**								**ASE**	**MISC**
DF	Pr>ChSq	**Parameter**			DF	Estimate	Pr>ChSq				
9	0.0051	Intercept			1	4.8806	0.3799			0.178	0.237
		AgeV1			1	0.1789	0.0261				
		AlcoholDrugs (0)			1	-0.5304	0.4713				
		DivisionCode (0)			1	-4.9235	0.9710				
		RanOff (0)			1	-0.2937	0.5539				
		SpeedLevel (0)			1	4.4550	0.9738				
		WBV1			1	-0.0047	0.1074				
		WTV1			1	0.0011	0.6687				
		WeatherCode (0)			1	0.7098	0.0485				
		ccV1			1	0.0025	0.0454				
Odds Ratio Estimates											
		Effect			Point Estimate						
		AgeV1			1.196						
		AlcoholDrugs 0 vs 1			0.346						
		DivisionCode 0 vs 1			<0.001						
		RanOff 0 vs 1			0.556						
		SpeedLevel 0 vs 1			999.000						
		WBV1			0.995						
		WTV1			1.001						
		WeatherCode 0 vs 1			4.136						
		ccV1			1.003						
Accuracy Performance											
Accuracy Rate with Training Sample (N=76)					Accuracy Rate with Original Population (N=500)					Accuracy Performance with 10 Stratified Random Samples	
FN[1]	TN[2]	FP[3]	TP[4]	% AR[5]	TPs[6]	FPs[7]	TNs[8]	FNs[9]	%AR[10]	%A.AR[11]	S.D.[12]
11	31	7	27	76.3	20	100	362	18	76.4	65.3	2.6

1 False Negative; 2 True Negative; 3 False Positive ; 4 True Positive; 5 Percentage of Accuracy Rate; 6 True Positives; 7 False Positives; 8 True Negavtives1; 9 False Negatives; 10 Percentage of Accuracy Rate; 11 Average of Accuracy Rate for the 10 stratified random samples; 12 Standard Deviation for the 10 stratified random samples.

The logistic regression equation developed to predict the probability of a FatalSIK in single-vehicle crashes, Model-IC-S is presented next.

$$\ln\left(\frac{P(Y=1)}{P(Y=0)}\right) = 4.8806 + 0.1789 * AgeV1 - 0.5304 * AlcoholDrugs(=0) - 4.9235$$
$$* DivisionCode(=0) - 0.2937 * RanOff(=0) + 4.4550 * SpeedLevel(=0) - 0.0047$$
$$* WBV1 + 0.0011 * WTV1 + 0.7098 * WeatherCode(=0) + 0.0025 * ccV1$$

Model-ID-S

Table Model-ID-S Characteristics for Single-Vehicle Crashes.

MODEL ID-S							
Fit Statistics							
Test for Global H_0		**Analysis of Maximum Likelihood Estimates**				**ASE**	**MISC**
DF	Pr>ChSq	**Parameter**	DF	Estimate	Pr>ChSq		
4	0.0243	Intercept	1	0.0985	0.0798	0.216	0.368
		BIN_AgeV1 low-5.75	1	-1.3175	0.0057		
		BIN_AgeV1 5.75-10.5	1	-0.2474	0.5718		
		BIN_AgeV1 10.5-15.25	1	0.5469	0.3361		
		DivisionCode (0)	1	-0.5421	0.0809		
Odds Ratio Estimates							
		Effect		Point Estimate			
		BIN_AgeV1 low-5.75 vs 15.25-high		0.097			
		BIN_AgeV1 5.75-10.5 vs 15.25-high		0.282			
		BIN_AgeV1 10.5-15.25 vs 15.25-high		0.624			
		DivisionCode 0 vs 1		0.338			

Accuracy Rate with Training Sample (N=76)					Accuracy Rate with Original Population (N=500)					Accuracy Performance with 10 Stratified Random Samples	
FN[1]	TN[2]	FP[3]	TP[4]	% AR[5]	TPs[6]	FPs[7]	TNs[8]	FNs[9]	%AR[10]	%A.AR[11]	S.D.[12]
12	22	16	26	63.2	11	77	385	27	79.2	56.6	1.9

1 False Negative; 2 True Negative; 3 False Positive ; 4 True Positive; 5 Percentage of Accuracy Rate; 6 True Positives; 7 False Positives; 8 True Negavtives1; 9 False Negatives; 10 Percentage of Accuracy Rate; 11 Average of Accuracy Rate for the 10 stratified random samples; 12 Standard Deviation for the 10 stratified random samples.

The logistic regression equation developed to predict the probability of a FatalSIK in single-vehicle crashes, Model-ID-S is presented next.

$$\ln\left(\frac{P(Y=1)}{P(Y=0)}\right) = 0.0985 - 1.3175 * BinAgeV1(low-5.75) - 0.2474 * BinAgeV1(5.75-10.5)$$
$$+ 0.5469 * BinAgeV1(10.5-15.25) - 0.5421 * DivisionCode(=0)$$

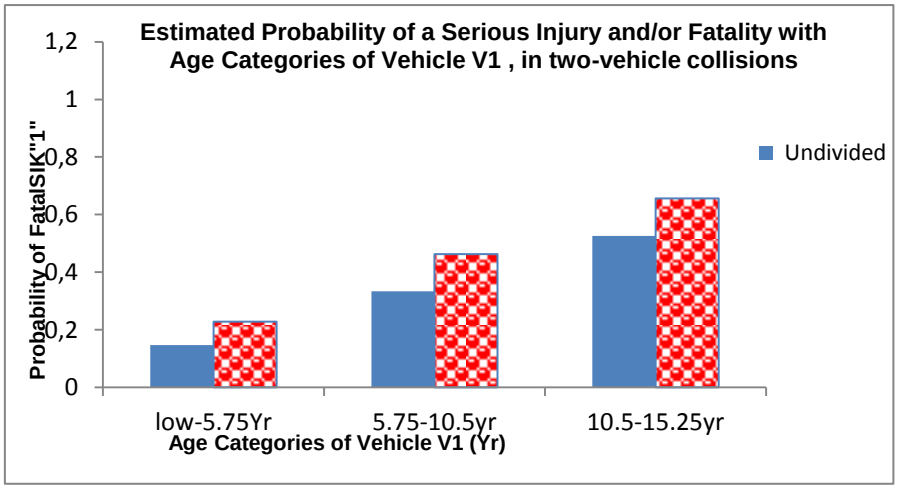


Figure 5 – Probability of a Serious Injury and/or killed in single-vehicle crashes, using Model-IC-S.

Appendix 9: Models for Crash Severity Prediction - Two

Model-I-A-T

```
   -*
 2 User:                  Guilhermina
 3 Date:                  02 de Junho de 2012
 4 Time:                  15H26m
 5 *------------------------------------------------------------
   -*
 6 * Training Output
 7 *------------------------------------------------------------
   -*
 8
 9
10
11
12 Variable Summary
13
14              Measurement    Frequency
15 Role             Level         Count
16
17 INPUT          INTERVAL           12
18 INPUT          NOMINAL             7
19 REJECTED       INTERVAL           18
20 REJECTED       NOMINAL            14
21 TARGET         BINARY              1
22
23
24
25
26 Model Events
27
28                                      Number
29                        Measurement      of
30  Target      Event         Level      Levels      Order
   Label
31
32 FatalSik       1          BINARY          2      Descending
```

Predicted and decision variables

Type	Variable	Label
TARGET	FatalSik	
PREDICTED	P_FatalSik1	Predicted: FatalSik=1
RESIDUAL	R_FatalSik1	Residual: FatalSik=1
PREDICTED	P_FatalSik0	Predicted: FatalSik=0
RESIDUAL	R_FatalSik0	Residual: FatalSik=0
FROM	F_FatalSik	From: FatalSik
INTO	I_FatalSik	Into: FatalSik

The DMREG Procedure

Model Information

Training Data Set	EMWS4.SMPL2_DATA.DATA
DMDB Catalog	WORK.REG3_DMDB
Target Variable	FatalSik
Target Measurement Level	Ordinal
Number of Target Categories	2
Error	MBernoulli
Link Function	Logit
Number of Model Parameters	20
Number of Observations	64

```
    ChiSq
295
296         1    AgeV1          1          1         4.4326         0
    .0353
297         2    HeadOn         1          2         5.7721         0
    .0163
298
299
300 The selected model is the model trained in the last step (St
    ep 2). It consists of the following effects:
301
302 Intercept  AgeV1  HeadOn
303
304
305      Likelihood Ratio Test for Global Null Hypothesis: BETA=
    0
306
307      -2 Log Likelihood          Likelihood
308   Intercept     Intercept &          Ratio
309        Only      Covariates     Chi-Square       DF     Pr >
     ChiSq
310
311      88.723          78.296        10.4269        2
    0.0054
312
313
314            Type 3 Analysis of Effects
315
316                                Wald
317 Effect             DF    Chi-Square    Pr > ChiSq
318
319 AgeV1                1        6.9495        0.0084
320 HeadOn              1        4.4624        0.0346
321
322
```

Fit Statistics

Target=FatalSik

Fit Statistics	Statistics Label	Train
AIC	Akaike's Information Criterion	84.296
ASE	Average Squared Error	0.211
AVERR	Average Error Function	0.612
DFE	Degrees of Freedom for Error	61.000
DFM	Model Degrees of Freedom	3.000
DFT	Total Degrees of Freedom	64.000
DIV	Divisor for ASE	128.000
ERR	Error Function	78.296
FPE	Final Prediction Error	0.232
MAX	Maximum Absolute Error	0.888
MSE	Mean Square Error	0.222
NOBS	Sum of Frequencies	64.000
_NW_	Number of Estimate Weights	3.000
RASE	Root Average Sum of Squares	0.460
RFPE	Root Final Prediction Error	0.482
RMSE	Root Mean Squared Error	0.471
SBC	Schwarz's Bayesian Criterion	90.773
SSE	Sum of Squared Errors	27.039
SUMW	Sum of Case Weights Times Freq	128.000
MISC	Misclassification Rate	0.328

Event Classification Table

Data Role=TRAIN Target=FatalSik

False Negative	True Negative	False Positive	True Positive
10	21	11	22

Model-IB-T

Table Model-IB-T results for FatalSIK prediction with logistic regression performed for a balanced dataset of two-vehicle collisions.

MODEL IB-T							
Fit Statistics							
Test for Global H_0		**Analysis of Maximum Likelihood Estimates**				**ASE**	**MISC**
DF	Pr<ChSq	**Parameter**	DF	Estimate	Pr>ChSq		
18	0.0058	Intercept	1	12.7651	0.8660	0.142	0.234
		AlcoholDrugs (0)	1	-19.9203	1.0000		
		BIN_AgeV2V1 low-4.75	1	3.3569	0.9646		
		BIN_AgeV2V1 4.75-9.5	1	3.0360	0.9680		
		BIN_AgeV2V1 9.5-14.25	1	-9.8549	0.9654		
		BIN_WBV2V1 low-419.25	1	-0.1480	1.0000		
		BIN_WBV2V1 419.25-837.5	1	0.2343	0.8242		
		BIN_WBV2V1 837.5-1255.75	1	-42.1122	0.9633		
		BIN_WTV2V1 low-718.75	1	-3.3030	1.0000		
		BIN_WTV2V1 718.75-1432.5	1	23.8886	0.9509		
		BIN_ccV2V1 low-626.5	1	8.3469	1.0000		
		BIN_ccV2V1 626.5-1253	1	7.8063	<0.0001		
		BIN_ccV2V1 1253-1879.5	1	-18.3480	0.9623		
		DivisionCode (0)	1	26.5263	<0.0001		
		HeadOn (0)	1	-1.2097	0.1159		
		RearEnd (0)	1	-0.2905	0.4956		
		Sideswipe (0)	1	-0.7549	0.1355		
		SpeedLevel (0)	1	-26.8774	1.0000		
		WeatherCode (0)	1	0.6189	0.1586		

Odds Ratio Estimates	
Effect	Point Estimate
AlcoholDrugs 0 vs 1	<0.001
BIN_AgeV2V1 low-4.75 vs 14.25-high	0.900
BIN_AgeV2V1 4.75-9.5 vs 14.25-high	0.653
BIN_AgeV2V1 9.5-14.25 vs 14.25-high	<0.001
BIN_WBV2V1 low-419.25 vs 1255.75-high	<0.001
BIN_WBV2V1 419.25-837.5 vs 1255.75-high	<0.001
BIN_WBV2V1 837.5-1255.75 vs 1255.75-high	<0.001
BIN_WTV2V1 low-718.75 vs 2146.25-high	999.000
BIN_WTV2V1 718.75-1432.5 vs 2146.25-high	999.000
BIN_ccV2V1 low-626.5 vs 1879.5-high	469.675
BIN_ccV2V1 626.5-1253 vs 1879.5-high	273.541
BIN_ccV2V1 1253-1879.5 vs 1879.5-high	<0.001
DivisionCode 0 vs 1	999.000
HeadOn 0 vs 1	0.089
RearEnd 0 vs 1	0.559
Sideswipe 0 vs 1	0.221
SpeedLevel 0 vs 1	<0.001
WeatherCode 0 vs 1	3.448

Accuracy Performance											
Accuracy Rate with Training Sample (N=64)					Accuracy Rate with Original Population (N=874)					Accuracy Performance with 10 Stratified Random Samples	
FN[1]	TN[2]	FP[3]	TP[4]	% AR[5]	TPs[6]	FPs[7]	TNs[8]	FNs[9]	%AR[10]	%A.AR[11]	S.D.[12]
10	27	5	22	76.6	19	139	703	13	82.6	72.5	1.3

1 False Negative; 2 True Negative; 3 False Positive ; 4 True Positive; 5 Percentage of Accuracy Rate; 6 True Positives; 7 False Positives; 8 True Negavtives1; 9 False Negatives; 10 Percentage of Accuracy Rate; 11 Average of Accuracy Rate for the 10 stratified random samples; 12 Standard Deviation for the 10 stratified random samples.

The final model has 18 explanatory variables which makes it very complex. The logistic regression Model-IB-T equation developed to estimate the probability of Y =FatalSIK).

$P\ (FatalSIK = 1) =$ (exp(12.7651-
19.9203*AlcoholDrugs(=0)+3.3569*BIN_AgeV2V1(=low-
4.75)+3.0360*BIN_AgeV2V1(=4.75-9.5) -9.8549*BIN_AgeV2V1(=9.5-14.25) -
0.1480*BIN_WBV2V1(=low-419.25)+0.2343*BIN_WBV2V1(=419.25-873.5) -
42.1122*BIN_WBV2V1(=873.5-1255.75) -3.3030*BIN_WTV2V1(=low-873.5)-
42.1122*BIN_WBV2V1(=873.5-1255.75)+8.3469*BIN_ccV2V1(= low-
626.5)+7.8063*BIN_ccV2V1(=626.5-1253)-18.3480*BIN_ccV2V1(=1253-1879.5)
+26.5263*DivisionCode(=0)-1.2097*HeadOn(=0)-0.2905*RearEnd(=0)-
0.7549*Sideswipe(=0)-26.8774*SpeedLevel(=0)+0.6189*WeatherCode(=0)))

/

(1+exp(12.7651-19.9203*AlcoholDrugs(=0)+3.3569*BIN_AgeV2V1(=low-
4.75)+3.0360*BIN_AgeV2V1(=4.75-9.5) -9.8549*BIN_AgeV2V1(=9.5-14.25) -
0.1480*BIN_WBV2V1(=low-419.25)+0.2343*BIN_WBV2V1(=419.25-873.5) -
42.1122*BIN_WBV2V1(=873.5-1255.75) -3.3030*BIN_WTV2V1(=low-873.5)-
42.1122*BIN_WBV2V1(=873.5-1255.75)+8.3469*BIN_ccV2V1(= low-
626.5)+7.8063*BIN_ccV2V1(=626.5-1253)-18.3480*BIN_ccV2V1(=1253-1879.5)
+26.5263*DivisionCode(=0)-1.2097*HeadOn(=0)-0.2905*RearEnd(=0)-
0.7549*Sideswipe(=0)-26.8774*SpeedLevel(=0)+0.6189*WeatherCode(=0)))

Model-IC-T

Table - Model-IC-T results for FatalSIK prediction with logistic regression performed for a balanced dataset of two-vehicle collisions.

MODEL IC-T							
Fit Statistics							
Test for Global H_0		**Analysis of Maximum Likelihood Estimates**				**ASE**	**MISC**
DF	Pr<ChSq	**Parameter**	DF	Estimate	Pr>ChSq		
9	0.003	Intercept	1	9.1217	0.9778	0.167	0.234
		AlcoholDrugs (0)	1	-6.2847	0.9728		
		BIN_WBV2 low-2347.5	1	-1.9391	0.9881		
		BIN_WBV2 2347.5-2883	1	-3.9748	0.9756		
		BIN_WBV2 2883-3418.5	1	-3.2695	0.9800		
		BIN_WTV2 low-1452.5	1	2.4477	0.9933		
		BIN_WTV2 1452.5-2135	1	3.8351	0.9895		
		BIN_WTV2 2135-2817.5	1	16.2879	0.9782		
		HeadOn (0)	1	-1.2460	0.0422		
		Sideswipe (0)	1	-0.8665	0.0358		
Odds Ratio Estimates							
		Effect		Point Estimate			
		AlcoholDrugs 0 vs 1		<0.001			
		BIN_WBV2 low-2347.5 vs 3418.5-high		<0.001			
		BIN_WBV2 2347.5-2883 vs 3418.5-high		<0.001			
		BIN_WBV2 2883-3418.5 vs 3418.5-high		<0.001			
		BIN_WTV2 low-1452.5 vs 2817.5-high		999.000			
		BIN_WTV2 1452.5-2135 vs 2817.5-high		999.000			
		BIN_WTV2 2135-2817.5 vs 2817.5-high		999.000			
		HeadOn 0 vs 1		0.083			
		Sideswipe 0 vs 1		0.177			
Accuracy Performance							
Accuracy Rate with Training Sample (N=64)					Accuracy Rate with Original Population (N=874)					Accuracy Performance with 10 Stratified Random Samples	
FN[1]	TN[2]	FP[3]	TP[4]	% AR[5]	TPs[6]	FPs[7]	TNs[8]	FNs[9]	%AR[10]	%A.AR[11]	S.D.[12]
5	22	10	27	76.6	13	131	711	19	82.8	60.6	5.6

1 False Negative; 2 True Negative; 3 False Positive ; 4 True Positive; 5 Percentage of Accuracy Rate; 6 True Positives; 7 False Positives; 8 True Negavtives1; 9 False Negatives; 10 Percentage of Accuracy Rate; 11 Average of Accuracy Rate for the 10 stratified random samples; 12 Standard Deviation for the 10 stratified random samples.

The logistic regression Model-IC-T equation to estimate the probability of Y (FatalSIK) is presented below.

$P\ (FatalSIK = 1) =$ (exp(9.1217-6.2847*AlcoholDrugs(=0)-1.9391*BinWBV2V1(=low-2347.5)-3.9748*BinWBV2V1(=2347.5-2883)-3.2695*BinWBV2V1(=2883-3418.5)+2.4477*BinWTV2V1(=low-1452.5)+3.8351*BinWTV2V1(=1452.5-2135)+16.2879*BinWTV2V1(=2135-2817.5)-1.2460*HeadOn(=0)-0.8665*Sideswipe(=0)))

/

(1+(exp(9.1217-6.2847*AlcoholDrugs(=0)-1.9391*BinWBV2V1(=low-2347.5)-3.9748*BinWBV2V1(=2347.5-2883)-3.2695*BinWBV2V1(=2883-3418.5)+2.4477*BinWTV2V1(=low-1452.5)+3.8351*BinWTV2V1(=1452.5-2135)+16.2879*BinWTV2V1(=2135-2817.5)-1.2460*HeadOn(=0)-0.8665*Sideswipe(=0)))

Model-II-T

SAS Code for FatalSIKV1

```
libname  models '.';

data models.twofatalSikV1;
   infile 'two.csv' dsd firstobs=2 missover lrecl=2000;

     Input Record $    RoadName $
     DivisionCode $  SpeedLimit $ SpeedLevel $
     WeatherCode $    AlcoholDrugs $

     RearEnd $ HeadOn $ Sideswipe $ Other $
     (PlateV1 BrandV1 ModelV1)(:$16.) CategoryV1 :$2. WTV1 ccV1 WBV1    FuelV1 :$1.    YrV1 AgeV1
     (PlateV2 BrandV2 ModelV2)(:$16.) CategoryV2 :$2. WTV2 ccV2 WBV2    FuelV2 :$1.    YrV2 AgeV2

     LightInjuryV1    SeriousInjuryV1 KilledV1
     LightInjuryV2    SeriousInjuryV2 KilledV2

     SUMLI    SUMSI    SUMK    SocietyCost ;

 run;
proc print data=models.twofatalSikV1;

 run;
```

```
data models.twofatalSikV1; set models.twofatalSikV1;


SIK = SUMSI + SUMK;
if (SUMSI > 0 or SUMK > 0) then FatalSikV1V2 = 1;
else
FatalSikV1V2 = 0;
if SUMK > 0 then FatalK = 1;
else FatalK = 0;
SIKRatio = SIK/(SIK + SUMLI);
If SIKRatio = 1 then FatalSIKRatio = 1;
Else FatalSIKRatio = 0;
```

```
if (SeriousInjuryV1 >0 or KilledV1 >0)  then FatalSik = 1;
else FatalSik = 0;

if (SeriousInjuryV2 >0  or KilledV2>0) then FatalSikV2 = 1;
else FatalSikV2 = 0;

WTV2V1 = WTV2-WTV1;
ccV2V1 = ccV2-ccV1;
WBV2V1 = WBV2-WBV1;
AgeV2V1 =AgeV2-AgeV1;
run;
```

```
proc print data=models.twofatalSikV1;

run;
```

Model-II-T (Cont.)

EM output

Likelihood Ratio Test for Global Null Hypothesis: BETA=0

-2 Log Likelihood Intercept Only	-2 Log Likelihood Intercept & Covariates	Likelihood Ratio Chi-Square	DF	Pr > ChiSq
58.224	54.670	3.5542	1	0.0594

Type 3 Analysis of Effects

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	P
ccV2	1	3.1444	

Odds Ratio

Odds Ratio Estimates

Effect	Point Estimate
ccV2	1.001

MISC	SSE
0.3571	9.3983

Analsyis of Maximum Likelihhod Estimates

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate	Exp(Est)	95% Confid Limits
Intercept	1	-2.0657	1.1961	2.98	0.0842		0.127	-4.4101
ccV2	1	0.00108	0.000610	3.14	0.0762	0.3454	1.001	-0.00011

Model-III-T

SAS Code for FatalSIKV2

```
libname  models '.';

data models.twofatalSikV2;
   infile 'two.csv' dsd firstobs=2 missover lrecl=2000;

      Input Record $    RoadName $
      DivisionCode $  SpeedLimit $ SpeedLevel $
      WeatherCode $    AlcoholDrugs $

      RearEnd $ HeadOn $ Sideswipe $ Other $
      (PlateV1 BrandV1 ModelV1)(:$16.) CategoryV1 :$2. WTV1 ccV1 WBV1    FuelV1 :$1.    YrV1 AgeV1
      (PlateV2 BrandV2 ModelV2)(:$16.) CategoryV2 :$2. WTV2 ccV2 WBV2    FuelV2 :$1.    YrV2 AgeV2

      LightInjuryV1    SeriousInjuryV1 KilledV1
      LightInjuryV2    SeriousInjuryV2 KilledV2

      SUMLI    SUMSI    SUMK    SocietyCost ;

run;
proc print data=models.twofatalSikV2;

run;

data models.twofatalSikV2; set models.twofatalSikV2;


SIK = SUMSI + SUMK;
if (SUMSI > 0 or SUMK > 0) then FatalSikV1V2 = 1;
else
FatalSikV1V2 = 0;
if SUMK > 0 then FatalK = 1;
else FatalK = 0;
SIKRatio = SIK/(SIK + SUMLI);
If SIKRatio = 1 then FatalSIKRatio = 1;
Else FatalSIKRatio = 0;


 if (SeriousInjuryV1 >0 or KilledV1 >0)  then FatalSikV1 = 1;
 else FatalSikV1 = 0;

 if (SeriousInjuryV2 >0  or KilledV2>0) then FatalSik = 1;
 else FatalSik = 0;

 WTV2V1 = WTV2-WTV1;
 ccV2V1 = ccV2-ccV1;
 WBV2V1 = WBV2-WBV1;
 AgeV2V1 =AgeV2-AgeV1;
 run;

proc print data=models.twofatalSikV2;

 run;
```

EM output for fatalSIKV2

Likelihood Ratio Test for Global Null Hypothesis: BETA=0

-2 Log Likelihood Intercept Only	-2 Log Likelihood Intercept & Covariates	Likelihood Ratio Chi-Square	DF	Pr > ChiSq
38.816	33.410	5.4061	1	0.0201

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
ccVl	1	4.2754	0.0387

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate	Exp(Est)	95% Confidence Limits	
Intercept	1	-3.5969	1.7800	4.08	0.0433		0.027	-7.0856	-0.1082
ccVl	1	0.00205	0.000992	4.28	0.0387	0.5563	1.002	0.000107	0.00400

Odds Ratio Estimates

Effect	Point Estimate
ccVl	1.002

Appendix 10: Models for Vehicles Emissions

MODEL CO_2 LPGV Emissions Estimation (N=817 Vehicles)

Model Fit Statistics

R-Square	0.9479	Adj R-Sq	0.9473
AIC	2405.5916	BIC	2407.8740
SBC	2452.6480	C(p)	8.6095

Type 3 Analysis of Effects

Effect	DF	Sum of Squares	F Value	Pr > F
Norm	4	19432.8927	258.83	<.0001
SpeedLimit	3	178939.681	3177.77	<.0001
cc	2	71582.6248	1906.84	<.0001

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept		1	172.2	0.7574	227.40	<.0001	170.7	173.7
Norm	ECE15-00/04	1	8.7158	0.6271	13.90	<.0001	7.4866	9.9450
Norm	Euro I	1	3.6156	0.5710	6.33	<.0001	2.4964	4.7348
Norm	Euro II	1	-5.5574	0.5576	-9.97	<.0001	-6.6502	-4.4646
Norm	Euro III	1	-3.2152	0.5598	-5.74	<.0001	-4.3124	-2.1179
SpeedLimit	100	1	-3.2140	0.6610	-4.86	<.0001	-4.5096	-1.9183
SpeedLimit	120	1	21.6372	0.4525	47.82	<.0001	20.7503	22.5240
SpeedLimit	50	1	-7.9800	1.1705	-6.82	<.0001	-10.2741	-5.6859
cc	1.4-2.	1	-8.3384	0.4431	-18.82	<.0001	-9.2069	-7.4700
cc	<1.4l	1	-23.7741	0.3923	-60.60	<.0001	-24.5430	-23.0053

ASE= 18.54

MODEL CO_2 LPDV Emissions Estimation (N= 344 Vehicles)

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	322790	80698	547.17	<.0001
Error	339	49996	147.482280		
Corrected Total	343	372787			

Model Fit Statistics

R-Square	0.8659	Adj R-Sq	0.8643
AIC	1722.7989	BIC	1724.8932
SBC	1742.0021	C(p)	6.7807

Type 3 Analysis of Effects

Effect	DF	Sum of Squares	F Value	Pr > F
Norm	2	24365.3576	82.60	<.0001
SpeedLimit	1	87093.5181	590.54	<.0001
cc	1	223388.516	1514.68	<.0001

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept		1	187.5	1.3421	139.67	<.0001	184.8	190.1
Norm	Conventional	1	17.3330	1.6170	10.72	<.0001	14.1637	20.5023
Norm	Euro III	1	-7.9800	1.2766	-6.25	<.0001	-10.4821	-5.4780
SpeedLimit	120	1	18.5736	0.7643	24.30	<.0001	17.0755	20.0716
cc	<2.01	1	-29.9821	0.7704	-38.92	<.0001	-31.4920	-28.4722

ASE=145.34

MODEL CO LPGV Emissions Estimation (N= 847 Vehicles)

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	1137.943396	162.563342	4968.61	<.0001
Error	839	27.450454	0.032718		
Corrected Total	846	1165.393850			

Model Fit Statistics

R-Square	0.9764	Adj R-Sq	0.9762
AIC	-2888.6323	BIC	-2886.4123
SBC	-2850.6987	C(p)	4.4895

Type 3 Analysis of Effects

Effect	DF	Sum of Squares	F Value	Pr > F
Norm	5	1066.1997	6517.50	<.0001
SpeedLimit	2	190.2079	2906.77	<.0001

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept		1	1.5339	0.0281	54.53	<.0001	1.4788	1.5890
Norm	ECE15-00/04	1	2.8646	0.0243	117.69	<.0001	2.8169	2.9123
Norm	Euro I	1	0.5828	0.0213	27.34	<.0001	0.5410	0.6246
Norm	Euro II	1	-0.8589	0.0206	-41.64	<.0001	-0.8994	-0.8185
Norm	Euro III	1	-0.2333	0.0207	-11.25	<.0001	-0.2739	-0.1926
Norm	Euro IV	1	-1.1593	0.0258	-45.00	<.0001	-1.2098	-1.1088
SpeedLimit	120	1	0.7568	0.0224	33.81	<.0001	0.7130	0.8007
SpeedLimit	50	1	-0.4766	0.0431	-11.07	<.0001	-0.5610	-0.3922

ASE=0.03

MODEL NO_x LPVD Emissions Estimation (N=769)

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	18.261962	2.282745	371.25	<.0001
Error	760	4.673134	0.006149		
Corrected Total	768	22.935096			

Type 3 Analysis of Effects

Type 3 Analysis of Effects

Effect	DF	Sum of Squares	F Value	Pr > F
Norm	5	6.3605	206.88	<.0001
SpeedLimit	3	12.4049	672.48	<.0001

Model Fit Statistics

Model Fit Statistics

R-Square	0.7962	Adj R-Sq	0.7941
AIC	-3906.4078	BIC	-3904.1827
SBC	-3864.6020	C(p)	8.4899

Analysis of Maximum Likelihood Estimates

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept		1	0.6569	0.0125	52.68	<.0001	0.6325	0.6813
Norm	Conventional	1	0.2069	0.0104	19.81	<.0001	0.1864	0.2273
Norm	Euro I	1	-0.0293	0.00900	-3.25	0.0012	-0.0469	-0.0116
Norm	Euro II	1	0.0197	0.00678	2.90	0.0038	0.00641	0.0330
Norm	Euro III	1	0.1205	0.00578	20.86	<.0001	0.1092	0.1319
Norm	Euro IV	1	-0.0398	0.00646	-6.15	<.0001	-0.0524	-0.0271
SpeedLimit	100	1	-0.0363	0.0152	-2.39	0.0169	-0.0660	-0.00659
SpeedLimit	120	1	0.2160	0.0122	17.76	<.0001	0.1922	0.2399
SpeedLimit	50	1	-0.1087	0.0342	-3.18	0.0015	-0.1757	-0.0417

Frequency Distribution of Input Class Variables

Class	Value	Total
Norm	Conventional	46
	Euro I	67
	Euro II	157
	Euro III	296
	Euro IV	187
	Euro V	16
SpeedLimit	100	35
	120	544
	50	3
	90	187
cc	<2.01	598
	>2.01	171

ASE= 0,006057

MODEL PM LPDP Emissions Estimation (N=731)

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	1.652972	0.275495	995.51	<.0001
Error	724	0.200357	0.000277		
Corrected Total	730	1.853330			

Type 3 Analysis of Effects

Type 3 Analysis of Effects

Effect	DF	Sum of Squares	F Value	Pr > F
Norm	5	1.6079	1162.03	<.0001
SpeedLimit	1	0.1282	463.24	<.0001

Model Fit Statistics

Model Fit Statistics

R-Square	0.8919	Adj R-Sq	0.8910
AIC	-5981.7101	BIC	-5979.5706
SBC	-5949.5492	C(p)	6.7822

Analysis of Maximum Likelihood Estimates

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept		1	0.0772	0.00101	76.70	<.0001	0.0752	0.0792
Norm	Conventional	1	0.1286	0.00226	56.89	<.0001	0.1242	0.1331
Norm	Euro I	1	0.0594	0.00195	30.48	<.0001	0.0555	0.0632
Norm	Euro II	1	-0.0206	0.00147	-13.96	<.0001	-0.0235	-0.0177
Norm	Euro III	1	-0.0304	0.00124	-24.43	<.0001	-0.0328	-0.0280
Norm	Euro IV	1	-0.0500	0.00140	-35.75	<.0001	-0.0527	-0.0473
SpeedLimit	120	1	0.0154	0.000714	21.52	<.0001	0.0140	0.0168

Frequency Distribution of Input Class Variables

Class	Value	Total
Norm	Conventional	44
	Euro I	64
	Euro II	146
	Euro III	284
	Euro IV	177
	Euro V	16
SpeedLimit	120	544
	90	187
cc	<2.01	571
	>2.01	160

ASE=0,000274