National Technical University of Athens
School of Civil Engineering
Department of Transportation Planning and Engineering

Machine-learning based road crash risk assessment fusing infrastructure, traffic and driver behaviour data

Dimitrios Nikolaou

Doctoral Dissertation

Supervising Committee:
George Yannis, Professor NTUA
Constantinos Antoniou, Professor TUM
Stergios Mavromatis, Associate Professor NTUA

March 2024

Contact Details

Colleagues, students, or anyone interested in obtaining further information about this PhD dissertation, please feel free to contact me using the following contact details:

Dimitrios Nikolaou

National Technical University of Athens (NTUA), Department of Transportation Planning and Engineering, 5 Heroon Polytechniou Str., Zografou Campus, GR-15773, Athens, Greece

Professional e-mail: dnikolaou@mail.ntua.gr
Personal e-mail: nikolaoudn@gmail.com

Phone (NTUA): +30.210.772.1155
Phone (mobile): +30.694.099.3136

Researcher website: https://www.nrso.ntua.gr/dnikolaou

Acknowledgements

Upon completion of this doctoral dissertation, I would like to thank my supervisor, Professor George Yannis, for giving me the opportunity to embark on this exciting research journey. He has been a constant source of inspiration and guidance ever since my undergraduate years. Our collaboration has been invaluable, not only for his insightful advice that contributed to my professional and research skills development but also for the profound life lessons he imparted along the way.

I would also like to sincerely thank my two co-supervisors, Professor Constantinos Antoniou and Associate Professor Stergios Mavromatis, for their contribution and support throughout all phases of this dissertation. In addition, I am very grateful to the remaining members of the examination committee, Professor Eleni Vlahogianni, Associate Professor Eleonora Papadimitriou, Assistant Professor Konstantinos Gkiotsalitis and Assistant Professor Athanasios Theofilatos regarding their constructive academic input and comments.

I am also grateful to OSeven Telematics for kindly providing the naturalistic driving data for this research. Additionally, I extend my thanks to Olympia Odos Operation SA, for kindly providing crash and traffic data essential for the motorway investigations.

I would also like to acknowledge the support and continuous guidance that I received from my colleagues Katerina Folla and Dr. Apostolos Ziakopoulos. From our earliest collaboration during my Diploma thesis, Katerina introduced me to the field of statistical analysis and instilled in me a dedication to detail and high-quality work. Apostolos is my mentor in statistical modelling and machine learning. His guidance not only helped me to overcome my initial hesitancy for coding at the beginning of my research journey but transformed it into a passion for exploring unknown paths and methods with determination and commitment to continuous learning.

I am also deeply grateful to Tassos Dragomanovits, whose expertise in geometric design was highly valuable to me during the investigation of available road geometry data in Greece, as well as in the collection of the road geometry data used for the motorway analyses. Likewise, I extend my gratitude to the rest of my colleagues from the great NRSO team and other research groups within the Department of Transportation Planning and Engineering for their support throughout these years.

Words cannot express my gratitude towards my family: my parents, Theodoros and Afroditi, and my brother, Nikos. Their love, guidance, support and continuous encouragement have been the guiding lights throughout my life.

Finally, I am also grateful to my beloved Theodora. Her love, faith, support and understanding have been my source of strength throughout this journey.

This PhD thesis has been performed within the framework of two research projects carried out by the Department of Transportation Planning and Engineering of the School of Civil Engineering of the National Technical University of Athens.

- “i-safemodels - International Comparative Analyses of Road Traffic Safety Statistics and Safety Modeling” co-financed by the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the Call “Bilateral and Multilateral R&T Cooperation between Greece and China” (project code: T7ΔΚΙ00253).
- “SmartMaps: Smart city mapping for safer and eco driver behaviour through smartphone sensor big data” co-financed by the European Union – European Regional Development Fund (ERDF) and Greek national funds through the Operational Program “Competitiveness, Entrepreneurship and Innovation” (EPAnEK) of the National Strategic Reference Framework (NSRF) (project code: T2EΔK-04388).

Table of Contents

Contact Details 3
Acknowledgements 4
Table of Contents 7
List of Figures 11
List of Tables 14
List of Abbreviations 15
Abstract 17
Περίληψη 19
Extended Abstract 21
Εκτεταμένη Περίληψη 31
1. Introduction 43
1.1 Road Safety Overview 43
1.1.1 Road Safety Globally 43
1.1.2 Road Safety in the European Union 45
1.1.3 Road Safety in Greece 46
1.1.4 Surrogate Safety Measures 48
1.2 Objective 50
1.3 Methodology of the Dissertation 51
1.4 Structure of the Dissertation 54
2. Literature Review 57
2.1 Introduction 57
2.2 Review Methodology 60
2.3 Review Findings 62
2.3.1 Types of Surrogate Safety Measures and Historical Crash Data 62
2.3.2 Modelling Approaches 69
2.3.3 Temporal Dimension 72
2.4 Discussion 75
2.4.1 Overall Findings and Trends from Reviewed Studies 75
2.4.2 Future Research Directions 76
2.5 Research Questions 79

3. Methodological Approach 81
3.1 General Methodological Framework 81
3.2 Theoretical Background 83
3.2.1 Descriptive Analysis 83
3.2.2 Linear Regression 83
3.2.3 Negative Binomial Regression 85
3.2.4 Zero-Inflated Negative Binomial Regression 86
3.2.5 Logistic Regression 88
3.2.6 Decision Tree 89
3.2.7 Random Forest 91
3.2.8 Support Vector Machines 93
3.2.9 K-Nearest Neighbours 95
3.2.10 Model Evaluation Metrics 96
3.2.11 SHapley Additive exPlanations (SHAP values) 100
3.2.12 Hierarchical Clustering 101
3.2.13 Detection of Spatial Dependence 102
3.2.14 Spatial Error and Lag Models 104
3.2.15 Spatial Random Forest 105
4. Investigation of Road Safety Modelling Data in Greece 107
4.1 Introduction 107
4.2 Crash Data 108
4.3 Traffic Data 113
4.4 Geometric Design Data 115
4.4.1 Potential Data Sources 115
4.4.2 Pilot Evaluation of Open GIS Road Geometry Data 116
4.4.2.1 Blender Software 117
4.4.2.2 GPS Visualizer platform and Shuttle Radar Topography Mission database 117
4.4.2.3 Comparison of Open GIS Data to topographic survey data 118
4.5 Smartphone Data 120
4.6 Discussion 122
5. Motorway Data Collection and Processing 125
5.1 Introduction 125

5.2 Crash Data 126
5.3 Traffic Data 127
5.4 Road Infrastructure Data 128
5.5 Driver Behaviour Data 130
5.6 Descriptive Statistics 131
6. Motorway Segment Analyses 141
6.1 Introduction 141
6.2 Crash Frequency Model 143
6.3 Definition of Crash Risk Levels 145
6.4 Comparing Machine Learning Techniques for Crash Risk Level Predictions 147
6.5 SHAP values for Crash Risk Level Classifier 152
6.6 Discussion 155
7. Urban and Interurban Road Network Data Collection and Processing 159
7.1 Introduction 159
7.2 Road Infrastructure Data 160
7.3 Driver Behaviour Data 163
7.4 Descriptive Statistics 166
8. Urban and Interurban Road Network Analyses 167
8.1 Introduction 167
8.2 Spatial Error and Lag Models 168
8.3 Spatial Zero-Inflated Negative Binomial Model 172
8.4 Spatial Random Forest 175
8.5 Discussion 180
9. Conclusions 183
9.1 Dissertation Overview 183
9.2 Main Findings for Motorway 187
9.3 Main Findings for Urban and Interurban Road Network 188
9.4 Innovative Contributions 189
9.4.1 Holistic Data Collection Approach 189
9.4.2 Multi-Dimensional Data Fusion for Segment-Level Analyses 190
9.4.3 Advanced and Innovative Combination of Modelling Techniques 191
9.4.4 Multi-factor Estimation of Crash Risk on Motorways 191

9.4.5 Surrogate Estimation of Crash Risk on Urban and Interurban Road Network 192
9.5 Further Challenges 193
References 195
Appendix I 209
List of publications produced within the framework of this dissertation 209
Publications in scientific journals with peer review 209
Publications in scientific conference proceedings (full papers with review) 209
Scientific awards for the publications produced within the framework of this dissertation 209
List of publications in other research thematic areas 210
Publications in scientific journals with peer review 210
Publications in scientific conference proceedings (full papers with review) 210
Scientific awards for the publications in other research thematic areas 213

List of Figures

Figure 1.1: Road fatalities per 100,000 population by continent and country-income level, 2021. 43
Figure 1.2:. Distribution of road fatalities by road user type and region for 2021. 44
Figure 1.3: Evolution of road fatalities in the EU, 2001-2022. 45
Figure 1.4: Road fatalities per million population in the EU, 2019-2022. 46
Figure 1.5: Evolution of basic road safety figures in Greece, 2010-2022. 47
Figure 1.6: Graphical representation of the overall methodological framework 53

Figure 2.1: PRISMA flow diagram 61
Figure 2.2: Time periods of historical road crash data and SSMs collected through smartphones, instrumented vehicles and connected vehicles 73
Figure 2.3: Time periods of historical road crash data and SSMs collected through video records and conflict surveys 73

Figure 3.1: Graphical representation of the overall methodological framework 82
Figure 3.2: Graphical explanation of box plot 83
Figure 3.3: Schematic diagram of simple linear regression 84
Figure 3.4: Example of Poisson distribution (mean=5, variance=5, sample=10,000) 85
Figure 3.5: Example of Negative Binomial distribution (mean=5, variance=5, *k*=1, sample=10,000) 86
Figure 3.6: Example of Zero-Inflated Negative Binomial distribution (mean=5, variance=5, *k*=1, sample=10,000, proportion of zeros = 35%) 87
Figure 3.7: The logistic function with example data 89
Figure 3.8: Typical hierarchical structure of a Decision Tree 90
Figure 3.9: Graphical illustration of the Random Forest algorithm 92
Figure 3.10: Graphical illustration of SVM classification (linear separable example) 93
Figure 3.11: Graphical illustration of the K-NN algorithm 95
Figure 3.12: ROC curve example 99
Figure 3.13: Example of a Dendrogram from Hierarchical Clustering 102
Figure 3.14: Spatial autocorrelation examples 103

Figure 4.1: Locations of available traffic data in the subregion of Viotia. (Source: Road management authority of Viotia subregion - field surveys in September 2014). 113
Figure 4.2: Blender and GPS Visualizer data assessment area 117
Figure 4.3: Coordinate systems of the smartphone and the vehicle 120

Figure 5.1: Olympia Odos motorway in Greece 125
Figure 5.2: Extract of the developed CAD drawing 129
Figure 5.3: Extract of the Google Earth .kmz file 129
Figure 5.4: Number of Total Road Crashes (Injury & PDO), 2018-2020 132

Figure 5.5: Length of motorway segment (km) 133
Figure 5.6: Average AADT of motorway segment (veh/day), 2018-2020 133
Figure 5.7: Posted speed limit (km/h) 134
Figure 5.8: Number of total road crashes by segment length, 2018-2020 134
Figure 5.9: Length of curve in segment (m) 135
Figure 5.10: Lane width (m) 135
Figure 5.11: Paved inside/outside shoulder width (m) 136
Figure 5.12: Distance from edge of inside/outside shoulder to barrier face (m) 136
Figure 5.13: Median width (m) 137
Figure 5.14: Number of recorded trips 137
Figure 5.15: Average speed (all trips) – km/h 138
Figure 5.16: Number of speeding events per trips 138
Figure 5.17: Number of harsh driving behaviour events (accelerations/brakings) per trips 139

Figure 6.1: Hierarchical Clustering Dendrogram 145
Figure 6.2: Confusion Matrix for the test dataset – Logistic Regression 148
Figure 6.3: Confusion Matrix for the test dataset – Decision Tree 149
Figure 6.4: Confusion Matrix for the test dataset – Random Forest 149
Figure 6.5: Confusion Matrix for the test dataset – Support Vector Machines 150
Figure 6.6: Confusion Matrix for the test dataset – K-Nearest Neighbours 150
Figure 6.7: SHAP values for the RF model and a representative motorway segment 153

Figure 7.1: Examined road network of the Eastern Macedonia and Thrace Region (in grey) 160
Figure 7.2: Length of the examined road segments 161
Figure 7.3: Linearity index (“efficiency”) of the examined road segments 161
Figure 7.4: Slope class (%) of the examined road segments 162
Figure 7.5: Histogram of trip duration frequencies in the examined road network 163
Figure 7.6: Speeding (secs) per segment trips 164
Figure 7.7: Mobile phone use (secs) per segment trips 165
Figure 7.8: Harsh braking events per segment trips 165
Figure 7.9: Harsh accelerations per segment trips 165
Figure 7.10: Histogram of harsh braking events in the examined road segments 166

Figure 8.1: Visualization of the SLM results on the examined road network 171
Figure 8.2: Zoomed-in view of the SLM results for the center of Xanthi 171
Figure 8.3: Visualization of the SZINB results on the examined road network 174
Figure 8.4: Zoomed-in view of the SZINB results for the center of Xanthi 174
Figure 8.5: Non-spatial RF model residuals 176
Figure 8.6: SRF model residuals 177
Figure 8.7: RMSE across 30 spatial folds 178

Figure 8.8: Permutation importance computed on the out-of-bag data 179

Figure 9.1: Graphical representation of the overall methodological framework..... 185
Figure 9.2: Innovative contributions of the dissertation 189

List of Tables

Table 1.1: Comparison of Greek and EU road crash statistics, 2019. 47

Table 2.1: Studies exploiting SSMs in historical crash record investigations 64

Table 4.1: Variables included in the national road crash database 109
Table 4.2: Road crashes with unknown road for the years 2011-2015 in the subregion of Viotia 110
Table 4.3: Crashes on known road and unknown station for years 2011-2015 in subregion of Viotia..... 110
Table 4.4: Crashes on known and codified roads and unknown station for the years 2011-2015 in the subregion of Viotia..... 111
Table 4.5: Crashes on known-codified roads and crashes with identical infrastructure characteristics 111
Table 4.6: Accuracy assessment of road centerline points - Blender software 118
Table 4.7: Accuracy assessment of road centerline points - GPS Visualizer platform 118

Table 5.1: Road crash, traffic, geometry, and driver behaviour variables per motorway segment..... 131

Table 6.1: Statistical model for crash frequency in motorway segments..... 143
Table 6.2: Descriptive statistics of the four crash risk levels of the examined motorway segments..... 146
Table 6.3: Basic elements of the five classification models' training 148
Table 6.4: Performance evaluation metrics per crash risk level and developed model 151
Table 6.5: Skewness, kurtosis, and median values of numeric predictors in the training dataset 152

Table 7.1: Geometric characteristics and driving behaviour metrics per examined road segment..... 166

Table 8.1: Log-linear regression (baseline), SEM and SLM results for harsh braking events..... 169
Table 8.2: Zero-inflated Negative Binomial and Spatial Zero-inflated Negative Binomial results for harsh braking events 172
Table 8.3: Key parameters and performance metrics of RF models 177

List of Abbreviations

AADT	Average Annual Daily Traffic
AIC	Akaike Information Criterion
AUC	Area Under the Curve
CAR	Conditional Autoregression
CNN	Convolutional Neural Network
CPM(s)	Crash Prediction Model(s)
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
DL	Deep Learning
DRAC	Deceleration Rate to Avoid the Crash
DT(s)	Decision Tree(s)
EU	European Union
EVT	Extreme Value Theory
FN	False Negatives
FP	False Positives
FPR	False Positive Rate
GLM(s)	Generalized Linear Model(s)
GLMM(s)	Generalized Linear Mixed Model(s)
GORP	Generalized Ordered Response Probit
HC	Hierarchical Clustering
HD	High-Definition
HSM	Highway Safety Manual
INLA	Integrated Nested Laplace Approximation
K-NN	K-Nearest Neighbours
LGM(s)	Latent Gaussian Model(s)
LIME	Local Interpretable Model-Agnostic Explanations
LR	Logistic Regression
ML	Machine Learning
MLP	Multi-layer Perception
mTTC	minimum Time-to-Collision
MTTC	Modified Time-to-Collision
MVCAR	Multivariate Conditional Autoregressive
NB	Negative Binomial
NTUA	National Technical University of Athens
OOB	Out-Of-Bag
OSM	OpenStreetMap
PDO	Property-Damage-Only
PET	Post Encroachment Time
PRISMA	Preferred Reporting Items for Systematic Reviews and the Meta-Analyses
PTW(s)	Powered Two-Wheeler(s)
RBF	Radial-Basis Function

RF	Random Forest
RMSE	Root Mean Square Error
ROC	Receiver Operating Characteristics
SDG(s)	Sustainable Development Goal(s)
SEM	Spatial Error Model
SHAP	SHapley Additive exPlanations
SLM	Spatial Lag Model
SPDE	Stochastic Partial Differential Equation
SRF	Spatial Random Forest
SRTM	Shuttle Radar Topography Mission
SSM(s)	Surrogate Safety Measure(s)
SVM(s)	Support Vector Machine(s)
SZINB	Spatial Zero-Inflated Negative Binomial
TA	Time-to-Accident
TC	Time-to-Crash
TN	True Negatives
TP	True Positives
TPR	True Positive Rate
TTC	Time-to-Collision
V2I	Vehicle-to-Infrastructure
V2V	Vehicle-to-Vehicle
V2X	Vehicle-to-Everything
VIF	Variance Inflation Factor
VRU(s)	Vulnerable Road User(s)
WGS84	World Geodetic System 1984
ZINB	Zero-Inflated Negative Binomial

Abstract

The objective of this doctoral dissertation is to assess road crash risk by fusing infrastructure, traffic, and driving behaviour data. For this reason, two distinct databases were developed. The first one concerned motorway segments and included road crash, traffic, road geometry and driver behaviour data, while the second database concerned urban and interurban road segments of a broader area for which crash and traffic data were unavailable.

The results of the negative binomial regression model for the motorway segments showed a positive and statistically significant relationship between road crash frequency and events of harsh driving behaviour. Subsequently, taking into account the number of road crashes per segment length and traffic volume, four crash risk levels of the motorway segments were formulated using hierarchical clustering. These four crash risk levels were used as the response variable in five machine learning classifiers that included predictors related to road geometry and risky driving behaviours. Among the five classification models, Random Forest demonstrated superior classification performance across all crash risk levels. Based on the SHAP values, it was revealed that harsh braking events serve as a more suitable Surrogate Safety Measure than harsh accelerations in terms of crash risk level prediction.

For this reason, harsh brakings were used as the dependent variable in the analyses for urban and interurban segments of the broader road network. In addition to developing non-spatial models, the identification of spatial autocorrelation led to the development of spatial modelling techniques to account for spatial dependencies. It was found that the number of trips per segment, segment length and linearity, speeding and mobile phone use are positively correlated with harsh brakings. Conversely, motorways exhibited fewer harsh braking events compared to other road types. Furthermore, the number of trips per examined road segment was found to be the most influential predictor, highlighting its importance as a proxy measure of risk exposure. In terms of model performance, the Spatial Lag Model outperformed both the log-linear model and the Spatial Error Model. Better fit was also observed for the spatial Zero-Inflated Negative Binomial model, compared to the corresponding non-spatial model. Finally, the Spatial Random Forest model reduced the absolute values of spatial autocorrelation in the residuals and showed a better fit to the observed data compared to the conventional Random Forest model.

Περίληψη

Ο στόχος της παρούσας διδακτορικής διατριβής είναι η αξιολόγηση του κινδύνου οδικού ατυχήματος συνδυάζοντας δεδομένα υποδομής, κυκλοφορίας και συμπεριφοράς οδηγού. Για τον σκοπό αυτό, αναπτύχθηκαν δύο βάσεις δεδομένων. Η πρώτη αφορούσε τμήματα αυτοκινητοδρόμου και περιλάμβανε δεδομένα οδικών ατυχημάτων, κυκλοφορίας, γεωμετρίας και συμπεριφοράς των οδηγών, ενώ η δεύτερη αφορούσε τμήματα αστικών και υπεραστικών οδών μιας ευρύτερης περιοχής, για τα οποία δεν υπήρχαν διαθέσιμα δεδομένα ατυχημάτων και κυκλοφορίας.

Τα αποτελέσματα του μοντέλου αρνητικής διωνυμικής παλινδρόμησης για τα τμήματα του αυτοκινητοδρόμου έδειξαν θετική και στατιστικά σημαντική συσχέτιση μεταξύ της συχνότητας οδικών ατυχημάτων και των συμβάντων απότομης συμπεριφοράς του οδηγού. Ακολούθως, λαμβάνοντας υπόψη τον αριθμό τον ατυχημάτων ανά μήκος τμήματος και τον κυκλοφοριακό φόρτο, διαμορφώθηκαν τέσσερα επίπεδα επικινδυνότητας των τμημάτων του αυτοκινητοδρόμου με χρήση της ιεραρχικής ομαδοποίησης. Τα τέσσερα επίπεδα επικινδυνότητας χρησιμοποιήθηκαν ως μεταβλητή απόκρισης σε πέντε ταξινομητές μηχανικής μάθησης που περιλάμβαναν προγνωστικούς παράγοντες σχετικά με τη γεωμετρία της οδού και επικίνδυνες συμπεριφορές οδήγησης. Μεταξύ των πέντε ταξινομητών που αναπτύχθηκαν, το μοντέλο Τυχαίων Δασών επέδειξε ανώτερες επιδόσεις ταξινόμησης σε όλες τις κατηγορίες επικινδυνότητας. Με βάση τις τιμές SHAP, προέκυψε ότι οι απότομες επιβραδύνσεις χρησιμεύουν ως καταλληλότερος Έμμεσος Δείκτης Ασφαλείας από τις απότομες επιταχύνσεις για την πρόβλεψη της επικινδυνότητας.

Για τον λόγο αυτό, οι απότομες επιβραδύνσεις αποτέλεσαν την εξαρτημένη μεταβλητή των αναλύσεων για τα αστικά και υπεραστικά τμήματα του ευρύτερου οδικού δικτύου. Πέραν της ανάπτυξης μη χωρικών μοντέλων, ο εντοπισμός χωρικής αυτοσυσχέτισης οδήγησε στην ανάπτυξη χωρικών τεχνικών μοντελοποίησης, ώστε να ληφθούν υπόψη οι χωρικές εξαρτήσεις. Προέκυψε ότι ο αριθμός των διαδρομών ανά τμήμα, το μήκος και η γραμμικότητα του τμήματος, η υπέρβαση των ορίων ταχύτητας και η απόσπαση προσοχής συσχετίζονται θετικά με τις απότομες επιβραδύνσεις. Αντιθέτως, οι αυτοκινητόδρομοι παρουσίασαν λιγότερες απότομες επιβραδύνσεις συγκριτικά με άλλους τύπους οδού. Επιπλέον, προέκυψε ότι ο αριθμός των διαδρομών ανά τμήμα είναι ο πιο σημαντικός παράγοντας πρόβλεψης, αναδεικνύοντας την σημασία του ως υποκατάστατο μέτρο έκθεσης στον κίνδυνο. Όσον αφορά την επίδοση των μοντέλων, το Χωρικό Μοντέλο Υστέρησης ξεπέρασε τόσο το λογαριθμογραμμικό μοντέλο όσο και το Χωρικό Μοντέλο διόρθωσης του Σφάλματος. Καλύτερη προσαρμογή παρατηρήθηκε και για το χωρικό μοντέλο Μηδενικά Διογκωμένης Αρνητικής Διωνυμικής Παλινδρόμησης, συγκριτικά με το αντίστοιχο μη χωρικό μοντέλο. Τέλος, το Χωρικό μοντέλο Τυχαίων Δασών μείωσε τις απόλυτες τιμές της χωρικής αυτοσυσχέτισης στα κατάλοιπα και παρουσίασε καλύτερη προσαρμογή στα παρατηρούμενα δεδομένα συγκριτικά με το συμβατικό μοντέλο Τυχαίων Δασών.

Extended Abstract

Recognizing road safety as a crucial public health issue with significant societal and economic consequences, it is essential to understand the **multifaceted nature of road crashes**. Road crashes are influenced by various parameters that can be divided into three distinct categories: (i) road users, (ii) vehicles, and (iii) road infrastructure and environment. Notably, a substantial percentage of road crashes, up to 94%, can be attributed to human factors and errors, either exclusively or partially.

Given the aforementioned context, the main objective of this dissertation is to **assess road crash risk by fusing infrastructure, traffic, and driving behaviour data**. This integration of data presents a promising avenue for research. Nevertheless, the practical implementation of this data fusion is frequently hindered by challenges such as insufficient availability or suboptimal quality of the data.

Within the framework of this dissertation, an extensive literature review was conducted. The aim of this literature review process was to provide a review of the scientific literature of studies exploiting Surrogate Safety Measures (SSMs) in historical crash record investigations. SSMs encompass a wide range of metrics and parameters, which are not directly derived from or rely on crash data. From the review process, it was concluded that **SSMs are steadily gaining ground in the road safety literature** as they are a sustainable way of gauging road safety and allow the conduction of analyses without necessarily requiring historical road crash records. These indicators can either be an alternative to road safety analyses or even complement analyses that are based on historical crash records. Moreover, the rapid and continuous progress in the field of technology makes it increasingly easier to collect such metrics. SSMs such as time-to-collision, harsh braking, post-encroachment time and so on, are widely proposed in transportation science and are particularly useful in order to evaluate driving risk and assess road crash risk.

Subsequently, the following **research questions** were formulated:

Question 1
How can infrastructure, traffic and driver behaviour data be fused and analyzed to derive meaningful conclusions for road crash risk assessment?

Question 2
a) Can harsh driving behaviour events be meaningfully considered reliable SSMs?
b) Is there a statistically significant positive correlation between harsh driving behaviour events and historical road crash records?

Question 3

Is it possible to predict the crash risk level of road segments by exploiting road geometry characteristics and driver-behaviour based SSMs, and, if so, which Machine Learning (ML) classifiers are the most appropriate?

Question 4

Are harsh braking events more pertinent than harsh accelerations in predicting the crash risk level of road segments?

Question 5

a) In the absence of highly detailed historical road crash data, how can harsh braking events be analyzed across various road environments?
b) Is there spatial autocorrelation present in harsh braking frequencies for road segments, and, if so, do spatial modelling approaches outperform their non-spatial counterparts?

Question 6

Which road infrastructure and driver behaviour parameters exhibit a statistically significant impact on the number of harsh braking events per road segment?

These research questions served as the driving force behind the entire research endeavor, exploring the integration and analysis of infrastructure, traffic, and driver behaviour data for meaningful conclusions in road crash risk assessment. In order to answer these research questions, an elaborate **methodological framework** was devised, which is shown in Figure I.

The core of the methodological framework involved a multi-step process, commencing with the **investigation of road safety modelling data in Greece**, laying the groundwork for subsequent directions. This investigation highlighted the constraints associated with conducting high-detailed crash prediction modelling in Greece. Such modelling is only feasible for motorways with high-quality crash data, specifically regarding crash location and traffic attributes per road segment. In response to this limitation, two distinct databases were developed.

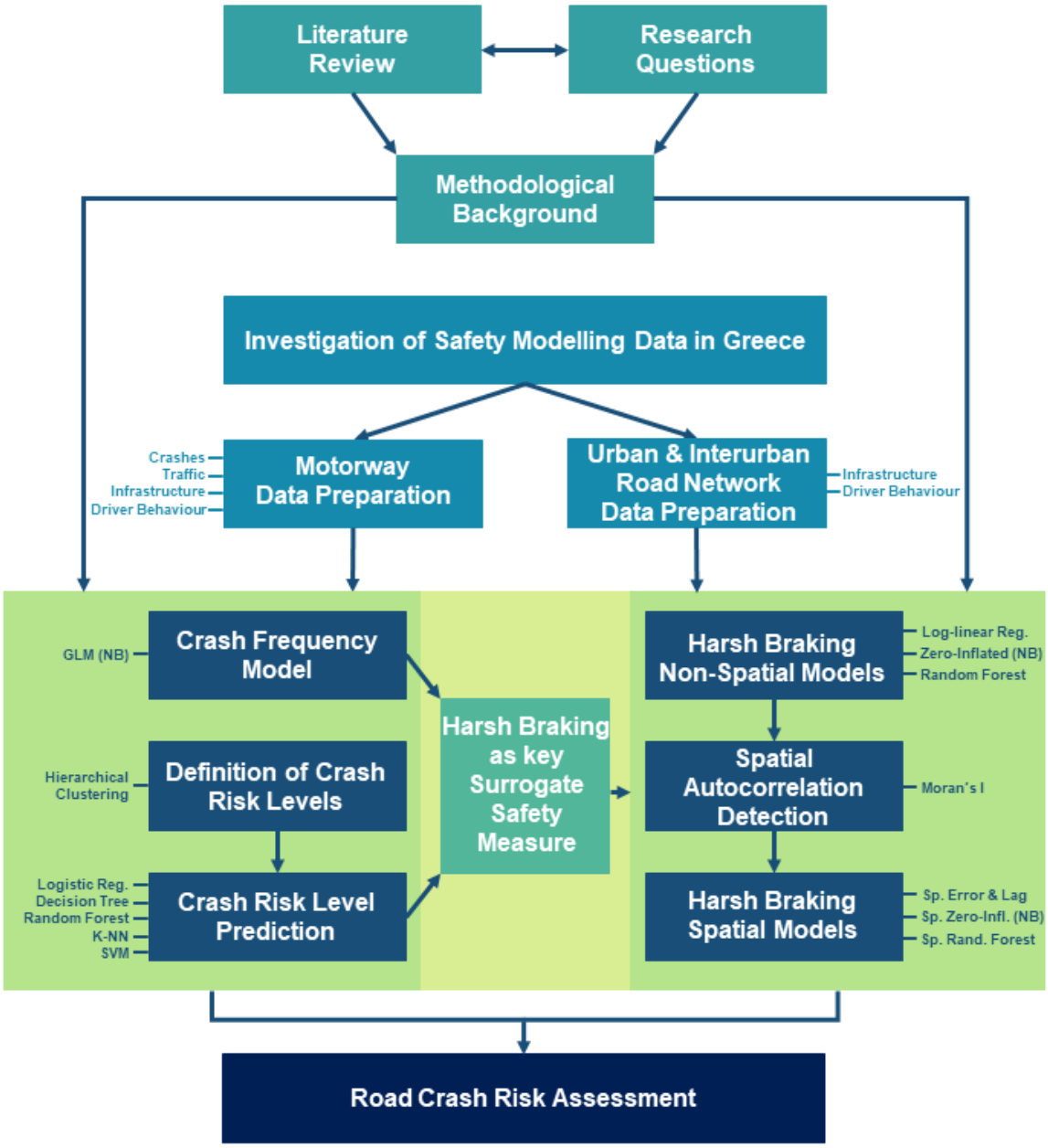


Figure I: Graphical representation of the overall methodological framework of the doctoral dissertation

The first one focused on **668 motorway segments** within the Olympia Odos motorway, containing comprehensive data on historical road crashes, traffic, road geometry characteristics, and naturalistic driver behaviour metrics. Specifically, crash data of all severity levels including property-damage-only (PDO) crashes for the years 2018-2020 were exploited. In parallel with the road crash data, Average Annual Daily Traffic (AADT) data for the same time period were included in the developed database.

Regarding the road infrastructure characteristics, a variety of sources, such as information from the road operator and the use of different software, including Open GIS, Google Earth and GoogleStreetView, were combined. The inclusion of these road infrastructure data and of reference drawings of the motorway also enabled the identification and isolation of naturalistic driver behaviour data from a smartphone application. Driver behaviour data were collected for the period from June 1, 2019, to December 31, 2020, from a sample of 327 drivers in 2019 and 330 drivers in 2020. The average number of trips per motorway segment over the entire study period was 769 trips.

The second one covered a **broader road network within the Region of Eastern Macedonia and Thrace**, including urban and interurban roads. Within this road network, an initial analysis was conducted on all road segments sourced from OpenStreetMap (OSM) to extract their geometric and network characteristics. Subsequently, naturalistic driving behaviour data that were extracted from a smartphone application were aligned with the corresponding OSM segments. The examined road network included **6,103 road segments**, with an average length of 288.8 meters, resulting in a total road network length of 1,763 kilometers. Regarding the naturalistic driver behaviour metrics, data from 5,129 trips during 2021 were utilized. The mean trip duration was 634 seconds, with a standard deviation of 556 seconds. However, the developed database for this road network lacked detailed crash and traffic data for the examined road segments.

Various **methodological tools** were applied for the road segments of Olympia Odos motorway. These included techniques such as Negative Binomial (NB) regression for developing a crash frequency model, Hierarchical Clustering (HC) to determine crash risk levels based on historical crash data and traffic attributes, and the utilization of ML classifiers such as Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), K-Nearest Neighbours (K-NN) and Support Vector Machine (SVM). These classifiers were used for crash risk level prediction, leveraging infrastructure and driver behaviour data. A critical focus was placed on evaluating the reliability of harsh driving behaviour events as SSMs.

Subsequently, the framework extended to include the road network data of Eastern Macedonia and Thrace Region, employing harsh braking events for road crash risk assessment. This involved applying both non-spatial and **spatial models** to identify significant road infrastructure and driver behaviour parameters influencing harsh braking events per road segment.

Ultimately, the synthesis of all the analyses carried out within the framework of this doctoral dissertation resulted in a **comprehensive road crash risk assessment** with numerous original and interesting results, which are discussed in more detail below.

For the motorway analyses, a unified database including data on historical injury and PDO crashes, traffic attributes, road geometry characteristics, and driver behaviour SSMs of 668 road segments of the Olympia Odos motorway was exploited. The results of the crash frequency model (NB regression) revealed that road crash frequency in the examined motorway segments is positively correlated with the traffic volume, the length of the segment, the number of harsh accelerations and the number of harsh brakings per segment trips. This finding contributes to existing road safety literature by establishing a **positive and statistically significant relationship between road crash frequency and events of harsh driving behaviour**. Consequently, it is inferred that these events can serve as a valid subcategory of naturalistic SSMs. Specifically, they can be used either to complement Crash Prediction Models (CPMs) or as dependent variables in diverse proactive road safety analyses, particularly in cases where detailed historical road crash data are lacking.

As a further phase of the motorway investigations, an endeavor was made to formulate **crash risk level clusters** of the motorway segments. This was achieved by considering the number of road crashes by segment length and the traffic volume of each segment using the agglomerative hierarchical clustering technique. Considering the influence of segment length and traffic volume, as indicated by the results of the negative binomial regression model, both variables were included into the clustering analysis due to their statistically significant impact on motorway segment crash frequency. The outcomes of this clustering process delineated four distinct crash risk levels with a clear pattern whereby the first risk level class presents high average numbers of traffic volume and road crashes by segment length, while these figures decrease progressively for each subsequent class.

Subsequently, these identified four levels were utilized as the response variable in five ML classification models (LR, DT, RF, SVM, and K-NN). The models included predictors encompassing road geometry characteristics and unsafe driving behaviours, such as rates of harsh brakings, harsh accelerations, and speeding duration per trips within the analyzed segments. Among the five classification models, **RF demonstrated superior classification performance** across all crash risk levels, consistently achieving scores exceeding 89% (overall accuracy: 89.9%, macro-averaged precision: 90.7%, macro-averaged recall: 89.9%, macro-averaged F1 score: 90.2%). This outcome reveals the potential of the developed RF model as a highly promising proactive road safety tool, capable of effectively identifying and prioritizing potentially hazardous motorway segments.

Finally, to enhance the interpretability of the RF model, which inherently operates as a black-box ML model, SHapley Additive exPlanations (SHAP) values were calculated for a typical motorway segment. Based on the SHAP values of the naturalistic driving behaviour predictors, it was revealed that **harsh braking events serve as a more suitable SSM than harsh accelerations** in terms of crash risk level prediction.

Within the broader road network of the Eastern Macedonia and Thrace Region, a spatial dataset consisting aggregated naturalistic driving behaviour metrics, as well as geometric and network characteristics on a segment level was analyzed. For the examined 6,103 road segments, and based on Moran's I index, statistically significant and positive **spatial autocorrelation in harsh braking event frequencies** was detected. Initially, non-spatial modelling techniques, such as log-linear, Zero-Inflated Negative Binomial (ZINB) and conventional RF regression models were employed on harsh braking events frequencies. However, the existence of spatial autocorrelation highlighted the need for the development of spatial models, such as Spatial Error Model (SEM), Spatial Lag Model (SLM), Spatial Zero-Inflated Negative Binomial (SZINB) and Spatial Random Forest (SRF), in order to take into account such spatial dependencies.

Consistent signs of the beta coefficients emerged across all models. Specifically, road segment length and the number of trips per segment were identified as proxy indicators of risk exposure, positively correlated with harsh braking events. Additionally, the efficiency index (statistically significant only in the log-linear model, SEM and SLM), related to the linearity of road segments, revealed a positive correlation with harsh braking events, suggesting that drivers exhibit more frequent harsh braking on road segments with fewer curves. Variables related to speeding and mobile phone use were also positively associated with harsh braking events, whereas motorways exhibited fewer harsh braking events compared to other road types.

In both RF models, the **number of trips per examined road segment was found to be the most influential predictor**, highlighting its significant relevance in predicting the frequency of harsh braking events, as it serves as a naturalistic driving exposure metric. On the other hand, the motorway variable exhibited the lowest importance, indicating that road type is relatively less valuable in predicting the number of harsh braking events. This finding may suggest that factors other than road type such as driver distraction and speeding, might play a more crucial role in influencing harsh braking events frequencies.

Regarding the performance of the developed models, **SLM surpassed** both the log-linear model and the SEM, with lower AIC values and absence of spatial autocorrelation in its residuals. Lower AIC values, indicating a better fit, were also observed for the SZINB model compared to the non-spatial ZINB model. Moreover, the **SRF reduced the absolute values of spatial autocorrelation in the residuals** compared to the respective values of the conventional RF. In addition, the SRF outperformed the non-spatial RF model in terms of model fit to observed data, but the non-spatial model performed better in terms of generalization to unseen data.

The results of the developed models for the examined road network of the Eastern Macedonia and Thrace Region are also **visualized** in maps. Indicatively, the results

of the SZINB model are presented in Figure II, whereas Figure III provides a zoomed-in view of Figure II, focusing specifically on the center of the regional capital city of Xanthi.

Figure II: Visualization of the SZINB results on the examined road network

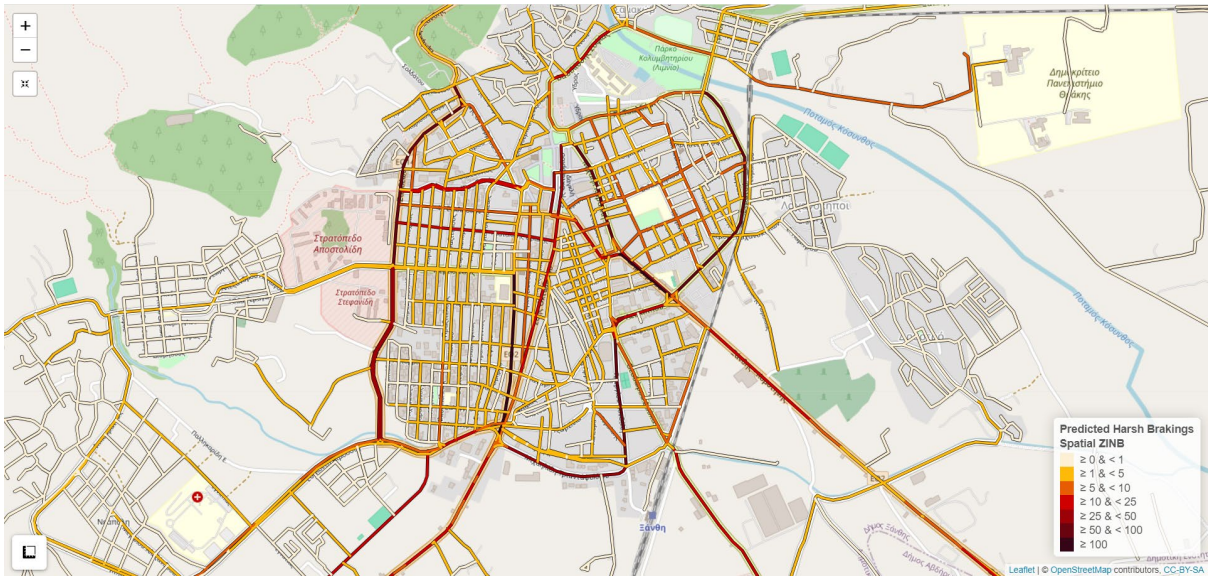


Figure III: Zoomed-in view of the SZINB results for the center of Xanthi

This doctoral dissertation offers significant **noteworthy contributions** in the field of road safety, as discussed below.

Holistic Data Collection Approach

In the context of this doctoral dissertation, a **holistic and comprehensive data collection** was conducted to investigate the impact of driver behaviour, road infrastructure characteristics and traffic attributes on road crash risk assessment. Technological advancements have significantly facilitated the collection of data from

various sources, opening up new research opportunities that were previously unexplored.

Specifically, this dissertation exploited **high-resolution naturalistic driving big datasets** collected from smartphone sensors to assess road crash risk on motorways and a broader road network, encompassing urban and interurban roads. For road infrastructure data on the examined motorway, a variety of sources were exploited, including data provided by the road operator and software such as Open GIS, Google Earth and GoogleStreetView. Geometric and network characteristics for the broader road network of the Eastern Macedonia and Thrace Region were derived using algorithms in the R programming language. Appropriate libraries were utilized to extract data from OSM and process them as simple spatial features. Concerning road crash and traffic data on the examined motorway, high-quality data from the road operator were employed. This included road crash data of all injury severities, including PDO crashes, with high accuracy in crash location, covering the period from 2018 to 2020. Additionally, AADT data derived from the motorway toll stations for the corresponding period were utilized.

Multi-Dimensional Data Fusion for Segment-Level Analyses

The collection of data from various sources and at different levels necessitates **appropriate processing for data integration**. The first database comprised 668 motorway segments ranging from 200 to 600 meters in length and was infrastructure-based. It included data on historical road crashes, traffic volumes and geometric characteristics. Subsequently, driver behaviour metrics derived from smartphone sensors had to be assigned to the examined road segments. This involved allocating driving behaviour metrics from naturalistic data, which are driver-based, to the examined motorway segments, which are infrastructure-based data. This allocation was achieved via isolating each trip portion to the corresponding segment within the internal recording of trips conducted in GIS by the smartphone data providers using ESRI polygons at 200m intervals.

For the broader urban and interurban network of the Eastern Macedonia and Thrace Region, which exclusively comprised infrastructure and driver behavior data, **a series of processing algorithms** were applied. Initially, a database was created for the considered road network, encompassing 6,103 road segments. This database contained key geometric characteristics such as length, curvature, road type, etc., for each segment. The data extraction from OSM and database creation involved exploiting R libraries specifically designed for these tasks. Next, the naturalistic driver behavior data, extracted from smartphone sensors and covering indicators like harsh braking events, speeding, distraction due to mobile phone use, etc., for every second of trips made in 2021 in the study area, had to be assigned to the corresponding road segments. This assignment was achieved through a spatial map-matching procedure. Initially, the centroid of each road segment line-string was identified using the

“st_centroid” function from the “sf” R library. It is noted that centroids are point-type quantities and represent the geometric center of each road segment. Subsequently, the aggregated driving behaviour metrics were assigned to the nearest road segment centroid based on the latitude and longitude coordinates for each trip-second. This process was executed using the "st_join" function and the "st_nearest_feature" geometry predicate function from the "sf" R library.

Overall, the algorithms utilized in this doctoral dissertation, especially for the broader urban and interurban road network, facilitate the **seamless transferability** of the methodological and data processing framework employed in this dissertation. With minimal modifications, spatial data frames can be generated for various regions, allowing for analyses using the same or different variables, study periods, and statistical methodologies.

Advanced and Innovative Combination of Modelling Techniques

The wealth of high-resolution multiparametric data and the robustness of data processing and fusion enabled the development of **advanced and innovate modelling techniques**.

Initially, a crash frequency model (NB regression) was developed. This model facilitated the investigation of the influence of various geometric characteristics, traffic attributes, and driver behaviour metrics on road crashes. Subsequently, agglomerative hierarchical clustering was employed to categorize crash risk levels for the analyzed road segments, which were then incorporated as the response variable in several ML classifiers. In addition to utilizing **ML techniques**, the analyses included the computation of **SHAP values**, a recent and potent addition in the field of explainable and interpretable ML. These values provided insights into the influential factors contributing to crash risk. This comprehensive approach enhances the sophistication of the modelling techniques and reinforces the interpretability of their results.

With regard to the broader road network of the Eastern Macedonia and Thrace Region, the analyses incorporated harsh braking events as the dependent variables for the developed models. Notably, the modelling techniques employed in this doctoral dissertation are, to the best of the author's knowledge, being **applied for the first time to harsh braking events**. Among these innovative modelling approaches are the SEM, SLM, SZINB, and SRF. It is worth emphasizing that the application of the SRF is particularly noteworthy, representing a novel modelling technique applicable not only to harsh braking events but also to various aspects of road safety analyses.

Multi-factor Estimation of Crash Risk on Motorways

Utilizing the high-quality and detailed database developed for the road segments of the motorway, aiming to address the research questions posed in this doctoral dissertation, valuable and innovative conclusions were drawn. Specifically, statistical

correlations from the road crash frequency model revealed a positive and statistically significant relationship between historical road crash data and the number of harsh driving behaviours. This applies to both the number of harsh accelerations and the number of harsh brakings per passed trips within the examined motorway segments. This indicates that these indicators of **harsh driving behaviour can be utilized as SSMs**, either complementing traditional crash frequency models or serving as dependent variables in road crash risk assessment models in areas where either road crash data are unavailable or the available crash data are of low quality.

Additionally, this thesis highlighted an innovative insight, emphasizing that the contribution of harsh brakings, compared to harsh accelerations, is higher in predicting the crash risk level for road segments. This makes **harsh brakings a more suitable SSM indicator** for proactive road safety analyses, enhancing the understanding of road crash risk and providing practical implications for targeted interventions.

Surrogate Estimation of Crash Risk on Urban and Interurban Road Network

The assessment of this dissertation's contributions would be inadequate without recognizing the broader implications of the developed models on the road network of the Eastern Macedonia and Thrace Region. In these models, the dependent variables were represented by the number of harsh braking events, serving as SSMs. The detection of statistically significant and positively correlated **spatial autocorrelation in harsh braking event frequencies** compelled the development of spatial modelling approaches. Pivotal to frequency analyses is the **measurement of exposure**, with this dissertation employing two primary exposure variables for the respective models: road segment length and the number of trips per segment. This research identifies the statistically significant influence of these exposure variables on the number of harsh braking events, quantifying their respective impacts. Additionally, it incorporates various indicators related to road environment and driver behaviour, contributing to a comprehensive assessment of road crash risk.

The creation of **comprehensive road safety maps and heatmaps** illustrating harsh braking events stands as a valuable tool for road management authorities, stakeholders and road users. These visualizations present complex data and model predictions in an easily comprehensible manner, facilitating communication and integration into diverse decision-making processes. Through these maps, the multifaceted efforts of this dissertation in road crash risk assessment are effectively communicated to both the scientific community and the public domain. Overall, SSMs, such as harsh braking events, offer significant potential for monitoring road safety, evaluating and enhancing countermeasures, and expanding road safety data coverage rapidly. In academia, SSM modelling exercises have emerged in recent years. Apart from contributing in that field, this doctoral dissertation demonstrated that with the necessary effort, **SSM-based spatial models can be used in scarcely-studied areas**.

Εκτεταμένη Περίληψη

Αναγνωρίζοντας την οδική ασφάλεια ως κρίσιμο ζήτημα δημόσιας υγείας με σημαντικές κοινωνικές και οικονομικές επιπτώσεις, είναι απαραίτητο να κατανοηθεί η **πολύπλευρη φύση των οδικών ατυχημάτων**. Τα οδικά ατυχήματα επηρεάζονται από διάφορες παραμέτρους που μπορούν να χωριστούν σε τρεις διακριτές κατηγορίες: (i) χρήστες της οδού, (ii) οχήματα και (iii) οδική υποδομή και οδικό περιβάλλον. Αξίζει να σημειωθεί ότι ένα σημαντικό ποσοστό των οδικών ατυχημάτων, έως και 94%, μπορεί να αποδοθεί, είτε αποκλειστικά είτε εν μέρει, στον ανθρώπινο παράγοντα και σε ανθρώπινα λάθη.

Λαμβάνοντας υπόψη το προαναφερθέν πλαίσιο, ο κύριος στόχος της παρούσας διδακτορικής διατριβής είναι η **αξιολόγηση του κινδύνου οδικού ατυχήματος συνδυάζοντας δεδομένα οδικής υποδομής, κυκλοφορίας και συμπεριφοράς του οδηγού**. Αυτός ο συνδυασμός των δεδομένων αποτελεί μια πολλά υποσχόμενη κατεύθυνση για έρευνα. Ωστόσο, η πρακτική εφαρμογή αυτού του συνδυασμού δεδομένων παρεμποδίζεται συχνά από δυσκολίες και προκλήσεις όπως η ανεπαρκής διαθεσιμότητα δεδομένων ή η μη βέλτιστη ποιότητά τους.

Στο πλαίσιο της παρούσας διατριβής, διεξήχθη εκτενής βιβλιογραφική ανασκόπηση. Σκοπός αυτής της διαδικασίας ήταν να παράσχει μια ανασκόπηση της επιστημονικής βιβλιογραφίας των μελετών που αξιοποιούν τους Έμμεσους Δείκτες Ασφαλείας (ΕΔΑ) σε διερευνήσεις ιστορικών οδικών ατυχημάτων. Οι ΕΔΑ περιλαμβάνουν ένα ευρύ φάσμα μετρήσεων και παραμέτρων, οι οποίες δεν προκύπτουν άμεσα από δεδομένα οδικών ατυχημάτων ή δεν βασίζονται σε αυτά. Από τη διαδικασία της βιβλιογραφικής ανασκόπησης προέκυψε το συμπέρασμα ότι **οι ΕΔΑ κερδίζουν συνεχώς έδαφος στην έρευνα για την οδική ασφάλεια**, καθώς αποτελούν έναν βιώσιμο τρόπο μέτρησης της οδικής ασφάλειας και επιτρέπουν τη διεξαγωγή αναλύσεων χωρίς να χρειάζονται απαραίτητα ιστορικά δεδομένα οδικών ατυχημάτων. Οι δείκτες αυτοί μπορούν είτε να αποτελέσουν εναλλακτική λύση για τις αναλύσεις οδικής ασφάλειας είτε ακόμη και να συμπληρώσουν τις αναλύσεις που βασίζονται σε ιστορικά δεδομένα ατυχημάτων. Επιπλέον, η ταχεία και συνεχής πρόοδος στον τομέα της τεχνολογίας καθιστά όλο και πιο εύκολη τη συλλογή τέτοιων δεικτών. Οι ΕΔΑ, όπως ο χρόνος για σύγκρουση με το προπορευόμενο όχημα, η απότομη επιβράδυνση κτλ., προτείνονται ευρέως στην επιστήμη των μεταφορών και είναι ιδιαίτερα χρήσιμοι προκειμένου να αξιολογηθεί ο κίνδυνος οδικών ατυχημάτων.

Στη συνέχεια, διατυπώθηκαν τα ακόλουθα **ερευνητικά ερωτήματα**:

Ερώτημα 1
Πώς μπορούν να συνδυαστούν και να αναλυθούν τα δεδομένα υποδομής, κυκλοφορίας και συμπεριφοράς των οδηγών ώστε να εξαχθούν χρήσιμα συμπεράσματα για την αξιολόγηση του κινδύνου οδικού ατυχήματος;

Ερώτημα 2

α) Μπορούν τα συμβάντα απότομης συμπεριφοράς του οδηγού να θεωρηθούν αξιόπιστοι ΕΔΑ;

β) Υπάρχει στατιστικά σημαντική και θετική συσχέτιση μεταξύ συμβάντων απότομης συμπεριφοράς του οδηγού και ιστορικών δεδομένων οδικών ατυχημάτων;

Ερώτημα 3

Είναι δυνατή η πρόβλεψη της επικινδυνότητας οδικών τμημάτων με την αξιοποίηση των γεωμετρικών χαρακτηριστικών της οδού και των ΕΔΑ που βασίζονται στη συμπεριφορά του οδηγού, και, αν ναι, ποιοι ταξινομητές μηχανικής μάθησης είναι οι καταλληλότεροι;

Ερώτημα 4

Είναι τα συμβάντα απότομων επιβραδύνσεων πιο σημαντικά από εκείνα των απότομων επιταχύνσεων για την πρόβλεψη της κατηγορίας επικινδυνότητας των οδικών τμημάτων;

Ερώτημα 5

α) Ελλείψει ιδιαίτερα λεπτομερών ιστορικών δεδομένων οδικών ατυχημάτων, πώς μπορούν να αναλυθούν οι απότομες επιβραδύνσεις σε διάφορα οδικά περιβάλλοντα;

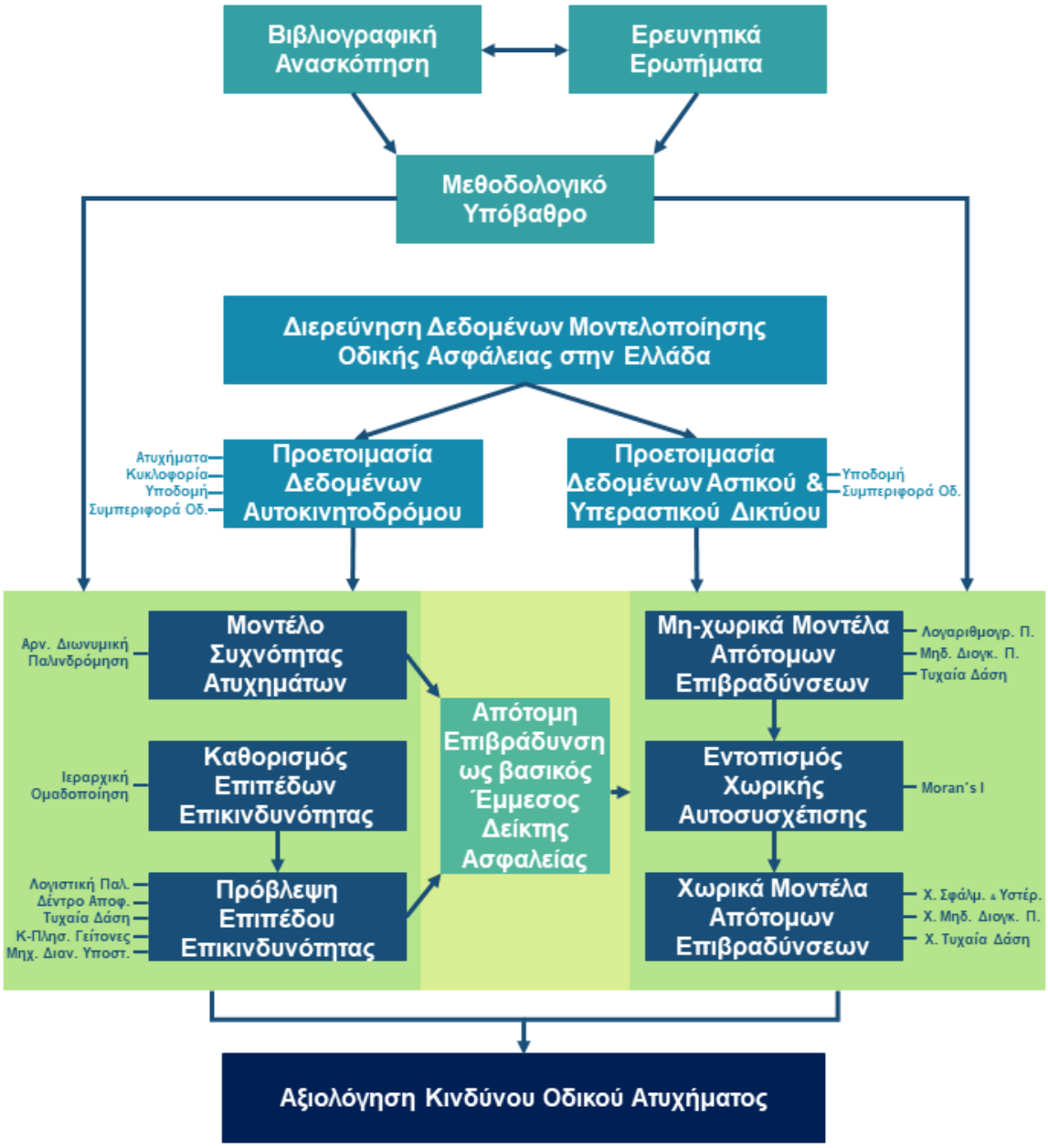
β) Υπάρχει χωρική αυτοσυσχέτιση στις συχνότητες απότομων επιβραδύνσεων για τα οδικά τμήματα και, αν ναι, οι προσεγγίσεις χωρικής μοντελοποίησης υπερτερούν έναντι των αντίστοιχων μη χωρικών προσεγγίσεων;

Ερώτημα 6

Ποιες παράμετροι της οδικής υποδομής και της συμπεριφοράς του οδηγού παρουσιάζουν στατιστικά σημαντική επιρροή στον αριθμό των απότομων επιβραδύνσεων ανά οδικό τμήμα;

Αυτά τα ερευνητικά ερωτήματα αποτέλεσαν την κινητήρια δύναμη πίσω από την παρούσα ερευνητική προσπάθεια, διερευνώντας τον συνδυασμό και την ανάλυση των δεδομένων υποδομής, κυκλοφορίας και συμπεριφοράς των οδηγών για την εξαγωγή ουσιαστικών συμπερασμάτων στην εκτίμηση του κινδύνου οδικών ατυχημάτων.

Προκειμένου να απαντηθούν αυτά τα ερευνητικά ερωτήματα, σχεδιάστηκε ένα σύνθετο **μεθοδολογικό πλαίσιο**, το οποίο παρουσιάζεται στο Σχήμα Ι.



Σχήμα Ι: Γραφική αναπαράσταση του γενικού μεθοδολογικού πλαισίου της διδακτορικής διατριβής

Ο πυρήνας του μεθοδολογικού πλαισίου περιλάμβανε μια διαδικασία πολλών σταδίων, η οποία ξεκίνησε με τη **διερεύνηση των διαθέσιμων δεδομένων μοντελοποίησης της οδικής ασφάλειας στην Ελλάδα**, θέτοντας τις βάσεις για τις επόμενες κατευθύνσεις. Η διερεύνηση αυτή ανέδειξε τους περιορισμούς που

σχετίζονται με την ανάπτυξη λεπτομερών μοντέλων πρόβλεψης ατυχημάτων στην Ελλάδα. Η ανάπτυξη τέτοιων μοντέλων είναι εφικτή μόνο για τους αυτοκινητοδρόμους καθώς για αυτούς υπάρχουν υψηλής ποιότητας διαθέσιμα δεδομένα ατυχημάτων, ειδικά όσον αφορά στην ακριβή θέση των ατυχημάτων και τα χαρακτηριστικά της κυκλοφορίας ανά οδικό τμήμα. Για να αντιμετωπιστεί αυτός ο περιορισμός, αναπτύχθηκαν δύο διαφορετικές βάσεις δεδομένων.

Η πρώτη βάση δεδομένων επικεντρώθηκε σε **668 οδικά τμήματα** του αυτοκινητοδρόμου της Ολυμπίας Οδού, για τα οποία υπήρχαν διαθέσιμα δεδομένα σχετικά με τα οδικά ατυχήματα, την κυκλοφορία, τα γεωμετρικά χαρακτηριστικά και διάφορους δείκτες συμπεριφοράς των οδηγών. Συγκεκριμένα, αξιοποιήθηκαν δεδομένα ατυχημάτων όλων των επιπέδων σοβαρότητας, συμπεριλαμβανομένων των ατυχημάτων με υλικές ζημιές μόνο, για τα έτη 2018-2020. Παράλληλα με τα δεδομένα οδικών ατυχημάτων, στη βάση δεδομένων που αναπτύχθηκε συμπεριλήφθηκαν δεδομένα Ετήσιας Μέσης Ημερήσιας Κυκλοφορίας (ΕΜΗΚ) για την ίδια χρονική περίοδο. Όσον αφορά τα χαρακτηριστικά της οδικής υποδομής, συνδυάστηκαν πληροφορίες από διάφορες πηγές, όπως δεδομένα από τον φορέα διαχείρισης του αυτοκινητοδρόμου και δεδομένα που προήλθαν από τη χρήση διαφόρων λογισμικών, συμπεριλαμβανομένων των Open GIS, Google Earth και GoogleStreetView. Η συμπερίληψη των δεδομένων οδικής υποδομής και των σχεδίων αναφοράς του αυτοκινητόδρομου επέτρεψε επίσης τον εντοπισμό και την απομόνωση των δεδομένων συμπεριφοράς του οδηγού υπό πραγματικές συνθήκες μέσω μιας εφαρμογής για έξυπνα κινητά τηλέφωνα. Τα δεδομένα συμπεριφοράς των οδηγών συλλέχθηκαν για την περίοδο από την 1η Ιουνίου 2019 έως τις 31 Δεκεμβρίου 2020, από δείγμα 327 οδηγών το 2019 και 330 οδηγών το 2020. Ο μέσος αριθμός διαδρομών ανά τμήμα αυτοκινητόδρομου καθ' όλη τη διάρκεια της περιόδου μελέτης ήταν 769 διαδρομές.

Η δεύτερη βάση δεδομένων κάλυψε ένα **ευρύτερο οδικό δίκτυο εντός της Περιφέρειας Ανατολικής Μακεδονίας και Θράκης**, συμπεριλαμβανομένων τόσο αστικών όσο και υπεραστικών οδών. Για το εξεταζόμενο δίκτυο, πραγματοποιήθηκε μια αρχική ανάλυση όλων των οδικών τμημάτων που προήλθαν από το OpenStreetMap (OSM) για την εξαγωγή των γεωμετρικών χαρακτηριστικών τους. Στη συνέχεια, τα δεδομένα συμπεριφοράς των οδηγών υπό πραγματικές συνθήκες, τα οποία εξήχθησαν από εφαρμογή για έξυπνα κινητά τηλέφωνα, αντιστοιχήθηκαν με τα αντίστοιχα οδικά τμήματα του OSM. Το εξεταζόμενο οδικό δίκτυο περιλάμβανε **6.103 οδικά τμήματα**, με μέσο μήκος 288,8 μέτρα, με αποτέλεσμα το συνολικό μήκος του οδικού δικτύου να ανέρχεται σε 1.763 χιλιόμετρα. Όσον αφορά τις μετρήσεις της συμπεριφοράς του οδηγού, χρησιμοποιήθηκαν δεδομένα από 5.129 ταξίδια κατά τη διάρκεια του 2021. Η μέση διάρκεια ταξιδιού ήταν 634 δευτερόλεπτα, με τυπική απόκλιση 556 δευτερόλεπτα. Ωστόσο, επισημαίνεται ότι η βάση δεδομένων που αναπτύχθηκε για το εν λόγω οδικό δίκτυο δεν περιείχε λεπτομερή δεδομένα ατυχημάτων και κυκλοφορίας για τα εξεταζόμενα οδικά τμήματα.

Εφαρμόστηκαν διάφορα **μεθοδολογικά εργαλεία** για τα οδικά τμήματα του αυτοκινητόδρομου της Ολυμπίας Οδού. Σε αυτά περιλαμβάνονταν τεχνικές όπως η αρνητική διωνυμική παλινδρόμηση (ΑΔΠ) για την ανάπτυξη ενός μοντέλου συχνότητας ατυχημάτων, η ιεραρχική ομαδοποίηση (ΙΟ) για τον προσδιορισμό των επιπέδων επικινδυνότητας των τμημάτων με βάση ιστορικά δεδομένα ατυχημάτων και χαρακτηριστικά της κυκλοφορίας, και η χρήση ταξινομητών μηχανικής μάθησης όπως η Λογιστική Παλινδρόμηση (ΛΠ), το Δέντρο Αποφάσεων (ΔΑ), τα Τυχαία Δάση (ΤΔ), οι Κ-Πλησιέστεροι Γείτονες (Κ-ΠΓ) και οι Μηχανές Διανυσμάτων Υποστήριξης (ΜΔΥ). Αυτοί οι ταξινομητές χρησιμοποιήθηκαν για την πρόβλεψη του επιπέδου επικινδυνότητας των τμημάτων, αξιοποιώντας δεδομένα υποδομής και συμπεριφοράς των οδηγών. Ιδιαίτερη έμφαση δόθηκε στην αξιολόγηση της αξιοπιστίας των συμβάντων απότομης συμπεριφοράς του οδηγού ως ΕΔΑ.

Στη συνέχεια, το πλαίσιο επεκτάθηκε για να συμπεριλάβει τα δεδομένα του οδικού δικτύου της Περιφέρειας Ανατολικής Μακεδονίας και Θράκης, αξιοποιώντας συμβάντα απότομων επιβραδύνσεων για την αξιολόγηση του κινδύνου οδικών ατυχημάτων. Αυτό περιλάμβανε την εφαρμογή τόσο μη χωρικών όσο και **χωρικών μοντέλων** για τον εντοπισμό σημαντικών παραμέτρων οδικής υποδομής και συμπεριφοράς του οδηγού που επηρεάζουν τον αριθμό των απότομων επιβραδύνσεων ανά οδικό τμήμα.

Τελικώς, η σύνθεση όλων των αναλύσεων που πραγματοποιήθηκαν στο πλαίσιο της παρούσας διδακτορικής διατριβής οδήγησε σε μια **ολοκληρωμένη αξιολόγηση του κινδύνου οδικών ατυχημάτων** με πολλά πρωτότυπα και ενδιαφέροντα αποτελέσματα, τα οποία αναλύονται με περισσότερη λεπτομέρεια παρακάτω.

Για τις αναλύσεις του αυτοκινητοδρόμου, αξιοποιήθηκε μια ενοποιημένη βάση δεδομένων που περιλάμβανε δεδομένα για ιστορικά οδικά ατυχήματα με τραυματισμούς και ατυχήματα με υλικές ζημιές, χαρακτηριστικά κυκλοφορίας, χαρακτηριστικά γεωμετρίας της οδού και ΕΔΑ συμπεριφοράς οδηγού για 668 οδικά τμήματα του αυτοκινητοδρόμου της Ολυμπίας Οδού. Τα αποτελέσματα του μοντέλου συχνότητας ατυχημάτων (ΑΔΠ) έδειξαν ότι η συχνότητα οδικών ατυχημάτων στα εξεταζόμενα τμήματα του αυτοκινητοδρόμου συσχετίζεται θετικά με τον κυκλοφοριακό φόρτο, το μήκος του εξεταζόμενου οδικού τμήματος, τον αριθμό των απότομων επιταχύνσεων και τον αριθμό των απότομων επιβραδύνσεων ανά διερχόμενες διαδρομές του κάθε τμήματος. Το εύρημα αυτό συμβάλλει στην υπάρχουσα βιβλιογραφία για την οδική ασφάλεια, καθώς διαπιστώνει **θετική και στατιστικά σημαντική σχέση μεταξύ της συχνότητας οδικών ατυχημάτων και των συμβάντων απότομης συμπεριφοράς του οδηγού**. Κατά συνέπεια, συνάγεται ότι αυτά τα συμβάντα μπορούν να χρησιμεύσουν ως έγκυρη υποκατηγορία των ΕΔΑ υπό πραγματικές συνθήκες οδήγησης. Συγκεκριμένα, μπορούν να χρησιμοποιηθούν είτε για τη συμπλήρωση των μοντέλων πρόβλεψης ατυχημάτων (ΜΠΑ) είτε ως εξαρτημένες μεταβλητές σε διάφορες προληπτικές αναλύσεις οδικής ασφάλειας, ιδίως

σε περιπτώσεις όπου δεν υπάρχουν λεπτομερή ιστορικά δεδομένα οδικών ατυχημάτων.

Ως περαιτέρω στάδιο των ερευνών για τον αυτοκινητόδρομο, έγινε προσπάθεια να διαμορφωθούν **επίπεδα επικινδυνότητας** των εξεταζόμενων οδικών τμημάτων. Αυτό επιτεύχθηκε λαμβάνοντας υπόψη τον αριθμό των οδικών ατυχημάτων ανά μήκος τμήματος και την κυκλοφορία κάθε τμήματος με τη χρήση της τεχνικής της συσσωρευτικής ιεραρχικής ομαδοποίησης. Λαμβάνοντας υπόψη την επιρροή του μήκους του οδικού τμήματος και του κυκλοφοριακού φόρτου, όπως προκύπτει από τα αποτελέσματα του μοντέλου ΑΔΠ, και οι δύο αυτές μεταβλητές συμπεριλήφθηκαν στην ανάλυση ομαδοποίησης λόγω της στατιστικά σημαντικής επίδρασής τους στη συχνότητα ατυχημάτων στα εξεταζόμενα οδικά τμήματα. Τα αποτελέσματα αυτής της διαδικασίας ομαδοποίησης καθόρισαν τέσσερα διακριτά επίπεδα επικινδυνότητας με ένα σαφές μοτίβο, σύμφωνα με το οποίο η πρώτη κατηγορία επικινδυνότητας παρουσιάζει υψηλό μέσο κυκλοφοριακό φόρτο και αριθμό οδικών ατυχημάτων ανά μήκος τμήματος, ενώ τα μεγέθη αυτά μειώνονται προοδευτικά για κάθε επόμενη κατηγορία επικινδυνότητας.

Στη συνέχεια, τα τέσσερα επίπεδα επικινδυνότητας χρησιμοποιήθηκαν ως εξαρτημένη μεταβλητή/ μεταβλητή απόκρισης σε πέντε μοντέλα ταξινόμησης μηχανικής μάθησης (ΛΠ, ΔΑ, ΤΔ, ΜΔΥ και Κ-ΠΓ). Οι ταξινομητές αυτοί, περιλάμβαναν προγνωστικούς παράγοντες σχετικά με τα γεωμετρικά χαρακτηριστικά της οδού και μη ασφαλείς συμπεριφορές οδήγησης, όπως δείκτες απότομων επιβραδύνσεων, απότομων επιταχύνσεων και διάρκεια υπέρβασης των ορίων ταχύτητας ανά διαδρομή εντός των εξεταζόμενων οδικών τμημάτων. Μεταξύ των πέντε μοντέλων, το **μοντέλο ΤΔ επέδειξε ανώτερες επιδόσεις ταξινόμησης** σε όλες τις κατηγορίες επικινδυνότητας, επιτυγχάνοντας σταθερά βαθμολογίες άνω του 89% (συνολική ορθότητα: 89,9%, μακρο-μεσοσταθμική ακρίβεια: 90,7%, μακρο-μεσοσταθμική ανάκληση: 89,9%, μακρο-μεσοσταθμική βαθμολογία F1: 90,2%). Το αποτέλεσμα αυτό αποκαλύπτει τις δυνατότητες του μοντέλου ΤΔ που αναπτύχθηκε ως ένα πολλά υποσχόμενο προληπτικό εργαλείο οδικής ασφάλειας, ικανό να εντοπίζει αποτελεσματικά και να ιεραρχεί δυνητικά επικίνδυνα τμήματα αυτοκινητοδρόμων.

Τέλος, για να διευκολυνθεί η ερμηνεία του μοντέλου ΤΔ, το οποίο λειτουργεί εγγενώς ως μοντέλο-μαύρο κουτί μηχανικής μάθησης, υπολογίστηκαν οι τιμές SHAP για ένα τυπικό τμήμα αυτοκινητόδρομου. Με βάση τις τιμές SHAP των προβλεπτικών παραγόντων συμπεριφοράς οδηγού, προέκυψε ότι οι **απότομες επιβραδύνσεις χρησιμεύουν ως πιο κατάλληλος ΕΔΑ από τις απότομες επιταχύνσεις** όσον αφορά την πρόβλεψη της επικινδυνότητας των οδικών τμημάτων.

Στο πλαίσιο του ευρύτερου οδικού δικτύου της Περιφέρειας Ανατολικής Μακεδονίας και Θράκης, αναλύθηκε ένα σύνολο χωρικών δεδομένων που αποτελείται από συγκεντρωτικούς δείκτες συμπεριφοράς του οδηγού υπό πραγματικές συνθήκες,

καθώς και γεωμετρικά χαρακτηριστικά και λοιπά χαρακτηριστικά του δικτύου σε επίπεδο οδικού τμήματος. Για τα εξεταζόμενα 6.103 οδικά τμήματα, και με βάση τον δείκτη Moran's I, εντοπίστηκε στατιστικά σημαντική και θετική **χωρική αυτοσυσχέτιση στις συχνότητες των απότομων επιβραδύνσεων**. Αρχικά, χρησιμοποιήθηκαν μη χωρικές τεχνικές μοντελοποίησης, όπως το λογαριθμογραμμικό μοντέλο, η Μηδενικά Διογκωμένη Αρνητική Διωνυμική (ΜΔΑΔ) παλινδρόμηση και το συμβατικό μοντέλο παλινδρόμησης ΤΔ στις συχνότητες των απότομων επιβραδύνσεων. Ωστόσο, η ύπαρξη χωρικής αυτοσυσχέτισης ανέδειξε την ανάγκη ανάπτυξης χωρικών μοντέλων, όπως το Χωρικό Μοντέλο διόρθωσης του Σφάλματος (ΧΜΣ), το Χωρικό Μοντέλο Υστέρησης (ΧΜΥ), το Χωρικό μοντέλο Μηδενικά Διογκωμένης Αρνητικής Διωνυμικής παλινδρόμησης (ΧΜΔΑΔ) και το Χωρικό μοντέλο Τυχαίων Δασών (ΧΤΔ), προκειμένου να ληφθούν υπόψη αυτές οι χωρικές εξαρτήσεις.

Σε όλα τα μοντέλα που αναπτύχθηκαν προέκυψαν **σταθερά πρόσημα στους συντελεστές των μεταβλητών**. Συγκεκριμένα, το μήκος του οδικού τμήματος και ο αριθμός των διαδρομών ανά τμήμα προσδιορίστηκαν ως υποκατάστατοι δείκτες της έκθεσης σε κίνδυνο, οι οποίοι συσχετίζονται θετικά με τις απότομες επιβραδύνσεις. Επιπλέον, ο δείκτης αποτελεσματικότητας (στατιστικά σημαντικός μόνο στο λογαριθμογραμμικό μοντέλο, στο ΜΧΣ και στο ΜΧΥ), που σχετίζεται με τη γραμμικότητα των οδικών τμημάτων, παρουσίασε θετική συσχέτιση με τα συμβάντα απότομων επιβραδύνσεων, υποδηλώνοντας ότι οι οδηγοί προβαίνουν συχνότερα σε απότομες επιβραδύνσεις σε οδικά τμήματα με λιγότερες καμπύλες. Οι μεταβλητές που σχετίζονται με την υπέρβαση των ορίων ταχύτητας και τη χρήση κινητού τηλεφώνου συσχετίστηκαν επίσης θετικά με τις απότομες επιβραδύνσεις, ενώ οι αυτοκινητόδρομοι παρουσίασαν λιγότερα συμβάντα απότομων επιβραδύνσεων σε σύγκριση με άλλους τύπους οδού.

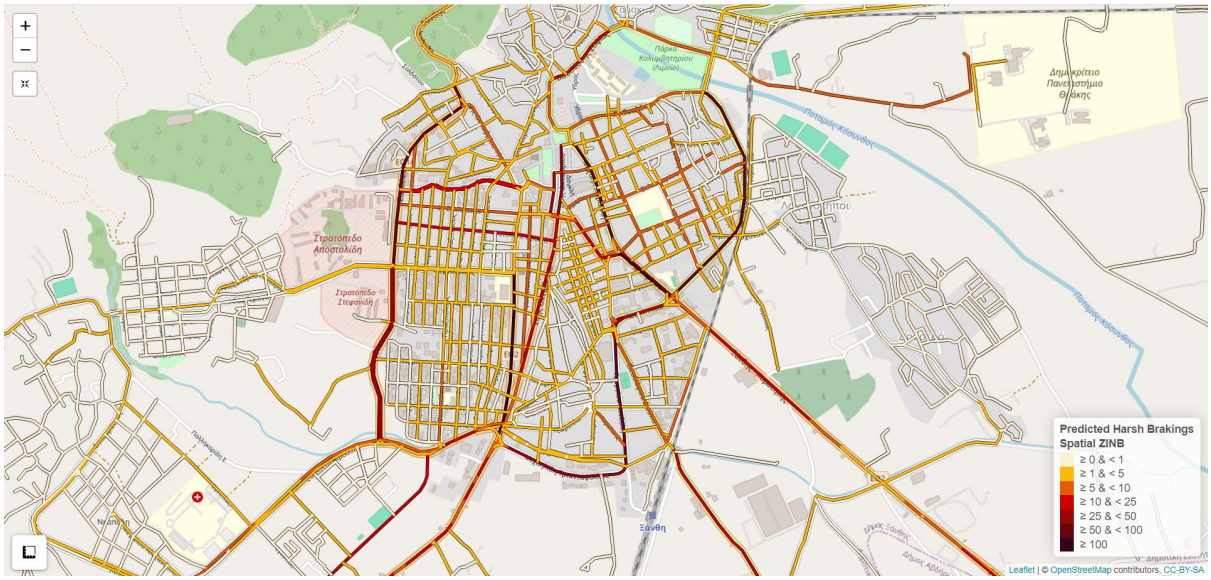
Και στα δύο μοντέλα ΤΔ, **ο αριθμός των διαδρομών ανά εξεταζόμενο οδικό τμήμα βρέθηκε να είναι ο πιο σημαντικός παράγοντας πρόβλεψης**, αναδεικνύοντας την υψηλή σημασία του στην πρόβλεψη της συχνότητας των απότομων επιβραδύνσεων, καθώς χρησιμεύει ως μέτρο έκθεσης στον κίνδυνο. Από την άλλη πλευρά, η μεταβλητή «αυτοκινητόδρομος» παρουσίασε τη χαμηλότερη επιρροή, υποδεικνύοντας ότι ο τύπος της οδού είναι σχετικά λιγότερο κρίσιμος για την πρόβλεψη του αριθμού των απότομων επιβραδύνσεων. Το εύρημα αυτό μπορεί να υποδηλώνει ότι άλλοι παράγοντες πέραν του τύπου της οδού, όπως η απόσπαση της προσοχής του οδηγού και η υπέρβαση των ορίων ταχύτητας, ενδεχομένως να κατέχουν σημαντικότερο ρόλο στην επιρροή της συχνότητας των απότομων επιβραδύνσεων.

Όσον αφορά την επίδοση των μοντέλων που αναπτύχθηκαν, το **ΧΜΥ ξεπέρασε** τόσο το λογαριθμογραμμικό μοντέλο όσο και το ΧΜΣ, με χαμηλότερες τιμές του δείκτη AIC και απουσία χωρικής αυτοσυσχέτισης στα κατάλοιπά του. Χαμηλότερες τιμές του δείκτη AIC, που υποδηλώνουν καλύτερη προσαρμογή, παρατηρήθηκαν επίσης για το μοντέλο ΧΜΔΑΔ σε σύγκριση με το μη χωρικό μοντέλο ΜΔΑΔ. Επιπλέον, τ**ο ΧΤΔ**

μείωσε τις απόλυτες τιμές της χωρικής αυτοσυσχέτισης στα κατάλοιπα σε σύγκριση με τις αντίστοιχες τιμές του συμβατικού μοντέλου ΤΔ. Επιπλέον, το ΧΤΔ υπερείχε του μη χωρικού μοντέλου ΤΔ όσον αφορά την προσαρμογή του μοντέλου στα παρατηρούμενα δεδομένα, αλλά το μη χωρικό μοντέλο είχε καλύτερες επιδόσεις όσον αφορά τη γενίκευση σε μη παρατηρούμενα δεδομένα.

Τα αποτελέσματα των μοντέλων που αναπτύχθηκαν για το εξεταζόμενο οδικό δίκτυο της Περιφέρειας Ανατολικής Μακεδονίας και Θράκης **απεικονίζονται επίσης σε χάρτες**. Ενδεικτικά, τα αποτελέσματα του μοντέλου ΧΜΔΑΔ παρουσιάζονται στο Σχήμα ΙΙ, ενώ το Σχήμα ΙΙΙ παρέχει μια μεγεθυμένη άποψη του Σχήματος ΙΙ, εστιάζοντας συγκεκριμένα στο κέντρο της πόλης της Ξάνθης.

Σχήμα ΙΙ: Απεικόνιση των αποτελεσμάτων του μοντέλου ΧΜΔΑΔ στο εξεταζόμενο οδικό δίκτυο



Σχήμα ΙΙΙ: Μεγεθυμένη άποψη των αποτελεσμάτων του μοντέλου ΧΜΔΑΔ για το κέντρο της Ξάνθης

Η παρούσα διδακτορική διατριβή προσφέρει **αξιοσημείωτες και καινοτόμες συνεισφορές** στον τομέα της οδικής ασφάλειας. Οι συνεισφορές αυτές παρουσιάζονται με περισσότερη λεπτομέρεια παρακάτω.

Ολιστική Προσέγγιση Συλλογής Δεδομένων

Στο πλαίσιο της παρούσας διδακτορικής διατριβής, πραγματοποιήθηκε μια **εκτενής συλλογή δεδομένων** για τη διερεύνηση της επιρροής της συμπεριφοράς του οδηγού, των χαρακτηριστικών της οδικής υποδομής και των χαρακτηριστικών της κυκλοφορίας στην αξιολόγηση του κινδύνου οδικών ατυχημάτων. Οι τεχνολογικές εξελίξεις έχουν διευκολύνει σημαντικά τη συλλογή δεδομένων από διάφορες πηγές, δημιουργώντας νέες ερευνητικές ευκαιρίες που προηγουμένως δεν είχαν διερευνηθεί.

Συγκεκριμένα, στην παρούσα διατριβή αξιοποιήθηκαν **βάσεις δεδομένων ευρείας κλίμακας με στοιχεία υψηλής ευκρίνειας** για την οδήγηση υπό πραγματικές συνθήκες που συλλέχθηκαν από αισθητήρες έξυπνων κινητών τηλεφώνων για την αξιολόγηση του κινδύνου οδικών ατυχημάτων σε αυτοκινητόδρομους και σε ένα ευρύτερο οδικό δίκτυο, που περιλαμβάνει τόσο αστικές όσο και υπεραστικές οδούς. Για τα δεδομένα οδικών υποδομών στον εξεταζόμενο αυτοκινητόδρομο, αξιοποιήθηκαν διάφορες πηγές, συμπεριλαμβανομένων δεδομένων που παρέχονται από την αρχή διαχείρισης και λειτουργίας του αυτοκινητοδρόμου και λογισμικών όπως το Open GIS, το Google Earth και το GoogleStreetView. Τα γεωμετρικά χαρακτηριστικά και τα χαρακτηριστικά του δικτύου για το ευρύτερο οδικό δίκτυο της Περιφέρειας Ανατολικής Μακεδονίας και Θράκης προέκυψαν με τη χρήση αλγορίθμων στη γλώσσα προγραμματισμού R. Συγκεκριμένα, χρησιμοποιήθηκαν κατάλληλες βιβλιοθήκες για την εξαγωγή δεδομένων από το OSM και την επεξεργασία τους ως απλά χωρικά στοιχεία. Όσον αφορά τα δεδομένα οδικών ατυχημάτων και κυκλοφορίας στον εξεταζόμενο αυτοκινητόδρομο, αξιοποιήθηκαν δεδομένα υψηλής ποιότητας που παραχωρήθηκαν από τον φορέα διαχείρισης και λειτουργίας της οδού. Αυτά περιλάμβαναν δεδομένα οδικών ατυχημάτων όλων των βαθμών σοβαρότητας, συμπεριλαμβανομένων των ατυχημάτων μόνο με υλικές ζημιές, με ακρίβεια στη θέση των ατυχημάτων, που καλύπτουν την περίοδο από το 2018 έως το 2020. Επιπλέον, χρησιμοποιήθηκαν δεδομένα ΕΜΗΚ που προέκυψαν από τους σταθμούς διοδίων του αυτοκινητοδρόμου για την αντίστοιχη χρονική περίοδο.

Πολυδιάστατος Συνδυασμός Δεδομένων για Αναλύσεις σε Επίπεδο Οδικού Τμήματος

Η συλλογή δεδομένων από διάφορες πηγές και σε διαφορετικά επίπεδα απαιτεί **κατάλληλη επεξεργασία για την ενοποίηση των δεδομένων**. Η πρώτη βάση δεδομένων ήταν σε επίπεδο οδικού τμήματος και περιλάμβανε 668 τμήματα αυτοκινητοδρόμου μήκους από 200 έως 600 μέτρα. Συγκεκριμένα, περιλάμβανε δεδομένα σχετικά με οδικά ατυχήματα, κυκλοφοριακούς φόρτους και γεωμετρικά χαρακτηριστικά. Στη συνέχεια, έπρεπε να αντιστοιχηθούν στα εξεταζόμενα οδικά τμήματα δείκτες συμπεριφοράς των οδηγών που προέκυψαν από αισθητήρες έξυπνων κινητών τηλεφώνων. Η αντιστοίχιση αυτή επιτεύχθηκε μέσω GIS και της

απομόνωσης κάθε μέρους των διαδρομών στο αντίστοιχο οδικό τμήμα από την εταιρεία που παρείχε τα δεδομένα με τη χρήση πολυγώνων ESRI σε διαστήματα των 200 μέτρων.

Για το ευρύτερο αστικό και υπεραστικό οδικό δίκτυο της Περιφέρειας Ανατολικής Μακεδονίας και Θράκης, το οποίο περιλάμβανε δεδομένα υποδομής και συμπεριφοράς των οδηγών, εφαρμόστηκε μια **σειρά αλγορίθμων επεξεργασίας**. Αρχικά, δημιουργήθηκε μια βάση δεδομένων για το εξεταζόμενο οδικό δίκτυο, η οποία περιλάμβανε 6.103 οδικά τμήματα. Αυτή η βάση δεδομένων περιείχε βασικά γεωμετρικά χαρακτηριστικά, όπως μήκος, καμπυλότητα, τύπος οδού κ.λπ. για κάθε τμήμα. Η εξαγωγή δεδομένων από το OSM και η δημιουργία της βάσης δεδομένων περιλάμβανε την αξιοποίηση βιβλιοθηκών της R που έχουν σχεδιαστεί ειδικά για αντίστοιχους σκοπούς. Στη συνέχεια, έπρεπε να αντιστοιχηθούν στα οδικά τμήματα τα δεδομένα συμπεριφοράς των οδηγών υπό πραγματικές συνθήκες οδήγησης, τα οποία εξήχθησαν από αισθητήρες έξυπνων κινητών τηλεφώνων και κάλυπταν δείκτες όπως οι απότομες επιβραδύνσεις, η υπέρβαση ορίων ταχύτητας, η απόσπαση προσοχής λόγω χρήσης κινητού τηλεφώνου κ.λπ. για κάθε δευτερόλεπτο των διαδρομών που πραγματοποιήθηκαν το 2021 στην περιοχή μελέτης. Η διαδικασία αυτή επιτεύχθηκε μέσω χωρικής αντιστοίχισης-χαρτών. Αρχικά, προσδιορίστηκε το κεντροειδές κάθε σειράς-γραμμών των οδικών τμημάτων με τη χρήση της συνάρτησης "st_centroid" από τη βιβλιοθήκη "sf" της γλώσσας προγραμματισμού R. Σημειώνεται ότι τα κεντροειδή είναι σημειακά μεγέθη και αντιπροσωπεύουν το γεωμετρικό κέντρο κάθε οδικού τμήματος. Στη συνέχεια, οι συγκεντρωτικοί δείκτες της συμπεριφοράς των οδηγών αντιστοιχήθηκαν στο πλησιέστερο κεντροειδές του οδικού τμήματος με βάση τις συντεταγμένες γεωγραφικού πλάτους και μήκους για κάθε δευτερόλεπτο διαδρομής. Η διαδικασία αυτή εκτελέστηκε με τη χρήση της συνάρτησης "st_join" και της συνάρτησης "st_nearest_feature" από τη βιβλιοθήκη της R "sf".

Συνολικά, οι αλγόριθμοι που χρησιμοποιήθηκαν στην παρούσα διδακτορική διατριβή, ιδίως για το ευρύτερο αστικό και υπεραστικό οδικό δίκτυο, διευκολύνουν την **απρόσκοπτη δυνατότητα μεταφοράς** του μεθοδολογικού πλαισίου και του πλαισίου επεξεργασίας δεδομένων που χρησιμοποιήθηκε στην παρούσα διατριβή. Με ελάχιστες τροποποιήσεις, μπορούν να δημιουργηθούν χωρικές βάσεις δεδομένων για διάφορες περιοχές, επιτρέποντας αναλύσεις με τη χρήση των ίδιων ή διαφορετικών μεταβλητών, περιόδων μελέτης και στατιστικών μεθοδολογιών.

Συνδυασμός Προηγμένων και Καινοτόμων Τεχνικών Μοντελοποίησης

Ο πλούτος των πολυπαραμετρικών δεδομένων υψηλής ευκρίνειας και η ακρίβεια της επεξεργασίας και του συνδυασμού των δεδομένων επέτρεψαν την ανάπτυξη **προηγμένων και καινοτόμων τεχνικών μοντελοποίησης**.

Αρχικά, αναπτύχθηκε ένα μοντέλο συχνότητας ατυχημάτων (ΑΔΠ). Το μοντέλο αυτό διευκόλυνε τη διερεύνηση της επιρροής διαφόρων γεωμετρικών χαρακτηριστικών,

χαρακτηριστικών της κυκλοφορίας και δεικτών της συμπεριφοράς του οδηγού στα οδικά ατυχήματα. Στη συνέχεια, χρησιμοποιήθηκε η συσσωρευτική ιεραρχική ομαδοποίηση για την κατηγοριοποίηση της επικινδυνότητας των οδικών τμημάτων που αναλύθηκαν, τα οποία στη συνέχεια ενσωματώθηκαν ως μεταβλητή απόκρισης/εξαρτημένη μεταβλητή σε διάφορους ταξινομητές μηχανικής μάθησης. Εκτός από τη χρήση **τεχνικών μηχανικής μάθησης**, οι αναλύσεις περιλάμβαναν τον υπολογισμό των **τιμών SHAP**, μια πρόσφατη και ισχυρή προσθήκη στον τομέα της ερμηνεύσιμης μηχανικής μάθησης. Οι τιμές αυτές παρείχαν πληροφορίες σχετικά με τους παράγοντες επιρροής που συμβάλλουν στο επίπεδο επικινδυνότητας. Αυτή η ολοκληρωμένη προσέγγιση αυξάνει την πολυπλοκότητα των τεχνικών μοντελοποίησης και ενισχύει την ερμηνεία των αποτελεσμάτων τους.

Όσον αφορά το ευρύτερο οδικό δίκτυο της Περιφέρειας Ανατολικής Μακεδονίας και Θράκης, οι απότομες επιβραδύνσεις χρησιμοποιήθηκαν ως εξαρτημένες μεταβλητές στα μοντέλα που αναπτύχθηκαν. Αξίζει να σημειωθεί ότι οι τεχνικές μοντελοποίησης που χρησιμοποιήθηκαν στην παρούσα διδακτορική διατριβή, εξ όσων γνωρίζει ο συγγραφέας, **εφαρμόζονται για πρώτη φορά σε συμβάντα απότομων επιβραδύνσεων**. Μεταξύ αυτών των καινοτόμων προσεγγίσεων μοντελοποίησης είναι τα ΜΧΣ, ΜΧΥ, ΧΜΔΑΔ και ΧΤΔ. Αξίζει να τονιστεί ότι η εφαρμογή του μοντέλου ΧΤΔ είναι ιδιαίτερα αξιοσημείωτη, καθώς αποτελεί μια πρωτότυπη και πολλά υποσχόμενη τεχνική μοντελοποίησης που μπορεί να εφαρμοστεί και σε άλλες αναλύσεις οδικής ασφάλειας πέραν εκείνων των απότομων επιβραδύνσεων.

Πολύ-παραγοντική Εκτίμηση Κινδύνου Ατυχήματος στους Αυτοκινητόδρομους

Αξιοποιώντας την υψηλής ποιότητας και λεπτομερή βάση δεδομένων που αναπτύχθηκε για τα οδικά τμήματα του αυτοκινητοδρόμου, με στόχο την απάντηση των ερευνητικών ερωτημάτων που τέθηκαν στην παρούσα διδακτορική διατριβή, εξήχθησαν πολύτιμα και καινοτόμα συμπεράσματα. Συγκεκριμένα, οι στατιστικές συσχετίσεις από το μοντέλο συχνότητας οδικών ατυχημάτων αποκάλυψαν μια θετική και στατιστικά σημαντική συσχέτιση μεταξύ των ιστορικών δεδομένων οδικών ατυχημάτων και του αριθμού των απότομων συμβάντων οδήγησης. Αυτό ισχύει τόσο για τον αριθμό των απότομων επιταχύνσεων όσο και για τον αριθμό των απότομων επιβραδύνσεων ανά διερχόμενο ταξίδι εντός των εξεταζόμενων τμημάτων. Αυτό το εύρημα φανερώνει ότι **οι δείκτες απότομης συμπεριφοράς του οδηγού μπορούν να αξιοποιηθούν ως ΕΔΑ**, είτε συμπληρώνοντας τα παραδοσιακά μοντέλα συχνότητας ατυχημάτων είτε χρησιμεύοντας ως εξαρτημένες μεταβλητές σε μοντέλα αξιολόγησης του κινδύνου οδικών ατυχημάτων σε περιοχές όπου είτε δεν υπάρχουν διαθέσιμα δεδομένα οδικών ατυχημάτων είτε τα διαθέσιμα δεδομένα ατυχημάτων είναι χαμηλής ποιότητας.

Επιπλέον, η παρούσα διατριβή ανέδειξε μια καινοτόμο διαπίστωση, τονίζοντας ότι η συμβολή των απότομων επιβραδύνσεων, σε σύγκριση με τις απότομες επιταχύνσεις, είναι υψηλότερη στην πρόβλεψη της επικινδυνότητας των οδικών τμημάτων. Αυτό

καθιστά τις **απότομες επιβραδύνσεις καταλληλότερο ΕΔΑ** για προληπτικές αναλύσεις οδικής ασφάλειας, ενισχύοντας την κατανόηση του κινδύνου πρόκλησης οδικών ατυχημάτων και παρέχοντας πρακτικές πληροφορίες για στοχευμένες παρεμβάσεις.

Έμμεση Εκτίμηση Κινδύνου Ατυχήματος στο Αστικό και Υπεραστικό Οδικό Δίκτυο

Η αξιολόγηση της συμβολής αυτής της διατριβής δεν θα ήταν πλήρης χωρίς την αναγνώριση των ευρύτερων συμπερασμάτων των μοντέλων που αναπτύχθηκαν για το ευρύτερο οδικό δίκτυο της Περιφέρειας Ανατολικής Μακεδονίας και Θράκης. Σε αυτά τα μοντέλα, οι εξαρτημένες μεταβλητές αντιπροσωπεύονταν από τον αριθμό των απότομων επιβραδύνσεων, που χρησιμεύουν ως ΕΔΑ. Ο εντοπισμός στατιστικά σημαντικής και θετικής **χωρικής αυτοσυσχέτισης στις συχνότητες των απότομων επιβραδύνσεων** επέβαλε την ανάπτυξη προσεγγίσεων χωρικής μοντελοποίησης. Κομβικό σημείο στις αναλύσεις συχνοτήτων είναι η **μέτρηση της έκθεσης στον κίνδυνο**, με την παρούσα διατριβή να χρησιμοποιεί δύο βασικές μεταβλητές έκθεσης για τα αντίστοιχα μοντέλα: το μήκος οδικού τμήματος και τον αριθμό διαδρομών ανά τμήμα. Η παρούσα έρευνα αναδεικνύει τη στατιστικά σημαντική επίδραση αυτών των μεταβλητών έκθεσης στον κίνδυνο, στον αριθμό των απότομων επιβραδύνσεων, ποσοτικοποιώντας τις αντίστοιχες επιρροές τους. Επιπλέον, ενσωματώνει διάφορους δείκτες που σχετίζονται με το οδικό περιβάλλον και τη συμπεριφορά του οδηγού, συμβάλλοντας σε μια ολοκληρωμένη αξιολόγηση του κινδύνου οδικών ατυχημάτων.

Η δημιουργία **ολοκληρωμένων χαρτών οδικής ασφάλειας** που απεικονίζουν τα συμβάντα απότομων επιβραδύνσεων αποτελεί πολύτιμο εργαλείο τόσο για τις αρχές διαχείρισης της οδικής κυκλοφορίας και τα ενδιαφερόμενους φορείς όσο και τους χρήστες της οδού. Οι οπτικοποιήσεις αυτές, παρουσιάζουν πολύπλοκα δεδομένα και προβλέψεις μοντέλων με απλό και κατανοητό τρόπο, διευκολύνοντας την επικοινωνία και την ενσωμάτωση τους σε διάφορες διαδικασίες λήψης αποφάσεων. Μέσω αυτών των χαρτών, οι πολύπλευρες προσπάθειες της παρούσας διατριβής για την αξιολόγηση του κινδύνου οδικών ατυχημάτων κοινοποιούνται αποτελεσματικά τόσο στην επιστημονική κοινότητα όσο και στο ευρύ κοινό. Συνολικά, οι ΕΔΑ, όπως οι απότομες επιβραδύνσεις, προσφέρουν σημαντικές δυνατότητες για την παρακολούθηση της οδικής ασφάλειας, την αξιολόγηση και την ενίσχυση των μέτρων και την ταχεία επέκταση της κάλυψης δεδομένων οδικής ασφάλειας. Στην ακαδημαϊκή κοινότητα, τα τελευταία χρόνια έχουν εμφανιστεί διάφορες προσπάθειες μοντελοποίησης των ΕΔΑ. Εκτός από τη συμβολή στον τομέα αυτό, η παρούσα διδακτορική διατριβή κατέδειξε ότι με την απαιτούμενη προσπάθεια, χωρικά μοντέλα με βάση τους ΕΔΑ μπορούν να χρησιμοποιηθούν σε περιοχές που έχουν μελετηθεί ελάχιστα από άποψη οδικής ασφάλειας.

1. Introduction

1.1 Road Safety Overview

1.1.1 Road Safety Globally

Road safety has been recognized as one of the most important public health issues, bearing an immense societal and economic burden. Despite significant efforts in recent years, road safety remains a substantial global challenge. Road crashes, leading to injuries and fatalities, rank as the 12th leading cause of death across all age groups worldwide, with young individuals aged 5-29 facing the highest risk. According to the latest data, 1.19 million road fatalities were recorded in 2021 globally (World Health Organization, 2023). Figure 1.1 depicts the number of road fatalities per 100,000 population in 2021 by country-income level across the six geographic regions defined by the World Health Organization. Within the continents, the same correlation between income level and fatality rates can be observed, with fatality rates highest in low-income countries and lowest in high-income countries in all continents. Specifically, the risk of fatal injury in a crash is three times higher in low-income countries compared to high-income countries. Additionally, it is observed that the African Region has the highest fatality rate (19 road fatalities per 100,000 inhabitants), whereas the European Region has the lowest fatality rate (7 road fatalities per 100,000 inhabitants).

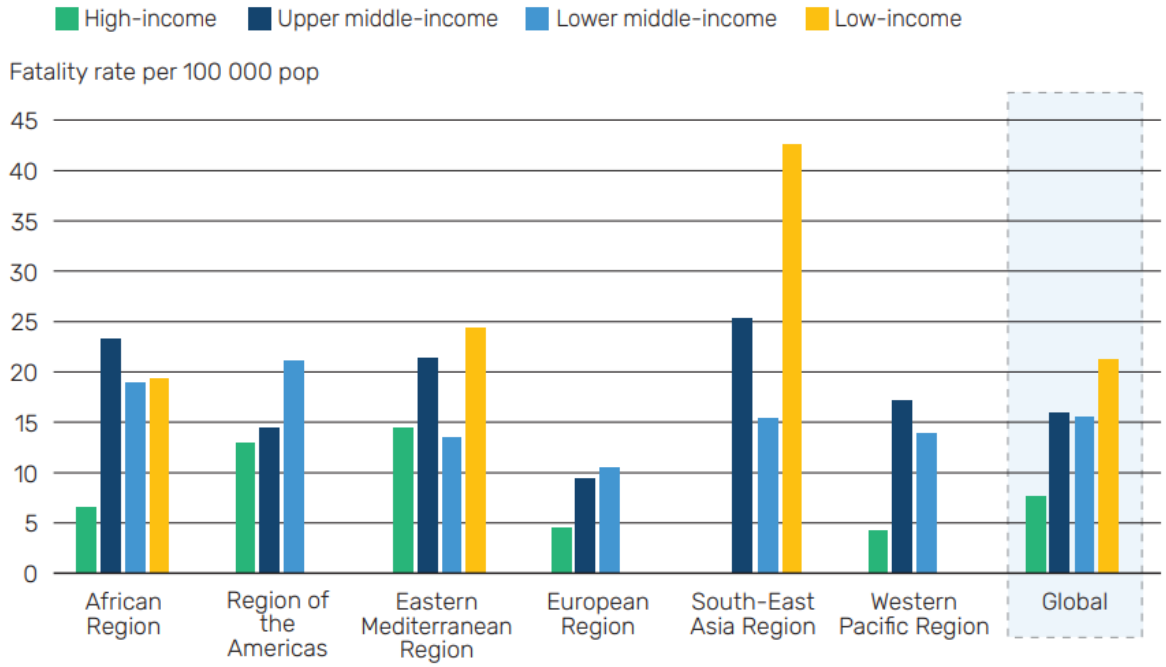


Figure 1.1: Road fatalities per 100,000 population by continent and country-income level, 2021. *(Source: World Health Organization, 2023)*

Moreover, profound differences are also evident in the analysis of road fatalities concerning different road user types. In Figure 1.2, the percentages of fatalities in road crashes per road user category within the six geographic regions of the World Health Organization for the year 2021 are presented. Globally, 30% of fatalities correspond to drivers and passengers of four-wheeled vehicles, 21% to drivers and riders of motorcycles and tricycles, 6% to cyclists, 23% to pedestrians, and 21% to unspecified road users. Notably, Southeast Asia and the Americas regions exhibit the highest proportion of road fatalities among users of motorcycles and tricycles, with percentages of 48% and 28%, respectively. In Europe, the rate of fatalities among users of four-wheeled vehicles is notably high at 49%, whereas Western Pacific and Africa record the highest pedestrian fatality rate at 29% and 27%, respectively.

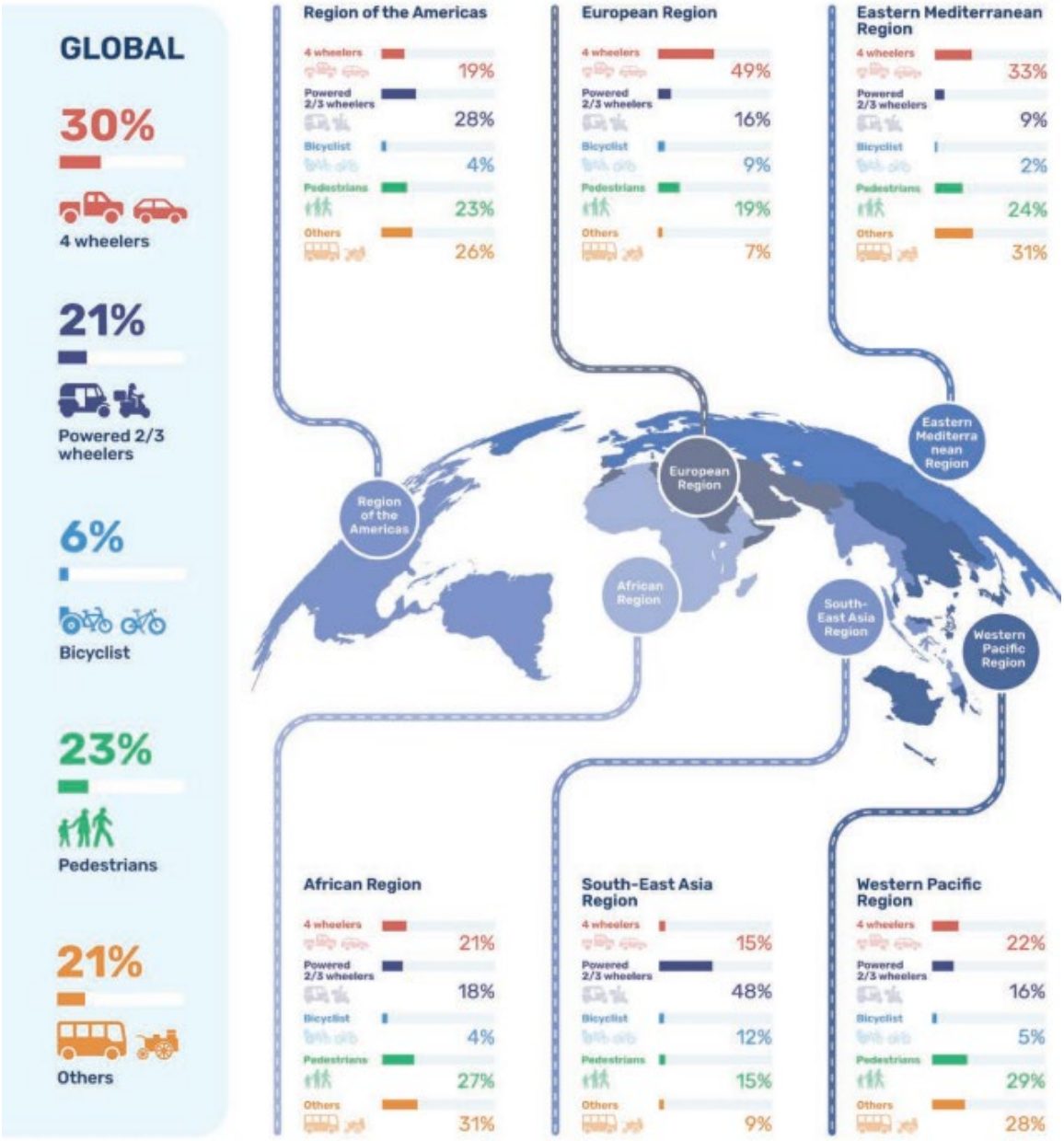


Figure 1.2:. Distribution of road fatalities by road user type and region for 2021.

(Source: World Health Organization, 2023)

1.1.2 Road Safety in the European Union

As previously discussed in the preceding subsection, Europe stands out as the continent with the best road safety performance globally. To delve deeper, the focus shifts to the European Union (EU), where 20,640 road fatalities were recorded in 2022. This figure reflects a 4% increase compared to 2021, attributed to the rebound in traffic levels after the Covid-19 pandemic. Although the long-term trend indicates a gradual decline (-9% in comparison to the pre-pandemic year), it is not decreasing at a fast enough pace to reach the EU target of halving the number of deaths by 2030, as illustrated in Figure 1.3.

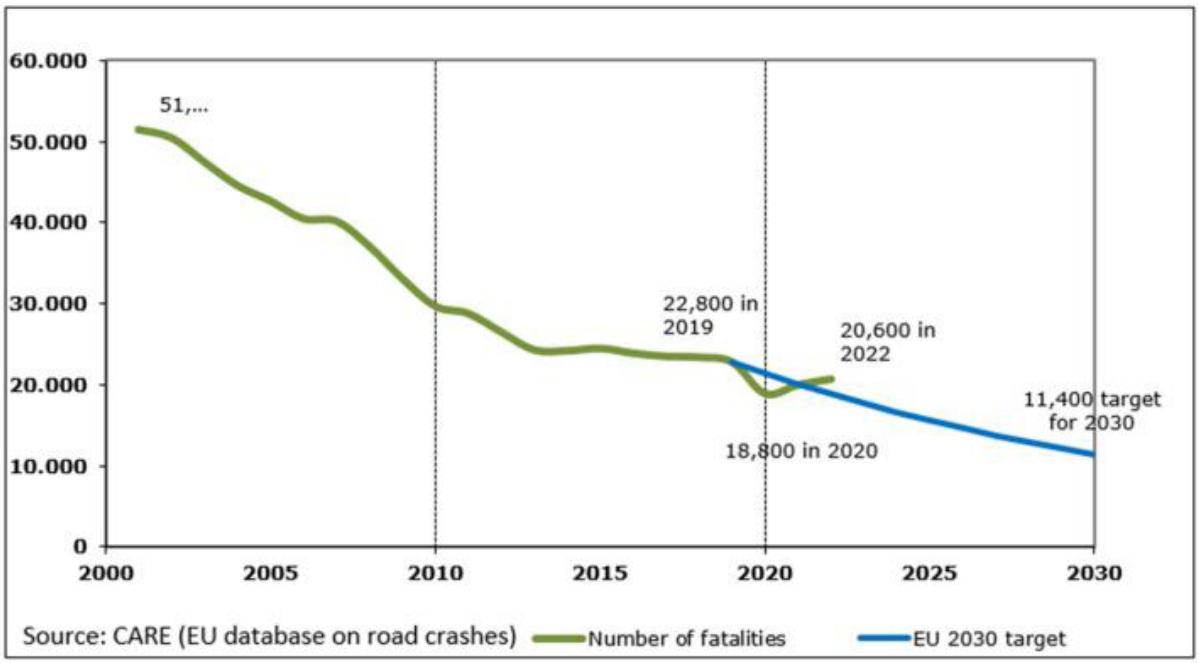


Figure 1.3: Evolution of road fatalities in the EU, 2001-2022.
(Source: European Commission, 2023)

Moreover, progress in this regard remains inconsistent among EU Member States. Notably, Lithuania and Poland reported the most substantial declines, exceeding 30%, between 2019 and 2022. However, Poland's fatality rate, while improved, remains above the EU average. Conversely, during the last three years, Ireland, Spain, France, Italy, the Netherlands, Slovakia, and Sweden have experienced either stagnation or an increase in the number of road deaths.

The overall ranking of countries based on fatality rates has remained relatively stable since the pre-pandemic period. Sweden (with 22 fatalities per million inhabitants) and Denmark (26) remain the countries with best road safety performance, while Romania

(86) and Bulgaria (78) reported the highest fatality rates in 2022. The EU's average fatality rate in 2022 was 46 road fatalities per million inhabitants (Figure 1.4).

Moreover, Figure 1.4 reveals that Greece, the author's native country and the location of the National Technical University of Athens, ranked 24th among the 27 EU Member States in 2022, with 61 road fatalities per million population, exceeding the EU average of 46. Further details regarding the national road safety state are available in the following subsection.

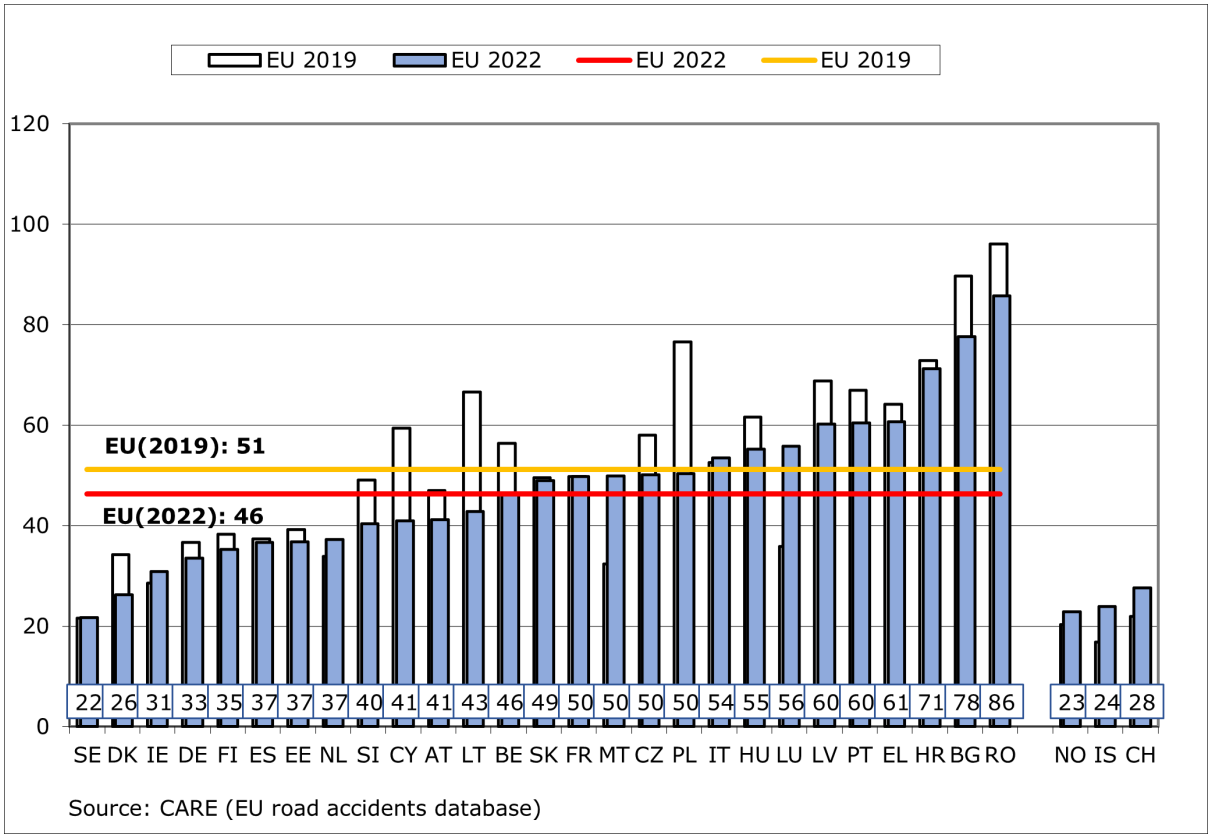


Figure 1.4: Road fatalities per million population in the EU, 2019-2022.
(Source: European Commission, 2023)

1.1.3 Road Safety in Greece

In 2022, 635 road fatalities (provisional data) were recorded in Greece (Hellenic Statistical Authority, 2023). This places Greece at the 24th position within the EU in terms of road safety performance. However, during the decade 2010-2020, Greece achieved the most remarkable improvement in road safety among EU Member States. As evident in Figure 1.5, depicting the evolution of key road safety figures in Greece, there was a 54% reduction in the number of road fatalities, surpassing the target of a 50% reduction.

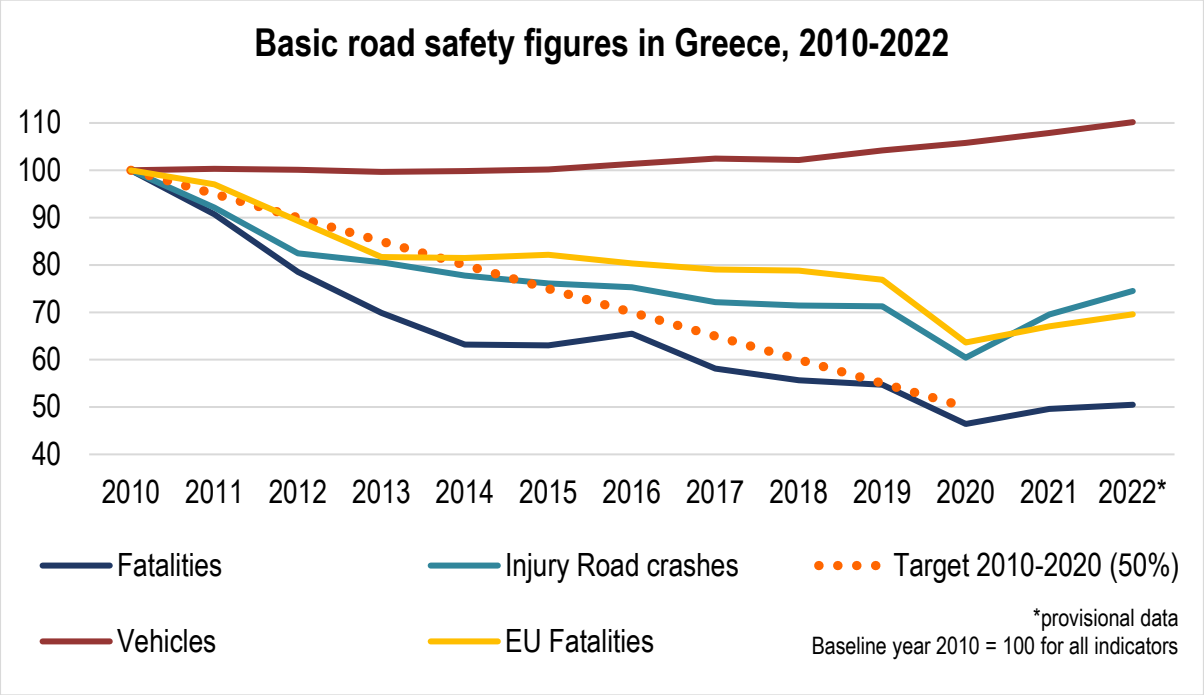


Figure 1.5: Evolution of basic road safety figures in Greece, 2010-2022.
(Sources: Hellenic Statistical Authority, 2023; European Commission, 2023)

Table 1.1: Comparison of Greek and EU road crash statistics, 2019.
(Sources: Hellenic Statistical Authority, 2023; CARE database)

	Greece			EU27
	2019	2010-2019 (%)	2019 (%)	2019 (%)
Total fatalities	***688***	***-45%***	*100%*	*100%*
Drivers	470	-44%	68%	65%
Passengers	73	-70%	11%	15%
Pedestrians	145	-19%	21%	20%
Inside built-up areas	370	-38%	**54%**	**39%**
Outside built-up areas	318	-52%	46%	61%
On motorways	50	-43%	7%	9%
Passenger Cars	202	-63%	29%	44%
Motorcycles/Mopeds	247	-55%	**36%**	**18%**
Bicycles	22	-4%	3%	9%
Young drivers (18-24)	61	-54%	9%	8%
Older drivers (65+)	99	-24%	14%	15%
Children (0-14)	12	-60%	2%	2%
Male drivers	441	-43%	64%	55%
Female drivers	29	-52%	4%	8%
In crashes with Heavy Goods Vehicles	40	-61%	6%	13%
Drivers/Passengers in single-vehicle crash	280	-44%	**41%**	**31%**

The comparison of Greek and EU road crash statistics for 2019, as presented in Table 1.1, reveals the most significant road safety problems in Greece. One of them is the particularly high rate of Powered Two-Wheeler (PTW) (motorcycles and mopeds) riders' fatalities (36%), which was twice the respective EU average (18%). In 2019, Greece also presented one of the highest rates (54%) of road fatalities inside built-up areas. Moreover, 41% of total road fatalities were vehicle occupants in single-vehicle road crashes (EU average 31%). Greece performs poorly in regards to road fatalities occurring inside built-up areas and in single vehicle crashes, which are both associated with the high traffic of motorcycles and related crashes, but also with significant deficiencies (e.g., high rates of speeding and driver distraction, low seatbelt and helmet use rates, poor enforcement of traffic violations, inadequate public transport network, etc.).

1.1.4 Surrogate Safety Measures

Road crashes are a complex phenomenon affected by several parameters that can be categorized into three distinct aspects: (i) road users (drivers, riders, passengers and pedestrians), (ii) vehicles and (iii) road infrastructure and environment. Among these main categories, it has been observed that the vast majority of road crashes can be attributed to human factors and human error, either exclusively or partially, accounting for rates as high as 94% (Singh, 2015). This particularly high percentage of responsibility for the human factor in the causal chain of road crashes points out the importance of studying and analysing driver behaviour.

A large array of methods has been used so far in the international literature to study driver behaviour. For instance, questionnaire surveys are a traditional way of collecting data of road users self-declared behaviour in traffic and general road safety attitudes or perceptions (Rowe et al., 2015; Pires et al., 2020). Furthermore, a common way of collecting data on driver behaviour (e.g., driver distraction, seatbelt use, traffic rule compliance etc.) is roadside observation (Yannis et al., 2011; Sullman, 2012; Prat et al., 2015). Another widely used family of methods is the exploitation of driving simulators. Driver simulators can be used to extract various metrics related to speed, reaction time, lane position, headway distance, distraction, fatigue and others in a safe and virtual road environment (Lenné et al., 1997; Calvi & D'amico, 2013; Papantoniou et al., 2019).

Apart from these methods, the recent swift technological development in naturalistic driver recording has led to a growing abundance of data from sensors in vehicles and smartphones, which can be used to assess driver behaviour (Ziakopoulos et al., 2020). More specifically, smartphone sensors like accelerometers, gyroscopes, magnetometers, and GPS enable the extraction of various driver performance metrics

and Surrogate Safety Measures (SSMs) through a low-cost and rapid manner, without the need for user interaction (Mantouka et al., 2018).

SSMs encompass a wide range of metrics and parameters, which are not directly derived from or rely on crash data. These measures possess various advantages when compared to historical crash data. Specifically, they serve as a proactive approach, enabling road safety analyses before the occurrence of road crashes (Tarko et al., 2009). In contrast, crash data collection relies heavily on manual methods, which can be related to limitations such as inaccurate data and under-reporting (Imprialou & Quddus, 2019; Yannis et al., 2014).

The exploitation of SSMs in the field of road safety facilitates the understanding of crash-leading factors and allows for evaluating the effectiveness of different countermeasures (Tarko, 2018). Wang et al. (2021) divided SSMs into two main groups: (i) SSMs and (ii) SSM-based models. The first group encompasses SSMs that are time-based, deceleration-based, or energy-based. It includes SSMs that use predefined thresholds for traffic conflicts' detection, such as Time-to-Collision (TTC), Post-Encroachment Time (PET), Time-to-Crash (TC) and Deceleration Rate to Avoid the Crash (DRAC) (Bonela & Kadali, 2022). On the other hand, the other group of SSMs focuses on establishing a direct link between each traffic conflict and either a crash or a non-crash outcome, by estimating its crash probability (Songchitruksa & Tarko, 2006; Wang & Stamatiadis, 2014). This kind of SSMs can be also derived from simulation processes (Gettman & Head, 2003).

In addition, the constant advancement in technology has made smartphones a key choice for collecting data on SSMs, particularly regarding harsh driving behaviour events such as harsh braking and harsh acceleration (Nikolaou et al., 2023b). It is important to mention that the two aforementioned harsh driving behaviour events are distinct occurrences that take place in different traffic situations and should not be analyzed together as a single phenomenon. Firstly, drivers who experience higher levels of anger, frustration, and anxiety tend to exhibit increased acceleration values and apply higher physical pressure on the accelerator pedal (Stephens & Groeger, 2009). On the other hand, drivers more typically engage in harsh braking events as a reaction to various potentially hazardous situations, aiming to prevent near misses or collisions (Ziakopoulos et al., 2022).

All in all, SSMs can either be an alternative to road safety analyses or even complement analyses that are based on historical crash records (Johnsson et al., 2018). SSMs such as time-to-collision, harsh braking, post-encroachment time and so on, are widely proposed in transportation science and are particularly useful in order to assess road safety when detailed crash data are not available. Such events are of particular importance in evaluating driving risk (Gündüz et al., 2017) since they are inherently associated with the likelihood of a road crash (Tselentis et al., 2017).

1.2 Objective

Taking the previous into consideration, the primary objective of this dissertation is to assess road crash risk by fusing infrastructure, traffic, and driving behaviour data. This combination of data outlines a highly promising research field. However, the practical integration of these data types is often hindered by inadequate availability or low quality of the data. Consequently, this dissertation initially explores the feasibility of developing comprehensive crash prediction models in Greece by leveraging these diverse data types. A pivotal aspect of this exploration is the availability of high-quality crash data especially in terms of crash location recording.

Hence, for roads where high-quality data can be obtained, statistical models are developed, and machine learning techniques are applied to investigate the influence of geometric characteristics, traffic attributes, and naturalistic driving behaviour metrics on road crash occurrence and corresponding crash risk per examined road segment. A critical aspect of this research entails thoroughly exploring the reliability of harsh driving behaviour events as SSMs and their utilization for assessing the safety levels of road segments across various road environments where detailed road crash data are unavailable.

To achieve these outlined objectives, a plethora of statistical tools, spatial analyses, and machine learning techniques were employed within the framework of this dissertation. These methodologies include the following:

- Generalized Linear Model (GLM) with Negative Binomial (NB) distribution,
- Hierarchical Clustering (HC),
- Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), K-Nearest Neighbours (K-NN), Support Vector Machines (SVM),
- SHapley Additive exPlanations (SHAP values),
- Log-linear Regression, Spatial Error Model (SEM), Spatial Lag Model (SLM),
- Zero-Inflated Negative Binomial (ZINB) Model, Spatial Zero-Inflated Negative Binomial (SZINB) Model,
- Spatial Random Forest (SRF).

These objectives are expected to lead to knowledge which will be useful for reducing crash occurrence, and increasing overall road safety levels.

1.3 Methodology of the Dissertation

To fulfill the scientific objectives outlined in this doctoral dissertation, a series of methodological steps were systematically executed. The steps are delineated within this subsection and visually represented in Figure 1.6. The methodological framework for this doctoral dissertation is designed to address a set of pertinent research questions through a systematic and comprehensive approach. The key elements of the methodological framework encompass a thorough Literature Review, precisely formulated Research Questions, and a structured sequence of steps.

The foundation of this overarching methodological framework rests upon an extensive literature review, providing contextualization and insights into the existing knowledge on road crash risk assessment through the utilization of SSMs collected under real road environment conditions. This review informs the subsequent research questions and guides the selection of appropriate methodologies. These research questions serve as the driving force behind the entire research endeavor, exploring the integration and analysis of infrastructure, traffic, and driver behaviour data for meaningful conclusions in road crash risk assessment.

The core of the methodological framework involves a multi-step process, starting with the investigation of road safety modelling data in Greece, laying the groundwork for subsequent directions. This investigation highlighted the limitation of conducting high-detailed crash prediction modelling in Greece, feasible only for motorways with high-quality crash data, in terms of crash location, and traffic attributes per road segment. To that end, two distinct databases were developed, one for motorway segments with comprehensive data on historical road crashes, traffic, road geometry characteristics, and naturalistic driver behaviour metrics, and the other for a road network including urban and interurban roads, which lacked detailed crash and traffic data, comprising only geometric characteristics and naturalistic driver behaviour metrics for the examined road segments.

With the problem under consideration as well as the scientific literature in mind, a methodological investigation explored the underlying theory of statistical models and Machine Learning (ML) techniques suitable for road crash risk assessment analysis. The analysis of motorway data entailed the application of several methodologies. These included utilizing a GLM with NB distribution to predict crash frequency, employing HC to establish crash risk levels for motorway segments based on historical road crash and traffic data, and deploying various ML classifiers—such as LR, DT, RF, K-NN, and SVM—for predicting crash risk levels by exploiting road geometry characteristics and driver behaviour metrics. Special emphasis was placed on assessing the reliability of harsh driving behaviour events as SSMs.

Upon evaluating the statistical significance and coefficients' signs of the NB crash frequency regression model, as well as considering SHAP values from the best-performing ML classifier for crash risk level prediction, it was deduced that harsh braking events could be meaningfully regarded as SSMs. These events are deemed suitable as dependent variables for both statistical and ML models, particularly when confronted with unavailable crash data or faced with issues related to the low-quality recording of crash locations.

The framework extended to urban and interurban road network data, where harsh braking events were examined for spatial autocorrelation using Moran's I and served as a key metric for road crash risk assessment. The subsequent analyses encompassed both non-spatial models (Log-linear, Zero-Inflated, Random Forest) and spatial models (SEM, SLM Zero-Inflated with spatial lag, SRF), aiming to identify statistically significant road infrastructure and driver behaviour parameters affecting the number of harsh braking events per road segment. Additionally, a performance comparison between spatial modelling approaches and their non-spatial counterparts was also conducted.

The final stage synthesized the findings from the aforementioned analyses, leading to a comprehensive road crash risk assessment. In summary, this methodological framework is a structured and logically sequenced process that combines statistical modelling, machine learning techniques, and spatial analyses to address the research questions of this doctoral dissertation and achieve the overarching objective of assessing road crash risk and enhancing road safety.

Further details on the methodological background and implementation of the techniques applied in this doctoral dissertation are presented in the subsequent sections.

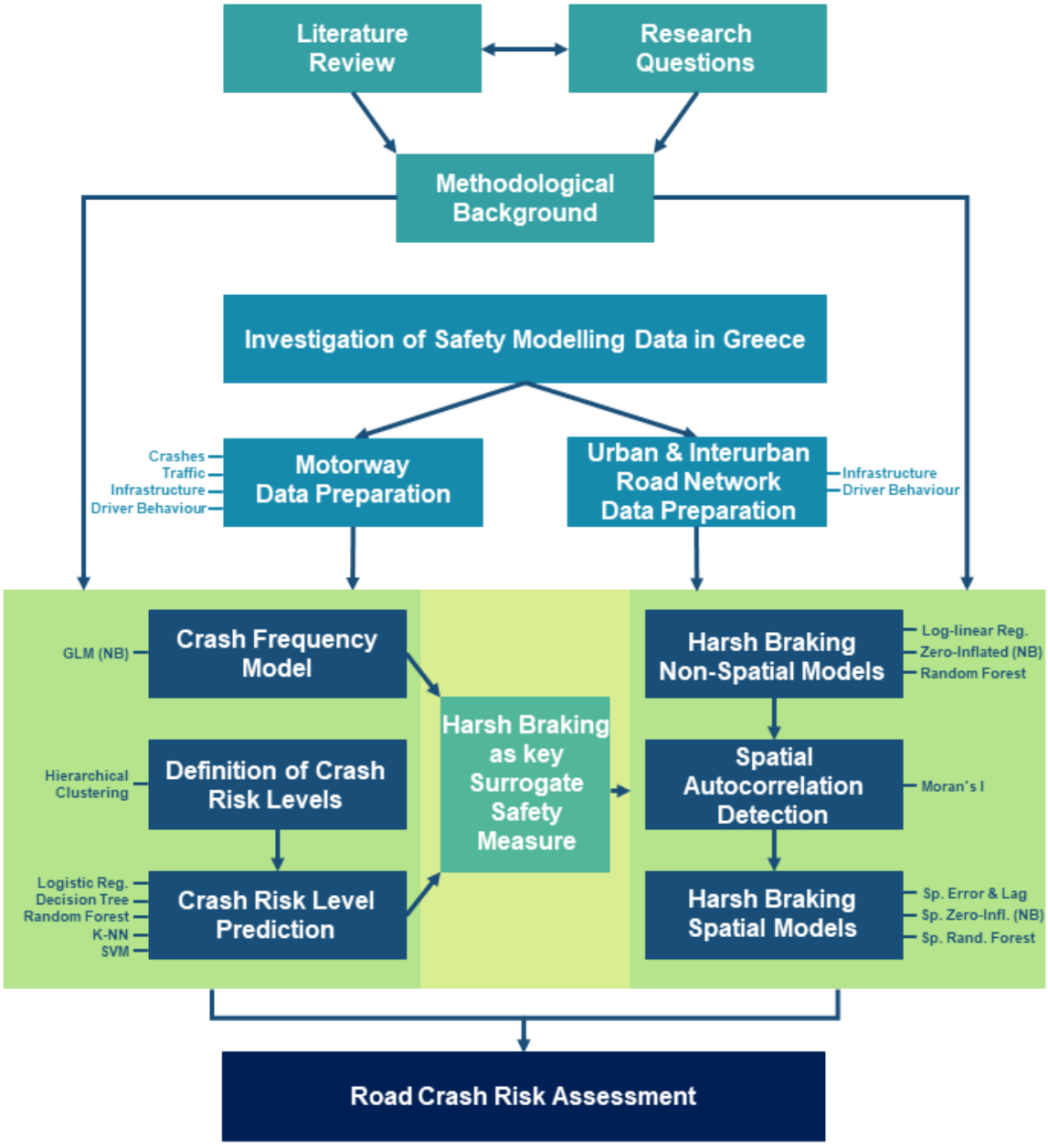


Figure 1.6: Graphical representation of the overall methodological framework of the doctoral dissertation

1.4 Structure of the Dissertation

The remainder of this doctoral dissertation is organized in nine sections which are briefly described within this subsection.

Section 2 provides a review of the scientific literature of studies exploiting SSMs in historical crash record investigations. It showcases the main review findings in terms of the different types of SSMs and crashes, modelling approaches, and the temporal dimension of the data used in the examined studies. Subsequently, it discusses overall findings and trends, future research directions, and outlines the specific research questions that this doctoral dissertation aims to address.

Section 3 describes the overall methodological framework employed to achieve the objectives of this doctoral dissertation and delves into the theoretical foundations of the analytical methods and models utilized throughout the dissertation.

Section 4 investigates and discusses the availability and accuracy of road safety modelling data in the primary rural road network of Greece, focusing on three types of data that are considered most critical: crash, traffic and road geometry data. The exploitation of smartphone data related to driver behaviour is also discussed.

Section 5 provides technical information on the process of data collection and descriptive statistics for the Olympia Odos motorway. The developed database includes data on road crashes, traffic, road geometry and driver behaviour per motorway segment. Detailed road crash and traffic data were kindly provided by the road operator. Road infrastructure data, sourced from tools like Open GIS software, Google Earth, and GoogleStreetView, were consolidated. Additionally, smartphone data were utilized for capturing naturalistic driver behaviour metrics.

Section 6 examines the relationship between road crash frequency in motorway segments and various explanatory variables based on road design characteristics and SSMs. Additionally, clusters representing crash risk levels of the examined motorway segments, based on crash and traffic data, are established. Furthermore, this section compares the classification performance of five well-known ML techniques that exploit road design data and SSMs to predict the crash risk level of motorway segments.

Section 7 describes the development of a database for the road network in the Eastern Macedonia and Thrace Region, including urban and interurban roads. As detailed traffic and crash data (in terms of geo-location) were unavailable for these roads, the resulting database includes only geometric characteristics and naturalistic driver behaviour metrics sourced from OpenStreetMap (OSM) and smartphone data, respectively. Key descriptive statistics for the considered variables are also provided.

Section 8 focuses on analysing harsh braking event frequencies per road segment within the Region of Eastern Macedonia and Thrace, and correlating them with various road network characteristics and driving behaviour metrics. To that end, various spatial modelling techniques, including SEM, SLM, SZINB and SRF are employed on harsh braking events frequencies.

Section 9 presents the conclusions of the thesis and discusses the contribution to knowledge, the limitations as well as the recommendations for further research.

Lastly, a complete list of the bibliographical references is provided.

2. Literature Review

2.1 Introduction

Road crashes and their related casualties constitute a major societal and public health problem as it is estimated that more than 1.19 million people are killed in road crashes and tens of millions are seriously injured annually (World Health Organization, 2023). Improving road safety is also included as a key component of the United Nations' Agenda, as manifested by Sustainable Development Goals (SGDs) 3.6 and 11.2, which aim to reduce road fatalities and injuries by half and provide sustainable and safe transport for road users of all age groups respectively (United Nations, 2022). Until now, the main indicator for measuring road safety outcomes has been historical crash data, considered to be hard evidence for the measurements of road safety performance. Even if it is natural to rely on road crash historical records for the assessment of the road safety level of an examined area or road, specific drawbacks of road safety analyses based on historical crash records have been determined as well.

In particular, a long period of time is typically required to collect a sufficient sample of road crash data that could allow for reliable estimates of the road safety level as road crashes are rare events by nature (Theofilatos et al., 2019). When examining large geographical areas, road crashes also face the typical issues inherent in all point data such as spatial dependence and spatial heterogeneity (Ziakopoulos & Yannis, 2020). Moreover, any before-and-after study based on historical crash records for the evaluation of the implementation of a road safety measure may be biased by the regression-to-the-mean phenomena (Elvik, 2008). In addition, significant discrepancies are found between the non-fatal road crash injury data provided by various data sources. This problem is known as under-reporting and several studies indicate that the Police Departments do not report an appreciable proportion of road crash injuries, whereas the extent of under-reporting may vary depending on the severity of the injuries or the road user types (Yannis et al., 2014; Janstrup et al., 2016). Apart from the aforementioned, it can be perceived that road safety analyses based on historical crash records are a reactive approach that forces road safety analysts to wait for road crashes to occur in order to examine measures that could prevent them and should rely on valid crash data, including accurate location data, which is not always the case (Imprialou & Quddus, 2019).

Therefore, over the past few years, significant efforts have been made in utilizing SSMs in order to address this issue (Wang et al., 2021). SSMs include all measures, parameters, or quantities, which do not stem directly from or rely on crash data. Such approaches are a sustainable way of gauging road safety and may be more preferable

as they allow for road safety analyses before the physical occurrence of road crashes. According to Tarko (2018), the use of SSMs in the field of road safety aids in the detection of road crashes' excessive risk, the knowledge improvement of crash-leading conditions, and the effectiveness estimation of various countermeasures. Wang et al. (2021) provide a comprehensive review of important SSMs and divide them into two key categories: (i) SSMs and (ii) SSM-based models. The first category includes key time-based, deceleration-based, and energy-based SSMs. These subcategories include predominant SSMs that use predefined thresholds for traffic conflicts' identification and are used widely across studies in the road safety literature such as Time-to-Collision (TTC), Post Encroachment Time (PET), Time-to-Crash/Accident (TC/TA) and Deceleration Rate to Avoid the Crash (DRAC) (Bonela & Kadali, 2022). On the other hand, the second category aims to directly associate each traffic conflict with either a crash or non-crash outcome, by estimating its crash probability (Songchitruksa & Tarko, 2006; Wang & Stamatiadis, 2014).

Initially, data collection of SSMs was based on roadside observation techniques (Sayed & Zein, 1999). As it can be intuitively perceived, such approaches were not accurate as they were based on subjective criteria (Shinar, 1984). In order to reduce such biases, video-based measurements were introduced many years ago (Hydén, 1987) and have been improving significantly since then. Recent, technological advancements have led to more advanced techniques that reduce human interventions and deploy computer vision and sensor techniques (Chen et al., 2017; Laureshyn et al., 2017; Wu et al., 2018). Moreover, several simulation-based analyses have been conducted aiming to derive SSMs from traffic simulation models (Gettman & Head, 2003; Mahmud et al., 2019). The rapid technological development in naturalistic driver recording has also brought about an increasing availability of data from sensors in vehicles and smartphones that can be used to extract various SSMs such as TTC, harsh braking events, and harsh acceleration events (Guido et al., 2012; Fazeen et al., 2012; Ziakopoulos et al., 2022). All in all, SSMs can either be an alternative to road safety analyses or even complement analyses that are based on historical crash records (Johnsson et al., 2018).

Within this framework, the aim of this literature review process is to provide a review of the scientific literature of studies exploiting SSMs in historical crash record investigations. More specifically, this review process focuses on studies that attempt either (i) to investigate the correlation of SSMs and historical crash records or (ii) to predict the number of expected road crashes through SSMs and then compare them with the historical crash records. The different types of SSMs, the manner in which they are collected, their connection with specific road crash types, and the type of the developed statistical models are examined and discussed. Particular emphasis is placed on the temporal periods dedicated to data collection for both the SSMs and road crash data, as uncertainties in the length of the data collection periods are a problem typically investigated in driver recording (Stavrakaki et al., 2020). In order to

achieve this aim, published scientific studies that are authored in English are critically examined. It should be mentioned that this literature review only includes relevant papers that concern SSMs collected under real road environment conditions, as opposed to studies that are based on traffic simulation and driver simulators.

During the review process, studies dealing with the use of traffic conflict techniques for use in-road safety assessments were also identified. Arun et al. (2021b) focused on mapping the concepts and methods related to surrogate safety assessment using traffic conflicts. Their study deals with specific topics such as the concept of crash surrogacy, the definition and identification of traffic conflicts, and the specification of the relationship between crashes and conflicts. In other studies, Arun et al. (2021a) assessed the different traffic conflict safety thresholds among various road environments and applications, while Zheng et al. (2021) discussed various conceptual and methodological issues related to traffic conflict modelling. However, this literature review presents novelty in different areas. Specifically, it (i) exclusively investigates studies that use both SSMs and historical crash records, (ii) extends beyond measures with predefined thresholds for traffic conflicts' identification to SSMs that can be extracted from smartphone sensors and instrumented vehicles related to harsh driving behaviour events, and (iii) sheds light on the temporal periods dedicated to data collection for both SSMs and crashes.

Following this Introduction, this section is organized as follows. Section 2.2 describes the methodological framework of this literature review, including the Preferred Reporting Items for Systematic Reviews and the Meta-Analyses (PRISMA) approach that was adopted. Section 2.3 showcases the main review findings in terms of the different types of SSMs and crashes, various modelling approaches, and the temporal dimension of the data used in the examined studies. Subsequently, a discussion of overall findings and trends from the reviewed studies and some future research directions are provided in Section 2.4. Lastly, Section 2.5 comprises the research questions that this doctoral dissertation seeks to address meaningfully, presenting substantial results and findings in the subsequent sections.

2.2 Review Methodology

The current review was carried out during June 2022 and adhered to the PRISMA guidelines (Moher et al., 2009). The search was undertaken in the Scopus, TRID and Web of Science databases; Figure 2.1 depicts the search terms and the study selection process. It should be noted that there was no specific search restriction on the publication date of the examined articles. Moreover, articles had to be peer-reviewed before publication and authored in English which is the predominant written language in the global scientific literature. Emphasis should be placed on the fact that the present review process aims to provide a review of the scientific literature regarding studies exploiting SSMs towards historical crash record investigations and thus includes only studies that were conducted under real road environment conditions (as opposed to simulators).

After the exclusion of some papers based on their titles and abstracts, a total of 52 articles were selected for full-text review. After the full-paper review, 18 studies were excluded for not meeting the inclusion criteria (e.g., absence of historical crash data or SSMs, separate statistical models for SSMs and road crashes, crash data available but not used in statistical modelling, etc.). Finally, 34 articles were identified and reviewed. The literature review findings are presented and discussed in detail in the following subsections.

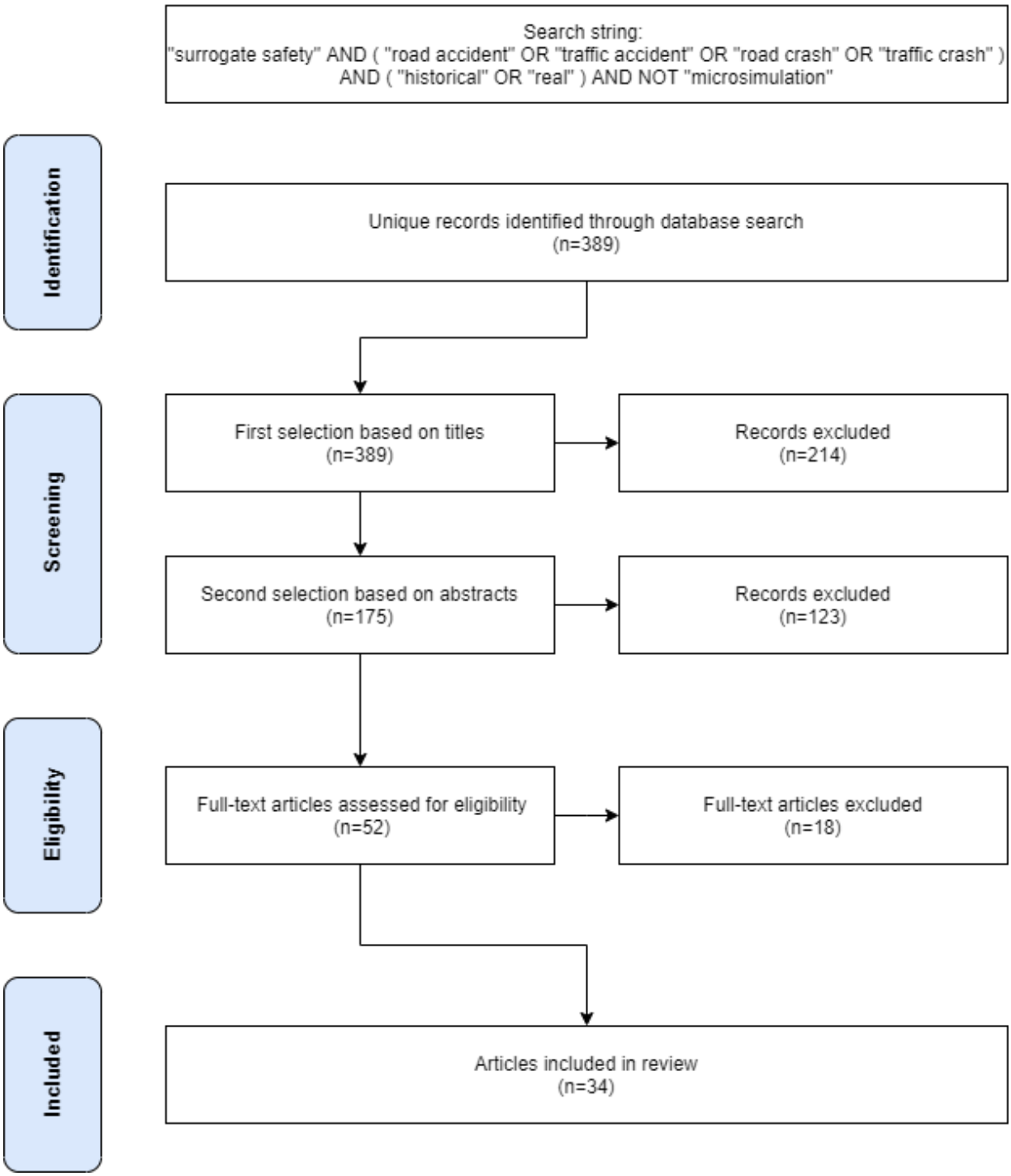


Figure 2.1: PRISMA flow diagram

2.3 Review Findings

2.3.1 Types of Surrogate Safety Measures and Historical Crash Data

As already pointed out in the introductory section of this literature review process, SSMs can be leveraged in road safety analyses in two ways. On one hand, they can provide an alternative to road safety analyses when road crash data are not available as a proactive approach. On the other hand, SSMs complement analyses based on historical crash records, which is also the main subject of this review process. The key information about the SSMs and historical crash records (types and temporal dimension), modelling approaches, the scale of analysis, and other considered variables used in the reviewed studies are summarized in Table 2.1, sorted by means of collection for SSMs. It should be noted that the column named “Temporal Ratio” of Table 2.1 has been calculated due to the observed discrepancies in data collection period lengths for road crashes and SSMs. The values of this column are dimensionless numbers as they have been calculated by converting the crash and SSMs data collection periods into the same time units.

Technological improvements during recent decades have led to the development of a wide array of sophisticated tools that provide more rich and rapid data acquisition in terms of various aspects of driving performance (Ziakopoulos et al., 2020). As can be observed from Table 2.1, during the last five years, the use of smartphone data has also begun to gain significant ground in studies featuring SSMs (Strauss et al., 2017; Paleti et al., 2017; Stipancic et al., 2018a; Stipancic et al., 2018b; Stipancic et al., 2019; Yang et al., 2019; Khorram et al., 2020; Guo et al., 2021). Exploiting smartphone sensors such as accelerometers, digital compasses, gyroscopes, and GPS allows the extraction of various driver performance metrics and SSMs through an inexpensive and rapid way, even without requiring user engagement (Mantouka et al., 2018).

The SSMs collected via smartphone sensors in the examined studies concern harsh driving behaviour events such as harsh braking and harsh acceleration. Harsh braking events are generated by drivers as a reaction to various possibly dangerous situations in order to avoid a near miss or even a road crash (Ziakopoulos et al., 2022). Moreover, harsh braking events are a critical element for the assessment of driving risk (Gündüz et al., 2017), as they are innately associated with crash occurrence probability (Tselentis et al., 2017). However, harsh acceleration events are different phenomena than harsh braking events, as they are mainly affected by drivers’ levels of anger, frustration, and anxiety (Stephens & Groeger, 2009). Based on previous studies, it is noted that the levels of deceleration and acceleration that define harsh braking and harsh acceleration events respectively may vary across different studies and transport modes (Kamla et al. 2019; Park et al. 2021).

Specifically, in a relevant summary table presented in a study by Kamla et al. (2019), the thresholds of harsh brakings are recorded, ranging from 1.96 m/s^2 for trucks (Blanco et al., 2011) to as high as 8.43 m/s^2 for passenger cars under dry surface conditions (Greibe, 2007). Regarding certain studies included in Table 2.1, the thresholds for harsh braking are as follows: 2 m/s^2 (Stipancic et al., 2018a), 2.67 m/s^2 (Desai et al., 2021; Hunter et al., 2021), 3.4 m/s^2 (Strauss et al., 2017), and 4 m/s^2 (Kim et al., 2016).

A frequent barrier encountered in studies exploiting harsh events is that they do not provide their specific thresholds and calculation methods for commercial reasons (Guo et al., 2021; Kontaxi et al., 2021; Zhao et al., 2022). Indicatively, the data provider for the analyses of the study by Yang et al. (2019) mentions that a harsh braking event is identified if a reduction in the speed is fast enough to thrust the driver and passengers' bodies forward hard enough to cause the seatbelt to lock.

As can be observed from Table 2.1, naturalistic driving experiments using instrumented vehicles are another frequently selected option for collecting SSMs. These experiments are a quite similar alternative to smartphone data but much more expensive as there are significant costs that depend on the equipment used (Ball & Ackerman, 2011) and the duration of the experiment (Regan et al., 2012). The majority of the SSMs collected through instrumented vehicles range in a similar concept to the data collected by smartphones and concern harsh driving behaviour events (Kim et al., 2016; Pande et al., 2017; Ambros et al., 2019; Kamla et al., 2019; Mousavi et al., 2019; Stipancic et al., 2021; Desai et al., 2021; Hunter et al., 2021; Li et al., 2021a; Park et al., 2021; Li et al., 2021b). Apart from these studies that focus on harsh driving behaviour events, traffic conflicts and related measures for rating their severity have also been examined in other naturalistic driving experiments using instrumented vehicles (Lu et al., 2011; He et al., 2018).

Table 2.1: Studies exploiting SSMs in historical crash record investigations

Reference	Surrogate Measures				Other Variables			Historical crash data		Temporal Ratio (Crash/SSM)	Modelling Approach	Scale of analysis
	Type	Sample	Collection	Period	Infrastructure	Traffic	Other	Period	Type			
Khorram et al. (2020)	harsh braking	176 bus drivers	smartphone	4 months	length	deceleration	driver age & experience	3 years	Bus driver at-fault	9	Pearson correlation, GLM (NB)	2 routes (13km, 10km)
Paleti et al. (2017)	harsh braking, harsh acceleration	11 drivers, 228 trips, 58h of driving (4-6pm)	smartphone	1 year	interchange, surface	traffic volume, avg speed, SD acceleration	-	1 year	4-6pm weekdays	1	random parameters Generalized Ordered Response Probit (GORP)	513 freeway segments
Stipancic et al. (2018a)	harsh braking	~22,000 trips, >4000 drivers	smartphone	21 days	length, class	congestion, avg speed, speed variation	-	11 years	Total	191	INLA Full Bayesian Latent Gaussian Model	1000 links and intersections
Stipancic et al. (2018b)	harsh braking, harsh acceleration	~22,000 trips, >4000 drivers	smartphone	21 days	class	-	-	5 years	Total	87	Spearman correlation and pairwise Kolmogorov-Smirnov test	20586 links and 10721 intersections
Stipancic et al. (2019)	harsh braking	~22,000 trips, >4000 drivers	smartphone	21 days	length, class	congestion, avg speed, speed variation	-	11 years	Total	191	INLA Full Bayesian Latent Gaussian Model, Fractional Multinomial Logit	4623 links and 4429 intersections
Strauss et al. (2017)	harsh braking	over 10,000 trips, ~1000 cyclists	smartphone	137 days	-	traffic volume	-	6 years	Cyclists	16	empirical Bayes (EB) estimates - Spearman correlation	13279 intersections and 19837 segments (aggregated also at corridors level)
Yang et al. (2019)	harsh braking, harsh acceleration	10,512 events	smartphone	6 months	bus & subway stations, intersections, length	traffic volume, truck flow, speeding	distraction, land use, population, unemployment, income, housing, commuting	6 months	Total	1	MVCAR, UCAR, two-sample Kolmogorov-Smirnov test, Wilcoxon signed-rank test	282 census tracts
Guo et al. (2021)	Harsh: braking, acceleration, turn, merge into lane	-	in-vehicle navigation software	2 months	-	traffic volume, congestion, avg speed,	-	2 months	Total	1	Random Forest, Logistic regression	40 freeway segments

Reference	Surrogate Measures				Other Variables			Historical crash data		Temporal Ratio (Crash/SSM)	Modelling Approach	Scale of analysis
	Type	Sample	Collection	Period	Infrastructure	Traffic	Other	Period	Type			
						speed variation						
Ambros et al. (2019)	harsh braking, harsh acceleration	1,172 company vehicles	instrumented vehicle	8 months	curve length & radius	traffic volume, acceleration	-	6 years	Single-vehicle	9	GLM (NB)	30 rural curves
Boonsiripant et al. (2011)	stop frequency, variation of stops, 90th percentile count of stops	36,724 trips, 408 drivers	instrumented vehicle	1 year	speed limits	traffic volume, speed variation,V85 , V95, V5, acceleration	-	4 years	Daytime, clear weather, motor vehicle	4	Regression tree and GLM	61 urban corridors
Desai et al. (2021)	harsh braking	196,215 events	instrumented vehicle	2 months	length	-	-	2 months	Injury and PDO	1	Linear regression	23 construction work zones (150 miles)
Guo et al. (2010)	near crash	100 cars, 2 million veh-miles, 43000h	instrumented vehicle	1 year	-	-	-	1 year	Total	1	GLM (Poisson)	Northern Virginia/Metro Washington, DC
He et al. (2018)	TTC, MTTC, DRAC, brake duration	100 vehicles	instrumented vehicle	2 months	length	avg speed	avg trip duration, extreme trip index	5 years	Rear-end mid-block	30	GLM (NB)	2772 links
Hunter et al. (2021)	harsh braking	10,000 events	instrumented vehicle	1 months	-	traffic volume	-	4.5 years	Rear-end	55	Spearman, Pearson & Kendall Cor., Sensitivity Analysis, GLM (Poisson)	8 intersections
Kamla et al. (2019)	harsh braking	8,000 trucks, 195,297 harsh braking events	instrumented vehicle	2 years	width, inscribed circle diameter	traffic volume, truck traffic	-	11 years	Total	6	GLM (NB) random/fixed-parameters	70 roundabouts
Kim et al. (2016)	harsh braking	20 vehicles, 150k seconds of data, 224 trips	instrumented vehicle	3 months	internal TMC, recurrent bottleneck	speed, acceleration, deceleration	-	4 years	Rear-end /veh-km	16	Correlation, Spatial distribution using GIS	60 segments (63 mile freeway)

Reference	Surrogate Measures				Other Variables			Historical crash data		Temporal Ratio (Crash/SSM)	Modelling Approach	Scale of analysis
	Type	Sample	Collection	Period	Infrastructure	Traffic	Other	Period	Type			
Li et al. (2021a)	harsh braking, harsh acceleration	300 buses, 6.7million GPS records	instrumented vehicle	3 months	-	-	number of buses	10 years	Pedestrian & bicycle	41	Spearman correlation, Bayesian NB, Bayesian NB-CAR	200m & 100m buffer circles
Li et al. (2021b)	harsh braking	16 participants	instrumented vehicle	2 weeks	length	traffic volume	-	3 years	Total / veh-miles	78	Line-constrained clustering method (combines DBSCAN with spatial selection functions)	156 quarter mile segments of 2 highways
Lu et al. (2011)	conflicts / vehicles	50 taxies, 2.25million km travelled	instrumented vehicle	6 months	-	-	-	3 years	Total / vehicles	6	Linear regression	city, country
Mousavi et al. (2019)	harsh braking	31 participants	instrumented vehicle	2 weeks	curvature	traffic volume	-	5 years	Total / traffic volume	130	GLM (NB)	31+21 quarter mile segments of 2 highways
Pande et al. (2017)	harsh braking	33 drivers	instrumented vehicle	10 days	curve(y/n), auxiliary lane(y/n)	traffic volume	-	10 years	Total	365	GLM (NB) random/fixed-parameters	39 freeway segments
Park et al. (2021)	Harsh: acceleration, braking, start, stop, lane change, overtaking, turning, U-turn	all commercial vehicles in Korea	instrumented vehicle	1 week	length	speeding	city	4 years	Total	209	Random Forest, GLM (NB)	38 segments in 4 cities
Stipancic et al. (2021)	harsh braking	~1.5 million trips	instrumented vehicle	30 days	length, class	congestion, avg speed, speed variation	-	5-11 years	Total	61	INLA Full Bayesian Latent Gaussian Model	123792 links
Hu et al. (2020)	harsh braking, harsh acceleration, wait-time	90 vehicles	connected vehicle	1 month	approaches, traffic light	-	traffic volume, speed, acceleration, deceleration	5 years	Total	61	Multi-layer perceptron (MLP), Convolutional Neural Network (CNN), Decision Tree	774 intersections
Xie et al. (2019)	TTC, DRAC, TTCD	90 vehicles, 15.7 million GPS points	connected vehicle	1 month	-	traffic volume	-	1 year	Rear-end / traffic volume	12	Pearson correlation	75 highway segments
Yang et al. (2021)	TTC, DRAC, TTCD	2.7 million trajectory points	connected vehicle	1 month	class, speed limit, lanes	traffic volume	GPS points	1 year	Rear-end	12	SEM-CAR-RP	220 road segments

Reference	Surrogate Measures				Other Variables			Historical crash data		Temporal Ratio (Crash/SSM)	Modelling Approach	Scale of analysis
	Type	Sample	Collection	Period	Infrastructure	Traffic	Other	Period	Type			
Alhajyaseen (2015)	kinetic energy, PET	-	video records	3 hours	-	-	-	6 years	Severe	17,520	Sensitivity Analysis, Exponential Relationships	5 urban intersections
Fu & Sayed (2021a)	DRAC	2,202 events	video records	15 hours	-	-	-	3 years	Rear-end, daytime	1,752	Bayesian hierarchical extreme value model	4 signalized intersections
Fu & Sayed (2021b)	TTC, MTTC, PET, DRAC	7,998 conflicts	video records	24 hours	-	traffic volume, shock wave area, platoon ration	-	3 years	Rear-end, daytime, good weather	1,095	Random Parameters Bayesian hierarchical extreme value model	4 signalized intersections
Johnsson et al. (2021)	mTTC, PET	-	video records	24 hours	-	traffic volume	country	7 years	Between cyclists and motor vehicles	2,555	GLM (NB)	9 signalized intersections
Mukherjee & Mitra (2020)	PET	187,174 crossing behaviours	video records	6 hours	pavement marking, night visibility street light	traffic volume, pedestrian traffic, overtaking tendency, speed	land use, zebra cross. following, cross/wait time, cross difficulty, population, attraction zone, residential area	6 years	Fatal Pedestrian	8,760	GLM (NB), GLM (Poisson)	110 intersections and 54 midblock segments
Wang et al. (2019)	TA, PET, mTTC, MaxD	-	video records (UAV)	4 hours x10 inters.	-	-	-	5 years	Angle, Rear-end	1,095	Bivariate extreme value model	10 urban signalized intersections
Zheng et al. (2019)	TTC, MTTC, PET, DRAC	-	video records	2 hours x 4 inters.	-	-	-	3 years	Rear-end, daytime	3,285	Bivariate extreme value model	4 signalized intersections
El-Basyouny & Sayed (2013)	TTC	-	conflict survey	8h x 2 days	class, right turn	traffic volume	-	3 years	Total	1,643	Two-phase model: Lognormal (conflicts) - GLM (NB) (crashes)	51 signalized intersections

The term traffic conflict denotes an observable event that would end in a road crash unless one of the involved road users slows down, changes lane, or accelerates to avoid a collision (Risser, 1985). Based on Table 2.1, it is demonstrated that the collection of traffic conflict-related SSMs under real road conditions in the majority of the examined studies is based on video recordings (Alhajyaseen, 2015; Zheng et al., 2019; Wang et al., 2019; Mukherjee & Mitra, 2020; Johnsson et al., 2021; Fu & Sayed, 2021a; Fu & Sayed, 2021b). Conflict surveys through field observations are another option for collecting such data (El-Basyouny & Sayed, 2013). When real vehicle trajectories and speeds are not available, simulation models are widely used (Gettman, & Head, 2003; Saccomanno et al., 2008). However, simulation studies fall outside the scope of this literature review research and are not discussed further.

Among the different traffic conflict-related SSMs used in the reviewed studies, it can be observed that PET, TTC, and DRAC are the most widely used. According to Gettman and Head (2003), PET is defined as the time elapsed between the encroachment's end of the turning vehicle and the time that the trough vehicle reaches the potential point of the crash, while TTC corresponds to the expected time for two vehicles to collide if they maintain their present speed and path. Various modifications of the TTC have been used in the examined studies such as the minimum TTC (mTTC) (Wang et al., 2019; Johnsson et al., 2021), which corresponds to the TTC's lowest values obtained, and the modified TTC (MTTC) proposed by Ozbay et al. (2008) that takes into account relative position, relative speed and relative acceleration of the conflicting vehicles (Zheng et al., 2019; Fu & Sayed, 2021b). Lastly, DRAC corresponds to the minimum deceleration rate required by the following vehicle to come to a timely stop (or match the leading vehicle's speed) and hence to avoid a crash (Zheng, & Sayed, 2019). However, a frequent issue encountered in such studies and also identified by a relevant study is that the safety thresholds of conflicts vary by traffic environment type and the application purposes of conflict measures (Arun et al., 2021b).

According to Lu et al. (2014), connected vehicles are the key to the evolution of next-generation intelligent transportation systems. In addition, they are expected to bring multiple benefits to driving behaviour monitoring tools as well (Ziakopoulos et al., 2020). Table 2.1 reveals that, when utilized, connected vehicles are an additional emerging option for studies exploiting SSMs for historical crash record investigations and can be a standardized, streamlined, and seamless collection source of both harsh event and traffic conflict data (Xie et al., 2019; Hu et al., 2020; Yang et al., 2021).

Regardless of how SSMs are collected, in most of the studies reviewed, the type of historical road safety data used is either the absolute number of total crashes or the number of total road crashes divided by a risk exposure indicator such as the number of vehicles or vehicle kilometers traveled (Lu et al., 2011; Mousavi et al., 2019; Li et al., 2021b). Furthermore, the severity of road crashes is not taken into account in the majority of the studies included in Table 2.1. However, there are certain studies that

focus on serious or fatal road crashes (Alhajyaseen, 2015; Mukherjee & Mitra, 2020). Several studies attempt to correlate SSMs with specific road crash types such as rear-end, angle and single-vehicle crashes (Wang et al., 2019; Ambros et al., 2019; Hunter et al., 2021; Yang et al., 2021). Other research studies focus on specific road crash characteristics such as the weather conditions, and the time or the day of the crash, which usually correspond to the conditions of SSM collection (Paleti et al., 2017; Fu & Sayed, 2021b). Moreover, the historical crash records of some other studies target specific road user types such as vulnerable road users (Strauss et al., 2017; Li et al. 2021a; Johnsson et al., 2021) and drivers of various transport modes (Khorram et al., 2020).

Lastly, in addition to the SSMs and historical crash data, most of the examined studies in Table 2.1 include some supplementary variables that are mainly related to road infrastructure and traffic. Among these variables, road length and road class prevail for infrastructure, while traffic volume and speed prevail for traffic parameters.

2.3.2 Modelling Approaches

This subsection of the review process gives a brief overview of the various modelling approaches implemented in the reviewed studies that are presented in Table 2.1 and exploit SSMs for historical crash records. Initially, it can be observed that some studies are only limited to different correlation methods, such as Pearson or Spearman correlation, which aim to measure the strength of association between SSMs and road crashes (Kim et al., 2016; Strauss et al., 2017; Stipancic et al., 2018b; Xie et al., 2019). Certainly, correlation matrices are also included in other studies as a preliminary step before the development of more advanced statistical models (Khorram et al., 2020; Stipancic et al., 2021; Hunter et al., 2021; Li et al., 2021a).

GLMs have been implemented widely in the road safety literature for many years, as they assume that crashes are independent, random, and sporadic countable events (Dobson & Barnett, 2018). Based on Table 2.1, it is observed that Poisson (Guo et al., 2010; Mukherjee & Mitra, 2020; Hunter et al., 2021) and NB models (El-Basyouny & Sayed, 2013; He et al., 2018; Ambros et al., 2019; Mousavi et al., 2019; Khorram et al., 2020; Park et al., 2021; Johnsson et al., 2021) are the most common forms of GLMs among studies exploiting SSMs for historical crash record investigations, with NB models being more prevalent than Poisson models. The key difference between these two GLM forms has to do with the fact that NB models relax the equal mean and variance assumption of the Poisson model, which can account for overdispersion resulting from unobserved heterogeneity and temporal dependency (Lord & Mannering, 2010). Specific research documents among the reviewed studies have also introduced random effects to GLMs in order to extend them to Generalized Linear Mixed Models (GLMMs) and account for unobserved heterogeneity (Pande et al., 2017; Kamla et al., 2019).

Several of the reviewed studies have also attempted to incorporate into their analyses the effects of various road safety indicators' spatial characteristics. Bayesian approaches are widely used to consider the spatial correlation for modelling crash frequencies. In that context, Li et al. (2021a) developed a Bayesian NB model with conditional autoregression (CAR) prior to accounting for spatial correlation between neighbouring bus stops. The results of this research indicated the necessity of considering spatial autocorrelation during the crash frequency model process as the developed Bayesian NB-CAR model outperformed the Bayesian model in terms of various model evaluation metrics. In another study, both the spatial and temporal dependence of crash observation were taken into account in a multivariate conditional autoregressive (MVCAR) model in the full Bayesian framework (Yang et al., 2019).

Yang et al. (2021) proposed a new safety measure termed Risk Status, which was modeled as a latent variable in a Structural Equation Model in the Bayesian framework that could account for both spatial autocorrelation through CAR spatial effect and unobserved heterogeneity through road segments random parameters (i.e., SEM-CAR-RP). Overall, SEM is a powerful multivariate tool for jointly modelling interrelationships among observed and latent variables (Washington et al., 2020). However, the proposed approach of SEM-CAR-RP extends the methodological frontier of SEM applications in the field of road safety as it was found to be superior compared to more traditional alternatives of SEMs that did not take into account CAR spatial effect and unobserved heterogeneity. This finding demonstrates that various fundamental methodological issues of crash data modelling such as spatial autocorrelation, unobserved heterogeneity, etc. need to be investigated when exploring data from new data sources similar to those that were presented in Section 2.3.1. Paleti et al. (2017) developed a random parameter Generalized Ordered Response Probit (GORP) model which is a type of model that can easily handle over or under-representation of multiple count outcomes at the same time without demanding a hurdle or zero-inflated model. The outcomes of this research revealed that the best-performing model was one including measurement error, random parameter heterogeneity, and spatial dependency.

In a more straightforward approach, Li et al. (2021b) utilized a line-constrained clustering method that combines Density-Based Spatial Clustering of Applications with Noise (DBSCAN) with spatial selection functions in order to identify individual-specific risky road segments. Latent Gaussian Models (LGMs) are a subcategory of structure additive models, in which the dependent variable for each subject follows a distribution from the exponential family and can introduce temporal or spatial dependence (Blangiardo & Cameletti, 2015). This spatial modelling approach using the Integrated Nested Laplace Approximation (INLA) technique has been chosen as an appropriate tool for road network screening (Stipancic et al., 2018a; Stipancic et al., 2019; Stipancic et al., 2021). The INLA approach was introduced by Rue et al. (2009) as a computationally efficient alternative to Markov chain Monte Carlo methods. INLA can be combined with the Stochastic Partial Differential Equation (SPDE) approach

proposed by Lindgren et al. (2011) in order to implement spatial and spatio-temporal models for point-reference data (Lindgren & Rue, 2015).

Extreme Value Theory (EVT) is a statistical approach that enables extrapolation from observed levels to unobserved levels (Coles, 2001), which is in alignment with the goal of predicting less frequent road crashes from more frequent traffic conflicts. EVT Models are becoming increasingly popular with substantial developments achieved recently. These models are mainly used to estimate the number of road crashes and then compare them to the observed historical crash records. Among studies presented in Table 2.1, bivariate EVT models have been proposed and it was found that this approach generated more accurate crash estimates than univariate models (Zheng et al., 2019; Wang et al., 2019). In a more recent study, Fu and Sayed (2021a) developed a Bayesian hierarchical extreme value model, which had three layers: the data layer, the process layer, and the prior layer. However, as also mentioned for different other model types and highlighted by Zheng et al. (2021), one important issue while developing such models is accounting for the unobserved heterogeneity across different observation locations. In order to deal with the issue, Fu and Sayed (2021b) propose a random parameters Bayesian hierarchical extreme value modelling approach.

As can be observed in Table 2.1, traditional modelling approaches such as linear or logistic regression models have been used in a few studies exploiting SSMs for historical crash record investigations, but are less preferred (Lu et al., 2011; Alhajyaseen, 2015; Desai et al., 2021; Guo et al. 2021). This is partly also due to the emergence of ML and Deep Learning (DL) approaches as powerful tools that are gaining more ground for road safety analyses due to their ability to handle large volumes of data, their heightened predictive capabilities, and the complex, non-linear relationships they can disclose. Indicatively, the random forest algorithm is a data-mining tool that has been used to determine the importance of the variables and includes in the statistical models the variables with the strongest impacts on road crashes (Guo et al., 2021; Park et al., 2021). Furthermore, Hu et al. (2020) exploited SSMs derived from connected vehicles' data such as harsh braking, harsh acceleration, and wait time in order to predict the crash risk at intersections using DL approaches. Their analyses revealed that the performance of two black-box DL models, Multi-Layer Perceptron (MLP) and convolutional neural network (CNN) was slightly better than the Decision Tree Model. However, in the context of the examined studies it can be perceived that ML/DL approaches are not among the most prevalent methods at present.

In summary, various modelling approaches have been implemented in the reviewed studies. However, the selection of an appropriate modelling framework depends highly on the research questions being asked, the available data, and the specific context of each study. Specifically, the type of crash data being analyzed (e.g., count data, rates

such as crashes divided by an exposure parameter, categorical/binary data, etc.), the level of spatial and temporal dependence, and the existence of unobserved heterogeneity are some factors that should be taken into consideration towards the selection of a suitable modelling methodology. While there are many different modelling approaches available in the literature, they should be treated as starting points for road safety practitioners, rather than definitive guides.

2.3.3 Temporal Dimension

When examining Table 2.1, no clear pattern can be observed with regard to the time periods of historical road crash data and SSMs collection. This is a constant topic, and researchers have to anticipate and plan accordingly in the study design process. Therefore, in this section, the authors attempt to shed light on this issue and identify potential hidden patterns through the visualization of the respective data in Table 2.1. As already mentioned in previous parts of the current review process, there are different ways that can be used to extract SSMs. It is observed that in studies using smartphones, instrumented vehicles, or connected vehicles the time period for which the SSMs were collected can vary from a few days (Pande et al., 2017; Park et al., 2021) to several months (Boonsiripant et al., 2011; Paleti et al., 2017; Kamla et al., 2019).

On the other hand, SSMs collected through video recordings or conflict surveys are collected for a few hours (Alhajyaseen, 2015; Wang et al., 2019). As per the aforementioned, this discrepancy was also one of the main incentives for calculating the “Temporal Ratio” column of Table 2.1. The difference in time periods between the collection of historical road crash data and SSMs is mainly attributed to the emergence of new technologies, which allow for the rapid collection of SSMs data and the conduction of analyses with shorter time periods. The “Temporal Ratio” column could be interpreted as by how much more time is needed to collect an equivalent sample of SSMs with road crash data. For this reason, as well as for readability reasons, two different graphs have been produced. Specifically, Figure 2.2 demonstrates the time periods of historical road crash data and SSMs collected through smartphones, instrumented vehicles, and connected vehicles, while Figure 2.3 presents the respective values for the studies that used video records or conflict surveys for the extraction of SSMs.

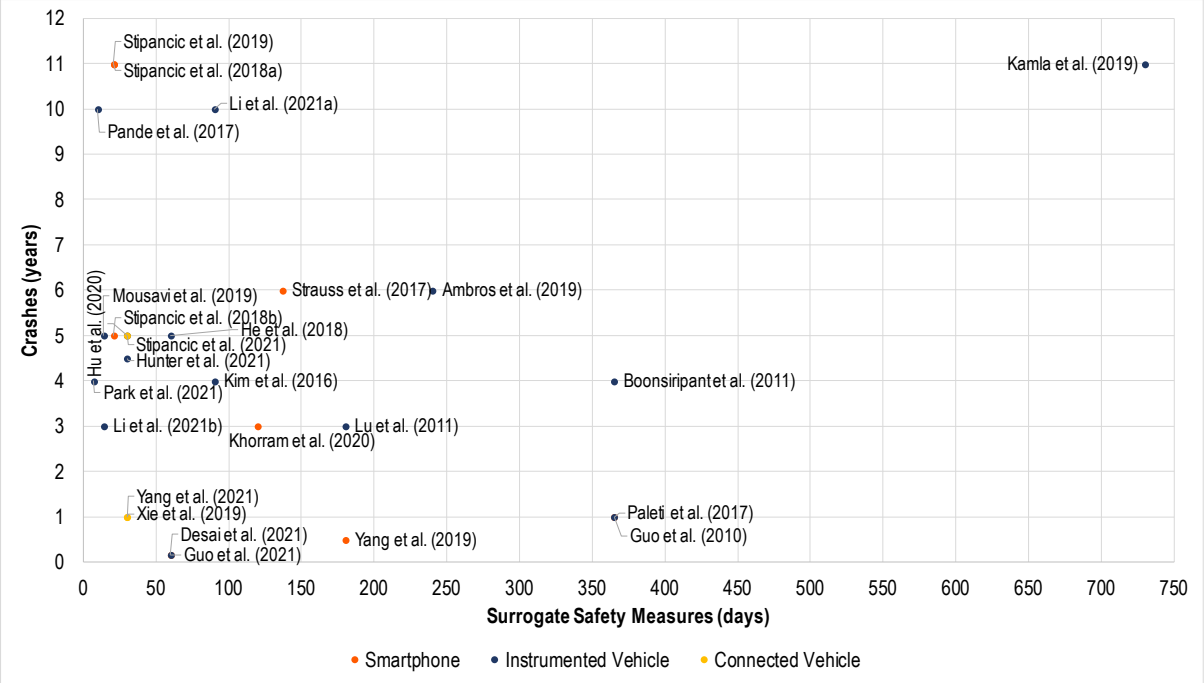


Figure 2.2: Time periods of historical road crash data and SSMs collected through smartphones, instrumented vehicles and connected vehicles

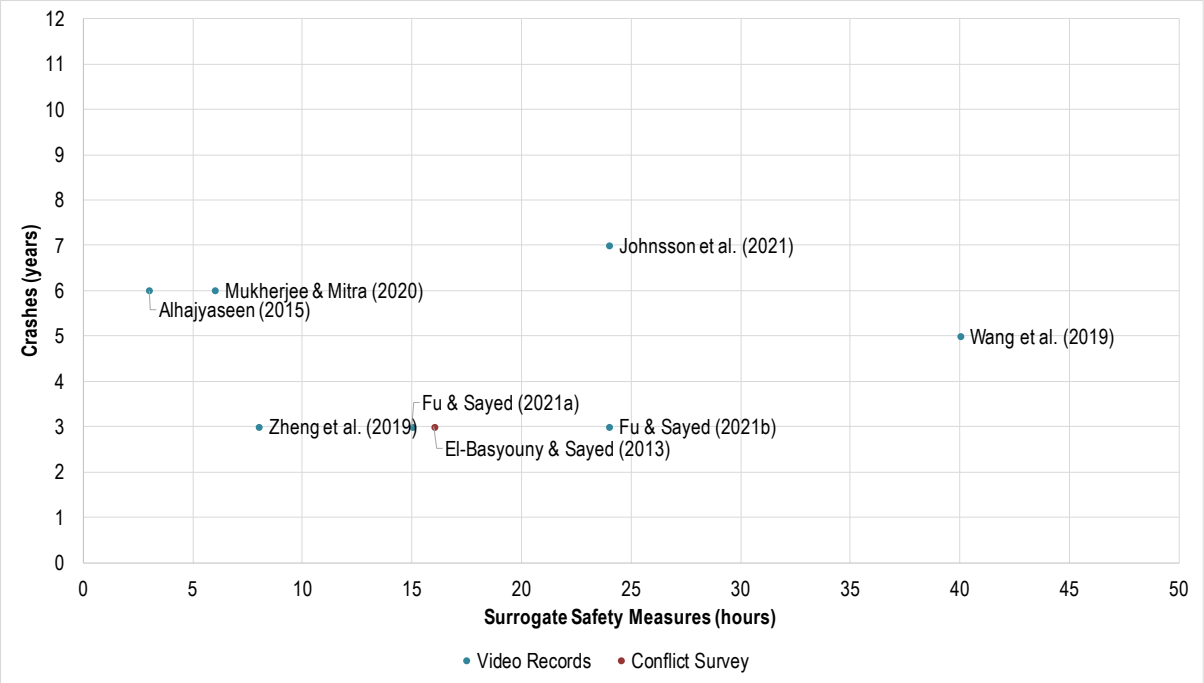


Figure 2.3: Time periods of historical road crash data and SSMs collected through video records and conflict surveys

Based on Figures 2.2 and 2.3, a general trend that can be observed is that among all the examined studies the time period of road crash data is always greater than or equal to the time period of collection of SSMs, as expected from the increased usability that SSMs provide. Furthermore, regardless of the manner in which SSMs are collected, it is observed that in the majority of the examined studies (21 out of 34), historical road crash data used correspond to periods of three to six years.

Only five of the examined studies, use exactly the same time periods of historical crash data and SSMs. These studies exploit smartphones (Paleti et al., 2017; Yang et al., 2019; Guo et al., 2021) and instrumented vehicles (Guo et al., 2010; Desai et al., 2021) for the extraction of SSMs. It can be observed that they are concentrated in the low spectrum of the Y-axis of Figure 2.2 as the crash data that they include in their analyses do not exceed one year. The highest ratio of road crash data time period to the time period of SSMs corresponds to the studies presented in the upper left part of Figure 2.2 (Pande et al., 2019; Stipancic et al., 2018a; Stipancic et al., 2019; Park et al., 2021). In particular, in these studies, the road crash data time period is calculated to be between 191 and 365 times longer (mean: 239, st.dev: 84.4) than the SSM time periods. The vast majority of the studies presented in Figure 2.2 are concentrated in the middle level of the Y-axis and towards the left side of the X-axis. In these studies, the time period of road crashes is estimated to be between 12 and 130 times longer (mean: 50, st.dev: 36.3) than that of the SSMs. In addition, there are also some studies located in the central and upper right part of Figure 2.2 for which the time period of road crashes is 4–9 times longer than that of SSMs (mean: 7, st.dev: 2.3) (Lu et al., 2011; Boonsiripant et al., 2011; Ambros et al., 2019; Kamla et al., 2019; Khorram et al., 2020).

Lastly, the comparison between Figure 2.2 and Figure 2.3 reveals that the ratio of road crash data time period to the time period of SSMs is much higher in the studies that collect SSMs through video records or conflict surveys compared to the other studies. This is due to the fact that the collection of SSMs through video recordings or conflict surveys requires only a few hours and the historical crash records correspond to time periods of at least three years, lending further credence to the utility of SSMs due to their rapid data collection.

2.4 Discussion

2.4.1 Overall Findings and Trends from Reviewed Studies

SSMs are steadily gaining ground in the road safety literature as they are a sustainable way of gauging road safety and allow the conduction of analyses without necessarily requiring historical road crash records. Moreover, the rapid and continuous progress in the field of technology makes it increasingly easier to collect such indicators. However, SSMs can also be combined with data from historical road crash records in order to complement and provide additional information to relevant road safety analyses. The present research focused on studies that exploit real-condition SSMs for historical crash record investigations.

The examination of the studies in the framework of this literature review has revealed some noteworthy conclusions for road safety analyses that combine SSMs and historical crash data. It appears that the technological development in recent years has significantly contributed to making smartphones a key choice for collecting SSMs (Strauss et al., 2017; Paleti et al., 2017; Stipancic et al., 2018a; Stipancic et al., 2018b; Stipancic et al., 2019; Yang et al., 2019; Khorram et al., 2020; Guo et al., 2021). The indicators collected through smartphones' sensors can be quite similar to those collected by instrumented vehicles (Ambros et al., 2019; Hunter et al., 2019; Stipancic et al., 2021). However, the cost of collecting SSMs via smartphones is significantly lower compared to that of instrumented vehicles. A fact that is also reflected by the increase in the use of smartphones in the relevant studies during the last five years.

The majority of SSMs collected through either smartphones or instrumented vehicles involve harsh driving behaviour events. Through these studies, it becomes clear that the most commonly exploited harsh driving behaviour events such as harsh braking and harsh acceleration events are positively correlated with various types of road crash counts (Pande et al., 2017; Stipancic et al., 2018a; Stipancic et al., 2019; Yang et al., 2019; Ambros et al., 2019; Mousavi et al., 2019; Khorram et al., 2020; Li et al., 2021a; Desai et al., 2021; Hunter et al., 2021) and road crash risk (Guo et al., 2021). As this relationship is verified by several studies, it can be deduced that harsh events could be used as dependent variables in statistical models as a proactive approach that does not require the collection of historical road crash data. Another approach used to collect SSMs is based on traffic conflicts. As for real road conditions, the collection of relevant indicators is mainly carried out through the analyses of video recordings (Alhajyaseen, 2015; Zheng et al., 2019; Wang et al., 2019; Mukherjee & Mitra, 2020; Johnsson et al., 2021; Fu & Sayed, 2021a; Fu & Sayed, 2021b). As with the SSMs collected through smartphones or instrumented vehicles, the reviewed studies based on traffic conflict indicators aimed either to investigate the relationship between the produced SSMs and historical crash counts or to predict the number of road crashes and then compare it with the observed crash counts.

Regarding the type of statistical analyses used in studies that combine SSMs and historical road crash data, GLMs including their various modifications dominate. There are also several studies that choose more specialized approaches to take into account unobserved heterogeneity and spatial dependence as they are among the most prevalent methodological issues typically faced when dealing with crash data modelling. Another common approach chosen by the reviewed studies concerns the different variants of EVT. Finally, it can be observed that ML techniques are not often used in the reviewed studies. Overall, the research questions, data type, and specific contextual factors of each study are critical to the choice of the respectively developed modelling framework.

Finally, a key finding of this literature review that could be also highlighted as its most significant contribution relates to the time periods for which both the historical road crash data and the SSMs are collected. Until recently, it was not clear if there was any particular pattern. This research sheds light on this topic by revealing that in most studies that collect SSMs via smartphones and instrumented or connected vehicles, road crash data correspond on average to time periods that are 50 times longer than the collection periods of the SSMs. In cases of collection of the alternative indicators through video recordings, the time period of crash data is significantly higher than the respective period of collection of SSMs.

2.4.2 Future Research Directions

This subsection outlines research directions that do not appear to be sufficiently investigated from the present literature of studies exploiting SSMs for historical crash record investigations and can form meaningful upcoming research endeavors. An important aspect of road safety analyses is the level of injury severity of road crashes. However, it is observed that in the majority of the studies, severity has not been adequately investigated as they mainly exploit the total number of all injury road crashes without taking into account the different severity levels. However, there are a small number of studies that focus on serious or fatal road crashes (Alhajyaseen, 2015; Mukherjee & Mitra, 2020). The inclusion of the level of injury severity in similar studies would be highly interesting for the quantification and the comparative assessment of the relationship between SSMs and different crash severity levels. Injury severity estimation using SSMs is also highlighted as a critical research need by Arun et al. (2021a). In that direction, a few recent research studies have attempted to estimate crashes by severity level using different SSMs (Goyani et al., 2021; Arun et al., 2022).

Furthermore, most of the reviewed studies focus on road crashes involving all road users without separating them. However, there are some specific types of road users such as pedestrians, pedal cyclists, and motorcyclists that are considered vulnerable road users (VRUs), as they are prone to injury in any vehicular collision, primarily

because there is little or no external protective device that could absorb the impact of a road crash (Yannis et al., 2020). It is estimated that VRUs account for half of all road fatalities globally (World Health Organization, 2023). Moreover, the noteworthy increase in the use of new micromobility transport modes such as e-scooters in many cities around the globe has raised particular concerns for the safety of these emerging types of VRUs (Karpinski et al., 2022). Therefore, more research is needed on the manner in which various SSMs could be exploited to enhance the safety of VRUs. Towards this direction, Ali et al. developed a Bayesian Generalized EVT model in order to estimate real-time pedestrian crash risks at signalized intersections using Artificial Intelligence-based video analytics (Ali et al., 2023).

Regarding the spatial scale of the analyses, it appears that the examined studies focus on the microscopic level as they mainly investigate road segments and intersections. Another promising research direction would be the application of analyses at a more macroscopic level such as regional areas (cities, metropolitan areas, local administrative units, etc.). In such cases, apart from different SSMs and road crash rates, various demographic, socioeconomic, and traffic exposure factors of the examined areas could be taken into consideration in the analyses. However, it is important to note that as the size of the examined area increases, capturing unobserved heterogeneity becomes more challenging (Wang et al., 2016). Apart from demographic and socioeconomic factors, key road safety performance indicators reflecting the safety of road users (seatbelt and helmet use, speeding, driving under the influence of alcohol, distraction), infrastructure, vehicles, and post-crash response in the examined regional areas could be also taken into account.

Over the last years, ML models have been proven to be very efficient prediction tools, making them also particularly popular in road safety analyses. ML and DL approaches have come to challenge the hitherto dominance of traditional modelling approaches by being implemented alongside or instead of them. Based on the results of this literature review research, it appears that these approaches have not found frequent application in studies that exploit SSMs for historical crash record investigations. This could be attributed to the major challenge of interpreting the results generated by the respective algorithms accurately. However, this issue could be tackled by using model agnostic methods such as the SHAP values and Local Interpretable Model-Agnostic Explanations (LIME) that would explain the interpretation of the model regardless of the model type. Furthermore, hybrid modelling approaches integrating both statistical and ML techniques could be considered in future research studies, as this framework represents a methodological advancement in traffic conflict-based crash estimation models (Hussain et al., 2022).

Lastly, the aforementioned future research directions can all be further augmented by the constant improvements in the technological field such as the further exploitation of smartphone data that can provide a vast amount of driving big data under real road conditions and connected vehicles that can be used for a more connected traffic

environment. The rollout of fifth-generation networks provides a unique opportunity for creating and exploiting innovative solutions to improve communication between all transport system components and reduce road crash casualties. The application of 5G in traffic environments could be a game changer over the next years as it enhances direct communication capabilities with very low latency such as Vehicle-to-Vehicle (V2V), Vehicle-to-Infrastructure (V2I) and Vehicle-to-Everything (V2X) (Hussein et al., 2021). This framework could assist in the collection of a wealth of real-time data that can be also used for the extraction of various SSMs that could be integrated into traditional road safety analysis.

2.5 Research Questions

Based on the results of the literature review, the following research questions are formulated:

Question 1
How can infrastructure, traffic and driver behaviour data be fused and analyzed to derive meaningful conclusions for road crash risk assessment?

Question 2
a) Can harsh driving behaviour events be meaningfully considered reliable SSMs?
b) Is there a statistically significant positive correlation between harsh driving behaviour events and historical road crash records?

Question 3
Is it possible to predict the crash risk level of road segments by exploiting road geometry characteristics and driver-behaviour based SSMs, and, if so, which ML classifiers are the most appropriate?

Question 4
Are harsh braking events more pertinent than harsh accelerations in predicting the crash risk level of road segments?

Question 5
a) In the absence of highly detailed historical road crash data, how can harsh braking events be analyzed across various road environments?
b) Is there spatial autocorrelation present in harsh braking frequencies for road segments, and, if so, do spatial modelling approaches outperform their non-spatial counterparts?

Question 6
Which road infrastructure and driver behaviour parameters exhibit a statistically significant impact on the number of harsh braking events per road segment?

The following sections of this doctoral dissertation are dedicated to delivering substantial results and findings that meaningfully address these research questions.

3. Methodological Approach

3.1 General Methodological Framework

This subsection delineates the methodology employed to accomplish the objectives of this doctoral dissertation, focusing on the road crash risk assessment. This was achieved through the integration of infrastructure, traffic, and naturalistic driver behaviour data. An overarching summary of the methodological framework is briefly provided here and visually depicted in Figure 3.1. Subsequent sections delve into the theoretical background and explanatory frameworks of specific methods utilized in this dissertation.

The general methodological framework commenced with an exhaustive literature review and the formulation of precise research questions, followed by a structured sequence of actions. Initially, a comprehensive exploration of available data for detailed road safety modelling in Greece was undertaken. This led to the establishment of two distinct databases: one encompassed comprehensive data for the Olympia Odos motorway, including detailed historical road crash records, traffic attributes, road geometry characteristics, and driver behaviour data on a segmental basis; the other covered a broader road network within the Region of Eastern Macedonia and Thrace, albeit lacking detailed crash location data and traffic attributes.

Various methodologies were applied for motorway segments. These included techniques such as NB regression for developing a crash frequency model, HC to determine crash risk levels based on historical crash data and traffic attributes, and the utilization of ML classifiers such as LR, DT, RF, K-NN, and SVM. These classifiers were used for crash risk level prediction, leveraging infrastructure and driver behaviour data. A critical focus was placed on evaluating the reliability of harsh driving behaviour events as SSMs. The analyses revealed that harsh braking events could serve as reliable SSMs and as dependent variables in road crash risk assessment models, particularly when dealing with unavailable or low-quality crash data.

Subsequently, the framework extended to include the road network data of Eastern Macedonia and Thrace Region, employing harsh braking events for road crash risk assessment. This involved applying both non-spatial and spatial models to identify significant road infrastructure and driver behaviour parameters influencing harsh braking events per road segment. Ultimately, the synthesis of all the analyses carried out within the framework of this doctoral dissertation resulted in a comprehensive road crash risk assessment.

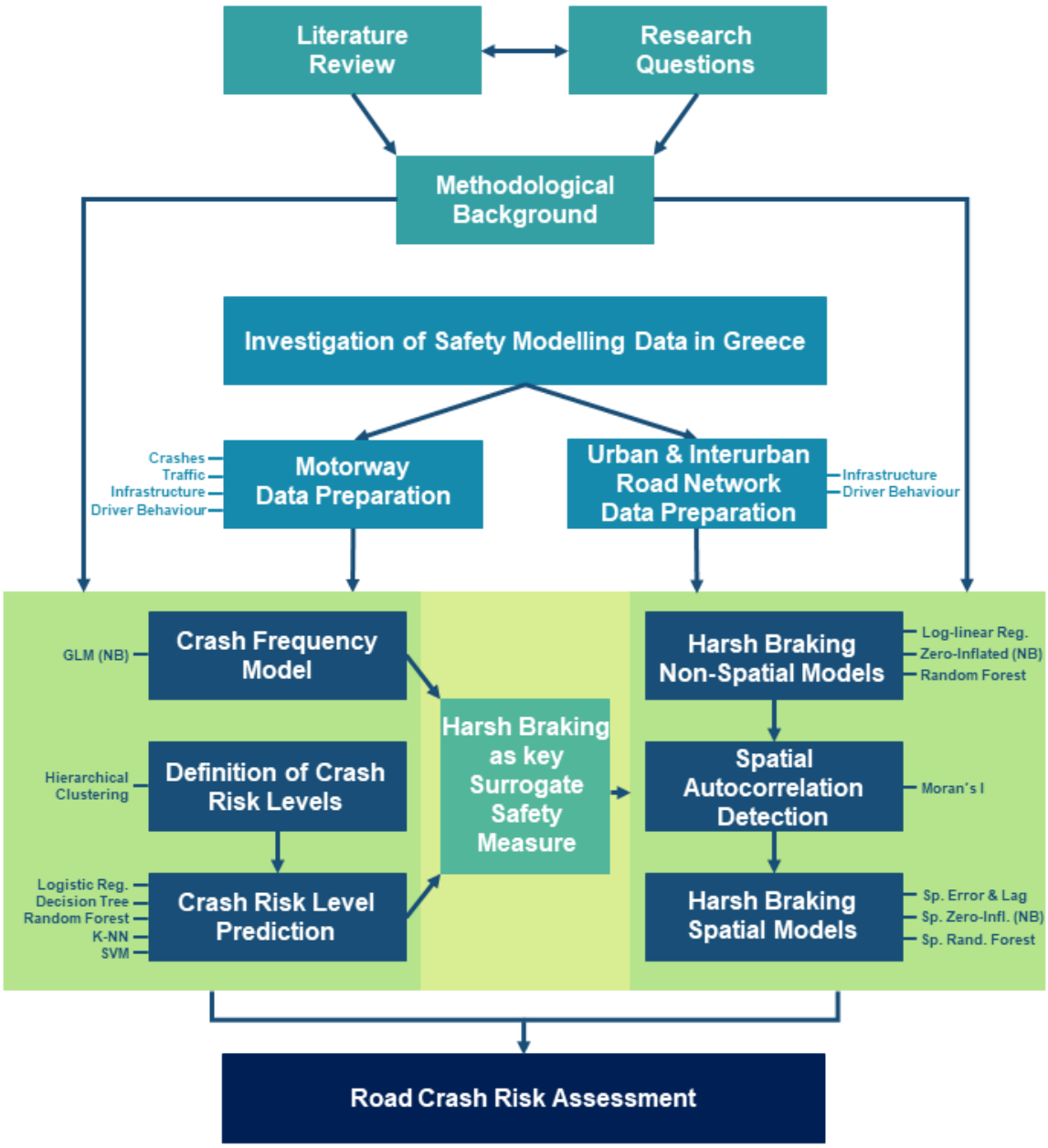


Figure 3.1: Graphical representation of the overall methodological framework of the doctoral dissertation

The above graphical illustration of the methodological approach is used to provide a broader and more comprehensive picture of the workflow that takes place and results to the better understanding on how infrastructure, traffic and naturalistic driver behaviour data can be combined for road crash risk assessment. Further details on the theoretical background and the implementation of the techniques applied in this dissertation are presented in the following sections.

3.2 Theoretical Background

3.2.1 Descriptive Analysis

This doctoral dissertation relies on big datasets, making it crucial to conduct descriptive analysis on a multitude of variables. Within this context, box plots (also known as box-and-whisker charts) offer a convenient means to illustrate numerical data groups, showcasing key parameters like minimum and maximum values, upper and lower quartiles, median values, as well as outliers and extreme values.

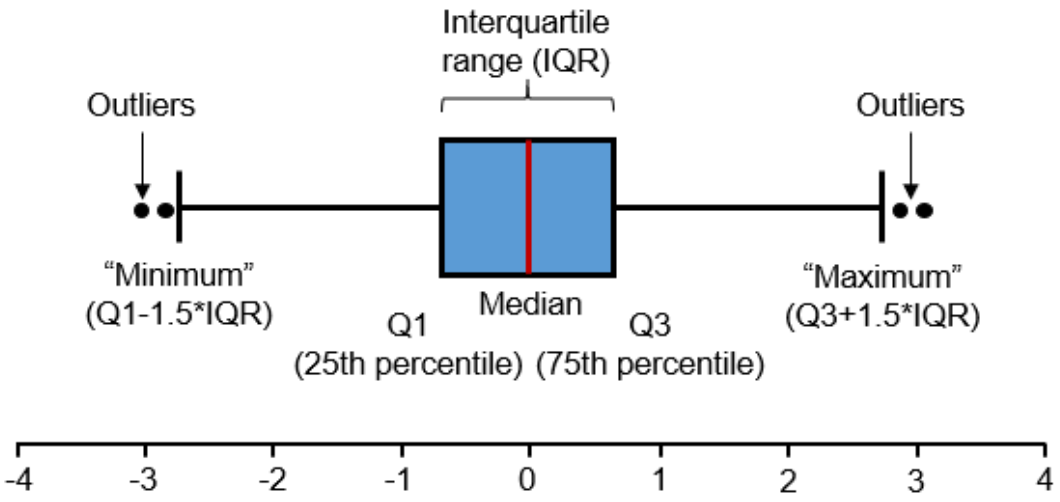


Figure 3.2: Graphical explanation of box plot

The spacing within the box plot signifies the data's dispersion and skewness, effectively pinpointing outliers. More specifically:

- The median is represented by the line in the middle of the boxes.
- The lower part of the box denotes the 25th percentile (25% of cases have values below the 25th percentile).
- The upper part of the box signifies the 75th percentile (25% of cases have values above the 75th percentile).

3.2.2 Linear Regression

In the field of statistical modelling, Linear Regression stands as a fundamental pillar, a cornerstone in understanding the relationships between variables. Linear regression is used to model a linear relationship between a continuous dependent variable and one or more independent variables. The case of one explanatory variable is called simple linear regression; for more than one, the process is called multiple linear regression (Freedman, 2009).

The simple linear regression model is given by Equation 3.1:

$$Y_i = \beta_0 + \beta_1 * X_{1i} + \varepsilon_i \qquad \text{Eq. (3.1)}$$

In this mathematical equation of the simple linear regression model, the dependent variable Y_i is a function of a constant term β_0 (the point where the line crosses the Y axis) and a constant β_1 times the value x_1 of independent variable X for observation i, plus a disturbance term ε_i. The subscript i corresponds to the individual or observation, where i = 1, 2, 3, …, n.

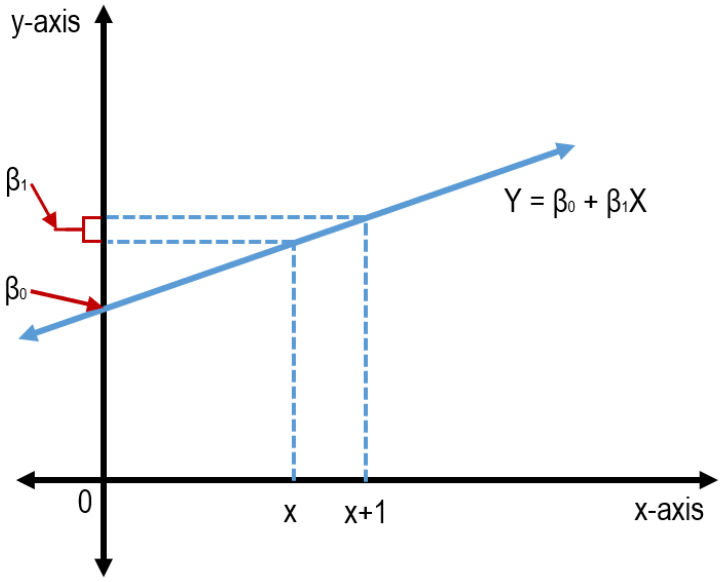


Figure 3.3: Schematic diagram of simple linear regression

Linear regression relies on several assumptions, the violation of which necessitates corrective measures or alternative modelling techniques. These key assumptions include:

- Continuous nature of the dependent or response variable.
- Inherent linearity in the relationship between variables.
- Disturbances exhibiting a mean of zero, indicating equivalence between model over-predictions and under-predictions.
- Homoscedasticity of disturbances, signifying a lack of systematic variation in model uncertainty across observations.
- Nonautocorrelation of disturbances, avoiding correlations stemming from repeated observations on individuals or spatial data exhibiting location-based dependencies.
- Exogeneity of regressors, implying the absence of correlation between the regressors and disturbance terms.
- While not mandatory for estimation, an approximately normal distribution of disturbance terms facilitates meaningful parameter inferences from the linear regression model.

For a deeper understanding and comprehensive insights, Washington et al. (2020) provide detailed elucidation on these concepts.

3.2.3 Negative Binomial Regression

However powerful in its simplicity, Linear Regression has limitations, primarily within scenarios where the outcome variable follows a normal distribution and exhibits a linear relationship with predictors. To extend regression techniques beyond these confines, Generalized Linear Models (GLMs) have emerged as a significant advancement. GLMs broaden Linear Regression's concept by accommodating various response variable types and employing a link function to establish a non-linear relationship between predictors and responses. This adaptation enables modelling diverse data types, including binary, count, and categorical outcomes.

Among GLMs, models like Poisson Regression and NB Regression offer specialized solutions for count data, where assumptions of normality or linearity might not hold. The Poisson regression makes the assumption that variance and mean are equal, which is not always the case for data such as road crashes. In many cases, such datasets have a mean that is lower than their variance meaning that some road segments concentrate more on crashes than others. To that end, Negative Binomial regression is another well-known approach that can be considered as a generalization of Poisson regression and is preferred when overdispersion exists in count data (Lord & Mannering, 2010).

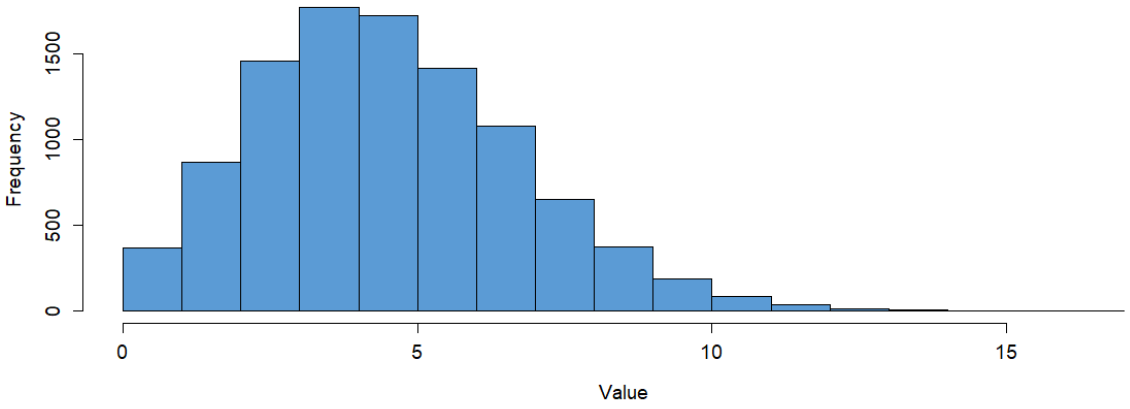


Figure 3.4: Example of Poisson distribution (mean=5, variance=5, sample=10,000)

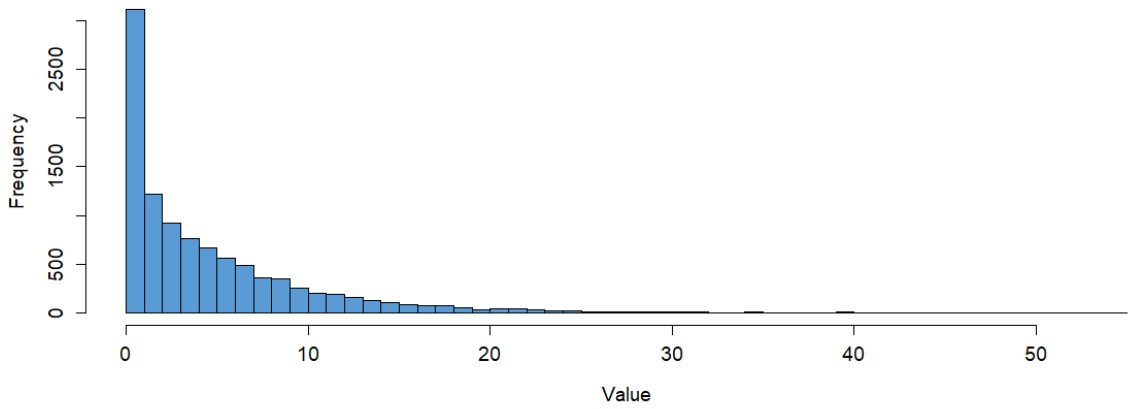


Figure 3.5: Example of Negative Binomial distribution (mean=5, variance=5, *k*=1, sample=10,000)

In the example of road crashes, based on a Poisson regression model, the probability of a road segment i having y_i crashes per some time period is given by:

$$P(y_i) = \frac{\exp(-\lambda_i)\lambda_i^{y_i}}{y_i!} \quad \text{Eq. (3.2)}$$

where λ_i is the Poisson parameter for segment i, which is equal to the expected number of crashes per period, $E[y_i]$ for segment i. The Poisson parameter λ_i needs to be defined as a function of independent variables. The most common functional form is:

$$\lambda_i = exp(\beta X_i) \qquad \text{Eq. (3.3)}$$

where X_i is a vector of independent variables and β is a vector of estimable parameters. In negative binomial distribution, the variance varies from the mean by adding the term $exp(\varepsilon_i)$to the equation (3.3):

$$\lambda_i = exp(\beta X_i + \varepsilon_\iota) \quad \text{Eq. (3.4)}$$

This extra term is a gamma-distributed error term with mean 1 and variance a that allows the variance to differ from the mean. For additional detailed explanations of the underlying statistical background, the reader can consult Washington et al. (2020).

3.2.4 Zero-Inflated Negative Binomial Regression

The ZINB regression is used for count data that exhibit overdispersion and excess zeros. The data distribution of the ZINB combines the negative binomial distribution and the logit distribution. The possible values of Y are non-negative integers such as 0, 1, 2, 3, and so on.

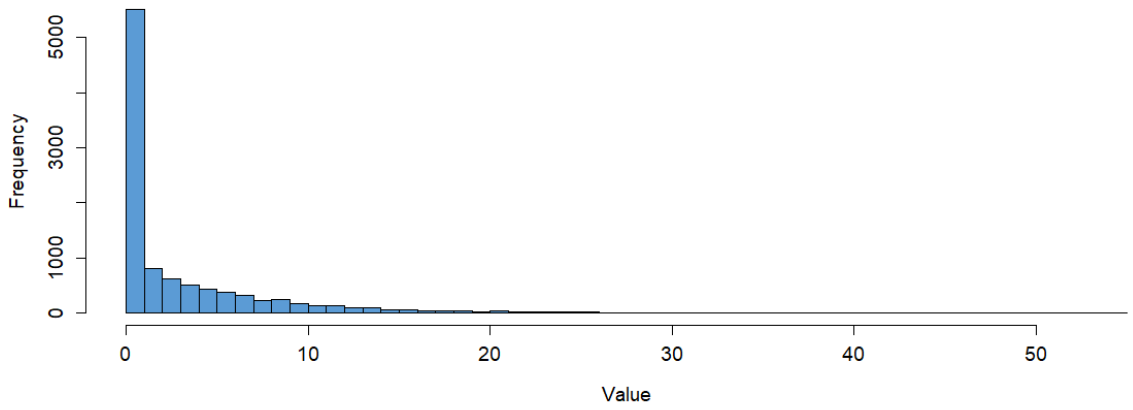


Figure 3.6: Example of Zero-Inflated Negative Binomial distribution (mean=5, variance=5, *k*=1, sample=10,000, proportion of zeros = 35%)

Suppose that for each observation, there are two possible cases. If the first case occurs, the count is zero. However, if the second case occurs, counts (including zeros) are generated according to the negative binomial distribution. Suppose that the first case occurs with probability π and the second case occurs with probability 1-π. Consequently, the probability distribution of the ZINB random variable y_i can be written:

$$\Pr(y_i = j) = \begin{cases} \pi_i + (1 - \pi_i)g(y_i = 0) & \text{if } j = 0 \\ (1 - \pi_i)g(y_i) & \text{if } j > 0 \end{cases} \qquad \text{Eq. (3.5)}$$

where π_i is the logistic link function defined below and $g(y_i)$ is the negative binomial distribution given by:

$$g(y_i) = \Pr(Y = y_i \mid \mu_i, \alpha) = \frac{\Gamma(y_i + \alpha^{-1})}{\Gamma(\alpha^{-1})\Gamma(y_i + 1)} \left(1/1 + a\mu_i\right)^{\alpha^{-1}} \left(a\mu_i/1 + a\mu_i\right)^{y_i} \qquad \text{Eq. (3.6)}$$

The negative binomial component can include an exposure time t and a set of k regressors variables (the $x's$).

The expression relating these quantities is the following:

$$\mu_i = \exp(\ln(t_i) + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki}) \qquad \text{Eq. (3.7)}$$

Often, $x_1 \equiv 1$, in which case β_1 is called intercept. The regression coefficients $\beta_1, \beta_2, \ldots, \beta_k$ are unknown parameters that are estimated from a set of data. Their estimates are symbolized as $b_1, b_2, \ldots, b_k$.

This logistic link function π_i is given by:

$$\pi_i = \frac{\lambda_i}{1+\lambda_i} \qquad \text{Eq. (3.8)}$$

where:

$$\lambda_i = \exp(\ln(t_i) + \gamma_1 z_{1i} + \gamma_2 z_{2i} + \cdots + \gamma_m z_{mi}) \qquad \text{Eq. (3.9)}$$

The logistic component can include an exposure time t and a set of m regressors variables (the $z's$). Note that the $z's$ and the $x's$ may or may not include terms in common.

For more in-depth explanations on the underlying background of the ZINB, the reader can refer to Cameron & Trivedi (2013) and Garay et al. (2011).

3.2.5 Logistic Regression

Logistic regression, despite its name, functions as a classification model, particularly suitable for analyzing data with a binary outcome variable. This model aims to estimate the probability (P) of an event occurring by considering various predictors. In logistic regression, the outcome variable denotes the presence or absence of a condition, often coded as 1 or 0.

The logistic regression equation incorporates a logit transformation, where the natural logarithm of the odds represents the relationship between the probability of an event (P) and the covariates. It is formulated as:

$$Y_i = logit\,(P_i) = LN\left(\frac{P_i}{1-P_i}\right) = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \cdots + \beta_K X_{K,i} \qquad \text{Eq. (3.10)}$$

and where β_0 is the model's constant and the $\beta_1, \ldots, \beta_K$ represent the unknown parameters associated with explanatory variables X_K. In Equation 3.10, the unknown binomial probabilities are a function of explanatory variables (which may include both continuous and discrete variables).

The estimation of unknown parameters in Equation 3.10 often employs maximum likelihood methods. Once these parameters are estimated, they're used to calculate the probability of the outcome being 1 based on the covariates:

$$P_i = \frac{EXP\,[\beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \cdots + \beta_K X_{K,i}]}{1 + EXP\,[\beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \cdots + \beta_K X_{K,i}]} \qquad \text{Eq. (3.11)}$$

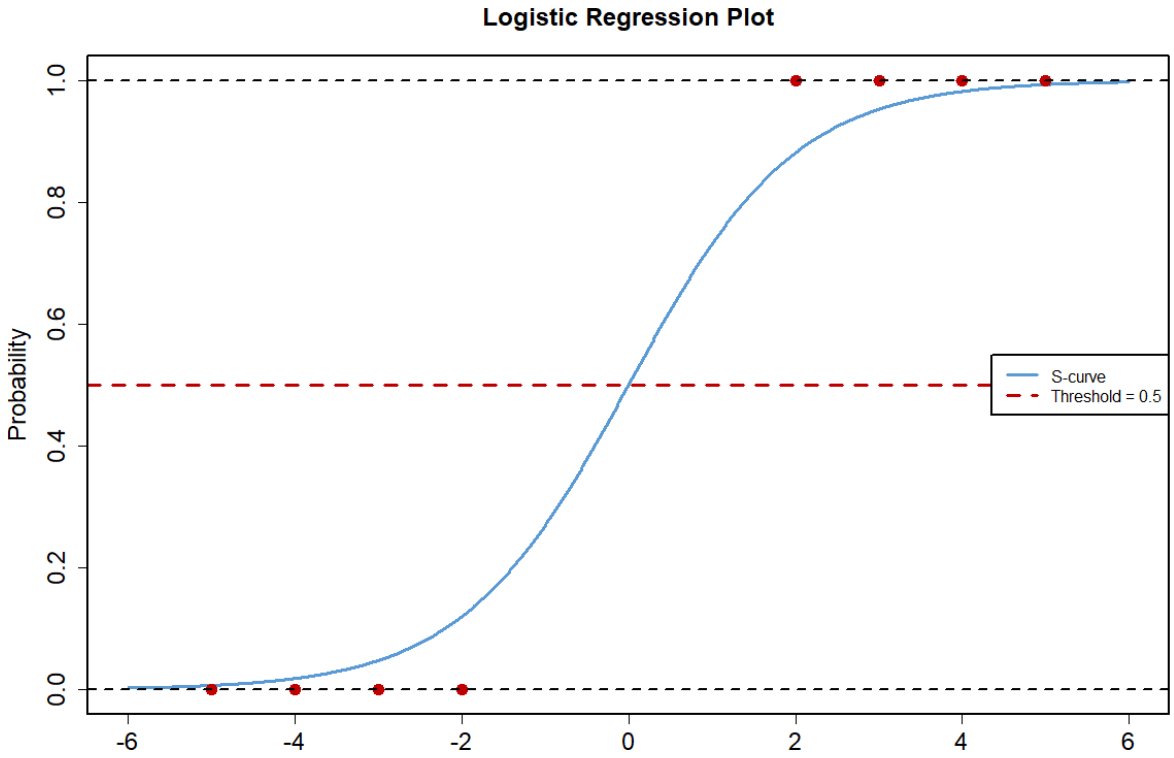


Figure 3.7: The logistic function with example data

For multiclass classification tasks, logistic regression expands its utility by employing strategies like One-vs-Rest or One-vs-All. For further information on its underlying theoretical background, the reader might consult Washington et al. (2020).

3.2.6 Decision Tree

DTs are a common class of non-parametric models that can be utilized for both regression and classification tasks and their concept was introduced by Quinlan (1986). It is noted that within the framework of this doctoral dissertation DTs were used for classification purposes. In terms of the underlying theoretical background, a DT classifier acts as a graphical representation where nodes encapsulate the features present in a dataset, branches denote potential values these features might assume, and leaves signify the resulting classification labels. These trees function on the fundamental principle of hierarchical decision-making, aiming to classify new data points by navigating through a series of decisions based on various features or attributes.

The construction of a DT involves iterative partitioning of the dataset into subsets based on the values of chosen features. This iterative process continues until specific stopping criteria are met, such as reaching a defined maximum depth or achieving a minimum reduction in impurity. The result is a structured tree that visually depicts the

decision-making process, enabling straightforward interpretation and understanding of how the data is classified.

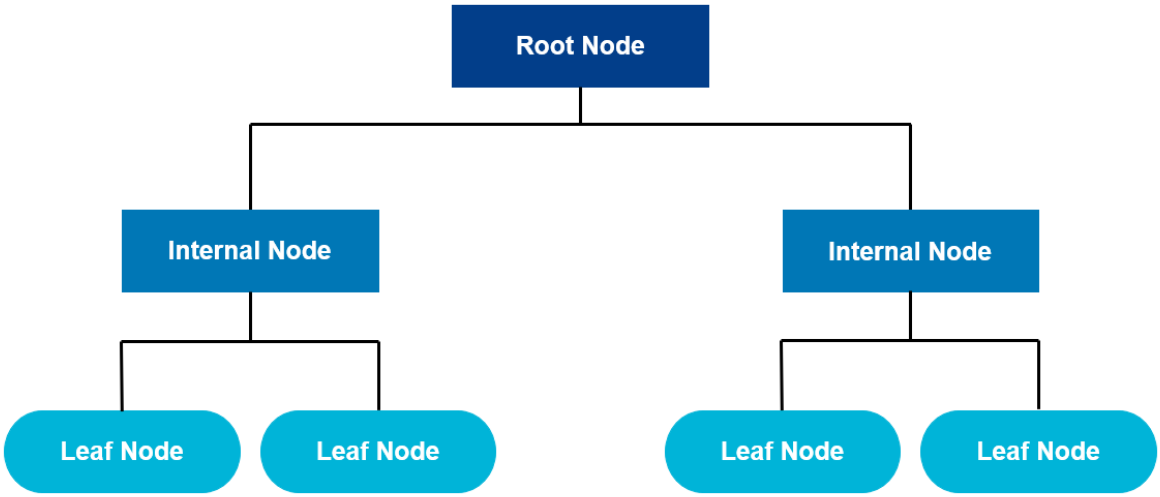


Figure 3.8: Typical hierarchical structure of a Decision Tree

Although there are several techniques to choose the best attribute at each node, information gain and Gini impurity are two approaches that are frequently used as splitting criteria for DT models. They aid in the assessment of each test condition's quality and its capacity to categorize samples into classes. Information gain is hard to describe without first talking about entropy. Entropy is a concept that stems from information theory, which measures the impurity of the sample values. The following formula defines it, where:

$$Entropy(S) = -\sum_{c \in C} p(c) log_2 p(c) \qquad \text{Eq. (3.12)}$$

- S represents the data set that entropy is calculated
- c represents the classes in set, S
- $p(c)$ represents the ratio of data points that belong to class c to the number of total data points in set, S.

Values of entropy can range from 0 to 1. When every sample in the data set S is a member of the same class, entropy is equal to zero Entropy will peak at 1 if half of the samples are categorized into one class and the other half into a different class. The attribute with the least level of entropy should be utilized to determine which feature is best to divide on and to identify the optimum DT. The difference in entropy before and after a split on a particular attribute is known as information gain. Since it is performing the best at categorizing the training data in accordance with its target classification, the attribute with the highest information gain will result in the best split. The following formula is typically used to describe information gain, where:

$$Information\ Gain\ (S, a) = Entropy(S) - \sum_{v \in vcalues(a)} \frac{|S_v|}{|S|} Entropy(S_v) \qquad \text{Eq. (3.13)}$$

- a represents a specific attribute or class label
- $Entropy(S)$ is the entropy of dataset, S
- $\frac{|S_v|}{|S|}$ represents the proportion of the values in S_v to the number of values in dataset, S
- $Entropy(S_v)$ is the entropy of dataset, S_v.

The probability that a random data point in a dataset would be incorrectly classified if its label were based on the class distribution of the dataset is represented by the Gini Impurity. Similar to entropy, the impurity of a set S is equal to zero when it is completely pure (belonging to a single class). The formula below is used to describe this concept:

$$Gini\ Impurity = 1 - \sum_i (p_i)^2 \qquad \text{Eq. (3.14)}$$

One critical challenge associated with DTs is the potential for overfitting, especially when the trees become too complex. Overfitting occurs when the model learns to fit the training data too precisely, resulting in reduced generalizability to new, unseen data. To mitigate this issue, various strategies are employed, including setting constraints on the tree's depth or complexity, employing pruning techniques to simplify the tree structure, or using ensemble methods that combine multiple trees to enhance predictive performance and reduce overfitting.

For a more comprehensive understanding of DTs, interested readers are directed to delve into Han et al. (2022).

3.2.7 Random Forest

RF, introduced by Ho (1995) and further improved by Breiman (2001), stands as a prominent ML algorithm known for its great skill in both regression and classification tasks. This algorithm operates on the foundation of DTs, an elemental component in its ensemble learning structure.

DTs, within the context of RF, function diversely for classification and regression tasks. In classification, these trees segment the dataset based on various attributes, thereby enabling the classification of instances into distinct classes or categories. On the other hand, regression trees facilitate the prediction of continuous numerical values by segmenting the dataset using specific feature thresholds.

The fundamental strength of RF lies in its ensemble learning approach. It embraces bagging, a process that involves generating numerous subsets of data by employing bootstrapping techniques from the original dataset. Subsequently, these diverse

subsets contribute to the creation of multiple DTs. The essence of RF lies in the aggregation of predictions from these varied trees, which collectively generate the final output.

Randomness assumes a pivotal role within the framework of RF. This algorithm introduces feature randomness by considering only a random subset of features at each node for the purpose of tree splitting. Furthermore, the utilization of bootstrapping ensures that each tree is trained on a distinct subset of the data, thereby enhancing the diversity and reducing the correlation between individual trees.

One more advantage of RF is that it makes use of an "out-of-bag" (OOB) estimating technique. For every tree, about one-third of the original dataset is removed during the bootstrapping procedure. Although these out-of-bag samples are not utilized for training the particular tree, they can be used to get an unbiased assessment of the model's effectiveness without requiring a different validation set. This technique provides an internal validation mechanism, offering insights into the model's generalization performance while optimizing computational resources.

In regression scenarios, the RF algorithm combines predictions from numerous trees by averaging their outputs. Consequently, this process yields a continuous prediction, ensuring a robust and reliable outcome. Meanwhile, in classification tasks, the algorithm relies on the aggregation of predictions from multiple trees to determine the mode, i.e., the most frequently occurring class prediction among the trees (majority voting).

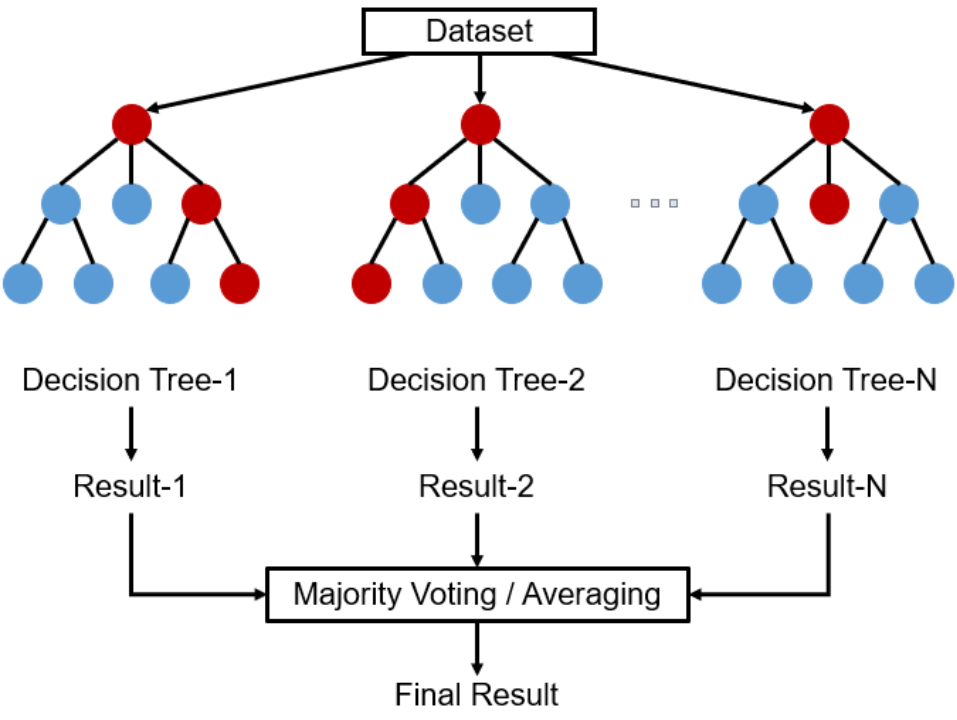


Figure 3.9: Graphical illustration of the Random Forest algorithm

The strengths of RF extend beyond its adaptability in handling diverse tasks. It boasts robustness by avoiding overfitting through the collective wisdom of multiple trees, accommodates missing data without necessitating imputation, offers insights into feature importance rankings, and exhibits scalability by efficiently processing large datasets through parallelization. However, the interpretability of complex ensembles can be challenging, hindering a comprehensive understanding of the model's decision-making process. Moreover, training multiple trees can be computationally demanding, especially when dealing with a substantial number of trees and features.

3.2.8 Support Vector Machines

SVMs stem from statistical learning theory (Vapnik, 1999) and were developed by Cortes & Vapnik (1995), primarily focusing on binary classification tasks. The key objective of constructing an SVM model is to establish an optimal dividing hyperplane between two classes by maximizing the margin, which refers to the distance between the closest points of each class (Meyer, 2001). Therefore, different classes are separated by the hyperplane:

$$\langle w, \Phi(x) \rangle + b = 0 \qquad \text{Eq. (3.15)}$$

which corresponds to the decision function

$$f(x) = sign(\langle \Phi(x_i), w \rangle + b) \qquad \text{Eq. (3.16)}$$

The support vectors encompass the points lying on the boundaries, whereas the optimum separating hyperplane is positioned at the center of the margin. The following Figure provides a graphical representation of a linear separable example of SVMs.

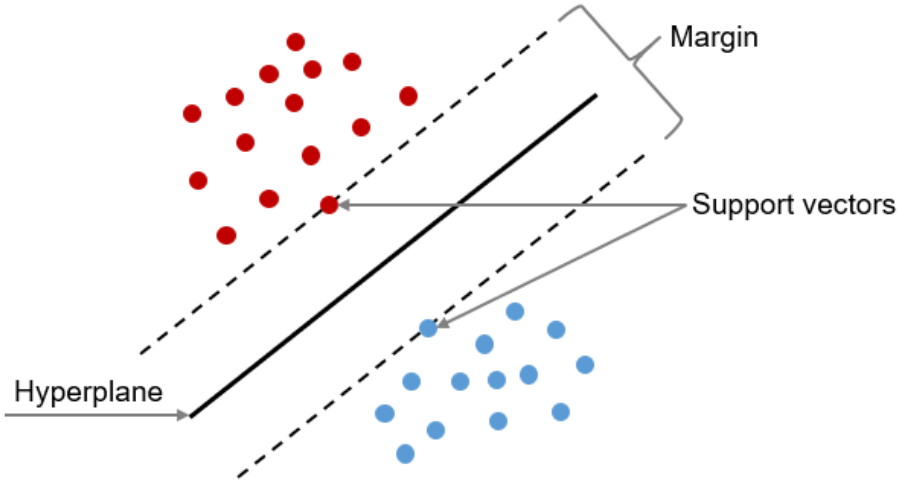


Figure 3.10: Graphical illustration of SVM classification (linear separable example)

Furthermore, SVMs can be extended to address nonlinear classification issues, regression tasks, and outlier detection. However, a significant drawback of SVMs is their inability to directly unveil the relationships between dependent and independent variables. Among the array of kernel-based algorithms (kernels) such as linear, polynomial, gaussian Radial-basis function, and sigmoid, this doctoral dissertation specifically focused on the gaussian Radial-basis function (Karatzoglou et al., 2005):

Radial-Basis Function kernel (RBF):

$$\mathrm{K}(x_i, x_j) = \exp\left(-\gamma \left\| x_i - x_j \right\|^2\right), \gamma > 0 \qquad \text{Eq. (3.17)}$$

where, γ is the kernel parameter.

Furthermore, two parameters (C, γ) of the SVM model (C-SVM) with the gaussian radial-basis kernel function need to be defined. According to Karatzoglou et al. (2006), the cost parameter C controls the penalty for incorrectly classifying a training point and, as a result, the prediction function's complexity. A complex prediction function will be produced by a high-cost value C in an effort to misclassify as few training data as feasible. Lower cost parameter C, on the other hand, leads to simpler prediction functions.

The primal form of the bound constraint C-SVM is the following:

minimize $\quad t(w, \xi) = \left(\frac{1}{2}\right) \|w\|^2 + \left(\frac{1}{2}\right) \beta^2 + \left(\frac{c}{m}\right) \sum_i^m \xi_i$

subject to $\quad y_i(\langle \Phi(x_i), w \rangle + b) \geq 1 - \xi_i \qquad$ Eq. (3.18)

where, i = 1, …, m, and $\xi_i \geq 0$.

The respective dual form of the bound constraint C-SVM is the following:

maximize $\quad W(\alpha) = \sum_{i=1}^m a_i - \frac{1}{2} \sum_{i,j=1}^m a_i a_j \left(y_i y_j + k\left(x_i, x_j\right)\right)$

subject to $\quad 0 \leq \alpha_i \leq \frac{c}{m}$, where, i = 1, …, m $\qquad$ Eq. (3.19)

and $\quad \sum_{i=1}^m a_i y_i = 0.$

For further information on the underlying background of the SVMs, the reader can also refer to Schölkopf & Smola (2002).

3.2.9 K-Nearest Neighbours

The K-NN algorithm is a non-parametric supervised learning classifier, which exploits proximity to make classifications or predictions regarding the grouping of individual data points. While applicable to both regression and classification problems, its primary use lies in classification, relying on the idea that similar points can be found very close to one another.

K-NN is a simple and intuitive classifier that assigns a label to a new data point based on the labels of its K nearest neighbours within the training set. The distance measure used to determine the nearest data points can be any of the standard related metrics, such as Euclidean distance or Manhattan distance. The value of K acts as a hyperparameter regulating model complexity, adjustable through cross-validation. K-NN can be used for both binary and multiclass classification tasks and can handle non-linear decision boundaries.

Within the scope of this doctoral dissertation, the K-NN algorithm serves classification tasks, employing the widely adopted Euclidean distance method for computing distances between data points. Euclidean distance is the most commonly used distance measure, and it is limited to real-valued vectors It quantifies a straight line between the query point and the point under measurement using the following formula:

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(y_i - x_i)^2} \qquad \text{Eq. (3.20)}$$

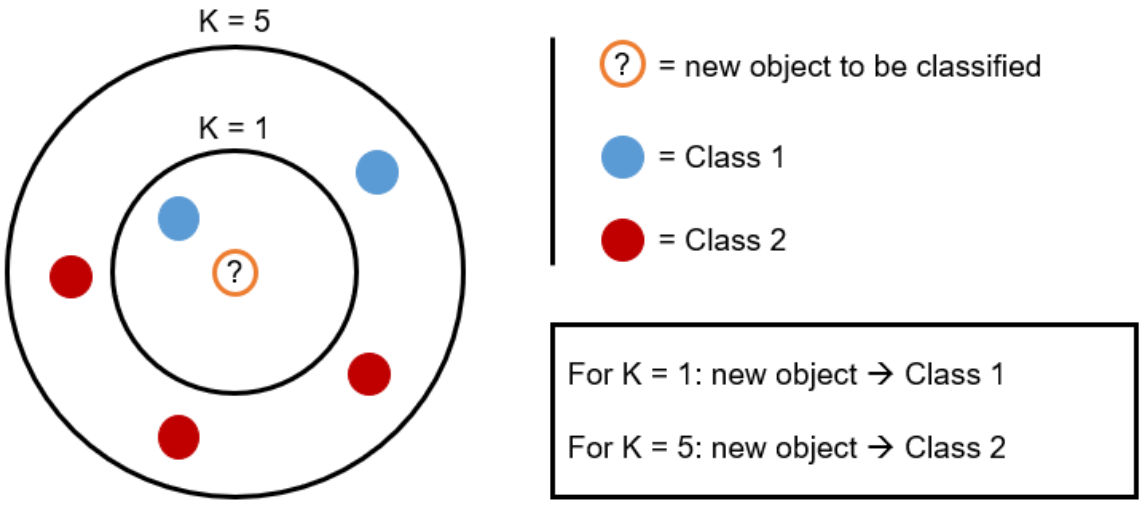


Figure 3.11: Graphical illustration of the K-NN algorithm

The reader is referred to Peterson (2009) for additional explanations on the theoretical background of the K-NN algorithm.

3.2.10 Model Evaluation Metrics

Model evaluation encompasses the utilization of diverse evaluation metrics to comprehend the performance of a machine learning or statistical model, along with identifying its strengths and weaknesses. This evaluation process holds significance in appraising a model's effectiveness during preliminary research phases and assumes a crucial role in ongoing model monitoring. Different key metrics offer insights into the model's performance, which vary depending on whether the model serves regression or classification purposes.

In regression analysis, R-Squared (R^2) or the coefficient of determination serves as a fundamental metric that determines the proportion of variance in a dependent variable predicted or explained by an independent variable. In simpler terms, R^2 indicates how well a regression model (independent variable) predicts the outcome of observed data (dependent variable). R^2 values range from 0 to 1. A value of 0 implies that the model explains or predicts 0% of the relationship between the dependent and independent variables, while a value of 1 indicates that the model predicts 100% of the relationship.

Mathematically, R^2 is calculated by dividing sum of squares of residuals (SS_{res}) by total sum of squares (SS_{tot}) and then subtract it from 1. In this case, SS_{tot} measures total variation. SS_{reg} measures explained variation and SS_{res} measures unexplained variation. As SS_{res} + SS_{reg} = SS_{tot}, R^2 = Explained variation / Total Variation.

$$R^2 = \frac{SS_{reg}}{SS_{tot}} \qquad \text{Eq. (3.21)}$$

Adjusted R^2 is a refinement that adjusts for model complexity. It measures the proportion of variation explained by only those independent variables that really help in explaining the dependent variable. It penalizes the inclusion of independent variables that do not significantly aid in predicting the dependent variable within regression analysis. The only difference between R-squared and Adjusted R-squared equation is degree of freedom.

$$R^2 adjusted = 1 - \frac{SS_{res}/df_e}{SS_{tot}/df_t} \qquad \text{Eq. (3.22)}$$

In the above equation, df_t is the degrees of freedom $n-1$ of the estimate of the population variance of the dependent variable, and df_e is the degrees of freedom $n-p-1$ of the estimate of the underlying population error variance. Adjusted R-squared value can be calculated based on value of R-squared, number of independent variables (predictors), total sample size.

$$R^2 adjusted = 1 - \frac{(1-R^2)(N-1)}{N-p-1} \qquad \text{Eq. (3.23),}$$

where R^2 = sample R-squared, p = number of predictors and N = total sample size.

The Root Mean Square Error (RMSE) is another widely employed measure to assess prediction quality. This metric indicates how far predictions fall from measured true values using Euclidean distance. To compute RMSE, one calculates the residual (difference between prediction and truth) for each data point, computes the norm of residual for each data point, computes the mean of residuals, and finally takes the square root of that mean. RMSE is commonly used in supervised learning applications, as RMSE uses and needs true measurements at each predicted data point. RMSE can be expressed as:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} \|y(i) - \hat{y}(i)\|^2}{N}} \quad \text{Eq. (3.24)}$$

where N is the number of the data points, $y(i)$ is the i-th measurement and $\hat{y}(i)$ is the corresponding prediction.

The Akaike Information Criterion (AIC) is a statistical metric widely used in regression model selection, balancing the trade-off between model complexity and goodness of fit. It serves as a tool to compare different models by considering both their performance and simplicity. AIC assigns a score to each model based on the balance between how well it fits the data and how many parameters it uses. The principle behind AIC is rooted in information theory, aiming to minimize the information loss between the model and the true underlying process it represents. Lower AIC values indicate a better trade-off between fit and complexity, suggesting a model that adequately represents the data without unnecessary complexity. It is calculated using the following formula:

$$AIC = 2k - 2\log(\hat{L}) \quad \text{Eq. (3.25)}$$

where k represents the number of parameters in the model and $\hat{L}$ denotes the maximum value of the likelihood function for the model (Akaike, 1970; Sakamoto et al., 1986).

For classification models, the first step for the evaluation of the classification performance is the development of the confusion matrix, which gives insights into the distribution of the predictions and targets. Confusion matrix is a performance measurement for machine learning classification problems where output can be two or more classes. For a binary classification scenario (two classes: positive and negative), the confusion matrix has four main components:

- True Positives (TP): These are cases where the model correctly predicted the positive class.

- True Negatives (TN): These are cases where the model correctly predicted the negative class.
- False Positives (FP): These are cases where the model predicted the positive class, but the actual class was negative (Type I error).
- False Negatives (FN): These are cases where the model predicted the negative class, but the actual class was positive (Type II error).

A core classification performance indicator is the overall classification accuracy, which is specified as the fraction of predictions that are rightly classified.

$$Accuracy = \frac{TN+TP}{TN+TP+FN+FP} \qquad \text{Eq. (3.26)}$$

While overall classification accuracy is an important measure, it may not be enough for classifiers with response variables that contain more than two classes. In such cases, precision, recall, and the F1 score are insightful per-class performance metrics that can be calculated (Grandini et al., 2020). These metrics are particularly helpful in cases of not uniformly distributed class labels. In such cases, relying solely on accuracy can be misleading because it is possible to achieve a high overall accuracy score by simply predicting the dominant class most of the time. However, this approach could lead to low precision and recall scores for the remaining categories.

Precision indicates the fraction of right predictions for a particular category, which is calculated by dividing the number of true positives by the sum of true positives and false positives.

$$Precision = \frac{TP}{FP+TP} \qquad \text{Eq. (3.27)}$$

Recall (or Sensitivity/ True Positive Rate) represents the fraction of cases of a category that were correctly predicted and is expressed by the number of true positives divided by the number of true positives plus the number of false negatives.

$$Recall = \frac{TP}{TP+FN} \qquad \text{Eq. (3.28)}$$

Specificity quantifies the proportion of true negative cases (correctly identified negatives) among all actual negative instances. It complements metrics like accuracy, precision, and recall, providing insight specifically into a model's performance regarding the true negative class.

$$Specificity = \frac{TN}{TN+FP} \qquad \text{Eq. (3.29)}$$

In addition to precision and recall, the F1 score, which is calculated as their harmonic mean, is also commonly provided.

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} \quad \text{Eq. (3.30)}$$

Finally, it is mentioned that the aforementioned per-class metrics can be averaged across all classes, resulting in the respective macro-averaged scores.

Apart from the aforementioned metrics, Receiver Operating Characteristics (ROC) curve is a graphical representation of the effectiveness of a binary classification model which plots the True Positive Rate (TPR) versus the False Positive Rate (FPR) at different classification thresholds.

$$FPR = \frac{FP}{TN + FP} = 1 - Specificity \quad \text{Eq. (3.31)}$$

Area Under the Curve (AUC) serves as a comprehensive metric for assessing the performance of a binary classification model. As both TPR and FPR range between 0 to 1, So, the area will always lie between 0 and 1, and a higher value of AUC indicates better model performance. The key objective is to maximize this area to achieve the highest TPR and lowest FPR at the given threshold. Essentially, the AUC measures the probability that the model will assign a randomly selected positive instance a higher predicted probability compared to a randomly chosen negative instance.

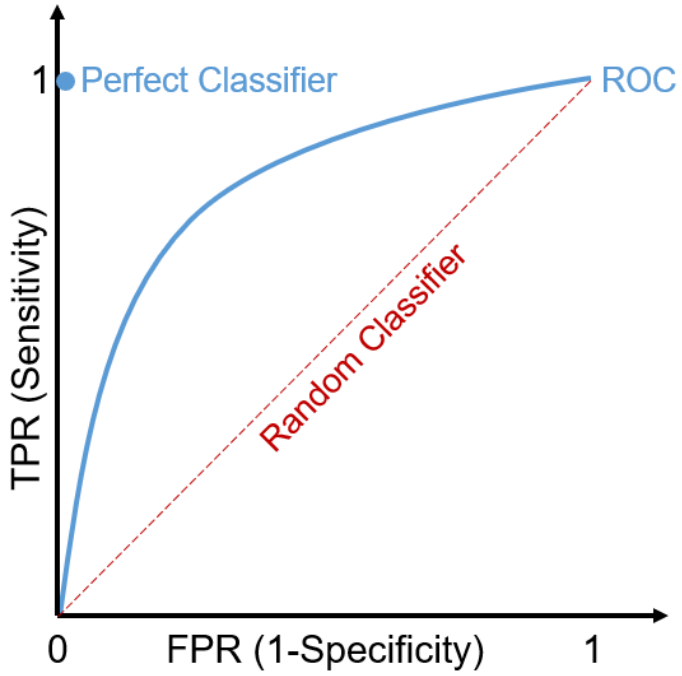


Figure 3.12: ROC curve example

For multiclass classification tasks, the One vs All methodology can be utilized, resulting in individual ROC curves for each class.

3.2.11 SHapley Additive exPlanations (SHAP values)

SHAP values are a recent addition to the field of explainable and interpretable ML, drawing from coalitional game theory (Shapley, 1953). These values provide a measure of contribution of each feature to the prediction of a particular instance in a model. The SHAP value for each feature is defined as the difference between the expected model output and the output when that feature is excluded. The SHAP values are a model-agnostic method, meaning it can be applied to explain the predictions of any machine learning model, including black-box models. In the case of multiclass classification models, SHAP values are calculated for each class separately as it allows the understanding of the contribution of each feature to the prediction of each class.

More specifically, SHAP values provide a solution to the problem wherein a group of individuals collaborates, resulting in an overall gain from their cooperation. Given that each player holds unique significance in the collaboration, determining how to distribute the surplus fairly among them becomes essential. By considering the distinct contributions of each player, Shapley values propose a potential equitable allocation of the generated surplus among the participants (Shapley, 1953).

Translating this issue into the context of a model's predictions involves regarding explanatory variables as the players and the model $f()$ as the coalition. The prediction made by the model represents the payoff from this coalition. The core challenge is determining the allocation of the model's prediction among specific variables. The concept of employing Shapley values to assess local variable importance was first introduced by Strumbelj & Kononenko (2010).

In a scenario involving a permutation J of p explanatory variables within model $f()$, $\pi(J,j)$ represents the set of indices that precede the j-th variable in permutation J. When the j-th variable is positioned first, $\pi(J,j) = \emptyset$. Considering a specific instance $\underline{x}_*$, the model's prediction $f(\underline{x}_*)$ defines the Shapley value as follows:

$$\varphi\left(\underline{x}_*, j\right) = \frac{1}{p!}\sum_J \Delta^{j|\pi(J,j)}(\underline{x}_*) \qquad \text{Eq. (3.32)}$$

where the sum is taken over all $p!$ possible permutations (orderings of explanatory variables) and $\Delta^{j|J}(\underline{x}_*)$ indicates the variable importance. Essentially, $\varphi\left(\underline{x}_*, j\right)$ is the average of the variable-importance measures across all possible orderings of explanatory variables.

It is worth mentioning that the value of $\Delta^{j|\pi(J,j)}(\underline{x}_*)$ remains constant for all permutations J that share the same subset $\pi(J,j)$. The previous equation can be expressed in an alternative form:

$$\varphi(\underline{x}_*, j) = \frac{1}{p!} \sum_{s=0}^{p-1} \sum_{\substack{S \subseteq \{1,\dots,p\} \setminus \{j\} \\ |S|=s}} \{s!\,(p-1-s)!\}\, \Delta^{j|S}(\underline{x}_*) \qquad \text{Eq. (3.33)}$$

where $|S|$ denotes the cardinal size of set S and the second sum is taken over all subsets S of explanatory variables, excluding the j-th one, of size s.

It is also noted that the number of all subsets of sizes from 0 to $p-1$ amounts to $2^p - 1$, significantly fewer than the permutations totaling $p!$. Nevertheless, when dealing with a large p, computing Shapley values using equations such as (3.32) or (3.33) might not be feasible. In such instances, employing an estimation based on a permutation sample becomes a viable option. Strumbelj & Kononenko introduced a Monte Carlo estimator for this purpose in (2014). Moreover, the SHAP package, developed by Lundberg & Lee (2017), presents an efficient implementation for computing Shapley values specifically tailored for tree-based models.

The properties of Shapley values in cooperative games extend to predictive models, granting them the following properties:

- Symmetry: if two explanatory variables j and k are interchangeable then their Shapley values are equal.
- Dummy feature: if an explanatory variable j does not contribute to any prediction for any set of explanatory variables, then its Shapley value is equal to 0.
- Additivity: if model $f()$ is a sum of two other models $g()$ and $h()$, then the Shapley value calculated for model $f()$ is a sum of Shapley values for models $g()$ and $h()$.
- Local accuracy: the sum of Shapley values is equal to the model's prediction, that is, $f(\underline{x}_*) - E_{\underline{X}}\{f(\underline{X})\} = \sum_{j=1}^{p} \varphi(\underline{x}_*, j)$, where $\underline{X}$ is the vector of explanatory variables (corresponding to $\underline{x}_*$) that are treated as random values.

3.2.12 Hierarchical Clustering

In data mining, hierarchical clustering is a type of clustering analysis that creates a hierarchy of clusters based on two key strategies: the agglomerative and the divisive. Agglomerative is a bottom-up approach where each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy. Divisive is a top-down approach where all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy. Within the framework of this doctoral dissertation, the agglomerative approach is used.

Hierarchical clustering has the distinct advantage that any valid measure of distance can be used. In fact, the observations themselves are not required: all that is used is a matrix of distances. In this dissertation, in order to determine which clusters should

be combined, the Euclidean distance between single observations of the dataset and Ward's minimum variance method as the linkage criterion were used.

The results of hierarchical clustering are usually presented in a dendrogram as in the example of the following figure.

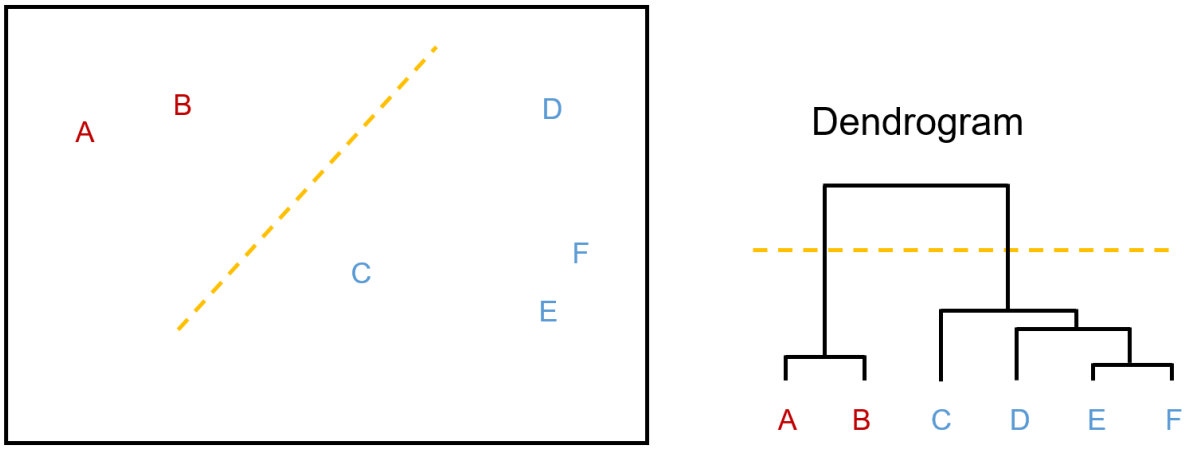


Figure 3.13: Example of a Dendrogram from Hierarchical Clustering

For further details on the theoretical background of hierarchical clustering, the reader is referred to Murtagh & Contreras (2012).

3.2.13 Detection of Spatial Dependence

The initial action in tackling spatial dependence involves identifying its extent within a specific phenomenon by observing its presence in a dataset. Moran's I coefficient, introduced by Moran in (1950), stands as the most commonly used metric for gauging spatial autocorrelation, and it was employed within the framework of this doctoral dissertation.

Global Moran's I is a measure of the overall clustering of the spatial data and it defined as:

$$I = \frac{N}{W} \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}(x_i-\bar{x})(x_j-\bar{x})}{\sum_{i=1}^{N}(x_i-\bar{x})^2} \qquad \text{Eq. (3.34)}$$

where:

- N is the number of spatial units indexed by i and j,
- x is the variable of interest,
- $\bar{x}$ is the mean of the variable of interest x,
- w_{ij} are the elements of a spatial weights' matrix with zeroes on the diagonal,
- and W is the sum of all w_{ij} so that:

$$W = \sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij} \quad \text{Eq. (3.35)}$$

The determination of I 's value can significantly hinge on the underlying assumptions embedded within the spatial weights matrix, denoted as w_{ij}. This matrix is pivotal as it provides a structured framework essential for addressing spatial autocorrelation and modelling spatial interaction by constraining the number of pertinent neighbours. The aim is to construct a matrix that accurately reflects one's presumptions regarding the specific spatial phenomenon under examination. Commonly, one approach involves assigning a weight of 1 to zones designated as neighbours, and 0 otherwise; however, the delineation of “neighbours” can vary. Alternatively, assigning a weight of 1 to the k nearest neighbours and 0 otherwise presents another prevalent method. Moreover, an option exists to employ a distance decay function for weight assignment. Occasionally, the length of a shared edge is utilized for assigning distinct weights to neighbours. The selection of the spatial weights’ matrix ought to be guided by the theoretical underpinnings of the phenomenon in focus. Notably, I 's value exhibits a high sensitivity to these weights and can significantly impact the conclusions drawn about a phenomenon, particularly when using distances.

The expected value of Moran's I under the null hypothesis of no spatial autocorrelation is:

$$E(I) = \frac{-1}{N-1} \quad \text{Eq. (3.36)}$$

As sample sizes expand, an anticipated outcome involves increased dispersion, leading $E(I)$ to converge towards 0. Moran's I values usually range from -1 to 1, but the coefficient can assume values outside this range, depending on the weighting function used. When I significantly surpasses $E(I)$, it signals positive spatial autocorrelation, while values notably lower than $E(I)$ indicate negative spatial autocorrelation. Intuitively, positive autocorrelation implies clustering, whereas negative autocorrelation suggests dispersion. To illustrate typical Moran's I values, commonplace patterns are often employed, as depicted in the following Figure.

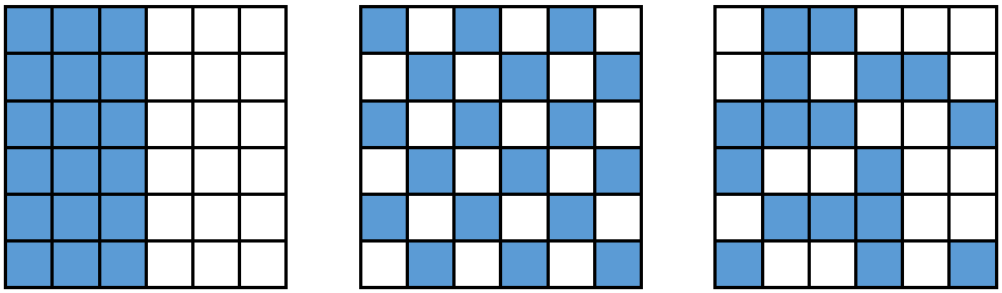


Figure 3.14: Spatial autocorrelation examples

Global spatial autocorrelation analysis yields only one statistic to summarize the whole study area. In other words, the global analysis assumes homogeneity. If that assumption does not hold, then having only one statistic does not make sense as the statistic should differ over space. Moreover, even if there is no global autocorrelation or no clustering, clusters can be found at a local level using local spatial autocorrelation analysis.

Anselin (1995) introduced Local Moran's I as part of the Local Indicators of Spatial Association framework, offering a per-observation coefficient I_i derived from the global Moran's I.

$$I_i = \frac{(x_i - \bar{x}) \sum_{j=1}^{N} w_{ij}(x_j - \bar{x})}{\sum_{i=1}^{N}(x_i - \bar{x})^2} \qquad \text{Eq. (3.37)}$$

Similar to the global Moran's I, the interpretation of Local Moran's I remains consistent. However, unlike its standardized global counterpart, Local Moran's I does not adhere strictly to the -1 to 1 range, allowing for values that significantly deviate from this range.

There are also several other metrics that were not used in the framework of this doctoral dissertation but they can be certainly used for the detection of spatial dependence such as the Geary's C (Geary, 1954) and Getis-Ord G_i^* tests (Ord & Getis, 1995), and the more recent Approximate Profile-Likelihood Estimator (Li et al., 2007).

3.2.14 Spatial Error and Lag Models

The SEM serves as an extension of the traditional linear regression modelling, which can be used to analyze spatially dependent data. In the linear regression model, it is assumed that the errors are independent and identically distributed, meaning there is no correlation or relationship between the error terms of different observations. However, this is not always the case in spatial datasets. The SEM considers and addresses spatial autocorrelation within the residuals. Essentially, this means that the errors resulting from regression analysis exhibit autocorrelation in a way that the error associated with a particular spatial feature can be represented as a weighted average of the errors observed in its neighbouring features. Mathematically, the SEM can be expressed as:

$$y = X\beta + u, \quad u = \lambda_{Err} W u + \varepsilon \qquad \text{Eq. (3.38)}$$

where y is a (N×1) vector of observations on a dependent variable taken at each of N locations, X is a (N×k) matrix of covariates, β is a (k×1) vector of parameters, u is a (N×1) spatially autocorrelated disturbance vector, ε is a (N×1) vector of independent and identically distributed disturbances and λ_{Err} is a scalar spatial parameter.

With regard to the SLM, this type of model can be used to address the spatial autocorrelation in the dependent variable and can be expressed as:

$$y = \rho_{Lag} W y + X\beta + \varepsilon \qquad \text{Eq. (3.39)}$$

where ρ_{Lag} is a scalar spatial parameter that indicates the degree to which a spatial feature is affected by its neighbours. For more in-depth explanations on the statistical background of the SEM and the SLM the reader can refer to Ward & Gleditsch (2018). It is also noted that the fit of the SEM and the SLM can be compared with the fit of the simple linear regression model by using the AIC, with lower values of this criterion indicating better statistical model quality.

3.2.15 Spatial Random Forest

SRF is a powerful ML algorithm that combines the principles of conventional RF with spatial analysis techniques. The mathematical background of conventional RF has been described in previous subsection of this dissertation and is hence omitted here for brevity. However, it should be noted that conventional RF models may fail to consider the spatial structure present in spatial datasets. Consequently, spatial relationships and autocorrelation in the residuals can lead to biased importance scores of non-spatial predictors and suboptimal model performance.

To overcome this limitation, one option is to generate spatial predictors. These predictors assist in taking into account the spatial structure of the training data, ultimately minimizing the spatial autocorrelation of the model residuals and providing accurate variable importance scores. One approach to accomplish this is by incorporating geographical proximity effects into the prediction process by adding the columns of the distance matrix of the examined road segments as explanatory variables, as suggested by Hengl et al. (2018). More specifically, Hengl et al. (2018) proposed the following generic SRF system:

$$Y(s) = f(X_G, X_R, X_P) \qquad \text{Eq. (3.40)}$$

where X_G represents covariates that consider the geographical proximity and spatial relations among observations:

$$X_G = \left(d_{p1}, d_{p2}, \dots, d_{pN}\right) \qquad \text{Eq. (3.41)}$$

where, d_{p1} is the buffer distance to the observed location pi from s and N stands for the total number of training points. X_R corresponds to surface reflectance covariates, while X_P are process-based covariates. For a more comprehensive elucidation, interested readers are directed to consult the detailed explanations provided by Hengl et al. (2018).

4. Investigation of Road Safety Modelling Data in Greece

4.1 Introduction

A road crash results from a combination of factors related to the components of the traffic system comprising roads, vehicles and road users, and the way they interact (Haddon Jr, 1980). Budgets for road safety policies and activities are not infinite. Therefore, decision makers and road safety stakeholders have to determine the optimal possible use of available funds. With regards to improvements in the existing road infrastructure, several quantitative methodologies have been developed over the years, to enhance evidence-based decision making. These methodologies include road crash analyses, road safety inspections, assessment of the "in-built" safety of roads, use of Crash Prediction Models (CPMs), etc. Probably the most detailed approach is offered through the application of CPMs, a practice well described in AASHTO Highway Safety Manual (HSM) (National Research Council, 2010). Yet, especially this methodology requires high quality data in order to predict crash frequency in specific road elements (segments, intersections, etc.) and produce reliable results. More specifically, the availability of detailed and good quality data on road crashes and related casualties, infrastructure geometric characteristics (e.g., curve radius, lane width, etc.) and traffic attributes consists a basic prerequisite for this kind of modelling (Ambros et al., 2018).

Within the above context, the aim of this section is to investigate and discuss the availability and accuracy of road safety modelling data in the primary rural road network of Greece, focusing on three types of data that are considered most critical: crash, traffic and road geometry data. This section is structured as follows: subsection 4.2 concerns crash data availability and presents a case study in the subregion of Viotia for the period 2011-2015. This analysis focuses on identifying the percentage of crashes that could be accurately geo-located and used for modelling purposes.

Then, concerning traffic data, an exploration of the coverage of the road network by spot traffic measurements also in the subregion of Viotia is performed and discussed. Subsection 4.4 focuses on geometric design data, which are generally not readily available in official databases in Greece. The investigation focuses on a section of Patras-Pyrgos two-lane highway, and compares the data that can be obtained through two common Open GIS Data Platforms ("Blender" software and Shuttle Radar Topography Mission (SRTM) data through the GPS Visualizer platform) with the actual data retrieved from a topographic survey of the highway. Moreover, the possibility of exploiting other types of data in road safety analyses, such as telematics data from smart mobile phone sensors, is discussed.

4.2 Crash Data

The Hellenic Statistical Authority maintains the official road crash database in Greece. This database includes road crashes in which at least one involved road user was injured (slightly/seriously) or killed. The case study that will be presented in this subsection is based on road crash data collected from the Police and codified into the National Road Crash Database by the Hellenic Statistical Authority. The Department of Transportation Planning and Engineering of the National Technical University of Athens has access to this National Road Crash Database.

More specifically, in Greece, Traffic Police officers attend the crash site and complete the road crash data in high detail in standardized templates, i.e., the Crash Data Collection Forms, immediately after the occurrence of a crash, providing information on crash conditions, as well as on characteristics related to the road, the involved persons or vehicles. The Crash Data Collection Forms are then forwarded to the Hellenic Statistical Authority, which is responsible for the final checking and codification into the official National Road Crash Database.

Copy files of the National Road Crash Database are provided to the Department of Transportation Planning and Engineering of the National Technical University of Athens (NTUA) (with personal identification removed), who developed a system of efficient queries to extract any combination of data. This NTUA database consists of disaggregated data for all road injury crashes in Greece for the period 1985-2021, is updated on an annual basis, and is also used for the purposes of this investigation.

The variables that are included in this database are presented in the following table grouped by crashes', involved road users' and vehicles' characteristics.

Table 4.1: Variables included in the national road crash database

Group	List of Variables
Crash characteristics	Year, Month, Location (geo-code), Area Type, Street number, Kilometer mark, Kilometrage direction, Road type, Road code, Road's in junction code, Motorway(Y/N), Week of the year, Day of week, Hour, Date, Number of fatalities, Number of serious injuries, Number of slight injuries, Number of vehicles involved, Pavement type, Weather conditions, Pavement conditions, Pavement state, Night lighting, Traffic directions, Number of lanes for each direction, Direction markings, Lane markings, Left edgeline markings, Right edgeline markings, Median, Central barrier, Left side barrier, Right side barrier, Left side shoulder, Right side shoulder, Pavement width, Straight, Narrowing, Lever crossing, Right turn, Left turn, Turn alteration, Ascent, Descent, Ascent / Descent alternation, Type of crash first impact, Maneuver of vehicle A which likely contributed to the crash, Pedestrian maneuver, Traffic control / signalization, Police / Port Authority, Hit and run crash
Involved road users' characteristics	Road user type, Gender, Age (in years), Nationality, Use of protective equipment, Injury severity, Position in vehicle, Purpose of trip
Involved vehicles' characteristics	Vehicle type and usage, Vehicle plates nationality, With trailer, Vehicle capacity, 1st year of registration, Vehicle Technical inspection, Number of drivers and passengers, Type of alcohol test, Result of alcohol test, Time of alcohol test, Place where alcohol test took place, Driving license, License category, Year of acquisition, Vehicle carried dangerous goods (ADR), Overweight vehicle, Load oversized

Data for all injury road crashes in the subregion of Viotia were considered for the five-year period 2011-2015. Firstly, a query was executed in the database in which all road crashes in the subregion of Viotia were searched by year, by type of area, by road code, by station, by infrastructure characteristics (intersection or not, curve or not) and by type of casualties (fatalities, serious injuries, slight injuries).

These data were used to investigate in which way they could be used for microscopic modelling analysis and identify if these data are appropriate for the development of CPMs. An important issue for consideration during the crash analysis is the treatment of road crashes with unknown location. In many cases of road crashes included in the database, there is no indication of the road on which the crash occurred and/or the specific location of the crash. The following table (Table 4.2) presents the number and the respective percentage of road crashes occurred on unknown roads during the period 2011-2015 in subregion of Viotia. Based on Table 4.2, it can be observed that 51% (232/451) of total injury road crashes in Viotia from 2011 to 2015 were coded as occurring on unknown road.

Table 4.2: Road crashes with unknown road for the years 2011-2015 in the subregion of Viotia

Year	Total Crashes	Unknown Road	Unknown Road (%)
2011	118	57	48%
2012	92	53	58%
2013	101	55	54%
2014	75	35	47%
2015	65	32	49%
Total	**451**	**232**	**51%**

Even for crashes on known roads, the specific location of some road crashes is unknown and is not included in the database. Table 4.3 demonstrates the number and the respective percentage of crashes that have occurred on known roads but there is no indication of the crash specific location. In a further 9% (42/451), although the road code was available, the specific location (road chainage) was unknown.

Table 4.3: Crashes on known road and unknown station for years 2011-2015 in subregion of Viotia

Year	Crashes – Known Road	Known Road – Unknown Station	Known Road – Unknown Station (%)
2011	61	9	15%
2012	39	14	36%
2013	46	8	17%
2014	40	8	20%
2015	33	3	9%
Total	**219**	**42**	**19%**

In a more detailed level of analysis, 14 rural roads were isolated and the geo-located crashes were analyzed in order to identify whether the infrastructure characteristics as recorded in the road crash database are identical to the actual infrastructure characteristics of the site. These roads are namely:

- National Road EO.03: Livadeia - I/C E.O. 3 (Livadeia) - Chaironia - Subregion limit (Fthiotida).
- National Road EO.29: Distomo - Steiri - Moni Osiou Louka.
- National Road EO 44: Subregion limit N. Evvoias (Ritsona) - I/C to Elaiona-Thiva.
- Regional Road Ep.5: Distomo - Paralia Distomou - Antikyra - Region limit (Fokida).
- Regional Road Ep.11: Kastro-Stroviki-Orchomenos - I/C E.O.3 (Livadeia).
- Regional Road Ep.17: I/C E.O. 3 (Aliartos) - Akraifnio – Kokkino.
- Regional Road Ep.21: Prodromos - Paralia Saranti.
- Regional Road Ep.23: I/C Kaparelli (Ep.24) - Plataies - Subregion limit N. Attikis (Erythres).

- Regional Road Ep.24: I/C E.O.03 (Thiva) - Loutoufi - Melissochori - I/C Kaparelli.
- Regional Road Ep.28: I/C E.O. 3 - Neochoraki - Asopia - Tanagra - I/C E.O. 1.
- Regional Road Ep.30: Subregion limit N. Attikis (Fyli) - Pyli – Dafni.
- Regional Road Ep.31: Subregion limit N. Attikis (Magoula) - Kokkini - Stefani - I/C Ep.30.
- Regional Road Ep.36: I/C E.O. 44 (Thiva) - Mouriki - Platanakia - Loukissia – Drosia.
- Regional Road Ep.37a & Ep.37b: Arachova - Kalyvia - Subregion limit N. Fokidas (Eptalofo) & I/C to Ski Centre - I/C to Eptalofo.

The following table (Table 4.4) presents the total number of road crashes on these roads and the number of road crashes on these roads with unknown crashes' specific location for the five-year period 2011-2015.

Table 4.4: Crashes on known and codified roads and unknown station for the years 2011-2015 in the subregion of Viotia

Year	Crashes – Known codified Road	Known codified Road – Unknown Station	Known codified Road – Unknown Station (%)
2011	16	1	6%
2012	14	2	14%
2013	12	1	8%
2014	9	0	0%
2015	4	1	25%
Total	**55**	**5**	**9%**

An additional table (Table 4.5) was created to identify whether the infrastructure characteristics as recorded in the road crash database match to the actual basic infrastructure characteristics retrieved from Google Earth aerial imagery. It was found that the basic geometric characteristics (e.g., intersection, curve or straight segment, presence of lighting) matched in only 54% of the cases.

Table 4.5: Crashes on known-codified roads and crashes with identical infrastructure characteristics

Year	Crashes – Known codified Road and known Station	Matching of infrastructure characteristics (crash database and road coding)	(%)
2011-2015	50	27	**54%**

As a conclusion, and taking into account the results of the four previous tables (Tables 4.2-4.5), out of a total of 451 recorded road crashes in the road network of Viotia, only for 177 (39%) is both the road code and the road station recorded. Furthermore, based on the detailed analysis of a sample of roads, it can be assumed that for approximately

half of these crashes (46%) there are obvious discrepancies between basic geometric characteristics of the crash location, as recorded in the database, compared to Google Earth data, leading to the deduction that no more than 21% of available injury crashes data is usable for purposes of crash analysis and modelling that requires precise road crash location.

For the purpose of this doctoral dissertation's analyses, Olympia Odos Operation, the firm operating the Elefsina – Korinthos – Patras motorway has kindly provided a fully detailed crash database for the period from January 1st, 2010 until December 31st, 2020, including road crashes with casualties as well as property-damage-only (PDO) crashes. Generally, motorway concessionaires in Greece usually maintain their own databases in which road crash data with exact location of crashes are recorded, commonly also including crashes with material damage only.

4.3 Traffic Data

In Greece, there is no official national database for traffic data, either traffic volumes or traffic synthesis. Regularly updated datasets exist only for urban areas (e.g., in Athens greater area) and on toll-operated motorways. However, even these datasets are usually not openly and readily available to researchers and practitioners. Traffic data on lower class rural roads (national and/ or regional) are usually collected on a per-case basis by regional road authorities, using spot traffic counts.

As a result, the lack of traffic data is also an important obstacle in microscopic road infrastructure safety research in Greece and in many cases, it actually defines the type and magnitude of research that can realistically be conducted. In order to gain an understanding of the extent of available data, a case study investigation of traffic data availability was performed in the national and regional road network of the subregion of Viotia. Contact with the road management authority of Viotia resulted in identifying a set of spot traffic count results, covering a 12h per day period (8am to 8pm) for a period of three days: Wednesday 10/9/2014, Friday 12/9/2014 and Saturday 13/9/2014, for only four locations, combined for both directions of travel: on Thiva-Livadeia Road, Livadeia-Lamia Road, Thiva Ring Road and Elefsina-Thiva Road (Figure 4.1).

Figure 4.1: Locations of available traffic data in the subregion of Viotia. (Source: Road management authority of Viotia subregion - field surveys in September 2014).

It can be expected that traffic data with a similar level of detail and extent can be obtained also for other sub regions of Greece. The above traffic data could be potentially useful for road safety analyses, after suitable elaboration to estimate the

Average Annual Daily Traffic (AADT). The available detailed information on traffic synthesis (passenger cars, buses, light trucks, 2-axes heavy trucks, 3-axes heavy trucks, and heavy trucks with trailers) may also provide qualitative information for the causes of road crashes during the road safety inspections. However, the data cover a very small fraction of the road network in Viotia subregion, thus, severely limiting the scope of the analyses.

Motorway concessionaires in Greece maintain traffic databases for the road axes they are responsible for. In general, on toll operated motorways, toll stations data can provide a very comprehensive and detailed dataset for traffic volumes and synthesis of traffic, that are fully appropriate for road safety analysis and modelling.

4.4 Geometric Design Data

The development and the application of road infrastructure CPMs is inherently related to the availability of data on the examined road infrastructure, including geometry (e.g. horizontal curvature, vertical curvature and slope), cross section elements (e.g. presence of central median, number of lanes, lane width, shoulder type and width, etc.), roadside conditions (e.g. distance of hazards, road safety barriers, etc.) and other road features and equipment (e.g. rumble strips, condition of markings and signs, road lighting, etc.). Not all types of road infrastructure data are necessary at all times; the selection of the parameters that need to be considered as independent variables in the models is probably the most critical decision that affects the robustness of the approach.

4.4.1 Potential Data Sources

Potential road geometric design data sources commonly include:

- National Road Authorities Databases: Road infrastructure and road design data are commonly collected and maintained in the asset management databases of National Road Authorities. In Greece however, the road registry maintained by the Ministry of Infrastructure and Transport includes mostly administrative data and there is no road geometry database exists with sufficient detail to be able to provide useful and meaningful data for road infrastructure analyses.

- Data from vehicle mounted cameras and road survey vehicles: Vehicle mounted cameras can be used for surveys of road infrastructure: a road is recorded in high resolution while driving at a constant speed appropriate for recording. Weather condition for this type of survey should be ideal, and it is typically performed during the day. The primary purpose of this type of survey is to collect geo-referenced images of road segments which can be used for road attribute coding. Furthermore, equipping the vehicle with various sensors enhances data collection and analysis of multiple data types such as road element data, operating data, and traffic volume data.

  An extensive use of such a road infrastructure data collection methodology took place during the period 2012-2015 by Egnatia Odos SA, in the framework of the Greek Road Rehabilitation and Safety Project. A large part of the national and regional rural road network of Greece (excluding motorways) was surveyed, including 4,200 km of national roads and 10,800 km of regional roads, covering the 13 regions of the country, in order to identify potential sections for road rehabilitation and safety works. In the data collection phase of the above project, vehicle mounted video cameras were used in conjunction with GPS and georeferenced AutoCAD drawings were developed with the

horizontal and vertical alignment of the examined roads and the respective road station. Satellite images were also used as a background of the horizontal alignment drawings. Using these drawings and the video footage, the following data were collected and coded in databases, on the basis of the start/ end station: road gutter, drainage ditch, pavement width, unsealed shoulders, high embankments, high cuts, additional traffic lanes, medians, sidewalks, technical works (culverts, bridges, etc.), traffic signs, road safety barriers, delineators, lighting posts, other posts, at-grade intersections, interchanges, access facilities, pavement deficits, bus stops, etc.

These data are adequately detailed and appropriate for road infrastructure analysis; yet they are somewhat outdated as road improvements have already taken place in some locations

- Data from High Definition (HD) maps: A high-definition map (HD map) is a highly accurate 3D map containing details not normally present on traditional maps. Such maps can be precise at a centimeter level. HD maps are captured using an array of sensors, such as LiDARs, radars, digital cameras and GPS. HD maps can also be constructed using aerial imagery. High-definition maps usually include map elements such as road shape, road marking, traffic signs and barriers. An example of HD mapping suppliers includes TomTom, Here, Navtech, MobilEyE etc.

- Open GIS road geometry data: A series of online utilities provides coordinates along the road network of many countries, including Greece. In order to investigate the potential and accuracy of Open GIS Data in effectively describing road geometry (horizontal elements and elevations) a pilot assessment study was performed as presented in the following subsection.

4.4.2 Pilot Evaluation of Open GIS Road Geometry Data

Data extraction and assessment was based on comparing road geometry data retrieved from OPEN GIS sources to the actual data for the road axis of Patras-Pyrgos National Road in the area "Vrachneika", as derived from a detailed topographic survey at scale 1: 500 (Figure 4.2).

The investigation included the use of:

- Blender software (free software available at: https://www.blender.org/) with GIS tracking of road data, and
- the GPS Visualizer platform that retrieves data from the SRTM database - same to OSM data that is also accessible via API.

Figure 4.2: Blender and GPS Visualizer data assessment area.

4.4.2.1 Blender Software

Blender is a free software released under the GNU General Public License. It supports the entirety of the 3D pipeline—modelling, rigging, animation, simulation, rendering, compositing and motion tracking, video editing and 2D animation pipeline. Using the add-on ‘’Blender GIS’’, Blender software can retrieve and process geographic information in standard GIS file formats e.g., shapefile vector, raster image, geotiff DEM, OpenStreetMap xml.

The steps followed using the Blender software were:

1. Using the add-on, the area where the topographic survey was made, was visually identified and the background map was retrieved (GIS tab → webgeodata → Basemap, source google – satellite level).
2. Digital model was retrieved (GIS tab → webgeodata → Get SRTM)
3. Open street Map (OSM) data was saved for the existing roads (GIS tab → webgeodata → Get OSM, highway level).
4. The highway level information was extracted (*.shp file) in the form of lines with elevation data (GIS tab → Export, feature: line).
5. Import of the *.shp file in AutoCAD software and comparison of the elevations of the imported lines to the topographic survey elevations.

4.4.2.2 GPS Visualizer platform and Shuttle Radar Topography Mission database

GPS Visualizer is an online utility that creates maps and profiles from geographic data (https://www.gpsvisualizer.com/). It is free and easy to use, yet powerful and extremely customizable. Input can be in the form of GPS data (tracks and waypoints), driving routes, street addresses, or simple coordinates.

The elevations of the same eight sampling points considered for the Blender software were also estimated by using the DEM database (https://www.gpsvisualizer.com/elevation). The procedure of converting the AutoCAD points (*.dwg file) to kml/kmz files was the following:

1. Export of the AutoCAD points to a *.shp file.
2. Import of the *.shp file in Google Earth and then export as *.kml file.
3. Import of the *.kml file in DEM database of GPS Visualizer site
4. Export in *.txt file.

4.4.2.3 Comparison of Open GIS Data to topographic survey data

The data extracted from Blender software and from GPS Visualizer platform were compared against the respective points on the topographic survey, with regards to their elevation as shown in Tables 4.6 and 4.7 that follow.

Table 4.6: Accuracy assessment of road centerline points - Blender software

Point no.	X (Easting)	Y (Northing)	Elevation (Blender)	Elevation (Survey)	Difference in elevations (m)
1	294999.85	4225789.11	29.18	25.98	3.20
2	295066.33	4225763.10	42.33	43.17	0.84
3	295230.16	4225760.95	48.94	49.60	2.36
4	295506.68	4225736.57	49.35	47.40	2.34
5	295867.39	4225772.21	68.33	71.20	2.94
6	295901.81	4225838.99	66.06	64.40	4.56
7	295917.74	4225759.28	82.82	87.10	13.55
8	296081.10	4225921.02	58.82	56.80	0.82

Table 4.7: Accuracy assessment of road centerline points - GPS Visualizer platform

Point no.	Latitude	Longitude	Elevation (GPS Visualizer)	Elevation (Survey)	Difference in elevations (m)
1	38.1593349	21.6618775	28.90	25.98	2.92
2	38.1591157	21.6626432	43.50	43.17	0.33
3	38.1591336	21.6645123	49.60	49.60	1.70
4	38.1589767	21.6676730	47.40	47.40	0.39
5	38.1593793	21.6717767	71.20	71.20	0.07
6	38.1599885	21.6721501	64.40	64.40	2.90
7	38.1592743	21.6723546	87.10	87.10	9.27
8	38.1607677	21.6741715	56.80	56.80	1.20

From the above analysis it is evident that no accurate information for vertical alignment and road elevations can be collected from both Open GIS data sources (Blender and GPS Visualizer). Specifically, street surface elevations obtained from Open GIS applications have very large deviations, both between applications (e.g., Blender data

compared to GPS Visualizer data) and (more importantly) when compared to actual surveyed elevations. In more than half of the randomly selected examined points (6 out of 8 for Blender data and 5 out of 8 for GPS Visualizer data), elevation differences from the survey exceed 1m. The problem seems to be is intensified in cases where the road is at a cut or fill section of considerable height (e.g., point 7), where differences up to 13.5m were observed.

On the other hand, with regards to the horizontal alignment, qualitative evaluation of data for the road centerline location from both Blender and GPS Visualizer reveals small differences compared to the surveyed road centerline and these data can potentially be used to build a road geometry database for the purpose of road safety analyses.

4.5 Smartphone Data

An alternative approach to road safety related data that will be exploited in parallel within the framework of this doctoral dissertation, is the use of crowdsourced smartphone data from OSeven Telematics (www.oseven.io). OSeven maintains and operates an innovative data collection scheme which records personalized driving behaviour analytics in real time, using smartphone sensors. An integrated system is used for the recording, collection, storage, evaluation and visualization of driving behaviour data, using smartphone applications and advanced ML algorithms. The system includes specially developed smartphone application for data collection and transmission, as well as for providing feedback to the participants on their driving behaviour.

The steps described below for data processing are exclusively performed by OSeven and do not constitute part of this dissertation. More details on the data processing steps cannot be provided since they are intellectual property of the company. However, the main features of the system are outlined below.

A smartphone app has been developed by OSeven to record driver behaviour using the sensors of the smartphone, and a variety of APIs is exploited to read sensor data and temporarily store them to the smartphone's database before transmitting them to the central (backend) database. The data recording is initiated automatically in the smartphone app when a driving status is recognized, and again it stops automatically when a non-driving status is recognized. The frequency of the data recording varies depending on the type of the sensor, with a minimum value of 1 Hz. Trip recording also continues if the vehicle is idled for five minutes, to consider the case that the driver resumes a trip after a few minutes stop. All extra information collected after the end of driving is discarded.

The recorded data come from various smartphone sensors and data fusion algorithms provided by Android (Google) and iOS (Apple). Indicatively, technology sensors integrated in the smartphone are the Accelerometer*, the Gyroscope*, the Magnetometer and the GPS (speed, course, longitude, latitude). Fusion Data provided by iOS and Android include yaw, pitch, roll, linear acceleration* and gravity* (elements marked with an asterisk "*" sign refer to x, y, z components).

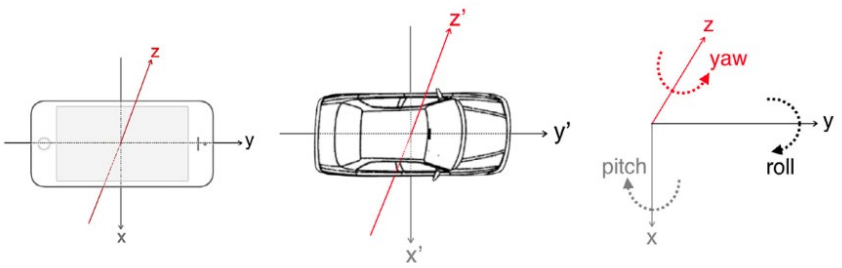


Figure 4.3: Coordinate systems of the smartphone and the vehicle

After the end of each trip, the application is transmitting all data recorded to the central database of the OSeven backend office via an appropriate communication channel, such as a Wi-Fi network or cellular network (upon user's selection) e.g., 4G (online options). The data collected are highly disaggregated in space and time. Once stored in the backend cloud server, they are converted into driving behaviour and safety indicators, using signal processing, ML algorithms, Data fusion and Big Data algorithms. ML methods (filtering, clustering and classification methods) are mainly used to clean the data from noise and errors, and to identify repeated patterns within the data.

Various forms of metadata are ultimately computed, including both exposure and driving behaviour indicators - such as trip duration, trip distance, driver speed, instances of speeding, the frequency of harsh braking and harsh acceleration incidents, and driver distraction from mobile phone use. The detection of harsh events is accomplished through the utilization of the proprietary OSeven algorithms, which are private and under intellectual property rights. Essentially, these algorithms utilize data from all axes of the accelerometer, along with inputs from GPS, magnetometer, and gyroscope sensors. The algorithms analyse the time series data throughout the entire trip in order to increase the accuracy of harsh events detection. Importantly, these algorithms do not rely on predefined thresholds to deem whether an event is harsh or not. Instead, ML techniques and data fusion are implemented to identify abrupt spikes in the sensor data (Kontaxi et al., 2021a) regardless of absolute values.

It is worth noting that all naturalistic driving data used in this doctoral dissertation were provided by OSeven Telematics in a fully anonymized format, complying with the relevant national and European personal data regulations, including the General Data Protection Regulation (GDPR).

OSeven smartphone application has been employed for research purposes pertaining mainly to driving behaviour, as extensively documented in prior studies (Papadimitriou et al., 2019; Kontaxi et al., 2021b; Ziakopoulos, 2021; Tarlochan et al., 2022; Fafoutellis et al., 2023). Therefore, it can be concluded that these indicators along with other data (e.g., from map providers) can be subsequently exploited to identify patterns and locate sections of the road network with above normal concentrations of harsh braking and harsh acceleration events, of speeding events and also provide estimations on average speed, to be used for road safety modelling purposes.

4.6 Discussion

Based on the results of the above pilot studies, for non-motorway rural roads in Greece the absence of traffic volume information and even more, of properly geo-located road crash data, seems to be an impermeable obstacle for detailed crash prediction modelling efforts. Specifically, for non-motorway rural roads it was found that:

- approximately 80% of the injury crashes recorded in the official National Road Crash Database has either missing or obviously inaccurate crash location information,
- existing traffic volume data on the rural road network are largely unavailable and derived from scarce spot counts performed several years ago,
- geometric (road design) data are available only in the deliverables of the Greek Road Rehabilitation and Safety Project performed on behalf of Egnatia Odos SA in 2012-2015. Limited data can be retrieved from Open GIS sources, keeping however in mind that elevation information is largely inaccurate.

On the other hand, road crash prediction modelling can potentially be performed on motorways, using crash and traffic data maintained by road operators, provided that an arduous and resource-consuming process is applied to collect and code missing geometric design data.

It should be acknowledged that this investigation exhibits certain limitations that need to be considered along with the findings and conclusions. Firstly, the investigation of this section focuses only on the rural road network; availability and accuracy of road safety data for urban roads may significantly differ. However, it can be expected that the lack of a registry for municipal roads (excluding those in Athens and Thessaloniki) in the National Road Crash Database is likely to exacerbate the issue of incomplete geolocation records for crashes on urban roads across the provinces of Greece.

Secondly, both the investigation of crash data reliability and location information with regards to the official Greek National Road Crash Database and of traffic data availability, is focused on a single prefecture of Greece. Although this prefecture (Viotia) is considered quite representative of average conditions, it may be true that different conditions may prevail in other prefectures (particularly in island prefectures). Lastly, the integration of naturalistic driver behaviour data from smartphones stands as an invaluable addition to road safety analyses.

The investigation of this section highlighted the limitation of conducting high-detailed crash prediction modelling in Greece, feasible only for motorways with high-quality crash data, in terms of crash location, and traffic attributes per road segment. This led to the establishment of two distinct databases: one encompassed comprehensive data for the Olympia Odos motorway, including detailed historical road crash records, traffic attributes, road geometry characteristics, and driver behaviour data on a segmental

basis; the other covered a broader road network within the Region of Eastern Macedonia and Thrace, albeit lacking detailed crash location data and traffic attributes. Additional details regarding the data collection and processing methods used for these two distinct databases, along with the outcomes of the statistical and ML analyses, are expounded upon in the subsequent sections of this doctoral dissertation.

5. Motorway Data Collection and Processing

5.1 Introduction

This section provides technical information on the process of data collection and descriptive statistics for the Olympia Odos motorway. Based on the experience and knowledge gained through Section 4 findings regarding the pilot study in Viotia subregion, the investigation of crash location data reliability and the pilot evaluation of Open GIS road geometry data, detailed road crash investigations of this doctoral dissertation focused on Olympia Odos motorway for which very detailed and accurate historical road crash and traffic data were kindly provided by the road operator. However, detailed road infrastructure and geometry data were not readily available. Therefore, the required dataset had to be developed exploiting available data from all potential sources. Olympia Odos motorway is located in Southern Greece and is a rural motorway from Athens to Patras that comprises 201.5 km of rural motorway in total, with two or three lanes per direction and 29 interchanges. Part of the motorway of 63 km (Elefsina-Korinthos) is in operation since 2010, whereas the rest (Korinthos-Patras) was fully operational since the summer of 2017.

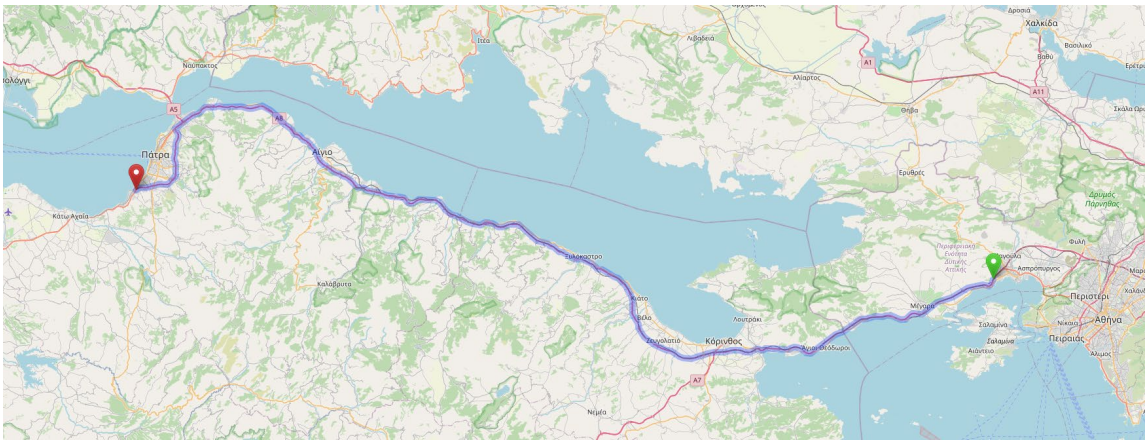


Figure 5.1: Olympia Odos motorway in Greece

The rest of this section is organised as follows. Subsections 5.2 and 5.3 provide brief descriptions of the main attributes of the collected crash and traffic data for Olympia Odos motorway. Following this, subsection 5.4 describes the activities performed to develop a road infrastructure database for Olympia Odos motorway, combining information from the road operator, Open GIS software, Google Earth and GoogleStreetView. The creation of this road infrastructure database and of reference drawings of the motorway also enabled the identification and isolation of naturalistic driver behaviour data from OSeven database (subsection 5.5). Concluding this section, subsection 5.6 presents a summary table containing the motorway-related variables that were ultimately analyzed in this doctoral dissertation, accompanied by their abbreviations and relevant descriptive statistics.

5.2 Crash Data

Crash data of all severity levels including PDO crashes were available for the period 2015–2020. As the entire motorway (i.e., from Athens to Patras) was finalized and started operating in summer 2017, crash data for the entire length were available for the years 2018-2020. Therefore, it was decided to focus on a smaller time period (2018-2020) but for a longer road network. The motorway is operated by a private road operator firm, Olympia Odos Operation SA, who kindly provided data for this doctoral dissertation. The following 28 variables (per crash) are included in the provided road crash database:

reference number, Crash Data Collection Form number, crash type, date, time, direction of travel, road station (chainage), interchange name, ramp name, tunnel name, toll station name, weather conditions, pavement conditions, lighting conditions, number of slightly injured, number of seriously injured, number of fatalities, crash severity, and ten variables on the type and number of involved vehicles: other/ powered-two-wheeler/ passenger car/ bus/ vehicle with trailer or caravan/ truck/ taxi/ truck with dangerous load/ truck <2.5 T/ truck>2.5 T.

5.3 Traffic Data

For the purpose of the analyses of this doctoral dissertation, Olympia Odos SA, the firm operating the Elefsina - Korinthos - Patras motorway (an identification of km posts using Google Street View) has provided traffic data as follows:

1. AADT for section of the motorway (28 sections defined according to the location of interchanges), for years 2015 to 2020. AADT is provided as a sum for both directions of travel, with an estimated equal distribution per direction, according to Olympia Odos.
2. Traffic composition, considering four different vehicle categories:
   Cat 1: Power-Two-Wheelers (PTWs).
   Cat 2: Passenger cars - light vehicles (may tow a trailer, height less than 2.2m).
   Cat 3: Heavy vehicles with maximum 3-wheel axes (may tow a trailer, height more than 2.2m).
   Cat 4: Heavy vehicles with 4 or more-wheel axes (may tow a trailer, height more than 2.2m).

Data from major toll stations are available (separate for each direction of traffic), dividing the motorway into five large sections. Traffic composition data are available for the same time period to AADT data. In parallel with the road crash data, the time period of traffic data that was examined within this doctoral dissertation corresponds to the same three-year time period (2018-2020).

5.4 Road Infrastructure Data

A road geometry database that focuses on the section from the toll station of Elefsina (CH.26 + 500) to the end of the motorway (CH.223 + 200) was developed through a multi-step process. As a first step, a draft centerline of Olympia Odos Motorway was preliminarily retrieved from Open GIS software, using the Blender application (https://www.blender.org/), as follows: the Open GIS polylines representing the existing road network in the vicinity of the motorway were exported in shapefile format and imported to CAD environment; then, all neighbouring road centerlines were removed and centerlines for the motorway, for transverse roads and entrance/exit ramps at interchanges were isolated. At this stage, the CAD drawing of the motorway was developed in the official national coordinate system in Greece (EGSA 87). It is noted that only horizontal alignment information on the road centerlines was retrieved.

Following the zone of the centerline defined in the first step, a series of high-detail satellite images (pixel size approximately 1.2 × 1.7m) were retrieved using the respective GIS module of the free online software HEC-RAS (https://www.hec.usace.army.mil/software/hec-ras/) and georeferenced as the background of the CAD drawing. Combining information from the Open GIS road centerline and the detailed satellite imagery, the centerline of the motorway was subsequently refined, as follows: the preliminary centerline from Open GIS software is a polyline with dense points. This was manually replaced in the CAD environment by a “road design equivalent” centerline, consisting of tangents, circular curves, and spiral (clothoid) curves. Spiral curves were introduced on the entrance and exit of all curves with a radius of less than R = 1000, assuming a clothoid curve parameter A ranging from R/3 to R (R/3 < A < R), according to Greek road design guidelines. In segments where the two directions of travel follow different paths, the main centerline was the direction from Elefsina to Patras and a secondary centerline was created for the opposite direction.

The refined CAD centerline was then imported into the Google Earth online platform (https://earth.google.com) and using the satellite views and the Google Street View imagery in conjunction, the location of km posts was determined. This location is of utmost importance for microscopic road safety analyses, as all elements of the analysis (crashes, speed limits, etc.) are recorded according to those locations (GPS use for crash location recording is not performed in Greece). The km posts, as identified in Google Earth, were subsequently imported into the base CAD drawing of the motorway and a road chainage system (stations) was established.

In the next step, all available road infrastructure data were imported into the CAD drawing as well as the Google Earth interface, mostly based on their respective road station (chainage) but also cross-checking their location against the Google Earth satellite imagery and Street View images. An important source of information at this stage was the motorway schematic provided by the road operator (Olympia Odos

Operation SA) with the exact locations (road station-chainage) of interchanges (with entrance/exit ramps), toll stations, motorway service stations, parking areas, tunnel and cut-and-cover entrance and exits, and speed limit signs.

The above procedure produced a CAD drawing, as presented in Figure 5.2, with georeferenced satellite images as the background, including motorway centerline geometry, chainage, speed limits, and visualization of other important road infrastructure elements: toll stations, interchanges (with transverse roads, entrance and exit ramps), km posts, location of lane addition or lane drop, weaving segments, etc., and a Google Earth Dataset in .kmz file form, presented in Figure 5.3, with several layers of information: center-line, chainage, tunnels, additional lane points (gore, start, and end), lane drops/additions, etc. These two powerful tools were utilized in order to code road infrastructure data for further analysis and create a database that forms the basis for subsequent motorway analyses.

Figure 5.2: Extract of the developed CAD drawing

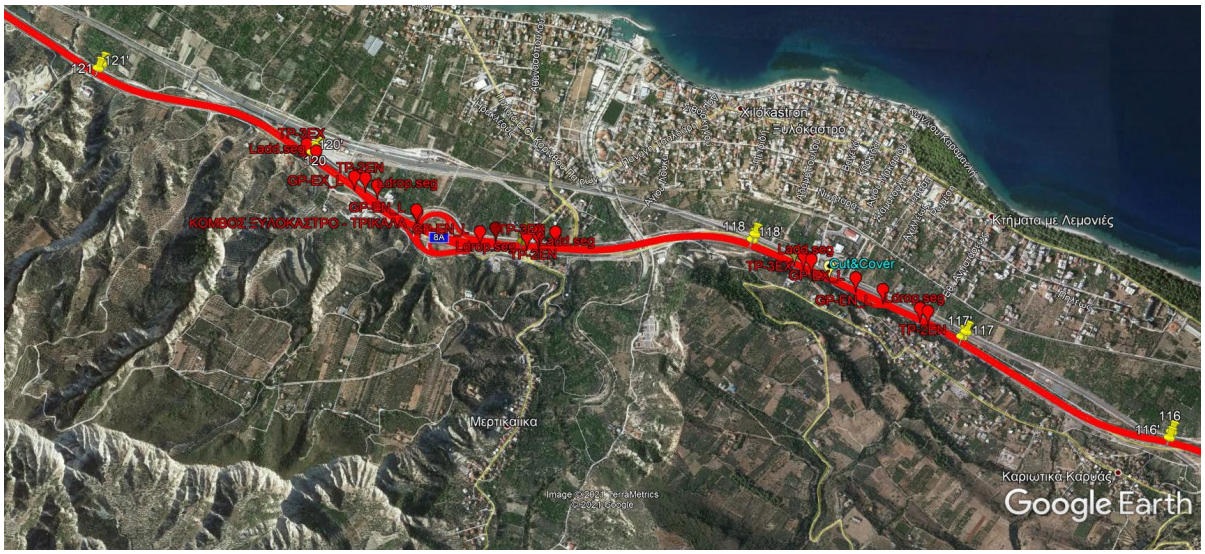


Figure 5.3: Extract of the Google Earth .kmz file

5.5 Driver Behaviour Data

Another database on road user behaviour data on Olympia Odos Motorway was developed, in order to be jointly investigated with the road infrastructure, crash, and traffic data. Naturalistic driver behaviour data were recorded via a smartphone application and processed in the platform, both developed by OSeven (https://oseven.io/). Drivers install the application developed by OSeven on their smartphones and subsequently engage in normal driving activities. The application engages automatically when driving is initiated and records different data types such as vehicle location, speed, acceleration, deceleration, duration of engagement with the phone, etc. These data are further processed to develop metrics to describe driver behaviour. Further details on the operation of this application have already provided in subsection 4.5 of this doctoral dissertation.

For the analyses of Olympia Odos motorway segments, OSeven has provided a representative dataset from its database in a completely anonymized format that corresponds to the period from 1 June 2019 to 31 December 2020. The data were recorded from a driver sample equal to 327 drivers for 2019 and 330 drivers for 2020. It is possible that some drivers were mindful that their driving behaviour was recorded through the application and were even more aware than usual. However, these effects have been reported to decrease over time as drivers gradually forget that they are being recorded (Tselentis, 2018). For the total considered time period the average number of recorded trips per motorway segment was 769 trips. Subsequently, driving behaviour metrics from naturalistic data, which are driver-based, needed to be assigned to the examined motorway segments, which are infrastructure-based data. This was achieved via isolating each trip portion to the corresponding segment within the internal recording of trips conducted in GIS by OSeven using ESRI polygons at 200m intervals.

5.6 Descriptive Statistics

At this point, it should be noted that the recording of driver behaviour through the smartphone app was not feasible within the tunnel road segments due to the loss of GPS signal. Furthermore, toll station segments are not typical motorway segments both in terms of geometric design and driver behaviour. Consequently, these two types of road segments were not included in the motorway segments' analyses of this doctoral dissertation. The motorway-related variables that were finally included and analyzed are presented in Table 5.1, along with their abbreviations and some key descriptive statistics. As also mentioned in Sections 2 and 4, the variables related to road design characteristics, traffic attributes, and road crashes are widely used in CPMs, whereas harsh driving behaviour events are SSMs that can complement road safety analyses.

Table 5.1: Road crash, traffic, geometry, and driver behaviour variables per motorway segment

Variable	Abbreviation	Descriptive Statistics
Number of Segment	no.	Count: 668
Direction	Direction	Frequencies: E: 337 T: 331
Segment Start (Chainage)	Seg_Start	-
Segment End (Chainage)	Seg_End	-
Number of through lanes	lanes	Frequencies: 2: 435, 3: 233
Length of motorway segment (km)	len_seg	Min.: 0.2000, Max.: 0.6000, Mean: 0.5284, Median: 0.6000
Average Annual Average Daily Traffic Volume of motorway segment (veh/day) 2018-2020	avg_AADT_18_20	Min.: 6,511, Max.: 22,079, Mean: 10,786, Median: 7,423
Posted speed limit (km/h)	speed_limit	Min.: 90.0, Max.: 130.0, Mean: 121.7, Median: 130.0
Number of Total Road Crashes (Injury & Property Damage Only) 2018-2020	TotCr18_20	Min.: 0.00, Max.: 13.00, Mean: 2.02, Median: 2.00
Number of Total Road Crashes (Injury & Property Damage Only) by segment length 2018-2020	TotCr18_20_len_seg	Min.: 0.00, Max.: 30.00, Mean: 3.88, Median: 3.33
Curve 1 - Radius R (m)	Curve1	Min.: 0, Max.: 50,000, Mean: 2,129, Median: 950
Curve 1 - Length of curve in segment (m)	Lcurve1_in_seg	Min.: 0.00, Max.: 600.00, Mean: 218.21, Median: 196.31
Lane width (m)	lane_width	Min.: 3.55, Max.: 3.95, Mean: 3.92, Median: 3.95
Paved inside shoulder width (m)	pav_ins_sh_width	Min.: 0.50, Max.: 1.75, Mean: 0.69, Median: 0.75
Median width (measured from near edges of traveled way in both directions) (m)	median_width	Min.: 2.25, Max.: 23.50, Mean: 4.96, Median: 4.88

Distance from edge of inside shoulder to barrier face (m)	dist_edginssh_barf	Min.: 0.00, Max.: 0.75, Mean: 0.04, Median: 0.00
Paved outside shoulder width (m)	pav_out_sh_width	Min.: 0.25, Max.: 4.50, Mean: 2.77, Median: 3.00
Distance from edge of outside shoulder to barrier face (m)	dist_edgoutsh_barf	Min.: 0.00, Max.: 3.25, Mean: 0.82, Median: 0.50
Number of recorded trips	rec_trips	Min.: 173, Max.: 1,689, Mean: 769, Median: 529
Average speed (all trips) (km/h)	avg_speed	Min.: 77.0, Max.: 153.0, Mean: 115.9, Median: 118.0
Number of harsh accelerations per trips	ha_per_trips	Min.: 0.0000, Max.: 0.1614, Mean: 0.0046, Median: 0.0020
Number of harsh brakings per trips	hb_per_trips	Min.: 0.0000, Max.: 0.1172, Mean: 0.0052, Median: 0.0022
Number of speeding events per trips	speeding_per_trips	Min.: 0.03, Max.: 2.56, Mean: 0.68, Median: 0.71

The histogram of Figure 5.4 presents the distribution of road crash frequencies in the examined motorway segments, while the boxplots of Figures 5.5 to 5.17 display the key descriptive statistics of the numeric variables of Table 5.1.

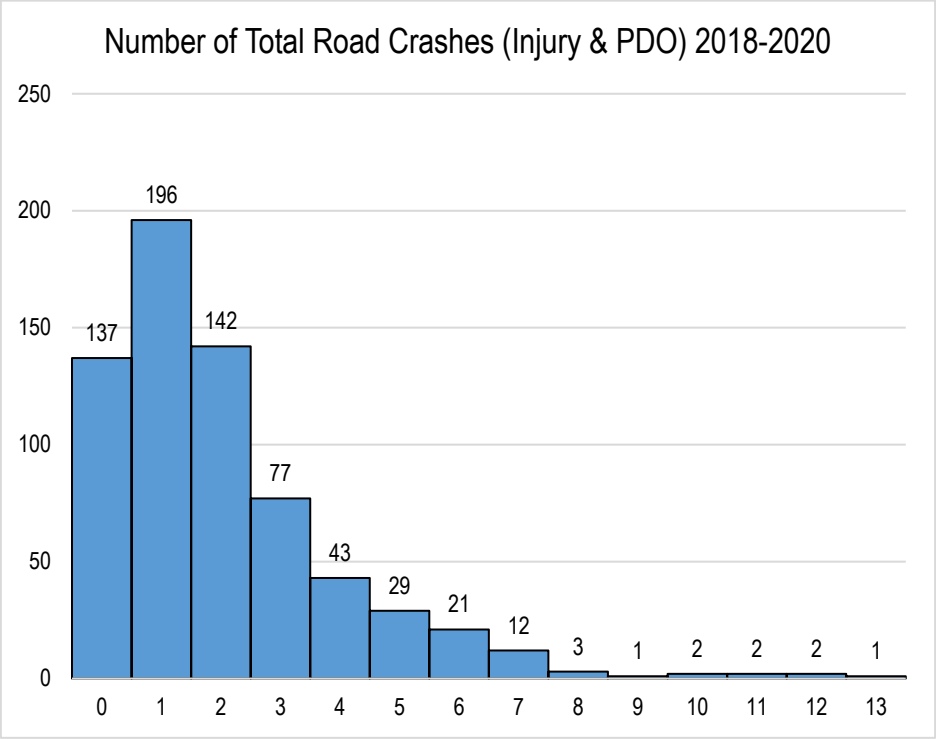


Figure 5.4: Number of Total Road Crashes (Injury & PDO), 2018-2020

Figure 5.5: Length of motorway segment (km)

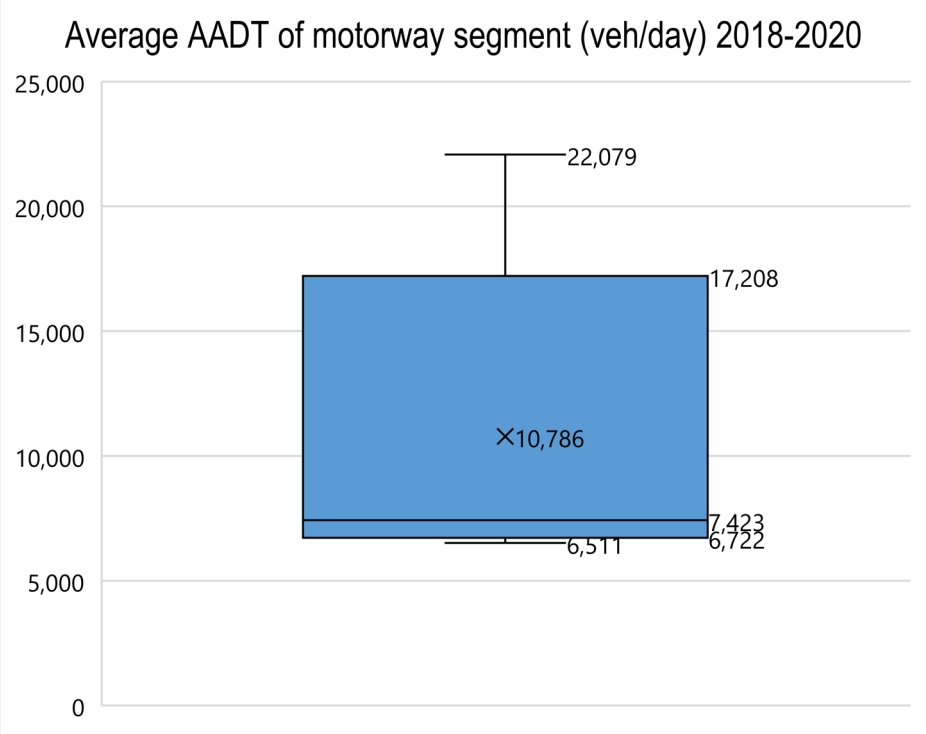


Figure 5.6: Average AADT of motorway segment (veh/day), 2018-2020

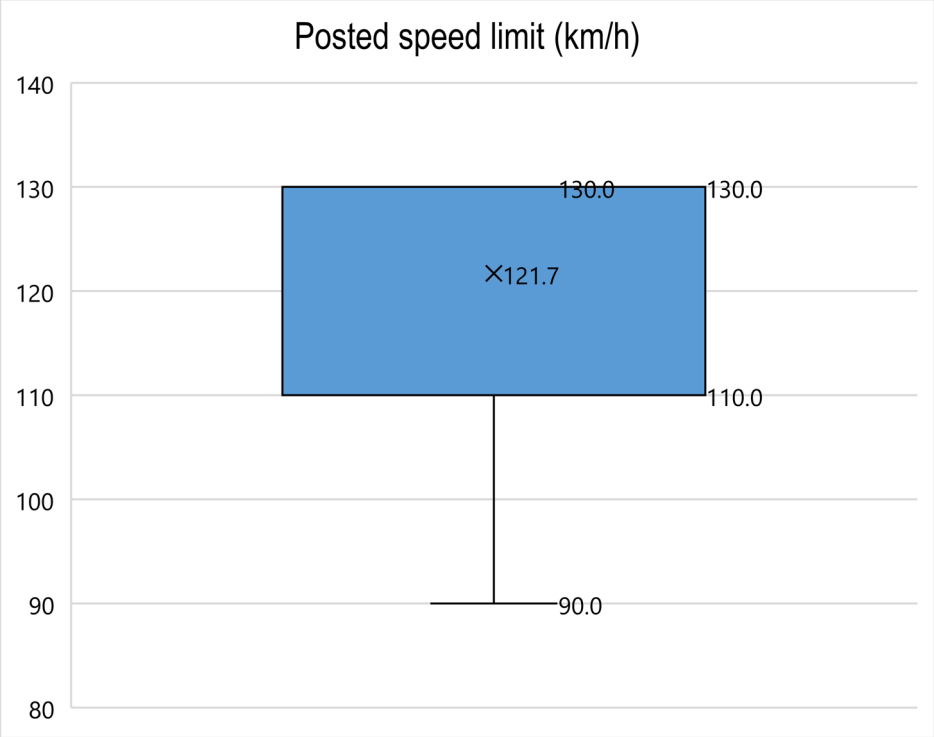


Figure 5.7: Posted speed limit (km/h)

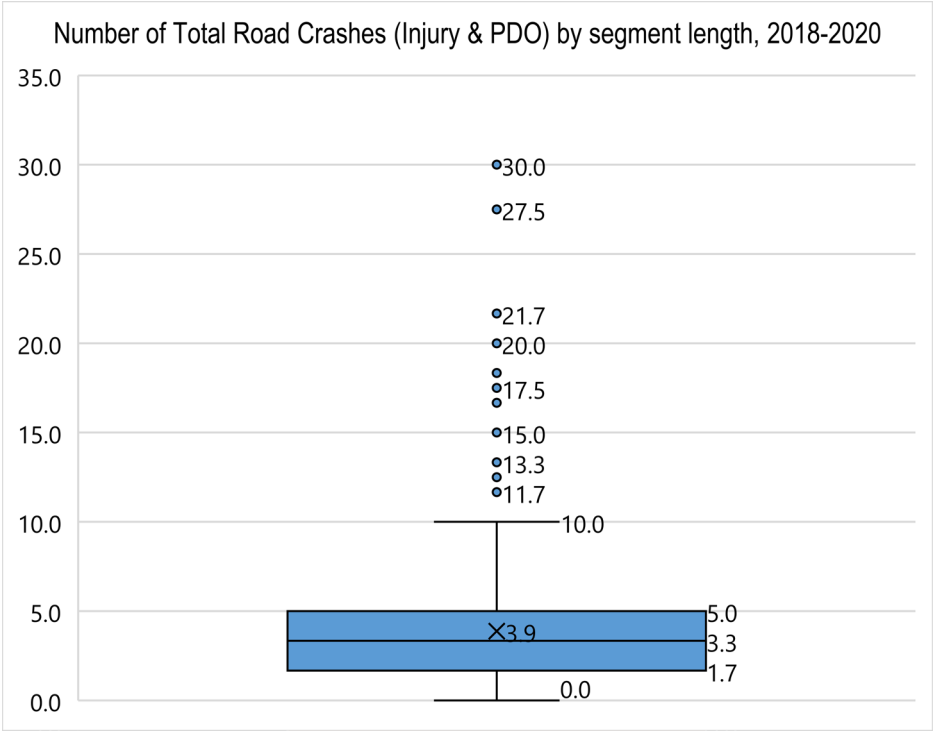


Figure 5.8: Number of total road crashes by segment length, 2018-2020

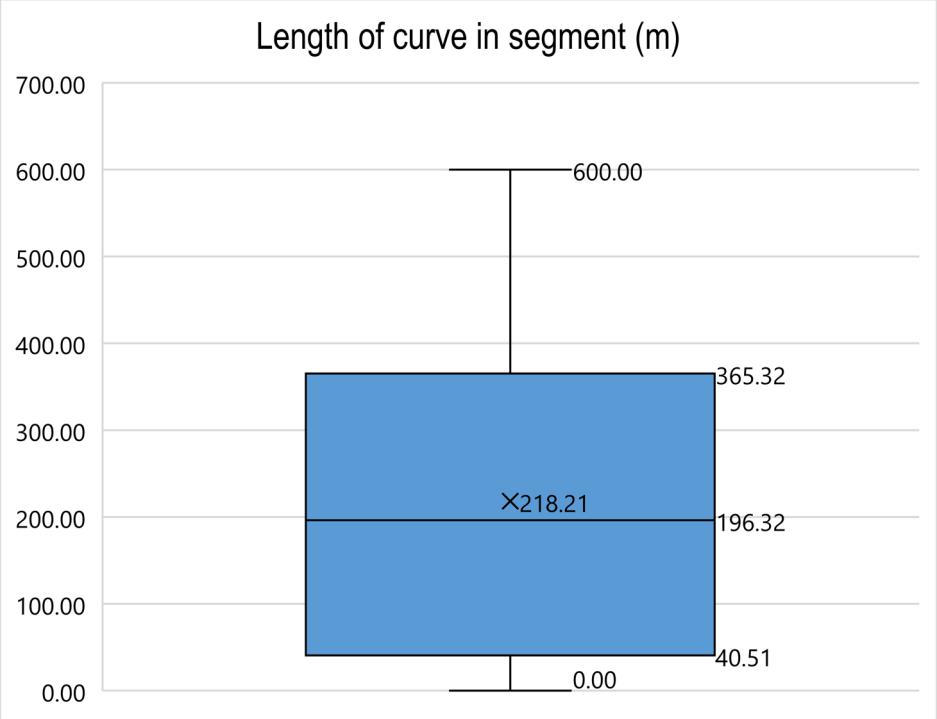


Figure 5.9: Length of curve in segment (m)

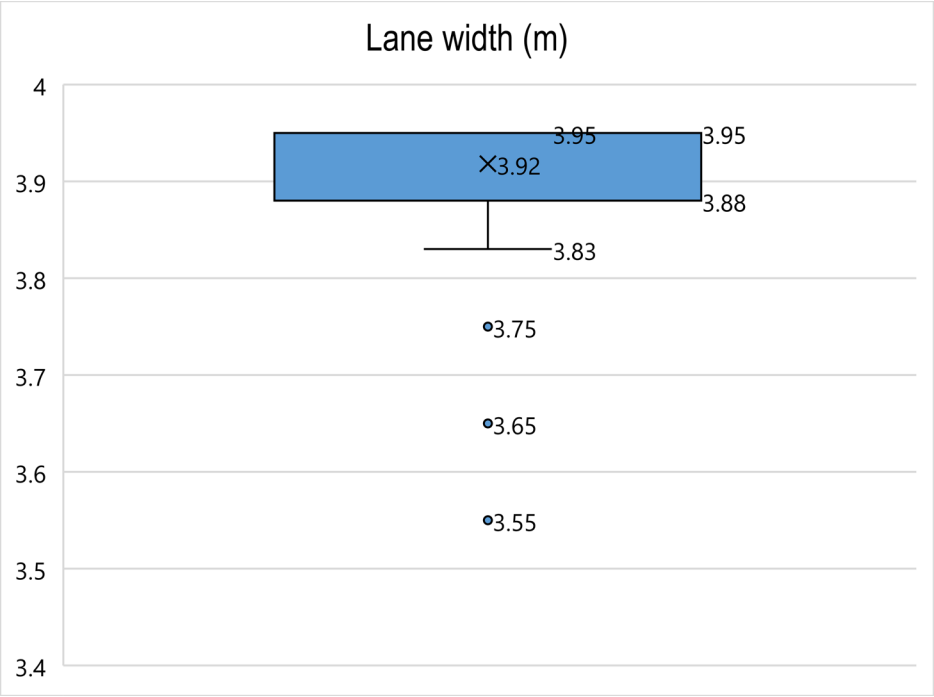


Figure 5.10: Lane width (m)

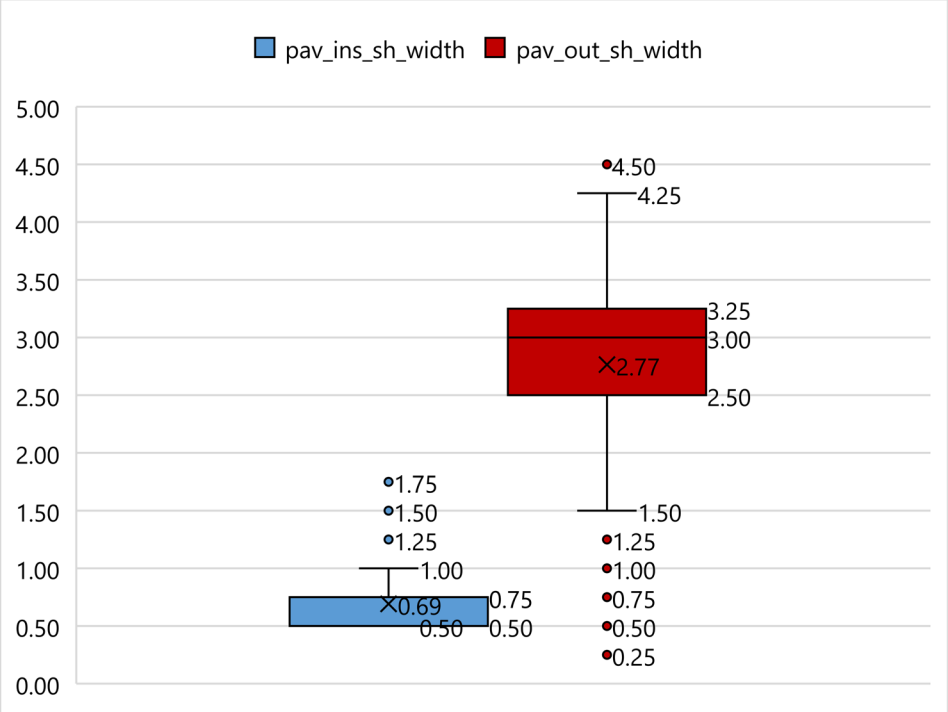


Figure 5.11: Paved inside/outside shoulder width (m)

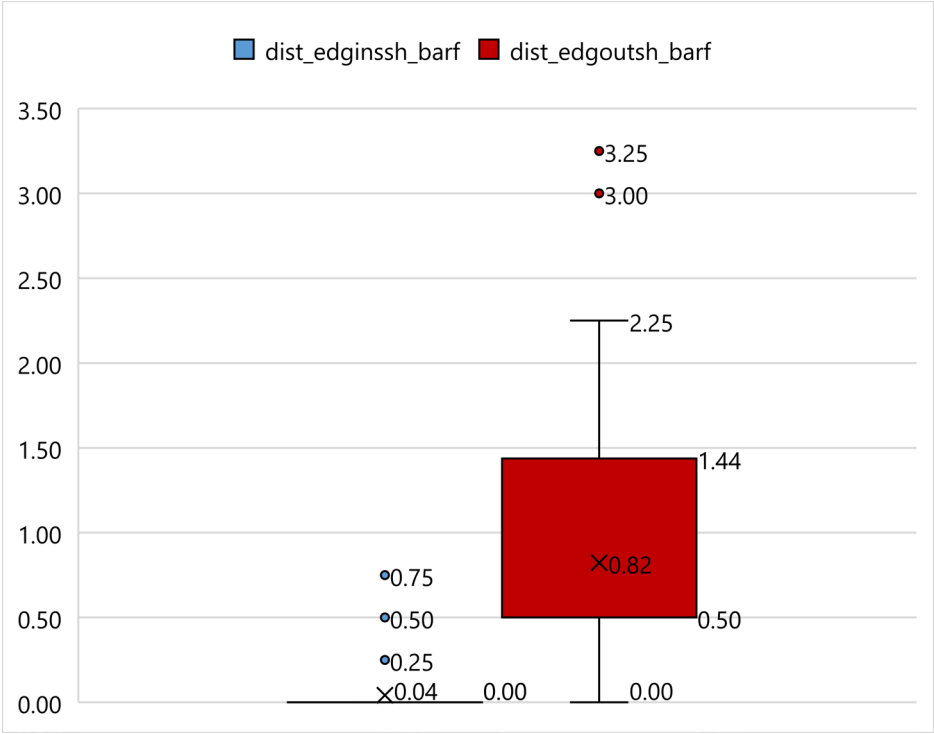


Figure 5.12: Distance from edge of inside/outside shoulder to barrier face (m)

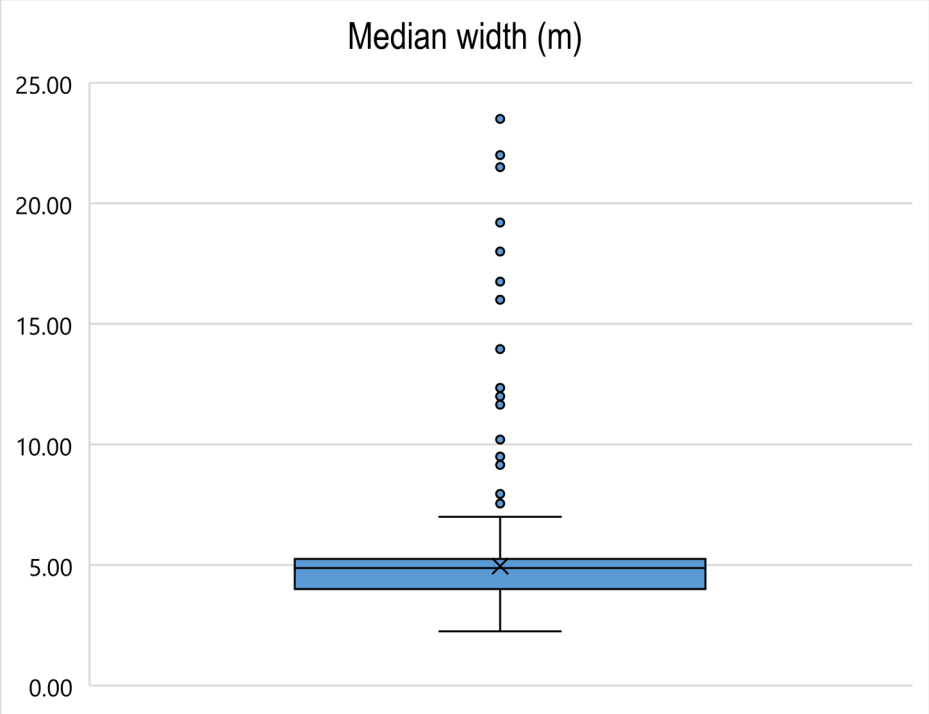


Figure 5.13: Median width (m)

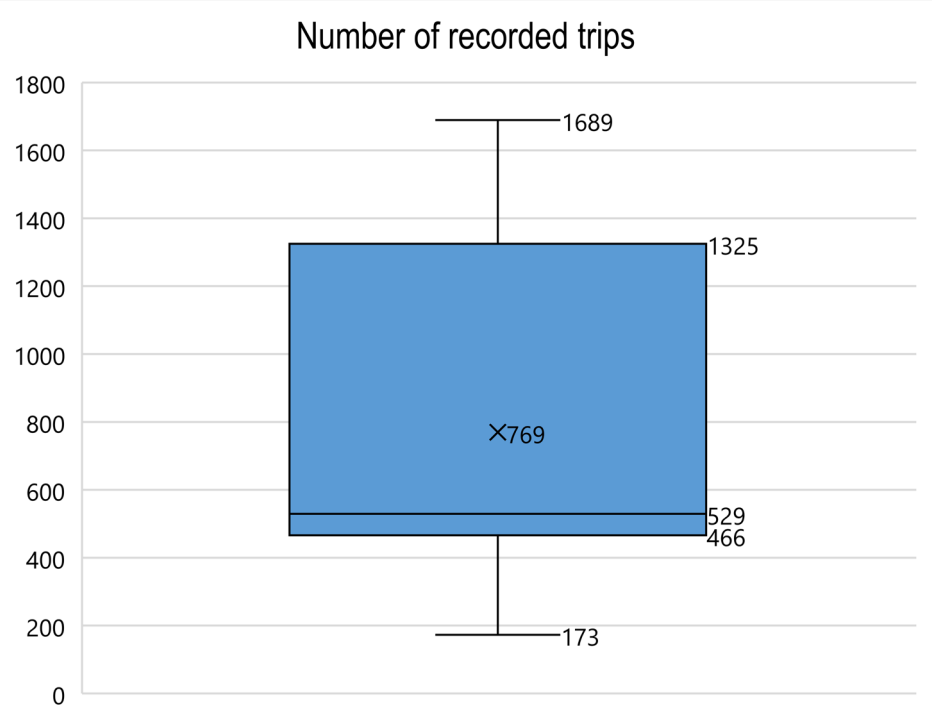


Figure 5.14: Number of recorded trips

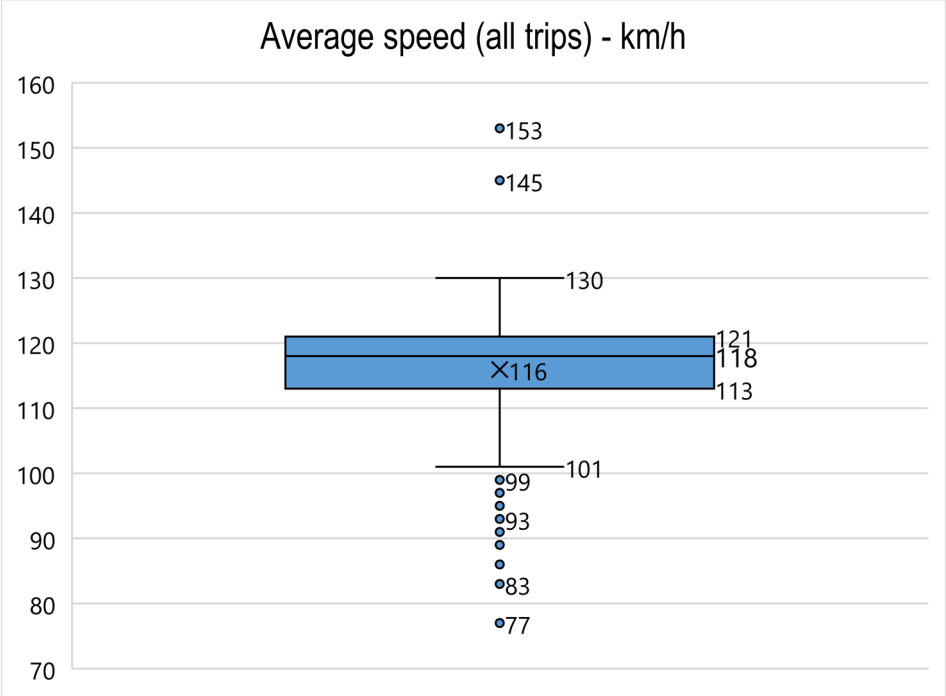


Figure 5.15: Average speed (all trips) – km/h

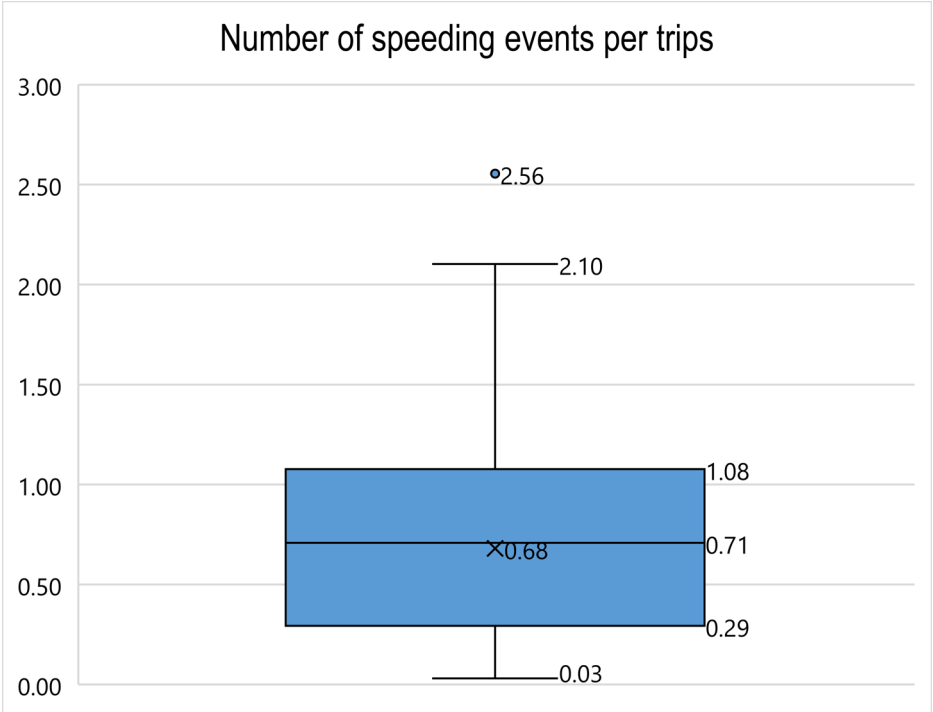


Figure 5.16: Number of speeding events per trips

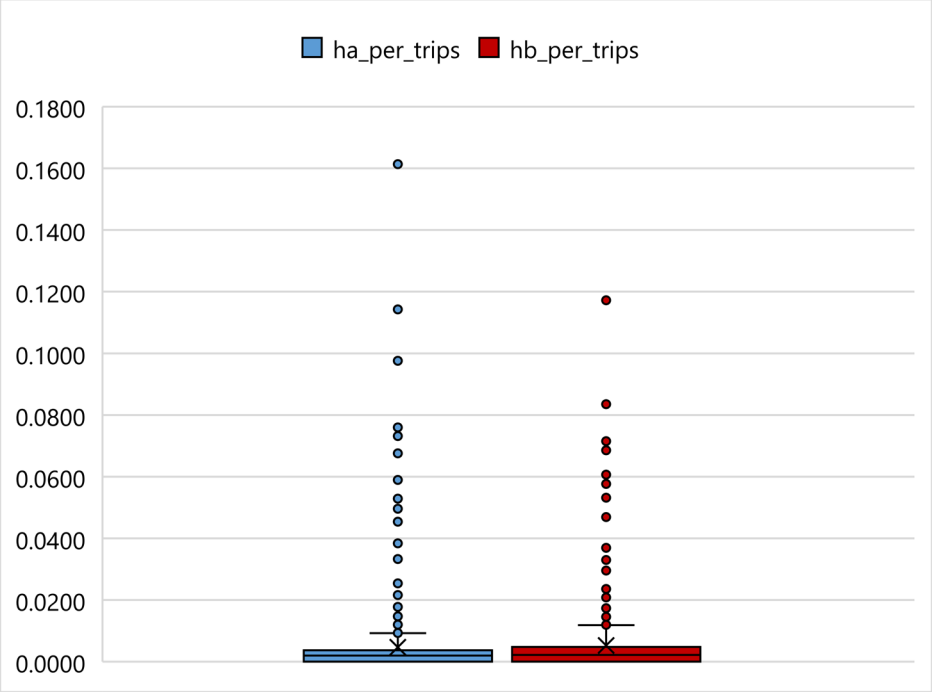


Figure 5.17: Number of harsh driving behaviour events (accelerations/brakings) per trips

6. Motorway Segment Analyses

6.1 Introduction

Motorways, also referred to as freeways, exhibit much lower crash rates, in terms of injury crashes per million vehicle kilometres, than other road types. Studies comparing motorways to standard rural and urban roads indicate 50% to 90% lower crash rates for motorways (European Commission, 2018). It has also been found that the extension of the motorway network is associated with a reduction in road fatality rates, while other road types do not present the same positive safety effects (Albalate & Bel, 2012). During the last few years, motorway length has increased substantially in many European countries (Papaioannou & Kokkalis, 2012). Elvik et al. (2017) evaluated the road safety effects of a new motorway in Norway through an Empirical Bayes before-after evaluation and found that injury severity was reduced markedly. In the case of Greece, the considerable improvement of its main road network from 750 km of motorways in 2007 to 2,200 km in 2018 was a key factor for the reduction in road fatalities by 54% during the period 2010-2020 (European Transport Safety Council, 2021).

Although motorways exhibit reduced crash rates compared to other road types, crashes still occur, and, due to high vehicle speeds, these crashes tend to be more severe. Therefore, there is still space for road safety improvements. In Greece, 50 road fatalities were recorded on motorways in 2019 and, towards this direction, a target of zero fatalities on motorways by 2030 has been set in the Greek Road Safety Strategic Plan for the period 2021-2030 (Yannis et al., 2023). Naturally, available funds for road safety interventions are not infinite. Consequently, decision-makers and stakeholders are forced to resolve their optimal allocation. Several quantitative techniques have been applied to enhance decision-making with regard to identification of segments' crash frequencies or risk levels and their prioritization in terms of potential upgrades.gt

Indicatively, Montella et al. (2008) developed two generalized linear CPMs with a negative binomial distribution error structure for estimating the safety of rural motorway segments in Italy. The first one considered total road crashes, while the second model considered only severe crashes. The key result of this research was that design consistency measures significantly affected road safety. La Torre et al. (2019) used a 5-year period dataset with fatal and injury crashes that occurred on 884 km of motorway segments in Italy, in order to develop two CPMs that could be applied and transferred to the entire Italian motorway network with proper calibration. That research provided a tool that enables the dealing with potential safety issues and helping in selecting treatments. Data Envelopment Analysis is another technique that

has been also used for the identification of hazardous motorway segments (Shah & Ahmad, 2020).

Regarding CPMs, they represent a reactive modelling approach primarily based on historical crash records collected within a long period of time (Theofilatos et al., 2019). Consequently, such approaches force road safety experts to wait for the occurrence of road crashes in order to identify the problems and examine measures for their prevention. Therefore, in recent years, researchers have increasingly started using indicators that are not based on historical crash data. As also mentioned in the literature review section of this doctoral dissertation, these indicators have been termed SSMs and can either be a proactive approach to road safety analyses (Wang et al., 2021) or even complement analyses that are based on historical road crashes (Johnsson et al., 2018). SSMs can be collected either through traffic simulation models (Gettman & Head, 2003; Mahmud et al., 2019) or under real driving conditions through smartphones (Paleti et al., 2017), equipped vehicles (Ambros et al., 2019), and video recordings (Johnsson et al., 2021). On one hand, SSMs can be time-based, deceleration-based, and energy-based. Among the most prevalent indicators of this subcategory of SSMs are PET, TTC, and DRAC (Bonela & Kadali, 2022). On the other hand, the recording of driving behaviour through sensors in vehicles and mobile phones has made harsh driving behaviour events an alternative subcategory of SSMs (Ziakopoulos et al., 2022; Stipancic et al., 2019).

Within this context, the objective of this section is threefold, specifically:

i. Investigate the relationship between road crash frequency in motorway segments and various explanatory variables based on road design characteristics and SSMs;
ii. Create risk-level clusters of the motorway segments based on crash and traffic data;
iii. Compare the classification performance of five well-known ML techniques which exploit road design data and SSMs for the prediction of the crash risk level of motorway segments.

A detailed description of the dataset that was exploited for the motorway segment analyses has been provided in the previous section of this dissertation.

6.2 Crash Frequency Model

As per the aforementioned, the first objective of this section was to develop a CPM in order to investigate the relationship between road crash frequency in road segments of the Olympia Odos motorway in Greece and various explanatory variables based on road design characteristics and SSMs. Since road crashes are count data, a count data modelling approach was selected.

As a first step, the variance and the mean of road crash frequency in the examined motorway segments were calculated in order to choose between Poisson regression and NB regression. In particular, it was estimated that the variance is equal to 3.98 and is higher than the mean which is equal to 2.02. For this reason, NB regression was chosen as the most appropriate modelling approach.

This analysis was conducted in R-studio (R Core Team, 2023) using the MASS R package (Ripley et al., 2013). A high number of regression model tests were conducted for different combinations of Table 5.1 variables. The optimal combination of variables was the one that had a sufficient number of statistically significant independent variables at a 95% confidence level (p-values ≤ 0.05) and the lowest possible AICc. Moreover, the independent variables were also checked for multicollinearity through the Variance Inflation Factor (VIF). A standard guideline is that VIF values higher than 10 indicate high multicollinearity. However, a threshold equal to 5 is also commonly used (Sheather, 2009). The dependent variable of the developed NB regression was the variable "TotCr18_20" of Table 5.1 and the results of the model are presented in the following Table.

Table 6.1: Statistical model for crash frequency in motorway segments

Independent Variables	Estimate	Std. Error	z value	Pr(|z|)	VIF
(Intercept)	-1.091	0.193	-5.667	**<0.001**	-
avg_AADT_18_20	$6.67 * 10^{-5}$	0.000	12.295	**<0.001**	1.014
ha_per_trips	7.604	2.174	3.499	**<0.001**	1.058
hb_per_trips	10.826	2.541	4.261	**<0.001**	1.066
len_seg	1.671	0.325	5.144	**<0.001**	1.012
AICc	*2,333.033*				

Based on this Table, it can be observed that all the explanatory variables are statistically significant at a 95% confidence level; there is no issue of multicollinearity as the VIF values are much lower than 5. With regard to the coefficients, it is revealed that road crash frequency in the examined motorway segments is positively correlated with the average AADT, showing that as traffic volume increases, the number of road crashes increases as well. This finding is also in alignment with the findings of a meta-analysis of 521 CPMs from more than one hundred studies (Høye & Hesjevoll, 2020).

Furthermore, it is demonstrated that both harsh accelerations and harsh braking have a positive relationship with the dependent variable, indicating that as the number of these two harsh driving behaviour events increases, crash frequency also increases. This is a noteworthy finding of the current doctoral dissertation as it confirms that harsh driving behaviour events present a statistically significant positive correlation with historical crash records. This conclusion means that these indicators can be meaningfully considered reliable SSMs that can be also used in proactive road safety analyses (Petraki et al., 2020; Ziakopoulos, 2021). Lastly, crash frequency is higher for motorway segments with higher length, as length serves as an exposure parameter.

6.3 Definition of Crash Risk Levels

The next stage of the statistical analysis carried out within the framework of this section focuses on the creation of crash risk level clusters of the examined motorway segments. For this purpose, agglomerative hierarchical clustering was applied through the “hclust” function of the stats R package (R Core Team, 2023).

As also mentioned in subsection 3.2.12, the Euclidean distance between single observations of the dataset and Ward’s minimum variance method as the linkage criterion were used. The variables considered for the formation of the risk level clusters of the motorway segments under consideration correspond to the number of total road crashes by segment length and the respective AADT of each segment. The selection of the number of clusters was based on the dendrogram illustrated in Figure 6.1.

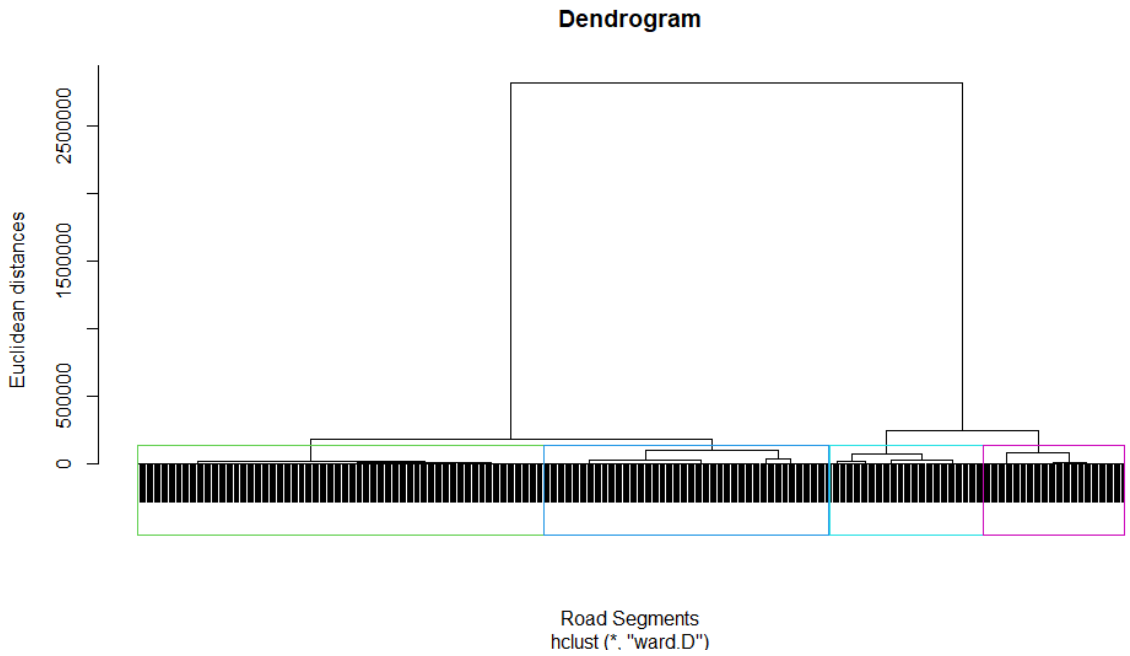


Figure 6.1: Hierarchical Clustering Dendrogram

As observed by Figure 6.1, and also based on the theoretical background of selecting the optimal number of clusters through the dendrogram, an appropriate choice of the number of clusters would be two. However, selecting only two clusters would lead to binary classification and to considerable detail and information loss. Therefore, in order to provide a more detailed description of the crash risk level of the examined road segments, four clusters were chosen as the next most appropriate option. Some basic descriptive statistics of the four crash risk levels are presented in the following Table.

Table 6.2: Descriptive statistics of the four crash risk levels of the examined motorway segments

Crash Risk Level	Count of Segments	Mean "TotCr18_20_len_seg"	Mean "avg_AADT_18_20"
1	96	7.57	20,876
2	104	4.55	17,218
3	193	3.25	8,086
4	275	2.76	6,726
Total	**668**	**3.87**	**10,786**

These numbers reveal a clear pattern whereby the first risk level class presents high average numbers of traffic volume and road crashes by segment length, while these figures decrease progressively for each subsequent class. It should be highlighted that these are subsample averages; hierarchical clustering does not readily include theoretical centroid calculations.

6.4 Comparing Machine Learning Techniques for Crash Risk Level Predictions

After defining the clusters of crash risk level, five ML classification models were developed in R-studio (R Core Team, 2023). The objective of these analyses was to identify the best performing model in terms of predicting crash risk level of the considered road segments. The response variable of these models was the multiclass variable “crash_risk_level” of Table 6.2. The independent predictors included in the models consisted of various road design characteristics and naturalistic driving behaviour metrics, represented by the following variables from Table 5.1:

- lanes: Number of through lanes,
- lane_width: Lane width (m),
- Curve1: Curve 1 - Radius R (m),
- Lcurve1_in_seg: Curve 1 - Length of curve in segment (m),
- median_width: Median width (measured from near edges of travelled way in both directions) (m),
- pav_ins_sh_width: Paved inside shoulder width (m),
- pav_out_sh_width: Paved outside shoulder width (m),
- dist_edginssh_barf: Distance from edge of inside shoulder to barrier face (m),
- dist_edgoutsh_barf: Distance from edge of outside shoulder to barrier face (m),
- speed_limit: Posted speed limit (km/h),
- avg_speed: Average speed (all trips) (km/h),
- speeding_per_trips: Number of speeding events per trips,
- ha_per_trips: Number of harsh accelerations per trips,
- hb_per_trips: Number of harsh brakings per trips.

The examined dataset was subsequently split into training and test subsets with a proportion of 75% and 25%, respectively. It is emphasized that the variable distributions were maintained to be similar during the splitting process. The training subset was used to train the classification models and included 501 segments, while the test subset was used to evaluate the classification performance of the models and amounted to 167 motorway segments. The core parts of the five models’ training, including the R packages that were used for their development, are demonstrated in the following Table.

Table 6.3: Basic elements of the five classification models' training

Classification Model	Key Elements
Logistic Regression	library(nnet), weights: 64 (45 variable)
Decision Tree	library(caret), Resampling: Cross-validated (5-fold), Method = rpart2, Maxdepth = 5
Random Forest	library(randomForest), Trees = 500, Variables tried at each split = 3, majority vote
Support Vector Machines	library(e1071), Type: C-classification, Kernel: radial, Cost:1, gamma = 0.0667
K-Nearest Neighbours	library(caret), Pre-processing: centred (14), scaled (14), Resampling: Cross-validated (10-fold, repeated 3 times), K = 5

As mentioned previously, the test subset was used to evaluate the performance of the developed models. The following Figures depict the confusion matrixes for the test dataset specifically, which reveal the distribution of predictions and targets for the different models.

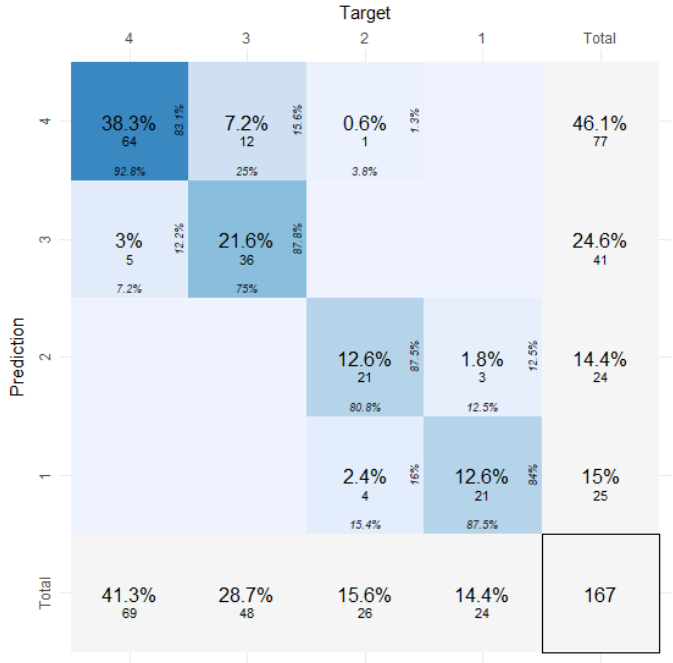


Figure 6.2: Confusion Matrix for the test dataset – Logistic Regression

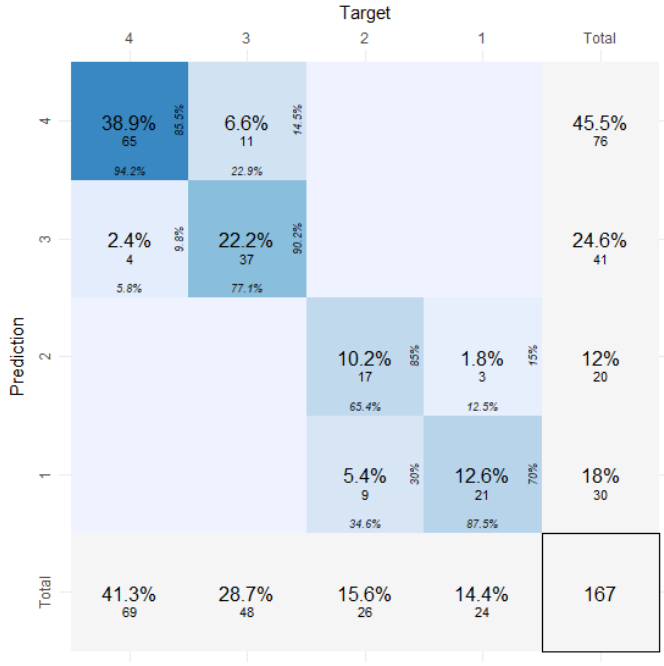


Figure 6.3: Confusion Matrix for the test dataset – Decision Tree

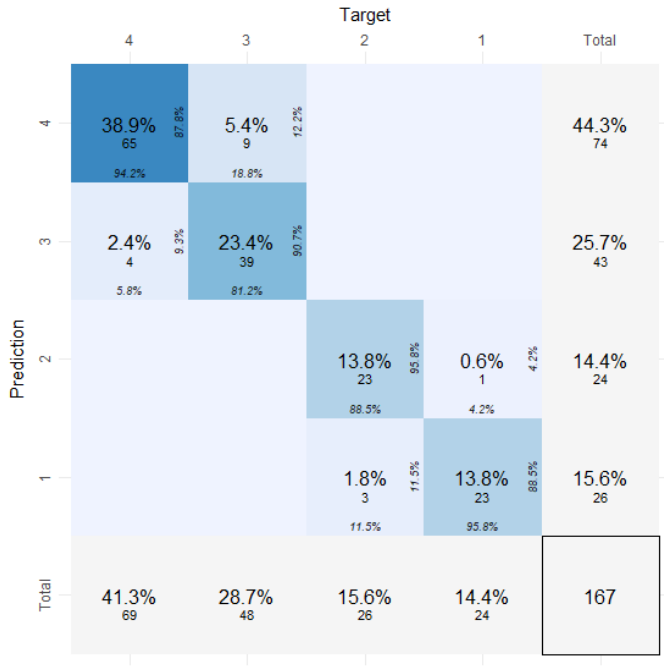


Figure 6.4: Confusion Matrix for the test dataset – Random Forest

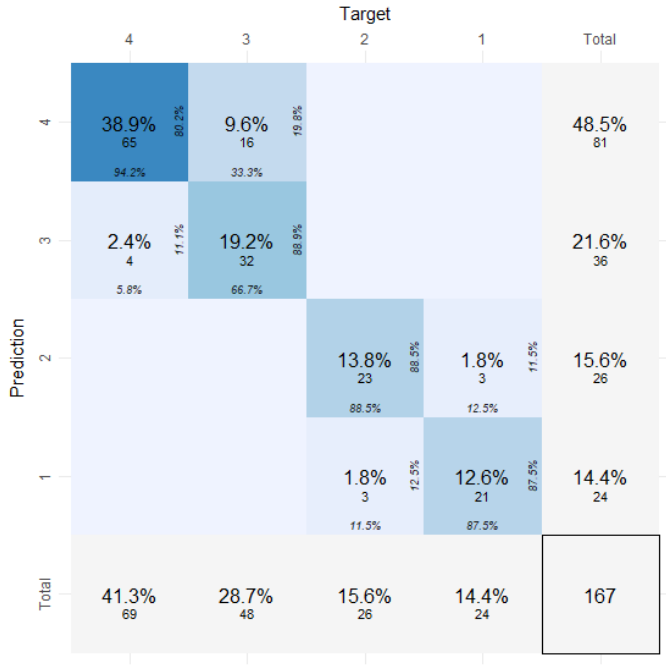


Figure 6.5: Confusion Matrix for the test dataset – Support Vector Machines

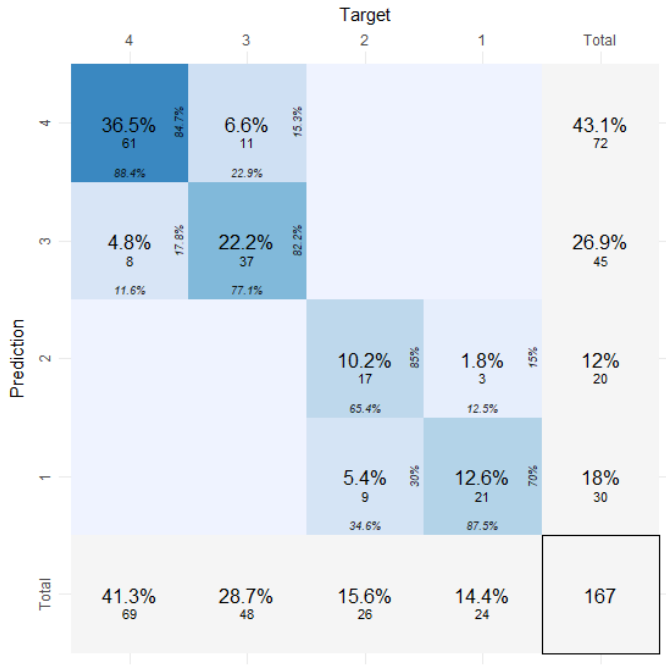


Figure 6.6: Confusion Matrix for the test dataset – K-Nearest Neighbours

As a first outcome, it can be gleaned that the diagonals of the matrices are highly populated. This is an indication that the proposed methodology allows for overall

accurate classification of crash risk levels without many losses due to misclassification in other categories (e.g., those on the secondary diagonal).

Regarding the quantification of accuracy performance, the sum of the cells of the diagonals indicates the overall accuracy of the five developed models. The resulting values in descending order are 89.9% for RF, 85.1% for LR, 84.5% for SVM, 83.9% for DT, and 81.5% for K-NN. However, in the developed classification models, the dependent variable includes four crash risk levels. Consequently, it is highly useful to investigate additional metrics for each particular category of the response variable, as overall accuracies may be misleading. To that end, Table 6.4 presents precision, recall, and the F1 score for each category, as well as the respective macro-averaged indicators for all the levels per developed ML classification model.

Table 6.4: Performance evaluation metrics per crash risk level and developed model

	LR	DT	RF	SVM	K-NN
Crash Risk Level	**Precision (%)**				
1	84.0	70.0	88.5	87.5	70.0
2	87.5	85.0	95.8	88.5	85.0
3	87.8	90.2	90.7	88.9	82.2
4	83.1	85.5	87.8	80.2	84.7
Macro-averaged	**85.6**	**82.7**	**90.7**	**86.3**	**80.5**
Crash Risk Level	**Recall (%)**				
1	87.5	87.5	95.8	87.5	87.5
2	80.8	65.4	88.5	88.5	65.4
3	75.0	77.1	81.2	66.7	77.1
4	92.8	94.2	94.2	94.2	88.4
Macro-averaged	**84.0**	**81.0**	**89.9**	**84.2**	**79.6**
Crash Risk Level	**F1 score (%)**				
1	85.7	77.7	92.0	87.5	77.8
2	84.0	73.9	92.0	88.5	73.9
3	80.9	83.1	85.7	76.2	79.6
4	87.7	89.7	90.9	86.7	86.5
Macro-averaged	**84.6**	**81.1**	**90.2**	**84.7**	**79.4**

Based on Table 6.4 performance metrics, it can be observed that the RF classification model was the best performing model for the classification of the crash risk level of motorway segments, with very satisfactory metrics for all levels. This outcome demonstrates the noteworthy value and utility of the developed RF model, as it can predict with high accuracy the crash risk level of a motorway segment, by using road design and naturalistic driving behaviour data. Therefore, this model could serve as a reliable method to identify the most hazardous motorway sections before road crashes occur and prioritize them. This model could also aid in the efficient allocation of available resources towards targeted road safety actions and measures.

6.5 SHAP values for Crash Risk Level Classifier

In this subsection, it was decided to calculate and provide SHAP values for the RF model as it demonstrated better classification performance than the other developed ML models. This approach was selected in order to overcome the difficult task of interpreting its outcomes. The DALEX R-package was used in order to calculate the SHAP values (Biecek, 2018). To create a representative instance of motorway segments, the median values of the continuous predictors were used. Medians were preferred instead of the mean values, as it can be concluded that the predictors are not normally distributed based on the outcomes of Shapiro-Wilk normality tests, skewness, and kurtosis values, which are presented in the Table 6.5 for each predictor.

Regarding the outcome of the Shapiro-Wilk test, it can be concluded that if the test is non-significant (p-value > 0.05), the distribution of the sample is not significantly different from a normal distribution (Thode, 2002). Moreover, a skewness value of 0 indicates a symmetric distribution, while positive or negative values indicate right or left skew, respectively. With regard to kurtosis, a value of 3 indicates a normal distribution, while higher or lower values indicate a more or less peaked distribution, respectively (Ho & Yu, 2015). With regard to categorical predictors, their most prevalent class from the training dataset was used. This approach ensured that the new instance was representative of the data and can be used to understand the model's prediction for similar instances.

Table 6.5: Skewness, kurtosis, and median values of numeric predictors in the training dataset

Abbreviation	Shapiro-Wilk (p-Value)	Skewness	Kurtosis	Median
lane_width	<0.001	-2.42	10.48	3.95
Curve1	<0.001	5.74	42.56	950.00
Lcurve1_in_seg	<0.001	0.49	2.27	197.65
median_width	<0.001	3.86	23.58	4.93
pav_ins_sh_width	<0.001	1.64	11.43	0.75
pav_out_sh_width	<0.001	-0.85	3.68	3.00
dist_edginssh_barf	<0.001	3.19	15.79	0.00
dist_edgoutsh_barf	<0.001	0.96	3.13	0.50
speed_limit	<0.001	-1.16	2.82	130.00
avg_speed	<0.001	-1.27	6.31	118.00
speeding_per_trips	<0.001	0.24	2.68	0.71511
hb_per_trips	<0.001	5.24	38.53	0.00215
ha_per_trips	<0.001	7.70	75.01	0.00197

Figure 6.7 presents the SHAP values plot for the multi-class RF classification model, which was determined as the best performing model among the developed five models. SHAP values for each feature are computed separately for each class and

the contribution of each feature to the model prediction for each class is displayed on the plot. The SHAP values can be positive (green bars) or negative (red bars) for each crash risk level, depending on whether the feature has a positive or negative contribution to the prediction for that class. It is noted that the purple boxplots of Figure 6.7 show the distribution of the attribution of a variable from every possible combination of variable layouts. It is also mentioned that Figure 6.7 demonstrates the SHAP values for a representative instance of motorway segments, which uses the median values of the numeric predictors.

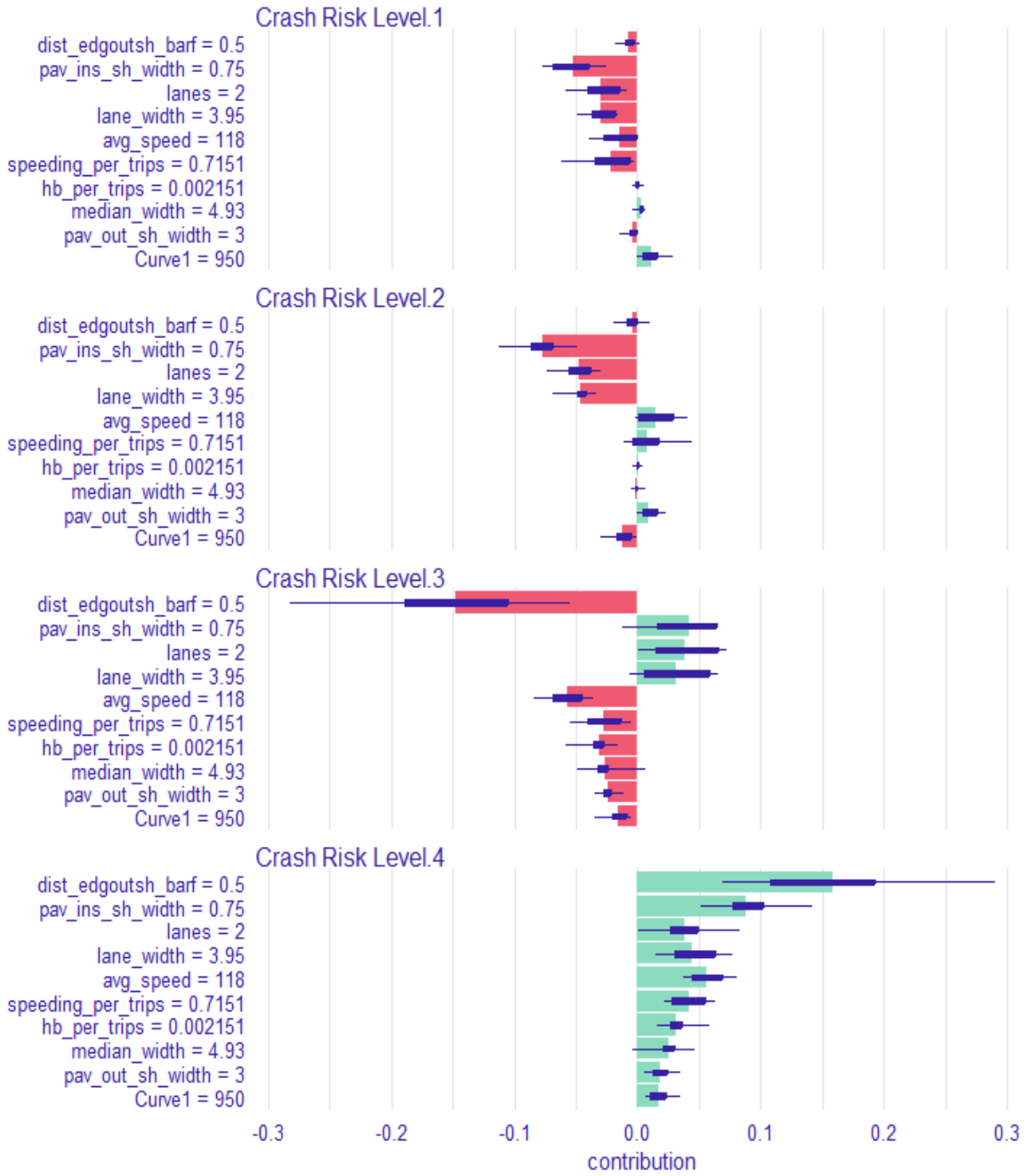


Figure 6.7: SHAP values for the RF model and a representative motorway segment

It can be observed that this representative motorway segment is more likely to belong to the lowest crash risk level, as denoted by the positive (green) bars of all predictors for this specific class. This crash risk level corresponds to overall safer locations with lower traffic volumes and road crashes by segment length than the motorway segments between the first and the third crash risk level (see Table 6.2).

It is worth noting that Figure 6.7 shows the contribution of only a subset of the variables that have been included in the multiclass classification RF model, as the other variables are not contributing much to the model's predictions and their contribution to the model's output can be, therefore, considered negligible.

A useful conclusion that can be drawn on this basis has to do with the fact that the harsh acceleration related variable does not make a significant contribution to the prediction of the segment crash risk level. Based on the literature, both harsh accelerations and harsh brakings constitute SSMs that can be used in various road safety analyses (Paleti et al., 2017; Stipancic et al., 2018b; Ziakopoulos et al., 2022; Nikolaou et al., 2023b). However, the results of this investigation suggest that harsh brakings may be more pertinent than harsh accelerations for predicting the crash risk level of motorway segments overall.

6.6 Discussion

The aim of this section was to exploit various road geometry data and SSMs for various road crash investigations in road segments of the Olympia Odos motorway in Greece. To that end, a unified database containing data on historical injury and PDO road crashes, road design characteristics, and SSMs of 668 motorway segments was utilized.

While the observational area and the data are singular for this study, they are viewed with three different approaches, each with a unique context. In particular, the first approach aimed to provide initial insights into the relationship significance and magnitude between road crash frequency and road geometry and SSM variables. However, since SSMs are still a new concept and their connection with hard road safety metrics such as crashes remains uncertain, it was fruitful to consider how these variables would perform for a clustering approach. To that end, the second model was applied as a first step, to reveal clusters that the segments can formulate based on crash and AADT data. The predictive power of road geometry and SSM variables was then tested on these clusters, having removed the variables used to obtain the clusters. Thus, in the present approach, the developed models contributed to prove that contextually, SSMs can be used to model crashes directly (negative binomial regression model – subsection 6.2), or indirectly, even without crashes, (ML classification models – subsection 6.4) when a type of safety categorization is established (clustering model – subsection 6.3).

To provide more detail, the negative binomial regression model was first developed to model motorway segment crash frequency. The results of this model pointed out that road crash frequency in the considered motorway segments is positively correlated with the traffic volume, the length of the segment, and the number of harsh accelerations and harsh brakings per segment trips. This analysis contributes to existing road safety literature by demonstrating a positive and statistically significant relationship between crash frequency and harsh driving behaviour events. Therefore, it can be concluded that such events can be a valid subcategory of naturalistic SSMs which can be used either to complement CPMs or as dependent variables of various road safety proactive analyses when detailed historical road crash data are not available.

As a further step of the statistical analysis, it was attempted to create crash risk level clusters of the motorway segments considering the number of road crashes by segment length and the traffic volume of each segment through the agglomerative hierarchical clustering technique. Segment length and traffic volume of each segment were taken into account in the clustering analysis, as the results of the negative binomial regression model revealed that these two variables have a statistically significant impact on the crash frequency of motorway segments. Based on the results of this clustering approach, four crash risk levels were defined.

Afterwards, these four levels formed the response variable of five ML classification models (LR, DT, RF, SVM, and K-NN). Data on road geometry characteristics and unsafe driving behaviours, such as rates of harsh brakings, harsh accelerations, and speeding duration, per trips in the considered segments were included as predictors in the developed models. Among these models, RF achieved the best overall and per crash risk level classification performance with very high and consistent scores of more than 89% (overall accuracy: 89.9%, macro-averaged precision: 90.7%, macro-averaged recall: 89.9%, macro-averaged F1 score: 90.2%). This finding is in alignment with previous studies, which report that RF is a promising modelling approach with high performance in either crash severity or crash risk prediction (Santos et al., 2022; Dimitrijevic et al., 2022). In addition, the SHAP values were calculated for a typical motorway segment in order to assist with the interpretation of the RF classification model, which is a black-box ML model. Based on the SHAP values of the naturalistic driving behaviour predictors, it was revealed that harsh brakings may serve as a more suitable SSM than harsh accelerations in terms of crash risk level prediction.

The findings of this section also suggest that the developed RF model could serve as a quite auspicious proactive road safety tool that could be used for the identification and prioritization of potentially hazardous motorway segments. Consequently, this approach could also assist to the best possible allocation of available resources for targeted interventions. Similar models could be applied to the rest of the motorway network in Greece, contributing to the achievement of the target of the Greek Road Safety Strategic Plan for the period 2021-2030, which aims at zero road fatalities on motorways by 2030 (Yannis et al., 2023). The inclusion of additional predictors that have not been considered in this research, such as the pavement conditions, may be beneficial towards the improved performance metrics of the ML models. Moreover, the prospect of extending the analyses included in this study to other types of road environments, such as urban and rural roads that are not motorways, is a quite challenging task that could be considered as well.

Naturally, this research is not without limitations. With regard to the extraction of road geometry data for Olympia Odos motorway, the results are obviously not an exact replication of the actual road design of the motorway and minor differences could be expected if a comparison with the as-built drawings of the project was made. Nevertheless, any differences would be minor and, although important from a designer’s point of view, they are not expected to be able to differentiate the study’s results. The negative binomial regression technique that was used for the development of the crash frequency regression model does not take into account unobserved heterogeneity and the effects of spatial characteristics of various road safety indicators. Another limitation of the current research is that tunnels and toll station segments were not considered in the analyses, leading to discontinuities in the research area.

However, these limitations can provide directions for future research efforts. Specifically, the inclusion of random effects in the crash frequency modelling approach could be considered in order to account for the unobserved heterogeneity. Moreover, spatial modelling approaches could be a promising alternative kind of modelling as it could consider the spatial dependency of road safety indicators.

Lastly, regarding the crash risk level classification models, it was found that RF outperformed the other developed classifiers in terms of predicting crash risk levels of the considered motorway segments. This is likely attributed to its ability to capture non-linear relationships, its robustness to hyperparameter choices, its ability to capture variable importance and its reduced risk of overfitting while remaining efficient. It should be mentioned that the performance of various ML models will probably vary across different datasets and the selection of the best performing approach that could serve as a proactive road safety approach should be completed with caution. The results of this research indicated that the RF classifier could be a strong candidate for this task. However, the development of additional classification models, such as Decision Jungle, which was found to outperform RF in a previous study (Ijaz et al., 2021), Gradient Boosting, and Linear Discriminant Analysis classifiers, could be considered in future research efforts.

7. Urban and Interurban Road Network Data Collection and Processing

7.1 Introduction

The investigation of road safety modelling data in Greece, as presented in Section 4, revealed that detailed crash prediction modelling is feasible only in motorways possessing high-quality crash data concerning crash locations and traffic attributes per road segment. However, based on the key findings of Section 6, it was concluded that harsh braking events could serve as a valid subcategory of naturalistic SSMs. These could be utilized as dependent variables in various road safety proactive analyses in cases where detailed historical road crash data are unavailable.

This section describes the development of a database for the road network in the Eastern Macedonia and Thrace Region, including urban and interurban roads. Detailed traffic and crash data (in terms of crash geo-location) were not available for the examined roads. Therefore, the developed database includes only geometric characteristics and naturalistic driver behaviour metrics for the examined road segments. Located in northeastern Greece, approximately 700 kilometers driving distance from Athens, this area was selected as a challenging location in terms of data availability, with the reasoning that if models converged in this area, they would be reasonably expected to converge in other regions of Greece.

The initial step of the data collection process involves the definition of a study road network within specific boundaries. Within this road network, an analysis is conducted on all road segments sourced from OSM to extract their geometric and network characteristics (Section 7.2). Subsequently, naturalistic driving behaviour data that were extracted from a smartphone application, including the number and location of harsh braking events and other metrics, are aligned with the corresponding OSM segments (Section 7.3).

This process leads to the development of a spatial dataset that encompasses aggregated behaviour metrics, as well as geometric and network characteristics on a segment level. Concluding this section, subsection 7.4 presents a summary table containing the segment-related variables that were ultimately analyzed in the subsequent section of this doctoral dissertation, accompanied by their abbreviations and relevant descriptive statistics.

7.2 Road Infrastructure Data

In this subsection of this doctoral dissertation, an analysis is conducted on all road segments sourced from OSM to extract their geometric and network characteristics. The OSM initiative is a collaborative effort, which offers user-generated street maps. About a decade ago, the accuracy of OSM data with regard to segment length was approximately 80% to 90%, with an error of ± 6 meters. Since then, OSM has undergone continuous enhancements (Haklay, 2010; Zhang & Malczewski, 2019).

It is also noted that the World Geodetic System 1984 (WGS84) which is widely utilized by GPS units and services, is also employed by OSM. All algorithms and analyses in this study have been conducted in R-studio (R Core Team, 2023) by using several packages. Specifically, the R library “osmdata” was used to extract the road segment data from OSM (Padgham et al., 2017). This library imports OSM data into R as simple features, which can be further processed with the R package “sf” (Pebesma, 2018).

The examined road network is illustrated in Figure 7.1 and consists of 6,103 road segments, with an average length of 288.8 meters, resulting in a total road network length of 1,763 kilometers. The distribution of road types is as follows: residential roads account for 67.8%, tertiary roads for 12.1%, secondary roads for 7.4%, motorways and motorway links for 3.8%, and the remaining 9% consists of other road types. It is noted that the maps presented in this dissertation were generated using the OSM/R-studio interface package and JavaScript library “leaflet” (Cheng et al., 2019).

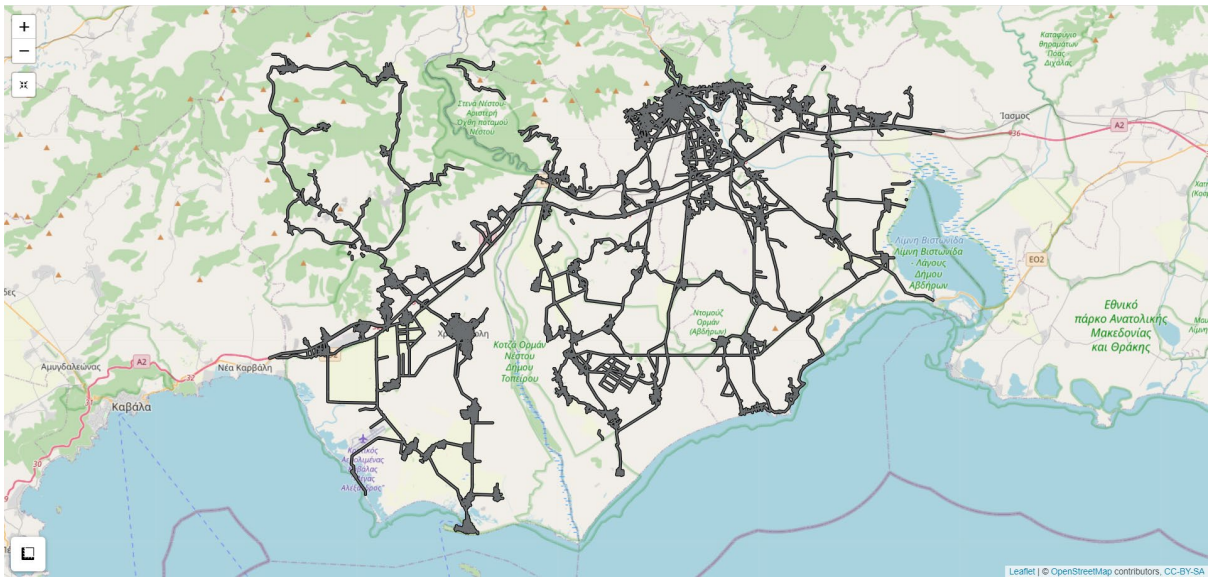


Figure 7.1: Examined road network of the Eastern Macedonia and Thrace Region (in grey)

Beyond road type, additional critical geometric characteristics, including road segment length, slope, and curvature, were also collected. Figure 7.2 illustrates the length of each analyzed road segment.

Figure 7.2: Length of the examined road segments

With regard to the curvature characteristics of the considered segments an index termed “efficiency” has been calculated. Specifically, it is a metric of segment linearity expressed by the ratio of the Euclidean distance between the start and end points of a road segment to the total segment length. It is a dimensionless ratio between 0 and 1, with higher values indicating a more linear road segment, while lower values indicate higher curvature.

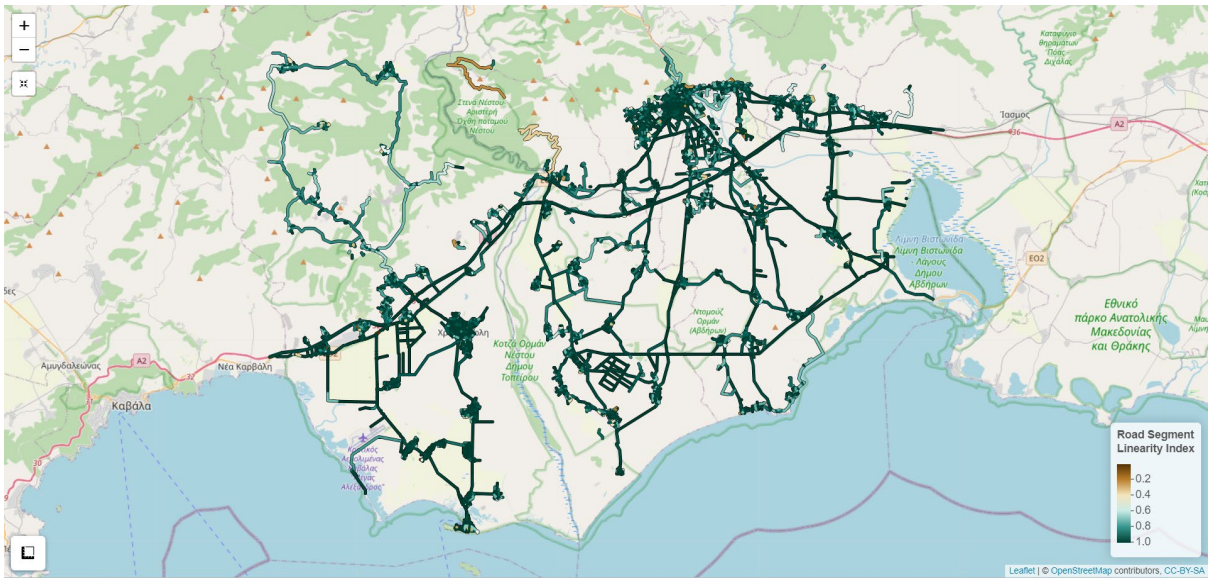


Figure 7.3: Linearity index (“efficiency”) of the examined road segments

Slope data were obtained using SRTM data. Figure 7.4 presents the slope class of the examined road segments. However, as detailed in subsection 4.4 of this doctoral dissertation, a notable disparity exists between these data and the surveyed elevations. Consequently, it was deemed more appropriate to exclude elevation data from the road crash risk assessment analyses of the subsequent section.

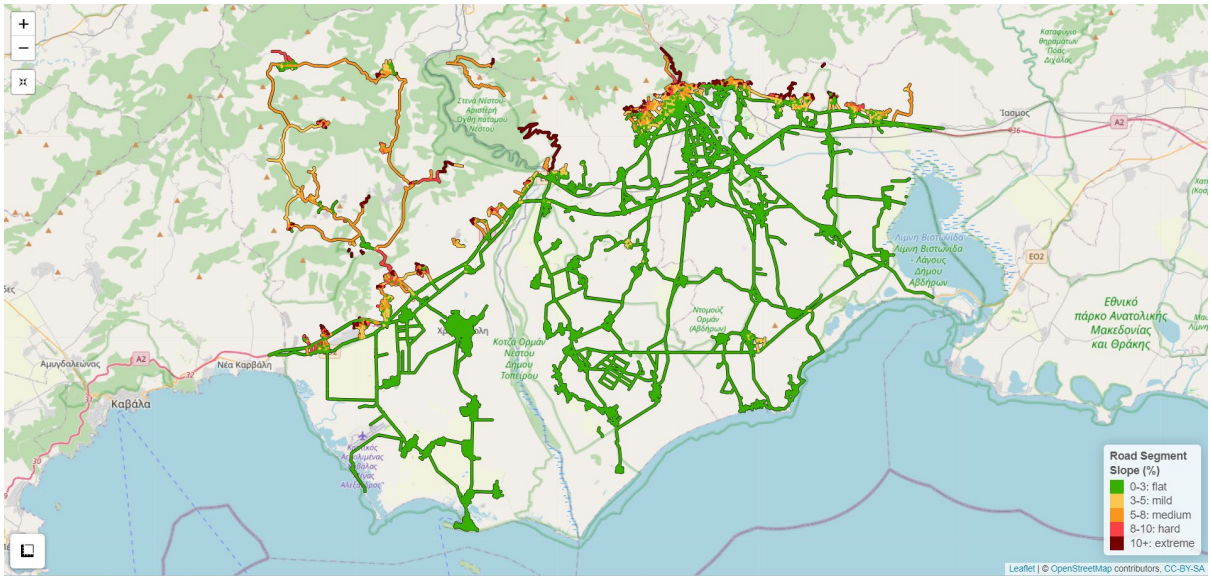


Figure 7.4: Slope class (%) of the examined road segments

7.3 Driver Behaviour Data

This doctoral dissertation utilizes naturalistic driving behaviour data obtained from an existing smartphone application developed by OSeven Telematics (https://oseven.io/), which is compatible with both Android and iOS devices. This application operates in the smartphone's background collecting sensors data without requiring any user initiation while driving or any other engagement. Sensors like the accelerometer, magnetometer, GPS and gyroscope are utilized to record the data. To clean and normalize the data, sophisticated ML algorithms, Data fusion and Big Data algorithms are implemented. Various forms of metadata are ultimately computed, including both exposure and driving behaviour indicators - such as trip duration, trip distance, driver speed, instances of speeding, the frequency of harsh braking and harsh acceleration incidents, and driver distraction from mobile phone use. Further details on the operation of this application have been provided in subsection 4.5 of this doctoral dissertation.

For the analyses of the road segments within the Eastern Macedonia and Thrace Region, data from 5,129 trips during 2021 were utilized. The mean trip duration was 634 seconds, with a standard deviation of 556 seconds. The histogram of trip durations is presented in Figure 7.5. Among these trips, a total of 2,889 harsh braking events were recorded.

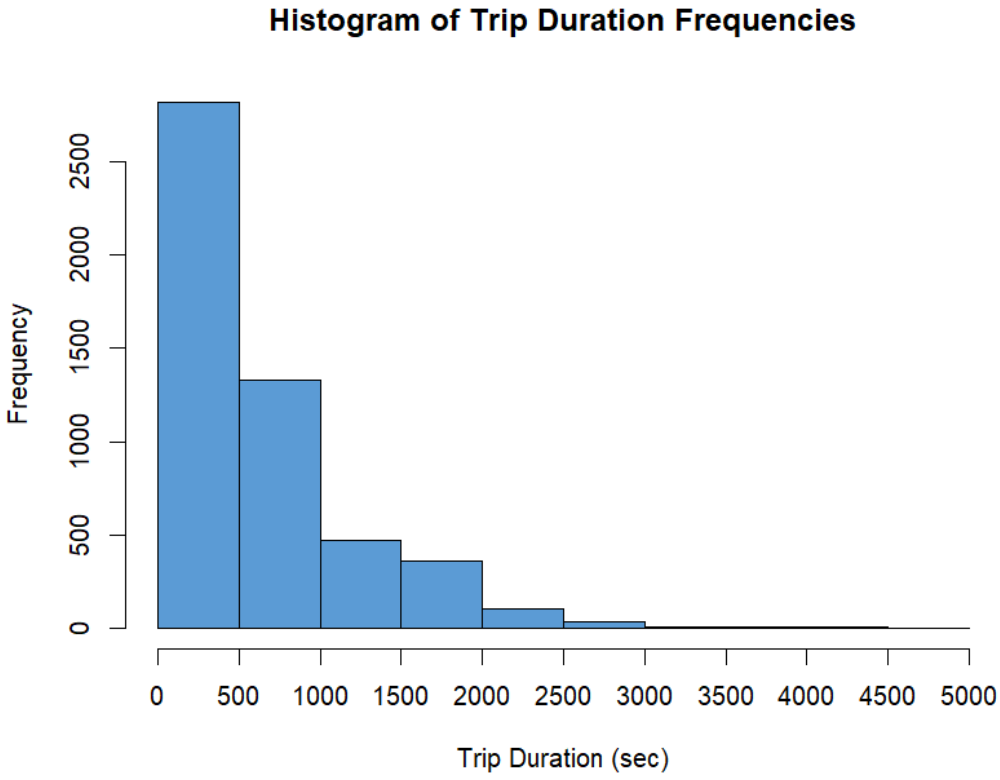


Figure 7.5: Histogram of trip duration frequencies in the examined road network

The data collection procedure aimed to create a spatial dataset that contains geometric and network characteristics, as well as aggregated driving behaviour metrics on the road segment level. The OSM segmentation was retained for the analysis, as there is a solid reasoning behind it that dictates that segments are separated when road/traffic conditions change (e.g., a lane is added or the speed limit changes). Smartphone data, which provide information for each second of a trip had to be associated with the corresponding road segment that each driver travelled through.

To this end, a spatial map-matching procedure was followed. Initially, the centroid of each road segment line-string was identified using the “st_centroid” function from the “sf” R library (Pebesma, 2018). Centroids are point-type quantities and represent the geometric center of each road segment. Next, the aggregated driving behaviour metrics were assigned to the nearest road segment centroid based on the latitude and longitude coordinates for each trip-second. This was accomplished using the "st_join" function and the "st_nearest_feature" geometry predicate function from the "sf" R library.

Figure 7.6 displays the duration (in seconds) of speeding per segment trips for the examined road segments, while Figure 7.7 illustrates the duration (in seconds) of mobile phone use during these trips. Additionally, Figures 7.8 and 7.9 present the number of harsh braking and harsh acceleration events, respectively, per segment trips for the examined road segments.

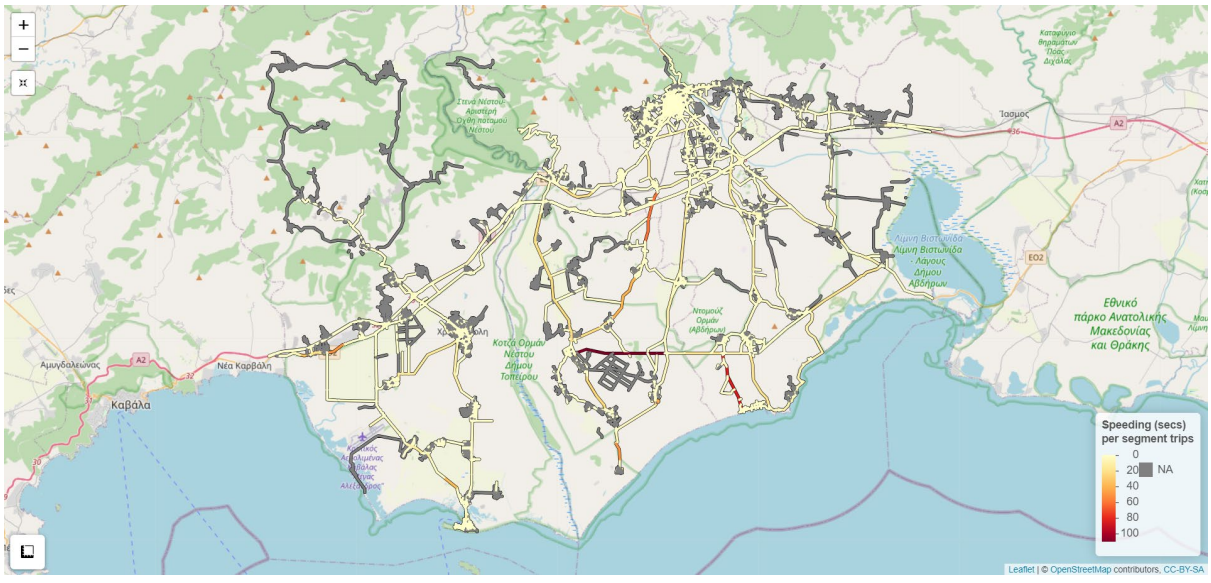


Figure 7.6: Speeding (secs) per segment trips

Figure 7.7: Mobile phone use (secs) per segment trips

Figure 7.8: Harsh braking events per segment trips

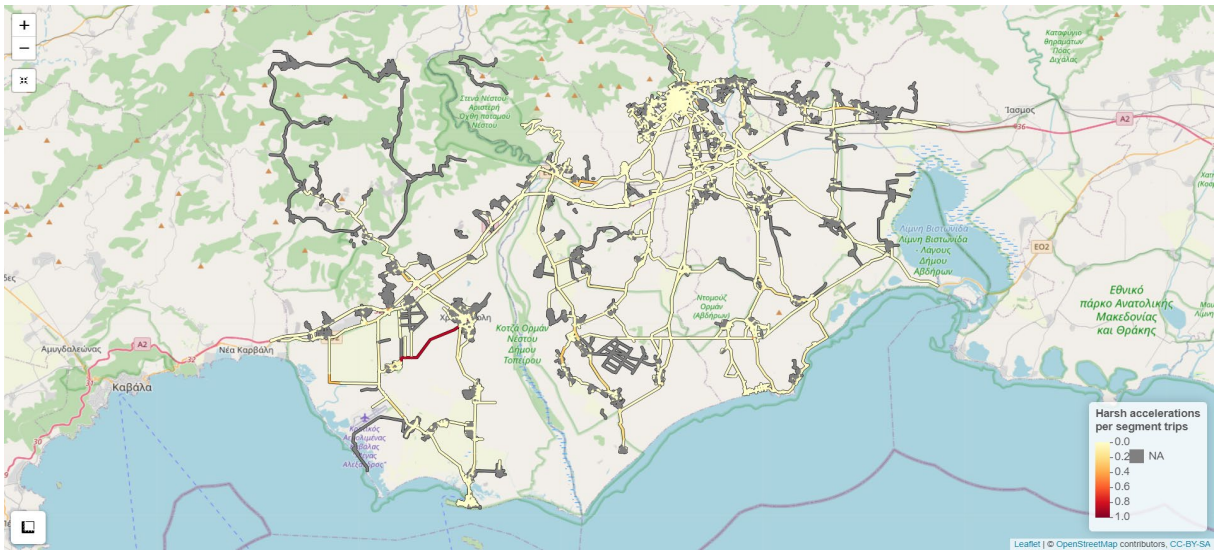


Figure 7.9: Harsh accelerations per segment trips

7.4 Descriptive Statistics

The procedure described in subsections 7.2 and 7.3 ultimately led to the development of a spatial dataset that includes aggregated behaviour metrics, as well as geometric and network characteristics at a segment level for 6,103 road segments in the Eastern Macedonia and Thrace Region. Table 7.1 presents the variables that were finally included and analyzed in the subsequent section of this doctoral dissertation, along with their abbreviations as well as key descriptive statistics.

Table 7.1: Geometric characteristics and driving behaviour metrics per examined road segment

Variable Description	Abbreviation	Descriptive Statistics
Number of trips [count]	trip_count	Min.: 0.00, Max.: 1,272.00, Mean: 32.10, Median: 1.00
Number of harsh braking events [count]	harsh_braking_count	Min.: 0.00, Max.: 117.00, Mean: 0.47, Median: 0.00
Duration of exceeding the speed limits [sec]	speeding_count	Min.: 0.00, Max.: 19,126.00, Mean: 16.05, Median: 0.00
Duration of mobile phone use [sec]	mobile_usage_count	Min.: 0.00, Max.: 2,461,00, Mean: 13,51, Median: 0.00
Segment length [m]	length	Min.: 2.05, Max.: 11,301.96, Mean: 288.84, Median: 123.07
Measure of segment linearity [dimensionless ratio]	efficiency	Min.: 0.01, Max.: 1.00, Mean: 0.94, Median: 1.00
Road type: motorway or motorway_link	motorway	Frequencies: No: 5,872, Yes: 231

The numeric values of Table 7.1, have been visually depicted on the examined roads in Figures 7.2-7.3 and 7.6-7.8. Additionally, Figure 7.10 illustrates the distribution of harsh braking frequencies among the examined segments. This variable also serves as the dependent variable for the models in the subsequent section of this thesis.

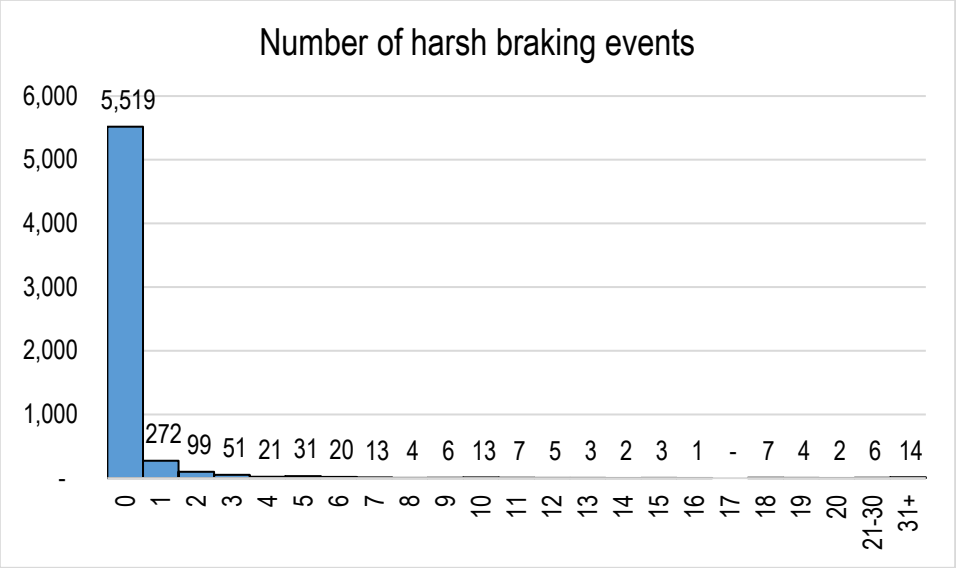


Figure 7.10: Histogram of harsh braking events in the examined road segments

8. Urban and Interurban Road Network Analyses

8.1 Introduction

The results of Section 6, focusing on 668 motorway segments in Greece, indicated that both the number of harsh braking events and harsh accelerations were positively correlated with the number of injury as well as property-damage only crashes (Nikolaou et al., 2023a). Moreover, it was also found that harsh brakings contribute significantly to predicting the crash risk level of the examined road sections, which is not the case for harsh accelerations. Therefore, and based also on the literature review findings of Section 2, it is concluded that harsh braking events are a plausible SSM that can be used either in various proactive road safety analyses before road crashes' occurrence or in cases of unavailable detailed road crash data (Nikolaou et al., 2023c).

Spatial autocorrelation often occurs when treating frequencies of harsh braking events as point-type data (Ziakopoulos, 2021; Ziakopoulos et al., 2022). In the case of the 6,103 road segments within the Eastern Macedonia and Thrace Region (Section 7), the observed Moran's I value is positive (0.0263) and statistically significant (p-value < 0.001), indicating that neighbouring road segments tend to have similar harsh braking counts. Consequently, it is important to consider spatial modelling techniques to capture spatial dependencies and enhance the reliability of the analyses, similar to studies that either exploit only historical road crashes or combine both SSMs and crash records (Ziakopoulos & Yannis, 2020; Aguero-Valverde & Jovanis, 2006; Stipancic et al., 2018b; Satria et al., 2021; Yang et al., 2021; Li et al., 2021a). However, spatial analysis of SSMs has not received significant attention in the road safety literature, making it a promising research direction in the field (Nikolaou et al., 2023b).

In light of this background, the objective of this section is to carry out spatial analysis of harsh braking events across various road environments within the Region of Eastern Macedonia and Thrace in Greece. This is achieved by exploiting smartphone driving behaviour data and OSM geometric data. The data collection and processing have been provided in Section 7 of this doctoral dissertation. This section focuses on analysing harsh braking event frequencies per road segment and correlating them with various road network characteristics and driving behaviour metrics. Spatial modelling techniques, including SEM, SLM, SZINB and SRF are employed on harsh braking events frequencies.

After this introduction, this section is organized as follows. Subsections 8.2 to 8.4 present and discuss the key results obtained from the developed models. Finally, Section 8.5 concludes the key findings and suggests potential avenues for future research.

8.2 Spatial Error and Lag Models

After completing the data collection process, which was described in Section 7, log-linear regression was selected as a suitable method for the preliminary analysis of the correlation between harsh braking events and the remaining variables of Table 7.1. Previous studies have established the advantages of utilizing GLMs for count data modelling (Lord & Mannering, 2010). However, considering harsh events as road segment attributes, particularly due to their substantially higher occurrence compared to road crashes within the same time period, linear regression can be explored in order to reveal potential linear relationships. This approach was also employed by Petraki et al. (2020) who demonstrated a robust relationship between harsh braking events and geometric and traffic characteristics, and was further supported by exploratory modelling, which highlighted that traditional count-based models (such as GLM - Negative Binomial Regression) were unsuitable for the current spatial dataset, probably due to the significant excess of zeros (Figure 7.10).

A series of various mathematical transformations of the independent variables such as logarithm usage were tested. It was also established that adding one harsh braking event to all segments allowed for better model fit, while enabling the inclusion of road segments with no events in the log-linear model with negligible numeric differences in the coefficients. The model was assessed for multicollinearity by means of the VIF. Overall, present results showed no multicollinearity as VIF values were lower than the established value of 5 (Sheather, 2009).

Moreover, as also pointed out in the introduction of Section 8, positive and statistically significant spatial autocorrelation was detected for the frequencies of harsh braking events among the examined road segments. Towards this direction, SEM and SLM have been developed in order to consider such spatial dependencies. The results of the log-linear model (baseline), SLM and SEM are presented in the following Table. These models were developed in R-studio (R Core Team, 2023) using packages “stats”, “spdep” (Bivand & Wong, 2018) and “spatialreg” (Bivand et al., 2021).

Table 8.1: Log-linear regression (baseline), SEM and SLM results for harsh braking events

Dependent variable: log (harsh_braking_count + 1)										
	Log-linear Model (baseline)				**Spatial Error Model (SEM)**			**Spatial Lag Model (SLM)**		
Independent variables	**Estimate**	**S.E.**	**P-value**	**VIF**	**Estimate**	**S.E.**	**P-value**	**Estimate**	**S.E.**	**P-value**
(Intercept)	-0.201	0.046	**<0.001**	-	-0.203	0.046	**<0.001**	-0.202	0.046	**<0.001**
trip_count	0.002	0.000	**<0.001**	1.484	0.002	0.000	**<0.001**	0.002	0.000	**<0.001**
log(1+length)	0.029	0.004	**<0.001**	1.152	0.029	0.004	**<0.001**	0.029	0.004	**<0.001**
log(1+speeding_count)	0.070	0.005	**<0.001**	1.260	0.071	0.005	**<0.001**	0.071	0.004	**<0.001**
log(1+efficiency)	0.126	0.056	**0.026**	1.084	0.127	0.056	**0.025**	0.123	0.056	**0.026**
mobile_usage_count	0.001	0.000	**<0.001**	1.499	0.001	0.000	**<0.001**	0.001	0.000	**<0.001**
motorway: yes	-0.071	0.022	**<0.001**	1.017	-0.069	0.022	**0.001**	-0.070	0.022	**0.002**
Lamda	-	-	-	-	0.021	0.010	**0.035**	-	-	-
Rho	-	-	-	-	-	-	-	0.020	0.008	**0.013**
Adjusted R^2	0.479	-	-	-	-	-	-	-	-	-
AIC	3,589.1	-	-	-	3,586.7	-	-	3,585.0	-	-
Residuals Moran's I	0.035	-	**0.018**	-	<0.001	-	0.496	0.003	-	0.418

As observed from the results presented in Table 8.1, the signs of the independent variables' coefficients remain consistent among the three models. In particular, both the length of the examined road segment and the number of trips per segment can be considered as proxy indicators of risk exposure and as expected, were found to be positively correlated with the number of harsh braking events, meaning that as either the length of the road segment increases or the number of trips taken on that segment rises, the number of instances where drivers perform harsh braking also tends to increase. These exposure metrics provide disjointed exposure dimensions for assessing the frequency of harsh braking events. More specifically, the road segment length represents geographical, infrastructure-based exposure, which is more fixed, while the number of trips per segment reflects naturalistic driving exposure, which depends on travel elements.

In addition, the positive sign of the beta coefficient of the efficiency index suggests that road segments with fewer curves have a higher number of harsh braking events. This implies that drivers perform more harsh brakings on straighter road segments, possibly due to the fact that they drive with higher speed or more aggressively when no curves are present. On the other hand, they tend to be more cautious in road curvature, which reduced visibility and introduces higher risk of run-off road instances. Moreover, the variables related to speeding and mobile phone use while driving, were found to be positively associated with the number of harsh braking events on road segments. In cases of exceeding speed limits, drivers are more likely to brake abruptly to avoid potential collisions or reduce excessive speed. Similarly, mobile phone use while driving can lead to distraction impacting reactions and increasing the likelihood of harsh braking. Lastly, the negative sign of the beta coefficient of the motorway variable indicates that the number of harsh braking events on motorways is lower than the respective number on other road types such as primary, secondary and residential

roads. This could be explained by smoother traffic flow, more lane options and longer visibility distances on motorways.

The Moran's I statistic indicates that the residuals from the baseline model have statistically significant positive spatial autocorrelation (Moran's I = 0.035 with p-value < 0.05). Based on this indication, the SEM was developed in order to address spatial autocorrelation. When considering the AIC values, it is observed that the SEM performs better than the baseline model. Moreover, the Lamda value of 0.021 is also statistically significant (p-value = 0.035), suggesting that the error term is spatially autoregressive. Based on the SEM residuals' Moran's I, it is also evident that there is no spatial autocorrelation in the residuals anymore as the Moran's I is close to zero and the p-value is higher than 0.05. The same is observed for the residuals of the SLM. Moreover, a statistically significant (p-value = 0.013) and positive spatial lag term "Rho" was obtained, indicating positive spatial autocorrelation. Finally, by comparing the values of the AIC criteria, it can be observed that the performance of the SLM outperforms the other two developed models.

The results of the SLM for the examined road network of the Eastern Macedonia and Thrace Region are visualized in Figure 8.1, whereas Figure 8.2 provides a zoomed-in view of Figure 8.1, focusing specifically on the center of the regional capital city of Xanthi.

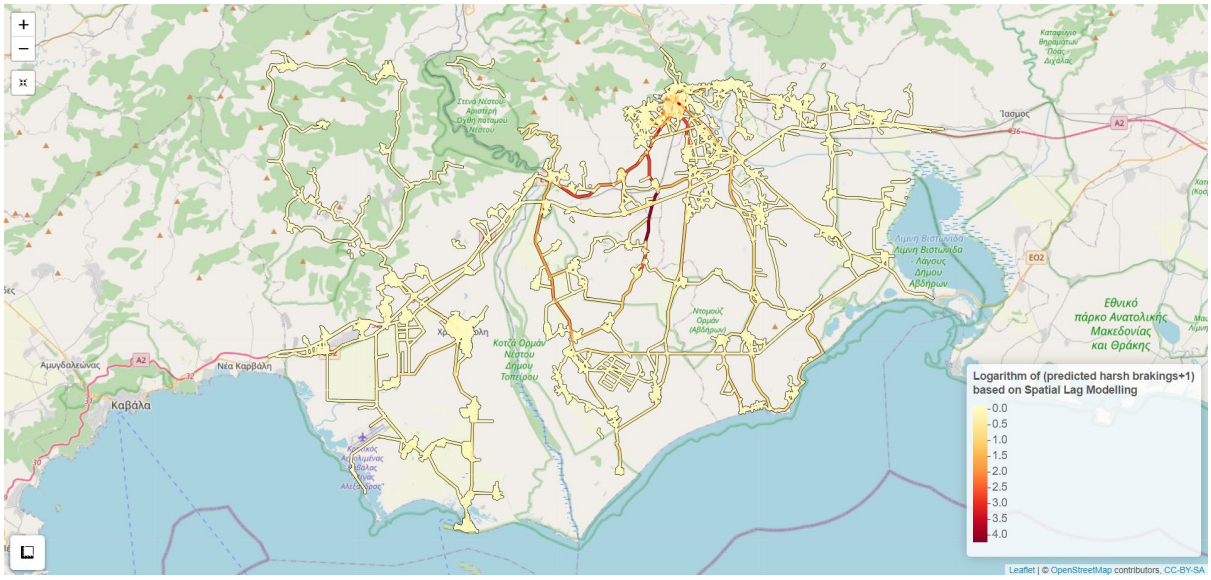


Figure 8.1: Visualization of the SLM results on the examined road network

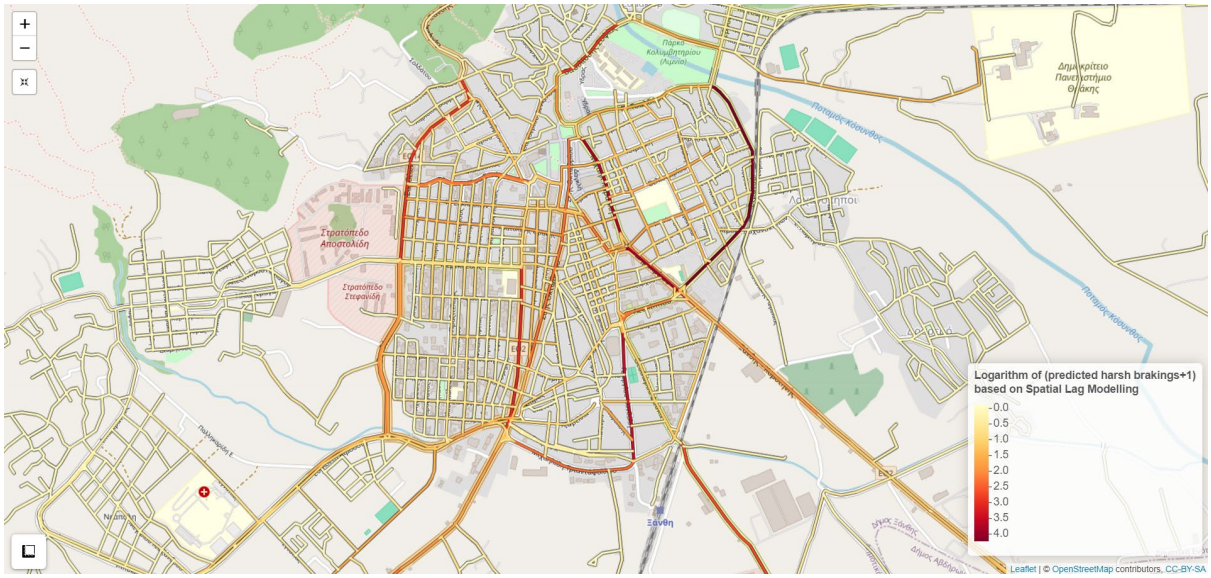


Figure 8.2: Zoomed-in view of the SLM results for the center of Xanthi

8.3 Spatial Zero-Inflated Negative Binomial Model

As noted in subsection 8.2, traditional GLM count modelling techniques, such as NB regression, could not be fitted to the dataset under consideration. This limitation likely stemmed from the significant excess of zero occurrences in harsh braking events across the analyzed road segments. Additionally, Figure 7.10's histogram depicting harsh braking frequencies on these road segments indicates a distribution following a ZINB pattern. Consequently, a ZINB model and a corresponding spatial model with spatial lag were constructed for this purpose. These models were developed using R-studio (R Core Team, 2023) and the "pscl" package (Zeileis et al., 2008; Jackman, 2020). ZINB models combine two components: one for modelling excessive zeros (using a logistic regression model) and another for modelling count data (using a NB regression model). The dependent variable of these two developed models is "harsh_braking_count" and the results are presented in the following Table.

Table 8.2: Zero-inflated Negative Binomial and Spatial Zero-inflated Negative Binomial results for harsh braking events

	Zero-Inflated Negative Binomial (ZINB)				**Spatial Zero-Inflated Negative Binomial (SZINB)**			
Count model coefficients (negbin with log link):								
Independent variables	**Estimate**	**Std. Error**	**z value**	**Pr(>|z|)**	**Estimate**	**Std. Error**	**z value**	**Pr(>|z|)**
(Intercept)	-1.527	0.112	-13.605	**<0.001**	-1.591	0.113	-14.111	**<0.001**
trip_count	0.004	0.000	9.192	**<0.001**	0.003	0.000	8.926	**<0.001**
log(1+speeding_count)	0.174	0.033	5.227	**<0.001**	0.191	0.032	5.869	**<0.001**
motorway: yes	-1.429	0.380	-3.758	**<0.001**	-1.359	0.367	-3.704	**<0.001**
length	0.0002	0.000	4.423	**<0.001**	0.0002	0.000	4.480	**<0.001**
log(1+mobile_usage_count)	0.273	0.038	7.242	**<0.001**	0.264	0.037	7.066	**<0.001**
spatial lag	-				0.109	0.032	3.436	**<0.001**
Log(theta)	-0.818	0.074	-11.017	**<0.001**	-0.794	0.074	-10.695	**<0.001**
Zero-inflation model coefficients (binomial with logit link):								
Independent variables	**Estimate**	**Std. Error**	**z value**	**Pr(>|z|)**	**Estimate**	**Std. Error**	**z value**	**Pr(>|z|)**
(Intercept)	4.209	0.364	11.551	**<0.001**	4.065	0.360	11.281	**<0.001**
trip_count	-0.434	0.104	-4.188	**<0.001**	-0.433	0.102	-4.258	**<0.001**
log(1+speeding_count)	-1.173	0.940	-1.248	0.212	-1.374	0.844	-1.628	0.103
motorway: yes	-1.763	2.267	-0.777	0.437	-1.355	2.019	-0.671	0.502
length	-0.0003	0.000	-0.864	0.388	-0.0003	0.000	-0.784	0.433
log(1+mobile_usage_count)	-0.402	0.172	-2.338	**0.019**	-0.421	0.177	-2.381	**0.017**
spatial lag	-				0.531	0.390	1.362	0.173
AIC	4,350.4				4,336.4			

The signs of the independent variables in the count component of the two ZINB models align with those of the three models (Log-linear, SEM, SLM) presented in Table 8.1. In particular, the results of Table 8.2 confirm that the variables "trip_count", "speeding_count", "length" and "mobile_usage_count" are positively correlated with the number of harsh brakings on the examined road segments, while the opposite is

the case for the variable “motorway”. Interpretations of these signs are covered in subsection 8.2 and are omitted here to prevent repetition. The only difference between the independent variables featured in Tables 8.1 and 8.2 lies in the absence of statistical significance for the “efficiency” variable in the ZINB models, leading to its exclusion.

Within the zero-inflated component of Table 8.2’s models, only “trip_count” and “mobile_usage_count” turned out to be statistically significant. In particular, their coefficients’ signs imply that an increase in these variables corresponds to decreased probabilities of zero harsh braking occurrences on the examined road segments.

Additionally, among the two models of Table 8.2, the SZINB model demonstrates superior data fit, evident from the AIC criterion values. Noteworthy is the positive and statistically significant (p-value < 0.001) spatial lag term in the count component, indicating positive spatial autocorrelation.

Visual representation of the SZINB model's results for the Eastern Macedonia and Thrace Region’s road network is displayed in Figure 8.3, with Figure 8.4 offering a more detailed view focused on the city of Xanthi.

Figure 8.3: Visualization of the SZINB results on the examined road network

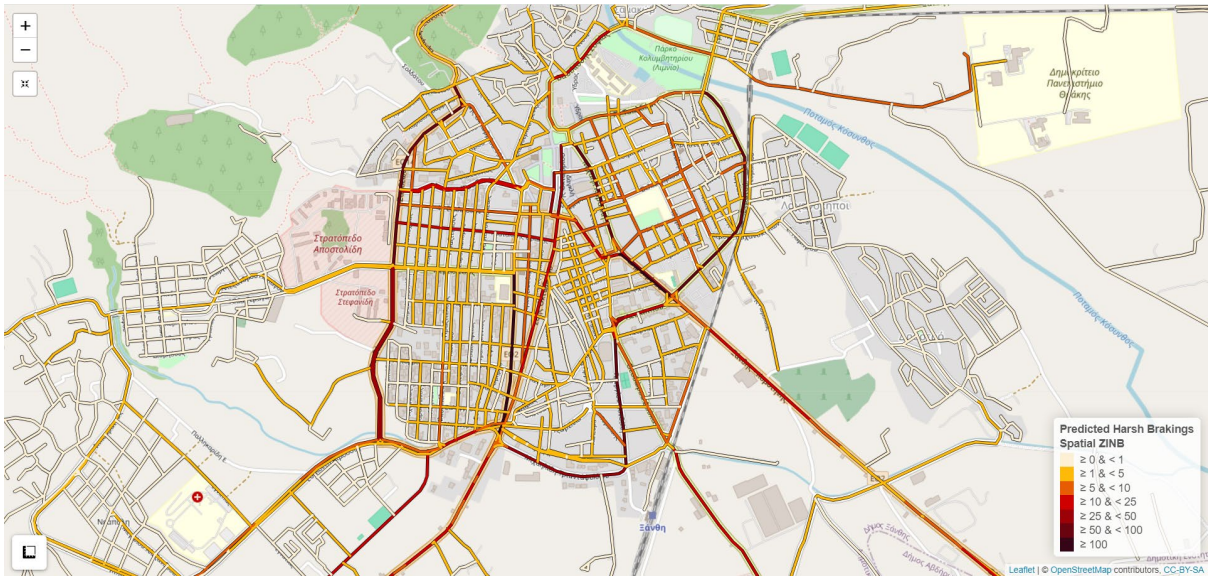


Figure 8.4: Zoomed-in view of the SZINB results for the center of Xanthi

8.4 Spatial Random Forest

The results of the five models, as seen in Tables 8.1 and 8.2, show a statistically significant relationship between the selected independent variables and the frequency of harsh braking events per segment. Thus, these variables were also selected as input to the SRF model. The R package “spatialRF” (Benito, 2021), which internally uses the R package “ranger” (Wright & Ziegler, 2015), was exploited for the development of the SRF model. It is noted that the dependent/response variable of the SRF model was “log (harsh_braking_count + 1)”.

Initially, a conventional non-spatial RF model is developed with defined distance thresholds for the examination of spatial autocorrelation in the residuals. If statistically significant and positive spatial autocorrelation exists, the SRF using spatial predictors is subsequently applied. These predictors are derived from the distance matrix of the considered road segments and are used as explanatory variables in the SRF model following Hengl et al. (2018).

It is noted that information overlap and over-parameterization due to excessive covariate usage are not problematic because RF has built-in protections against overfitting, allowing for the fitting of models with a large number of covariates, even surpassing the number of observations (Biau & Scornet, 2016; Hengl et al., 2018). By including spatial predictors, the SRF manages to enhance its capability so as to minimize the spatial autocorrelation in the residuals and provide more precise variable importance scores. Moreover, the inclusion of spatial predictors in the model can indirectly address some aspects of unobserved heterogeneity in the data, which pertains to variations in the response variable that are not accounted for by the remaining observed predictors included in the model.

Figure 8.5 illustrates information on the non-spatial RF model residuals. Specifically, its upper panels demonstrate the results of the normality test, while the middle panel indicates the relationship between the residuals and the fitted values and the lower panel shows the Moran’s I of the residuals across distance thresholds and their respective p-values (positive and statistically significant for distances between 0 and 2000 meters).

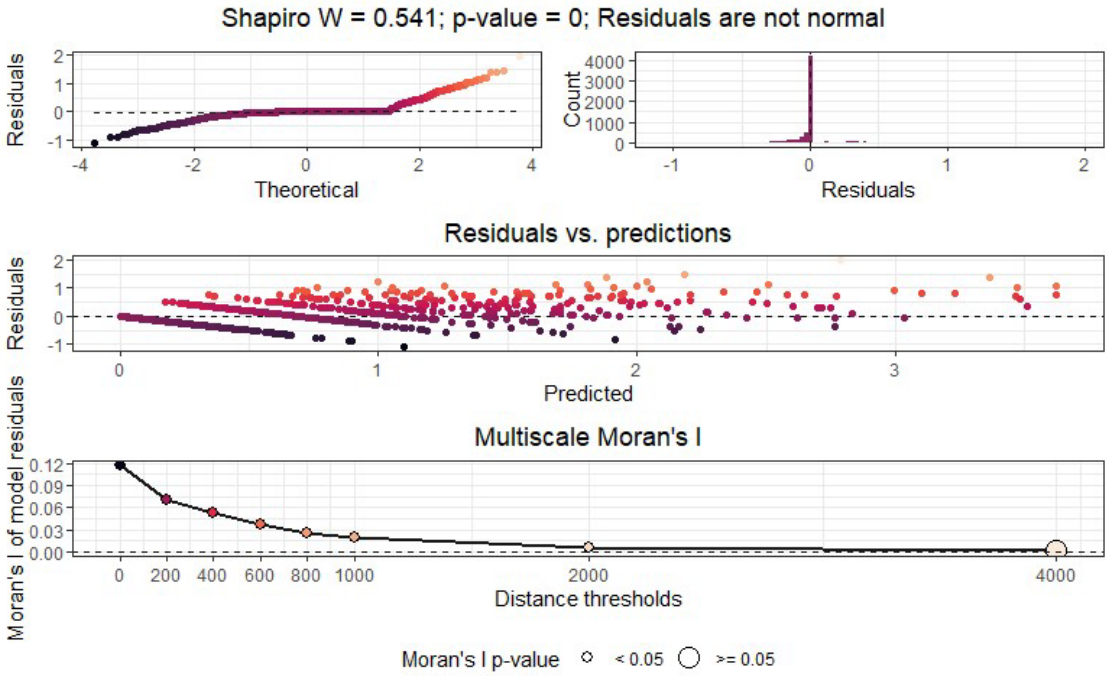


Figure 8.5: Non-spatial RF model residuals

The presence of spatial autocorrelation in the residuals, based on the high Moran's I residuals as indicated by the large y-values in the lower distances of the bottom plot, indicates that the non-spatial RF model did not fully capture the spatial dependencies in the data. In order to minimize the spatial autocorrelation of the residuals, the non-spatial RF model was transformed into a SRF model by adding the columns of the distance matrix of the road segments as spatial predictors (Hengl et al., 2018). Figure 8.6 presents the Moran's I of the residuals of the SRF model.

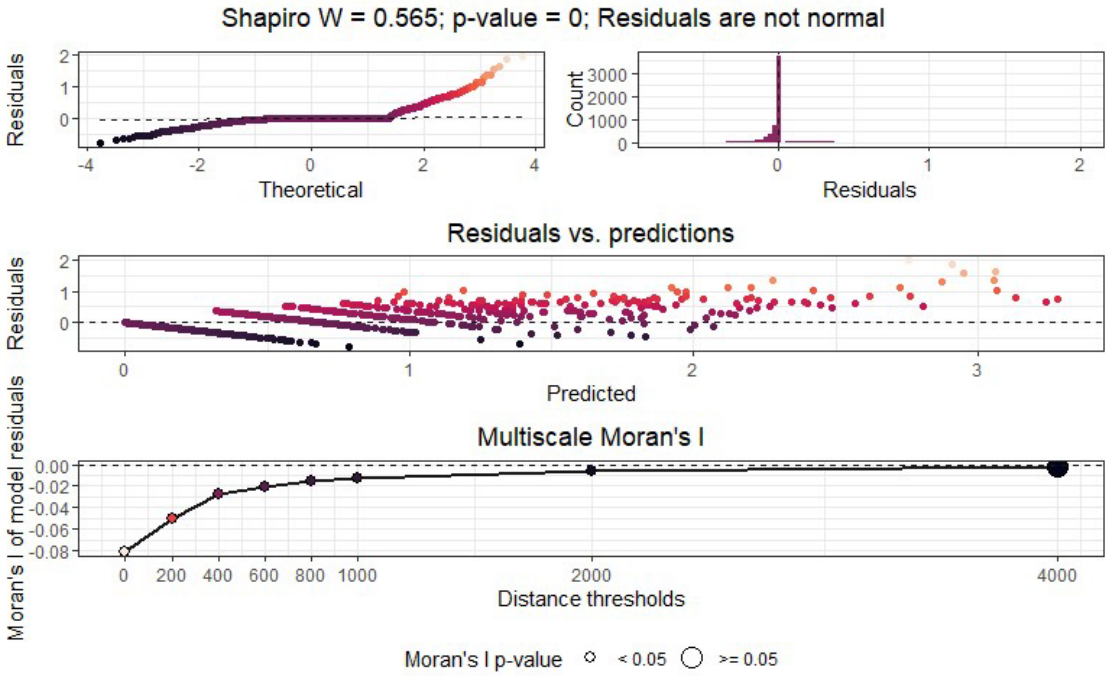


Figure 8.6: SRF model residuals

On the basis of the lower panels of Figures 8.5 and 8.6, it can be observed that the SRF was able to reduce the absolute values of the Moran's I statistics. A likely explanation for the change from positive to negative spatial autocorrelation (in the residuals) could be the inclusion of the additional spatial explanatory variables in the SRF model that incorporates spatial dependence structures as these additions help capture the spatial relationships among the observations.

However, it should be noted that the absolute values of the Moran's I index can provide some insight into the strength of spatial autocorrelation, but it is not the sole criterion for model evaluation. To that end, Table 8.3 provides the key parameters of each RF model, along with some key model performance metrics.

Table 8.3: Key parameters and performance metrics of RF models

	Non-spatial RF	SRF
Number of trees	500	500
Sample size	6,103	6,103
Number of predictors	6	6,109
Mtry	2	78
Minimum node size	5	5
R^2 (out-of-bag)	0.526	0.440
R^2 (cor (observed, predicted)^2)	0.900	0.928
Pseudo R^2 (cor (observed, predicted))	0.949	0.964
RMSE (out-of-bag)	0.309	0.336
RMSE	0.156	0.150

When examining typical metrics (not out-of-bag metrics), for instance, R^2, Pseudo R^2 and RMSE, it is observed that the SRF outperforms the non-spatial RF model. A spatial model can capture spatial dependencies among the considered data points leading to a better fit to the observed data compared to non-spatial model. However, based on the out-of-bag performance metrics, it is found that non-spatial RF model outperforms the SRF, declaring that the non-spatial model is likely performing better in terms of generalization on unseen data. This conclusion can be also enhanced by Figure 8.7, which compares the predictive performance of the two RF models across thirty spatial folds. It is noted here that spatial folds are subsamples of the initial data that are separated in location clusters, a concept known as spatial cross-validation (Lovelace et al., 2019). Thus, the localized spatial aspects and unobserved traits are retained through the cross-validation process as opposed to traditional random cross-validation.

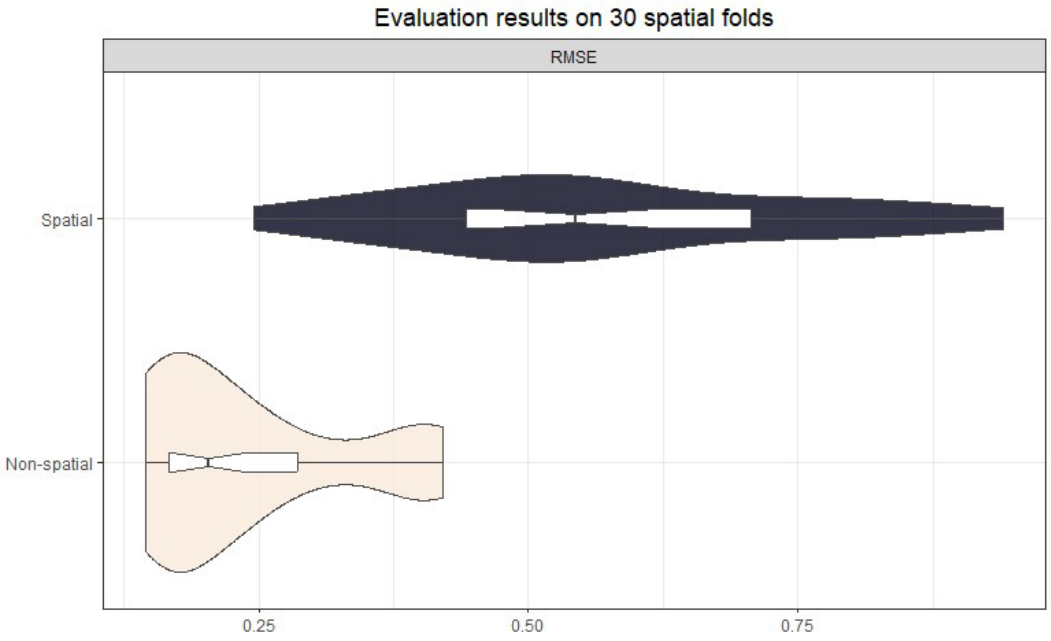


Figure 8.7: RMSE across 30 spatial folds

Within the framework of the two RF models' development, the permutation variable importance technique was also employed to assess and rank individual predictors on the basis of their relative importance. Variable importance scores are visualized in Figure 8.8, demonstrating the increase in mean error (computed on the out-of-bag data) observed across trees when a predictor is permuted. This approach provides valuable insights into the relative contributions of predictors in both spatial and non-spatial RF models. As a result, the SRF has an additional set of variable importance scores for the spatial predictors, with the maximum importance of a few of these spatial predictors matching the importance of the second and third most important predictors.

In both RF models, the number of trips per examined road segment (which serves as a naturalistic driving exposure metric), was found to be the most influential predictor,

highlighting its significant relevance in predicting the frequency of harsh braking events. On the other hand, the motorway variable exhibited the lowest importance in both RF models, indicating that road type is relatively less valuable in predicting the number of harsh braking events. This finding may suggest that factors other than road type such as driver distraction and speeding, might play a more crucial role in influencing harsh braking events frequencies.

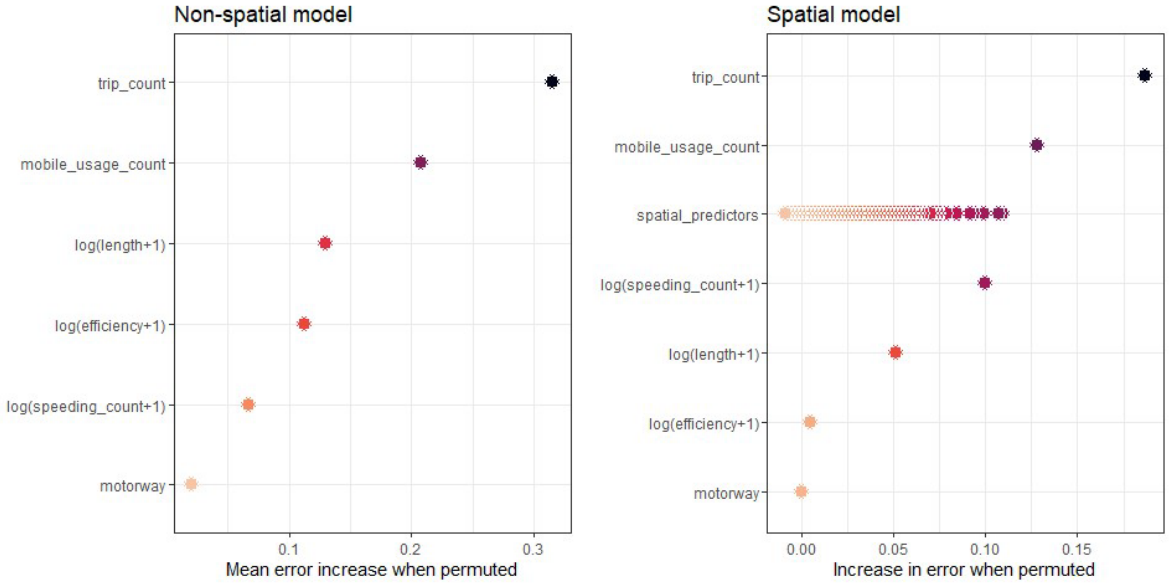


Figure 8.8: Permutation importance computed on the out-of-bag data

8.5 Discussion

The objective of this section was to conduct spatial analyses of harsh braking events by exploiting smartphone driving behaviour data and OSM road network characteristics, aiming to enhance the current SSM knowledge. The examined road network consists of 6,103 road segments located in the Region of Eastern Macedonia and Thrace in Greece. A spatial dataset consisting aggregated naturalistic driving behaviour metrics, as well as geometric and network characteristics on a segment level was analyzed. Initially, non-spatial modelling techniques, such as log-linear, ZINB and conventional RF regression models were employed on harsh braking events frequencies. However, the existence of statistically significant spatial autocorrelation highlighted the need for the development of spatial models, such as SEM, SLM, SZINB and SRF, in order to take into account such spatial dependencies.

The results of the log-linear regression model, SLM, SEM, ZINB and SZINB showed consistent signs of the beta coefficients of the considered variables across all models. In specific, road segment length and the number of trips per segment were identified as proxy indicators of risk exposure, positively correlated with harsh braking events. Furthermore, the efficiency index (statistically significant only in the log-linear model, SEM and SLM), related to the linearity of road segments, showed a positive correlation with harsh braking events, indicating that drivers tend to brake harshly more often on road segments with fewer curves. Variables related to speeding and mobile phone use were also positively associated with harsh braking events, while motorways exhibited fewer harsh braking events compared to other road types. It was also found that the SLM surpassed both the log-linear model and the SEM, with lower AIC values and absence of spatial autocorrelation in its residuals. Lower AIC values, indicating a better fit, were also observed for the SZINB model compared to the non-spatial ZINB model.

Moreover, the SRF reduced the absolute values of spatial autocorrelation in the residuals compared to the respective values of the conventional RF. In addition, the SRF outperformed the non-spatial RF model in terms of model fit to observed data, but the non-spatial model performed better in terms of generalization to unseen data. This is a typically expected finding, as spatial structures would be very challenging to transfer to completely unexamined areas in a manner that is informative and that would provide an edge in forecasting. Regarding variable importance ranking, the number of trips per examined road segment emerged as the most influential predictor in both models, highlighting its significance in predicting harsh braking events. A key takeaway is that for causal or exploratory ML analysis in a given area, spatial cross-validation would be reasonably more fruitful than its random counterpart. This would apply especially in cases where few variables are present in the data, as the unobserved spatial effects would be more pronounced then.

Overall, SSMs have immense potential for road safety monitoring, countermeasure assessment and improvement, and rapid expansion of road safety data coverage. In

academia, SSM modelling exercises have emerged in recent years. Apart from contributing in that field, this section demonstrated that with the necessary effort, SSM-based spatial models can be used in scarcely-studied areas. Aided by technological developments such as telematics, which enable scalable and expedient data collection, high-quality data applications and monitoring in such areas is possible and can even be converted to the norm in the short term.

Despite the valuable insights gained from this section, a significant limitation that needs to be acknowledged is the lack of available traffic data (AADT or real-time) per examined segment, which could have provided additional insights into the influence of traffic volume on harsh braking events. However, the absence of AADT was attempted to be tackled by using the number of trips and the segment length as substitute risk exposure metrics. Several microscopic and mesoscopic spatial analysis studies have been shown to include more disjointed parameters such as land use (Ziakopoulos & Yannis, 2020), however in this investigation it was considered that SSM models warrant more directly related variables.

In summary, this section provides valuable insights into the relationship between independent variables and harsh braking events, highlighting the relevance of exposure metrics and the impact of spatial autocorrelation in the models' development. Authorities can organize public awareness campaigns to educate drivers about the dangers of speeding and distracted driving, emphasizing on the positive correlation between such behaviours and harsh braking events, as revealed in this doctoral dissertation. Furthermore, leveraging such spatial modelling techniques, authorities can identify high-risk areas for harsh braking events and deploy targeted enforcement efforts to address specific road safety issues.

9. Conclusions

9.1 Dissertation Overview

Recognizing road safety as a crucial public health issue with significant societal and economic consequences, it is essential to understand the multifaceted nature of road crashes. Road crashes are influenced by various parameters that can be divided into three distinct categories: (i) road users, (ii) vehicles, and (iii) road infrastructure and environment. Notably, a substantial percentage of road crashes, up to 94%, can be attributed to human factors and errors, either exclusively or partially.

Given the aforementioned context, the main objective of this dissertation is to assess road crash risk by fusing infrastructure, traffic, and driving behaviour data. This integration of data presents a promising avenue for research. Nevertheless, the practical implementation of this data fusion is frequently hindered by challenges such as insufficient availability or suboptimal quality of the data.

Within the framework of this dissertation, an extensive literature review was conducted. The aim of this literature review process was to provide a review of the scientific literature of studies exploiting SSMs in historical crash record investigations. SSMs encompass a wide range of metrics and parameters, which are not directly derived from or rely on crash data. From the review process, it was concluded that SSMs are steadily gaining ground in the road safety literature as they are a sustainable way of gauging road safety and allow the conduction of analyses without necessarily requiring historical road crash records. These indicators can either be an alternative to road safety analyses or even complement analyses that are based on historical crash records. Moreover, the rapid and continuous progress in the field of technology makes it increasingly easier to collect such metrics. SSMs such as time-to-collision, harsh braking, post-encroachment time and so on, are widely proposed in transportation science and are particularly useful in order to evaluate driving risk and assess road crash risk.

Subsequently, the following research questions were formulated:

Question 1
How can infrastructure, traffic and driver behaviour data be fused and analyzed to derive meaningful conclusions for road crash risk assessment?

Question 2
a) Can harsh driving behaviour events be meaningfully considered reliable SSMs?
b) Is there a statistically significant positive correlation between harsh driving behaviour events and historical road crash records?

Question 3

Is it possible to predict the crash risk level of road segments by exploiting road geometry characteristics and driver-behaviour based SSMs, and, if so, which ML classifiers are the most appropriate?

Question 4

Are harsh braking events more pertinent than harsh accelerations in predicting the crash risk level of road segments?

Question 5

a) In the absence of highly detailed historical road crash data, how can harsh braking events be analyzed across various road environments?
b) Is there spatial autocorrelation present in harsh braking frequencies for road segments, and, if so, do spatial modelling approaches outperform their non-spatial counterparts?

Question 6

Which road infrastructure and driver behaviour parameters exhibit a statistically significant impact on the number of harsh braking events per road segment?

These research questions served as the driving force behind the entire research endeavor, exploring the integration and analysis of infrastructure, traffic, and driver behaviour data for meaningful conclusions in road crash risk assessment. In order to answer these research questions, an elaborate methodological framework was devised, which is replicated on Figure 9.1.

The core of the methodological framework involved a multi-step process, commencing with the investigation of road safety modelling data in Greece, laying the groundwork for subsequent directions. This investigation highlighted the constraints associated with conducting high-detailed crash prediction modelling in Greece. Such modelling is only feasible for motorways with high-quality crash data, specifically regarding crash location and traffic attributes per road segment. In response to this limitation, two distinct databases were developed.

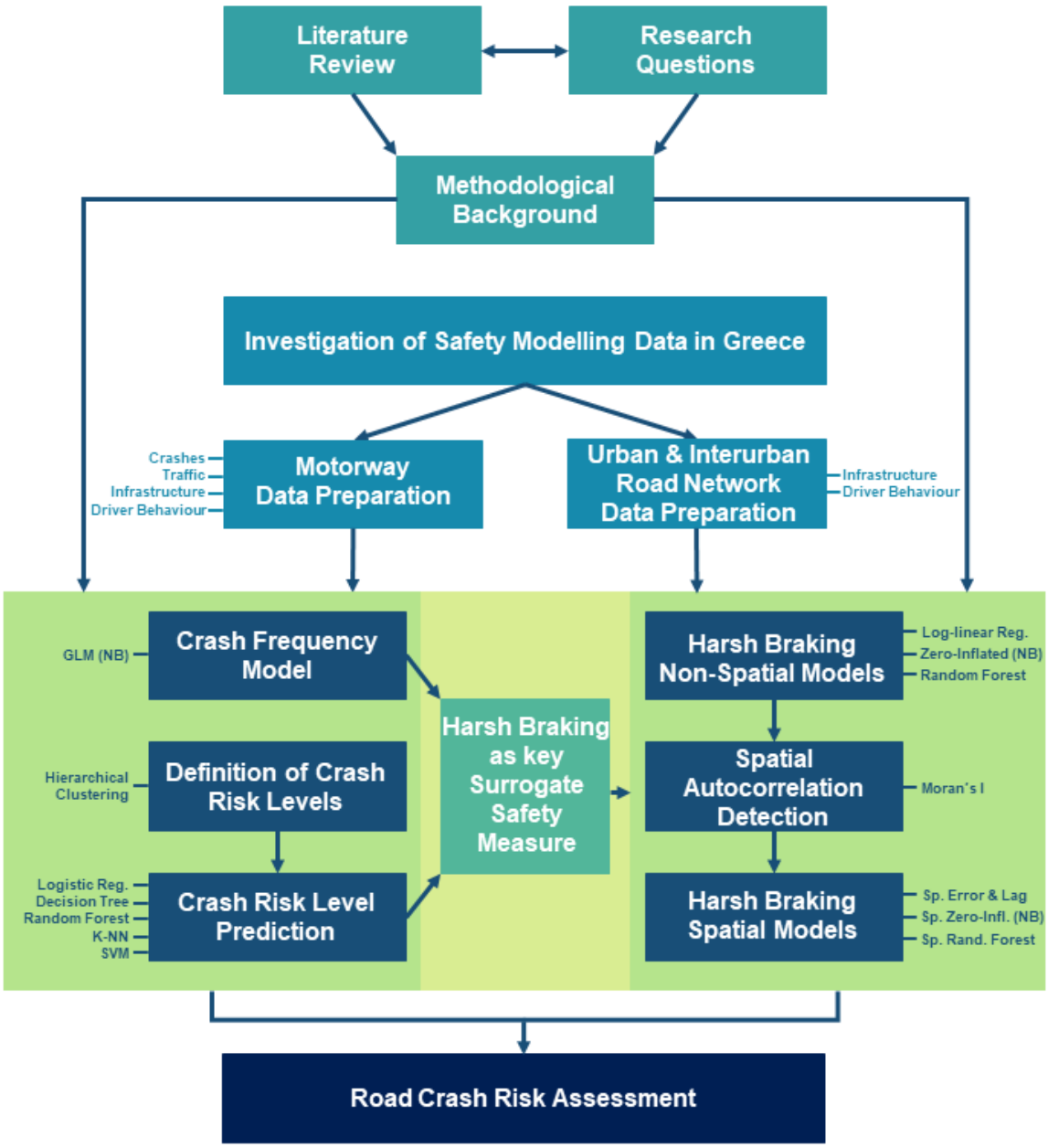


Figure 9.1: Graphical representation of the overall methodological framework of the doctoral dissertation

The first one focused on 668 motorway segments within the Olympia Odos motorway, containing comprehensive data on historical road crashes, traffic, road geometry characteristics, and naturalistic driver behaviour metrics. Specifically, crash data of all severity levels including PDO crashes for the years 2018-2020 were exploited. In parallel with the road crash data, AADT data for the same time period were included in the developed database. Regarding the road infrastructure characteristics, a variety of sources, such as information from the road operator and the use of different

software, including Open GIS, Google Earth and GoogleStreetView, were combined. The inclusion of these road infrastructure data and of reference drawings of the motorway also enabled the identification and isolation of naturalistic driver behaviour data from a smartphone application. Driver behaviour data were collected for the period from June 1, 2019, to December 31, 2020, from a sample of 327 drivers in 2019 and 330 drivers in 2020. The average number of trips per motorway segment over the entire study period was 769 trips.

The second one covered a broader road network within the Region of Eastern Macedonia and Thrace, including urban and interurban roads. Within this road network, an initial analysis was conducted on all road segments sourced from OSM to extract their geometric and network characteristics. Subsequently, naturalistic driving behaviour data that were extracted from a smartphone application were aligned with the corresponding OSM segments. The examined road network included 6,103 road segments, with an average length of 288.8 meters, resulting in a total road network length of 1,763 kilometers. Regarding the naturalistic driver behaviour metrics, data from 5,129 trips during 2021 were utilized. The mean trip duration was 634 seconds, with a standard deviation of 556 seconds. However, the developed database for this road network lacked detailed crash and traffic data for the examined road segments.

Various methodologies were applied for the road segments of Olympia Odos motorway. These included techniques such as NB regression for developing a crash frequency model, HC to determine crash risk levels based on historical crash data and traffic attributes, and the utilization of Machine Learning classifiers such as LR, DT, RF, K-NN and SVM. These classifiers were used for crash risk level prediction, leveraging infrastructure and driver behaviour data. A critical focus was placed on evaluating the reliability of harsh driving behaviour events as SSMs.

Subsequently, the framework extended to include the road network data of Eastern Macedonia and Thrace Region, employing harsh braking events for road crash risk assessment. This involved applying both non-spatial and spatial models to identify significant road infrastructure and driver behaviour parameters influencing harsh braking events per road segment.

Ultimately, the synthesis of all the analyses carried out within the framework of this doctoral dissertation resulted in a comprehensive road crash risk assessment with numerous original and interesting results, which are discussed in the following concluding subsections.

9.2 Main Findings for Motorway

For the motorway analyses, a unified database including data on historical injury and PDO crashes, traffic attributes, road geometry characteristics, and driver behaviour SSMs of 668 road segments of the Olympia Odos motorway was exploited. The results of the crash frequency model (NB regression) revealed that road crash frequency in the examined motorway segments is positively correlated with the traffic volume, the length of the segment, and the numbers of harsh accelerations and harsh brakings per segment trips. This finding contributes to existing road safety literature by establishing a positive and statistically significant relationship between crash frequency and events of harsh driving behaviour. Consequently, it is inferred that these events can serve as a valid subcategory of naturalistic SSMs. Specifically, they can be used either to complement CPMs or as dependent variables in proactive road safety analyses, particularly in cases where detailed historical crash data are lacking.

As a further phase of the motorway investigations, an endeavor was made to formulate crash risk level clusters of the motorway segments. This was achieved by considering the number of road crashes by segment length and the traffic volume of each segment using the agglomerative hierarchical clustering technique. Considering the influence of segment length and traffic volume, as indicated by the results of the negative binomial regression model, both variables were included into the clustering analysis due to their statistically significant impact on motorway segment crash frequency. The outcomes of this clustering process delineated four distinct crash risk levels with a clear pattern whereby the first risk level class presents high average numbers of traffic volume and road crashes by segment length, while these figures decrease progressively for each subsequent class.

Subsequently, these identified four levels were utilized as the response variable in five ML classification models (LR, DT, RF, SVM, and K-NN). The models included predictors encompassing road geometry characteristics and unsafe driving behaviours, such as rates of harsh brakings, harsh accelerations, and speeding duration per trips within the analyzed segments. Among the five classification models, RF demonstrated superior classification performance across all crash risk levels, consistently achieving scores exceeding 89% (overall accuracy: 89.9%, macro-averaged precision: 90.7%, macro-averaged recall: 89.9%, macro-averaged F1 score: 90.2%). This outcome reveals the potential of the developed RF model as a highly promising proactive road safety tool, capable of effectively identifying and prioritizing potentially hazardous motorway segments.

Finally, to enhance the interpretability of the RF model, which inherently operates as a black-box ML model, SHAP values were calculated for a typical motorway segment. Based on the SHAP values of the naturalistic driving behaviour predictors, it was revealed that harsh braking events serve as a more suitable SSM than harsh accelerations in terms of crash risk level prediction.

9.3 Main Findings for Urban and Interurban Road Network

Within the broader road network of the Eastern Macedonia and Thrace Region, a spatial dataset consisting aggregated naturalistic driving behaviour metrics, as well as geometric and network characteristics on a segment level was analyzed. For the examined 6,103 road segments, and based on Moran's I index, statistically significant and positive spatial autocorrelation in harsh braking event frequencies was detected. Initially, non-spatial modelling techniques, such as log-linear, ZINB and conventional RF regression models were employed on harsh braking events frequencies. However, the existence of spatial autocorrelation highlighted the need for the development of spatial models, such as SEM, SLM, SZINB and SRF, in order to take into account such spatial dependencies.

Consistent signs of the beta coefficients emerged across all models. Specifically, road segment length and the number of trips per segment were identified as proxy indicators of risk exposure, positively correlated with harsh braking events. Additionally, the efficiency index (statistically significant only in the log-linear model, SEM and SLM), related to the linearity of road segments, revealed a positive correlation with harsh braking events, suggesting that drivers exhibit more frequent harsh braking on road segments with fewer curves. Variables related to speeding and mobile phone use were also positively associated with harsh braking events, whereas motorways exhibited fewer harsh braking events compared to other road types.

In both RF models, the number of trips per examined road segment was found to be the most influential predictor, highlighting its significant relevance in predicting the frequency of harsh braking events, as it serves as a naturalistic driving exposure metric. On the other hand, the motorway variable exhibited the lowest importance, indicating that road type is relatively less valuable in predicting the number of harsh braking events. This finding may suggest that factors other than road type such as driver distraction and speeding, might play a more crucial role in influencing harsh braking events frequencies.

Regarding the performance of the developed models, SLM surpassed both the log-linear model and the SEM, with lower AIC values and absence of spatial autocorrelation in its residuals. Lower AIC values, indicating a better fit, were also observed for the SZINB model compared to the non-spatial ZINB model. Moreover, the SRF reduced the absolute values of spatial autocorrelation in the residuals compared to the respective values of the conventional RF. In addition, the SRF outperformed the non-spatial RF model in terms of model fit to observed data, but the non-spatial model performed better in terms of generalization to unseen data.

9.4 Innovative Contributions

This doctoral dissertation offers significant noteworthy contributions in the field of road safety, as illustrated in Figure 9.2. These contributions are discussed in detail in the following subsections.

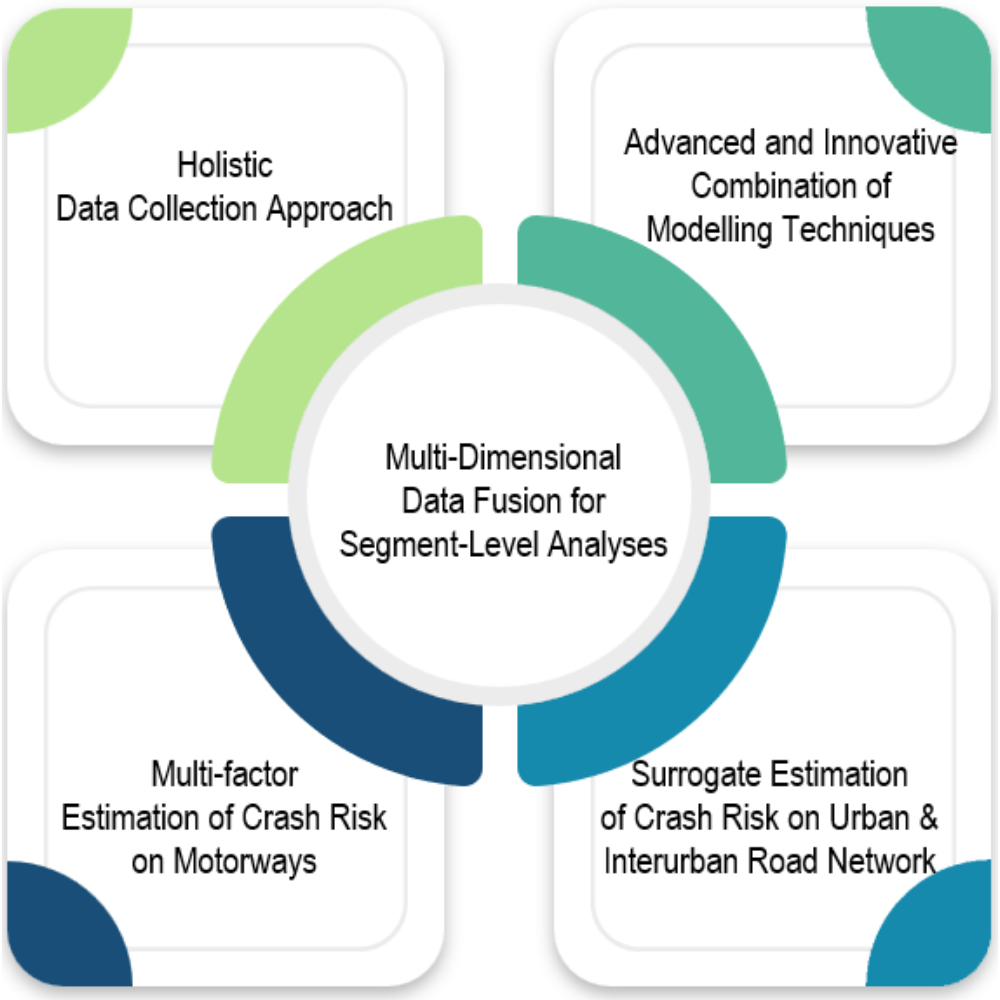


Figure 9.2: Innovative contributions of the dissertation

9.4.1 Holistic Data Collection Approach

In the context of this doctoral dissertation, a holistic comprehensive data collection was conducted to investigate the impact of driver behaviour, road infrastructure characteristics and traffic attributes on road crash risk assessment. Technological advancements have significantly facilitated the collection of data from various sources, opening up new research opportunities that were previously unexplored.

Specifically, this dissertation exploited high-resolution naturalistic driving big datasets collected from smartphone sensors to assess road crash risk on motorways and a broader road network, encompassing urban and interurban roads. For road infrastructure data on the examined motorway, a variety of sources were exploited,

including data provided by the road operator and software such as Open GIS, Google Earth and GoogleStreetView. Geometric and network characteristics for the broader road network of the Eastern Macedonia and Thrace Region were derived using algorithms in the R programming language. Appropriate libraries were utilized to extract data from OSM and process them as simple spatial features. Concerning road crash and traffic data on the examined motorway, high-quality data from the road operator were employed. This included road crash data of all injury severities, including PDO crashes, with high accuracy in crash location, covering the period from 2018 to 2020. Additionally, AADT data derived from the motorway toll stations for the corresponding period were utilized.

9.4.2 Multi-Dimensional Data Fusion for Segment-Level Analyses

The collection of data from various sources and at different levels necessitates appropriate processing for data integration. The first database comprised 668 motorway segments ranging from 200 to 600 meters in length and was infrastructure-based. It included data on historical road crashes, traffic volumes and geometric characteristics. Subsequently, driver behaviour metrics derived from smartphone sensors had to be assigned to the examined road segments. This involved allocating driving behaviour metrics from naturalistic data, which are driver-based, to the examined motorway segments, which are infrastructure-based data. This allocation was achieved via isolating each trip portion to the corresponding segment within the internal recording of trips conducted in GIS by the smartphone data providers using ESRI polygons at 200m intervals.

For the broader urban and interurban network of the Eastern Macedonia and Thrace Region, which exclusively comprised infrastructure and driver behavior data, a series of processing algorithms were applied. Initially, a database was created for the considered road network, encompassing 6,103 road segments. This database contained key geometric characteristics such as length, curvature, road type, etc., for each segment. The data extraction from OSM and database creation involved exploiting R libraries specifically designed for these tasks. Next, the naturalistic driver behavior data, extracted from smartphone sensors and covering indicators like harsh braking events, speeding, distraction due to mobile phone use, etc., for every second of trips made in 2021 in the study area, had to be assigned to the corresponding road segments. This assignment was achieved through a spatial map-matching procedure. Initially, the centroid of each road segment line-string was identified using the “st_centroid” function from the “sf” R library. It is noted that centroids are point-type quantities and represent the geometric center of each road segment. Subsequently, the aggregated driving behaviour metrics were assigned to the nearest road segment centroid based on the latitude and longitude coordinates for each trip-second. This process was executed using the "st_join" function and the "st_nearest_feature" geometry predicate function from the "sf" R library.

Overall, the algorithms utilized in this doctoral dissertation, especially for the broader urban and interurban road network, facilitate the seamless transferability of the methodological and data processing framework employed in this dissertation. With minimal modifications, spatial data frames can be generated for various regions, allowing for analyses using the same or different variables, study periods, and statistical methodologies.

9.4.3 Advanced and Innovative Combination of Modelling Techniques

The wealth of high-resolution multiparametric data and the robustness of data processing and fusion enabled the development of advanced and innovate modelling techniques.

Initially, a crash frequency model (NB regression) was developed. This model facilitated the investigation of the influence of various geometric characteristics, traffic attributes, and driver behaviour metrics on road crashes. Subsequently, agglomerative hierarchical clustering was employed to categorize crash risk levels for the analyzed road segments, which were then incorporated as the response variable in several ML classifiers. In addition to utilizing ML techniques, the analyses included the computation of SHAP values, a recent and potent addition in the field of explainable and interpretable ML. These values provided insights into the influential factors contributing to crash risk. This comprehensive approach enhances the sophistication of the modelling techniques and reinforces the interpretability of their results.

With regard to the broader road network of the Eastern Macedonia and Thrace Region, the analyses incorporated harsh braking events as the dependent variables for the developed models. Notably, the modelling techniques employed in this doctoral dissertation are, to the best of the author's knowledge, being applied for the first time to harsh braking events. Among these innovative modelling approaches are the SEM, SLM, SZINB, and SRF. It is worth emphasizing that the application of the SRF is particularly noteworthy, representing a novel modelling technique applicable not only to harsh braking events but also to various aspects of road safety analyses.

9.4.4 Multi-factor Estimation of Crash Risk on Motorways

Utilizing the high-quality and detailed database developed for the road segments of the motorway, aiming to address the research questions posed in this doctoral dissertation, valuable and innovative conclusions were drawn. Specifically, statistical correlations from the road crash frequency model revealed a positive and statistically significant relationship between historical road crash data and the number of harsh driving behaviours. This applies to both the number of harsh accelerations and the number of harsh brakings per passed trips within the examined motorway segments.

This indicates that these indicators of harsh driving behaviour can be utilized as SSMs, either complementing traditional crash frequency models or serving as dependent variables in road crash risk assessment models in areas where either road crash data are unavailable or the available crash data are of low quality.

Additionally, this thesis highlighted an innovative insight, emphasizing that the contribution of harsh brakings, compared to harsh accelerations, is higher in predicting the crash risk level for road segments. This makes harsh brakings a more suitable SSM indicator for proactive road safety analyses, enhancing the understanding of road crash risk and providing practical implications for targeted interventions.

9.4.5 Surrogate Estimation of Crash Risk on Urban and Interurban Road Network

The assessment of this dissertation's contributions would be inadequate without recognizing the broader implications of the developed models on the road network of the Eastern Macedonia and Thrace Region. In these models, the dependent variables were represented by the number of harsh braking events, serving as SSMs. The detection of statistically significant and positively correlated spatial autocorrelation in harsh braking event frequencies compelled the development of spatial modelling approaches. Pivotal to frequency analyses is the measurement of exposure, with this dissertation employing two primary exposure variables for the respective models: road segment length and the number of trips per segment. This research identifies the statistically significant influence of these exposure variables on the number of harsh braking events, quantifying their respective impacts. Additionally, it incorporates various indicators related to road environment and driver behaviour, contributing to a comprehensive assessment of road crash risk.

The creation of comprehensive road safety maps and heatmaps illustrating harsh braking events stands as a valuable tool for road management authorities, stakeholders and road users. These visualizations present complex data and model predictions in an easily comprehensible manner, facilitating communication and integration into diverse decision-making processes. Through these maps, the multifaceted efforts of this dissertation in road crash risk assessment are effectively communicated to both the scientific community and the public domain. Overall, SSMs, such as harsh braking events, offer significant potential for monitoring road safety, evaluating and enhancing countermeasures, and expanding road safety data coverage rapidly. In academia, SSM modelling exercises have emerged in recent years. Apart from contributing in that field, this doctoral dissertation demonstrated that with the necessary effort, SSM-based spatial models can be used in scarcely-studied areas.

9.5 Further Challenges

This doctoral dissertation addressed various composite issues related to the data collection, processing and integration, and advanced modelling for the examined road segments. Consequently, it is inevitable that limitations emerged during the entire research process, and open challenges remain, which need to be acknowledged.

With regard to the multisource-based extraction of road geometry data for Olympia Odos motorway, the results are obviously not an exact replication of the actual road design of the motorway and minor differences could be expected if a comparison with the as-built drawings of the project was made. The same is probably true for the geometric characteristics of the road segments of the Eastern Macedonia and Thrace Region extracted via OSM. Nevertheless, any differences would be minor and, although important from a designer's point of view they are not expected to be able to differentiate the results of this dissertation.

Another limitation related to the motorway segments is that the analyses did not include tunnels and toll station segments, resulting in discontinuities in the research area. Moreover, the motorway segments analyses did not take into account unobserved heterogeneity and the effects of spatial characteristics of various road safety indicators. However, this limitation provided directions for the research efforts in the broader road network of Eastern Macedonia and Thrace Region, where spatial modelling approaches were followed.

Despite the valuable insights gained from these spatial analyses, a significant limitation that needs to be acknowledged is the lack of available traffic data (AADT or flow conditions) per examined segment, which could have provided additional insights into the influence of traffic on harsh braking events. The absence of AADT was attempted to be tackled by using the number of trips and the segment length as substitute risk exposure metrics. Moreover, harsh driving events essentially represent behavioural variables. Consequently, despite the sample size of drivers and trips analyzed in this dissertation being substantial and meeting the standards of the literature, there still remains a possibility that the observed driving behaviour diverged from the norm, leading to a frequency of harsh braking events either exceeding or falling below the anticipated levels.

Upon concluding this dissertation, the author believes that the current research findings lead to several research issues that demand further scientific investigation. Indicatively, a promising avenue for research involves exploring temporal patterns, which would capture seasonal cyclical trends in both road crash and harsh braking hotspots.

It is also evident that this dissertation did not comprehensively cover all aspects of the road environment. While the existing analysis delves into certain factors, the inclusion

of additional independent variables, such as slopes, pavement conditions, the presence of roadworks, land use, weather conditions and more, can significantly enrich the depth of understanding and offer unexplored insights.

While this doctoral dissertation has employed a comprehensive set of statistical and ML models, the ever-expanding nature of data science and transportation research opens avenues for further exploration. Future investigations may benefit from the exploration of additional models that could contribute further insights. For instance, advanced deep learning architectures such as neural networks or recurrent neural networks could be explored for crash frequency modelling. Ensemble methods like gradient boosting machines and XGBoost might offer enhanced predictive performance for crash risk level classification tasks. Additionally, the integration of spatiotemporal models, considering both spatial and temporal dimensions simultaneously, could provide a better understanding of the factors influencing harsh braking events.

Finally, the scope of harsh braking analyses can be expanded by extending its application to include additional geographical regions, potentially encompassing other countries. This extension has the potential to transform the research into a digital twin, offering a comprehensive road crash risk assessment. This transformation is further facilitated by technological developments, such as telematics, which enable scalable and expedient data collection. Consequently, high-quality data applications and monitoring in scarcely-studied areas become possible and can even be converted to the norm in the short term.

References

Aguero-Valverde, J., & Jovanis, P. P. (2006). Spatial analysis of fatal and injury crashes in Pennsylvania. Accident Analysis & Prevention, 38(3), 618-625.

Akaike, H. (1970). Statistical predictor identification. Annals of the institute of Statistical Mathematics, 22(1), 203-217.

Albalate, D., & Bel, G. (2012). Motorways, tolls and road safety: evidence from Europe. SERIEs, 3, 457-473.

Alhajyaseen, W. K. (2015). The integration of conflict probability and severity for the safety assessment of intersections. Arabian Journal for Science and Engineering, 40(2), 421-430.

Ali, Y., Haque, M. M., & Mannering, F. (2023). A Bayesian generalised extreme value model to estimate real-time pedestrian crash risks at signalised intersections using artificial intelligence-based video analytics. Analytic methods in accident research, 38, 100264.

Ambros, J., Altmann, J., Jurewicz, C., & Chevalier, A. (2019). Proactive assessment of road curve safety using floating car data: An exploratory study. Archives of Transport, 50.

Ambros, J., Jurewicz, C., Turner, S., & Kieć, M. (2018). An international review of challenges and opportunities in development and use of crash prediction models. European transport research review, 10, 1-10.

Anselin, L. (1995). Local indicators of spatial association—LISA. Geographical analysis, 27(2), 93-115.

Arun, A., Haque, M. M., Bhaskar, A., & Washington, S. (2022). Transferability of multivariate extreme value models for safety assessment by applying artificial intelligence-based video analytics. Accident Analysis & Prevention, 170, 106644.

Arun, A., Haque, M. M., Bhaskar, A., Washington, S., & Sayed, T. (2021a). A systematic mapping review of surrogate safety assessment using traffic conflict techniques. Accident Analysis & Prevention, 153, 106016.

Arun, A., Haque, M. M., Washington, S., Sayed, T., & Mannering, F. (2021b). A systematic review of traffic conflict-based safety measures with a focus on application context. Analytic methods in accident research, 32, 100185.

Ball, K.K., & Ackerman, M.L. (2011). The Older Driver (Training and Assessment: Knowledge, Skills, and Attitudes). In Handbook of Driving Simulation for Engineering, Medicine and Psychology, 2011, CRC Press.

Benito, M. (2021). R package spatialRF: Easy Spatial Regression with Random Forest. doi: 10.5281/zenodo.4745208.

Biau, G., & Scornet, E. (2016). A random forest guided tour. Test, 25, 197-227.

Biecek, P. (2018). DALEX: Explainers for complex predictive models in R. Journal of Machine Learning Research, 19(84), 1-5.

Bivand, R., & Wong, D. W. (2018). Comparing implementations of global and local indicators of spatial association. Test, 27(3), 716-748.

Bivand, R., Millo, G., & Piras, G. (2021). A review of software for spatial econometrics in R. Mathematics, 9(11), 1276.

Blanco, M., Hanowski, R. J., Olson, R. L., Morgan, J. F., Soccolich, S. A., & Wu, S. C. (2011). The Impact of Driving, Non-driving Work, and Rest Breaks on Driving Performance in Commercial Vehicle Operations.

Blangiardo, M., & Cameletti, M. (2015). Spatial and spatio-temporal Bayesian models with R-INLA. John Wiley & Sons.

Bonela, S. R., & Kadali, B. R. (2022). Review of traffic safety evaluation at T-intersections using surrogate safety measures in developing countries context. IATSS research.

Boonsiripant, S., Rodgers, M. O., & Hunter, M. P. (2011). Speed profile variation as a road network screening tool. Transportation research record, 2236(1), 83-91.

Breiman, L. (2001). Random forests. Machine learning, 45, 5-32.

Calvi, A., & D'amico, F. (2013). A study of the effects of road tunnel on driver behavior and road safety using driving simulator. Advances in Transportation Studies, (30).

Cameron, A. C., & Trivedi, P. K. (2013). Regression analysis of count data (Vol. 53). Cambridge university press.

Chen, P., Zeng, W., Yu, G., & Wang, Y. (2017). Surrogate safety analysis of pedestrian-vehicle conflict at intersections using unmanned aerial vehicle videos. Journal of advanced transportation, 2017.

Cheng, J., Karambelkar, B., Xie, Y., et al. (2019). Package 'leaflet'. R package version 2.1.1.

Coles, S. (2001). An introduction to statistical modelling of extreme values. London: Springer.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine learning, 20, 273-297.

Desai, J., Li, H., Mathew, J. K., Cheng, Y. T., Habib, A., & Bullock, D. M. (2021). Correlating hard-braking activity with crash occurrences on interstate construction projects in Indiana. Journal of Big Data Analytics in Transportation, 3(1), 27-41.

Dimitrijevic, B., Khales, S. D., Asadi, R., & Lee, J. (2022). Short-term segment-level crash risk prediction using advanced data modeling with proactive and reactive crash data. Applied Sciences, 12(2), 856.

Dobson, A. J., & Barnett, A. G. (2018). An introduction to generalized linear models. Chapman and Hall/CRC.

El-Basyouny, K., & Sayed, T. (2013). Safety performance functions using traffic conflicts. Safety science, 51(1), 160-164.

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

Elvik, R., Ulstein, H., Wifstad, K., Syrstad, R. S., Seeberg, A. R., Gulbrandsen, M. U., & Welde, M. (2017). An Empirical Bayes before-after evaluation of road safety effects of a new motorway in Norway. Accident Analysis & Prevention, 108, 285-296.

European Commission. (2018). Motorways, European Commission, Directorate General for Transport, February 2018.

European Commission. (2023). Road safety: 20,640 people died in a road crash last year – progress remains too slow. Available online: https://transport.ec.europa.eu/news-events/news/road-safety-20640-people-died-road-crash-last-year-progress-remains-too-slow-2023-10-19_en. (Accessed on 07 November 2023).

European Transport Safety Council. (2021). 15th Annual Road Safety Performance Index (PIN) Report; ETSC: Brussels, Belgium, 2021.

Fafoutellis, P., Mantouka, E. G., Vlahogianni, E. I., & Fortsakis, P. (2023). Investigating the impacts of the COVID-19 pandemic on Eco-driving behavior. Safety Science, 106251.

Fazeen, M., Gozick, B., Dantu, R., Bhukhiya, M., & González, M. C. (2012). Safe driving using mobile phones. IEEE Transactions on Intelligent Transportation Systems, 13(3), 1462-1468.

Freedman, D. A. (2009). Statistical models: theory and practice. Cambridge university press.

Fu, C., & Sayed, T. (2021a). Comparison of threshold determination methods for the deceleration rate to avoid a crash (DRAC)-based crash estimation. Accident Analysis & Prevention, 153, 106051.

Fu, C., & Sayed, T. (2021b). Random parameters Bayesian hierarchical modelling of traffic conflict extremes for crash estimation. Accident Analysis & Prevention, 157, 106159.

Garay, A. M., Hashimoto, E. M., Ortega, E. M., & Lachos, V. H. (2011). On estimation and influence diagnostics for zero-inflated negative binomial regression models. Computational Statistics & Data Analysis, 55(3), 1304-1318.

Geary, R. C. (1954). The contiguity ratio and statistical mapping. The incorporated statistician, 5(3), 115-146.

Gettman, D., & Head, L. (2003). Surrogate safety measures from traffic simulation models. Transportation Research Record, 1840(1), 104-115.

Goyani, J., Paul, A. B., Gore, N., Arkatkar, S., & Joshi, G. (2021). Investigation of crossing conflicts by vehicle type at unsignalized t-intersections under varying roadway and traffic conditions in India. Journal of transportation engineering, Part A: Systems, 147(2), 05020011.

Grandini, M., Bagli, E., & Visani, G. (2020). Metrics for multi-class classification: an overview. arXiv preprint arXiv:2008.05756.

Greibe, P. (2007). Braking distance, friction and behaviour. Trafitec, Scion-DTU.

Guido, G., Vitale, A., Astarita, V., Saccomanno, F., Giofré, V. P., & Gallelli, V. (2012). Estimation of safety performance measures from smartphone sensors. Procedia-Social and Behavioral Sciences, 54, 1095-1103.

Gündüz, G., Yaman, Ç., Peker, A. U., & Acarman, T. (2017). Prediction of risk generated by different driving patterns and their conflict redistribution. IEEE Transactions on Intelligent Vehicles, 3(1), 71-80.

Guo, M., Zhao, X., Yao, Y., Yan, P., Su, Y., Bi, C., & Wu, D. (2021). A study of freeway crash risk prediction and interpretation based on risky driving behavior and traffic flow data. Accident Analysis & Prevention, 160, 106328.

Haddon Jr, W. (1980). Advances in the epidemiology of injuries as a basis for public policy. Public health reports, 95(5), 411.

Haklay, M. (2010). How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. Environment and planning B: Planning and design, 37(4), 682-703.

Han, J., Pei, J., & Tong, H. (2022). Data mining: concepts and techniques. Morgan kaufmann.

He, Z., Qin, X., Liu, P., & Sayed, M. A. (2018). Assessing surrogate safety measures using a safety pilot model deployment dataset. Transportation research record, 2672(38), 1-11.

Hellenic Statistical Authority. (2023). Road Traffic Accidents: December 2022. Available online: https://www.statistics.gr/en/statistics?p_p_id=documents_WAR_publicationsportlet_INSTANCE_qDQ8fBKKo4lN&p_p_lifecycle=2&p_p_state=normal&p_p_mode=view&p_p_cacheability=cacheLevelPage&p_p_col_id=column-2&p_p_col_count=4&p_p_col_pos=1&_documents_WAR_publicationsportlet_INSTANCE_qDQ8fBKKo4lN_javax.faces.resource=document&_documents_WAR_publicationsportlet_INSTANCE_qDQ8fBKKo4lN_ln=downloadResources&_documents_WAR_publicationsportlet_INSTANCE_qDQ8fBKKo4lN_documentID=491416&_documents_WAR_publicationsportlet_INSTANCE_qDQ8fBKKo4lN_locale=en. (Accessed on 07 November 2023).

Hengl, T., Nussbaum, M., Wright, M. N., Heuvelink, G. B., & Gräler, B. (2018). Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. PeerJ, 6, e5518.

Ho, A. D., & Yu, C. C. (2015). Descriptive statistics for modern test score distributions: Skewness, kurtosis, discreteness, and ceiling effects. Educational and psychological measurement, 75(3), 365-388.

Ho, T. K. (1995, August). Random decision forests. In Proceedings of 3rd international conference on document analysis and recognition (Vol. 1, pp. 278-282). IEEE.

Høye, A. K., & Hesjevoll, I. S. (2020). Traffic volume and crashes and how crash and road characteristics affect their relationship–A meta-analysis. Accident Analysis & Prevention, 145, 105668.

Hu, J., Huang, M. C., & Yu, X. (2020). Efficient mapping of crash risk at intersections with connected vehicle data and deep learning models. Accident Analysis & Prevention, 144, 105665.

Hunter, M., Saldivar-Carranza, E., Desai, J., Mathew, J. K., Li, H., & Bullock, D. M. (2021). A proactive approach to evaluating intersection safety using hard-braking data. Journal of Big Data Analytics in Transportation, 3(2), 81-94.

Hussain, F., Li, Y., Arun, A., & Haque, M. M. (2022). A hybrid modelling framework of machine learning and extreme value theory for crash risk estimation using traffic conflicts. Analytic methods in accident research, 36, 100248.

Hussein, H., Radwan, M. H., Elsayed, H. A., & Abd El-Kader, S. M. (2021). Depth-first-search-tree based D2D power allocation algorithms for V2I/V2V shared 5G network resources. Wireless Networks, 27, 3179-3193.

Hydén, C. (1987). The development of a method for traffic safety evaluation: The Swedish Traffic Conflicts Technique. Bulletin Lund Institute of Technology, Department, (70).

Ijaz, M., Zahid, M., & Jamal, A. (2021). A comparative study of machine learning classifiers for injury severity prediction of crashes involving three-wheeled motorized rickshaw. Accident Analysis & Prevention, 154, 106094.

Imprialou, M., & Quddus, M. (2019). Crash data quality for road safety research: Current state and future directions. Accident Analysis & Prevention, 130, 84-90.

Jackman, S. (2020). pscl: Classes and Methods for R Developed in the Political Science Computational Laboratory. United States Studies Centre, University of Sydney, Sydney, New South Wales, Australia. R package version 1.5.5.1.

Janstrup, K. H., Kaplan, S., Hels, T., Lauritsen, J., & Prato, C. G. (2016). Understanding traffic crash under-reporting: linking police and medical records to individual and crash characteristics. Traffic Injury Prevention, 17(6), 580-584.

Johnsson, C., Laureshyn, A., & Dágostino, C. (2021). Validation of surrogate measures of safety with a focus on bicyclist–motor vehicle interactions. Accident Analysis & Prevention, 153, 106037.

Johnsson, C., Laureshyn, A., & De Ceunynck, T. (2018). In search of surrogate safety indicators for vulnerable road users: a review of surrogate safety indicators. Transport Reviews, 38(6), 765-785.

Kamla, J., Parry, T., & Dawson, A. (2019). Analysing truck harsh braking incidents to study roundabout accident risk. Accident Analysis & Prevention, 122, 365-377.

Karatzoglou, A., Meyer, D., Hornik, K. (2006). Support Vector Machines in R. Journal of Statistical Software 15(9), 1-28.

Karatzoglou, A., Smola, A., Hornik, K., Zelies, A. (2005). Kernlab-Kernel Methods. R package, Version 0.6-2. Retrieved from: http://CRAN.R-project.org/.

Karpinski, E., Bayles, E., & Sanders, T. (2022). Safety analysis for micromobility: Recommendations on risk metrics and data collection. Transportation research record, 2676(12), 420-435.

Khorram, B., af Wåhlberg, A. E., & Tavakoli Kashani, A. (2020). Longitudinal jerk and celeration as measures of safety in bus rapid transit drivers in Tehran. Theoretical Issues in Ergonomics Science, 21(5), 577-594.

Kim, S., Song, T. J., Rouphail, N. M., Aghdashi, S., Amaro, A., & Gonçalves, G. (2016). Exploring the association of rear-end crash propensity and micro-scale driver behavior. Safety science, 89, 45-54.

Kontaxi, A., Ziakopoulos, A., & Yannis, G. (2021). Trip characteristics impact on the frequency of harsh events recorded via smartphone sensors. IATSS research, 45(4), 574-583.

Kontaxi, A., Ziakopoulos, A., & Yannis, G. (2021a). Trip characteristics impact on the frequency of harsh events recorded via smartphone sensors. IATSS research, 45(4), 574-583.

Kontaxi, A., Ziakopoulos, A., Yannis, G., (2021b). Investigation of the speeding behavior of motorcyclists through an innovative smartphone application. Traffic Injury Prevention, 22 (6), 460–466.

La Torre, F., Meocci, M., Domenichini, L., Branzi, V., & Paliotto, A. (2019). Development of an accident prediction model for Italian freeways. Accident Analysis & Prevention, 124, 1-11.

Laureshyn, A., De Ceunynck, T., Karlsson, C., Svensson, Å., & Daniels, S. (2017). In search of the severity dimension of traffic events: Extended Delta-V as a traffic conflict indicator. Accident Analysis & Prevention, 98, 46-56.

Lenné, M. G., Triggs, T. J., & Redman, J. R. (1997). Time of day variations in driving performance. Accident Analysis & Prevention, 29(4), 431-437.

Li, H., Calder, C. A., & Cressie, N. (2007). Beyond Moran's I: testing for spatial dependence based on the spatial autoregressive model. Geographical analysis, 39(4), 357-375.

Li, P., Abdel-Aty, M., & Yuan, J. (2021a). Using bus critical driving events as surrogate safety measures for pedestrian and bicycle crashes based on GPS trajectory data. Accident Analysis & Prevention, 150, 105924.

Li, X., Mousavi, S. M., Dadashova, B., Lord, D., & Wolshon, B. (2021b). Toward a crowdsourcing solution to identify high-risk highway segments through mining driving jerks. Accident Analysis & Prevention, 155, 106101.

Lindgren, F., & Rue, H. (2015). Bayesian spatial modelling with R-INLA. Journal of statistical software, 63, 1-25.

Lindgren, F., Rue, H., & Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 73(4), 423-498.

Lord, D., & Mannering, F. (2010). The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives. Transportation research part A: policy and practice, 44(5), 291-305.

Lovelace, R., Nowosad, J., & Muenchow, J. (2019). Geocomputation with R. CRC Press.

Lu, G., Cheng, B., Kuzumaki, S., & Mei, B. (2011). Relationship between road traffic accidents and conflicts recorded by drive recorders. Traffic Injury Prevention, 12(4), 320-326.

Lu, N., Cheng, N., Zhang, N., Shen, X., & Mark, J. W. (2014). Connected vehicles: Solutions and challenges. IEEE internet of things journal, 1(4), 289-299.

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in neural information processing systems, 30.

Mahmud, S. S., Ferreira, L., Hoque, M. S., & Tavassoli, A. (2019). Micro-simulation modelling for traffic safety: A review and potential application to heterogeneous traffic environment. IATSS research, 43(1), 27-36.

Mantouka, E. G., Barmpounakis, E. N., & Vlahogianni, E. I. (2018). Mobile sensing and machine learning for identifying driving safety profiles (No. 18-01416).

Meyer, D. (2001). Support vector machines. R News 1(3), 23-26.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & Prisma Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Annals of internal medicine, 151(4), 264-269.

Montella, A., Colantuoni, L., & Lamberti, R. (2008). Crash prediction models for rural motorways. Transportation Research Record, 2083(1), 180-189.

Moran, P. A. (1950). Notes on continuous stochastic phenomena. Biometrika, 37(1/2), 17-23.

Mousavi, S. M., Zhang, Z., Parr, S. A., Pande, A., & Wolshon, B. (2019, August). Identifying high crash risk highway segments using jerk-cluster analysis. In International Conference on Transportation and Development 2019: Smarter and Safer Mobility and Cities (pp. 112-123). Reston, VA: American Society of Civil Engineers.

Mukherjee, D., & Mitra, S. (2020). Comprehensive study of risk factors for fatal pedestrian crashes in urban setup in a developing country. Transportation research record, 2674(8), 100-118.

Murtagh, F., & Contreras, P. (2012). Algorithms for hierarchical clustering: an overview. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2(1), 86-97.

National Research Council. (2010). Highway Safety Manual. American Association of State Highway and Transportation Officials, Washington, DC.

Nikolaou, D., Dragomanovits, A., Ziakopoulos, A., Deliali, A., Handanos, I., Karadimas, C., ... & Yannis, G. (2023a). Exploiting Surrogate Safety Measures and Road Design Characteristics towards Crash Investigations in Motorway Segments. Infrastructures, 8(3), 40.

Nikolaou, D., Ziakopoulos, A., & Yannis, G. (2023b). A Review of Surrogate Safety Measures Uses in Historical Crash Investigations. Sustainability, 15(9), 7580.

Nikolaou, D., Ziakopoulos, A., Dragomanovits, A., Roussou, J., & Yannis, G. (2023c). Comparing machine learning techniques for predictions of motorway segment crash risk level. Safety, 9(2), 32

Ord, J. K., & Getis, A. (1995). Local spatial autocorrelation statistics: distributional issues and an application. Geographical analysis, 27(4), 286-306.

Ozbay, K., Yang, H., Bartin, B., & Mudigonda, S. (2008). Derivation and validation of new simulation-based surrogate safety measure. Transportation research record, 2083(1), 105-113.

Padgham, M., Lovelace, R., Salmon, M., & Rudis, B. (2017). osmdata. Journal of Open Source Software, 2(14).

Paleti, R., Sahin, O., & Cetin, M. (2017). Modelling the impact of latent driving patterns on traffic safety using mobile sensor data. Accident Analysis & Prevention, 107, 92-101.

Pande, A., Chand, S., Saxena, N., Dixit, V., Loy, J., Wolshon, B., & Kent, J. D. (2017). A preliminary investigation of the relationships between historical crash and naturalistic driving. Accident Analysis & Prevention, 101, 107-116.

Papadimitriou, E., Argyropoulou, A., Tselentis, D. I., & Yannis, G. (2019). Analysis of driver behaviour through smartphone data: The case of mobile phone use while driving. Safety Science, 119, 91-97.

Papaioannou, D., & Kokkalis, A. (2012). Motorway safety in Europe and Greece: A comparative analysis. Procedia-Social and Behavioral Sciences, 48, 3428-3440.

Papantoniou, P., Yannis, G., & Christofa, E. (2019). Which factors lead to driving errors? A structural equation model analysis through a driving simulator experiment. IATSS research, 43(1), 44-50.

Park, S., Son, S. O., Park, J., Oh, C., & Hong, S. (2021, June). Using vehicle data as a surrogate for highway accident data. In Proceedings of the Institution of Civil Engineers-Municipal Engineer (Vol. 174, No. 2, pp. 67-74). Thomas Telford Ltd.

Pebesma, E. (2018). Simple features for R: Standardized support for spatial vector data, R J., 10, 439–446.

Peterson, L. E. (2009). K-nearest neighbor. Scholarpedia, 4(2), 1883.

Petraki, V., Ziakopoulos, A., & Yannis, G. (2020). Combined impact of road and traffic characteristic on driver behavior using smartphone sensor data. Accident Analysis & Prevention, 144, 105657.

Pires, C., Torfs, K., Areal, A., Goldenbeld, C., Vanlaar, W., Granié, M. A., ... & Meesmann, U. (2020). Car drivers' road safety performance: A benchmark across 32 countries. IATSS research, 44(3), 166-179.

Prat, F., Planes, M., Gras, M. E., & Sullman, M. J. M. (2015). An observational study of driving distractions on urban roads in Spain. Accident Analysis & Prevention, 74, 8-16.

Quinlan, J. R. (1986). Induction of decision trees. Machine learning, 1, 81-106.

R Core Team. (2023). R: a Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria https://www.R-project.org/.

Regan, M. A., Williamson, A., Grzebieta, R., & Tao, L. (2012, August). Naturalistic driving studies: literature review and planning for the Australian naturalistic driving study. In Australasian college of road safety conference 2012, Sydney, New South Wales, Australia.

Ripley, B., Venables, B., Bates, D. M., Hornik, K., Gebhardt, A., Firth, D., & Ripley, M. B. (2013). Package ‘mass’. Cran r, 538, 113-120.

Risser, R. (1985). Behavior in traffic conflict situations. Accident Analysis & Prevention, 17(2), 179-197.

Rowe, R., Roman, G. D., McKenna, F. P., Barker, E., & Poulter, D. (2015). Measuring errors and violations on the road: A bifactor modeling approach to the Driver Behavior Questionnaire. Accident Analysis & Prevention, 74, 118-125.

Rue, H., Martino, S., & Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. Journal of the royal statistical society: Series b (statistical methodology), 71(2), 319-392.

Saccomanno, F. F., Cunto, F., Guido, G., & Vitale, A. (2008). Comparing safety at signalized intersections and roundabouts using simulated rear-end conflicts. Transportation Research Record, 2078(1), 90-95.

Sakamoto, Y., Ishiguro, M., & Kitagawa, G. (1986). Akaike information criterion statistics. Dordrecht, The Netherlands: D. Reidel, 81(10.5555), 26853.

Santos, K., Dias, J. P., & Amado, C. (2022). A literature review of machine learning algorithms for crash injury severity prediction. Journal of safety research, 80, 254-269.

Satria, R., Aguero-Valverde, J., & Castro, M. (2021). Spatial analysis of road crash frequency using Bayesian models with Integrated Nested Laplace Approximation (INLA). Journal of Transportation Safety & Security, 13(11), 1240-1262.

Sayed, T., & Zein, S. (1999). Traffic conflict standards for intersections. Transportation Planning and Technology, 22(4), 309-323.

Schölkopf, B., & Smola, A. J. (2002). Learning with kernels: support vector machines, regularization, optimization, and beyond. MIT press.

Shah, S. A. R., & Ahmad, N. (2020). Accident risk analysis based on motorway exposure: an application of benchmarking technique for human safety. International journal of injury control and safety promotion, 27(3), 308-318.

Shapley, L. S. (1953). A value for n-person games.

Sheather, S. (2009). A modern approach to regression with R. Springer Science & Business Media.

Shinar, D. (1984). The traffic conflict technique: A subjective vs. objective approach. Journal of Safety Research, 15(4), 153-157.

Singh, S. (2015). Critical reasons for crashes investigated in the national motor vehicle crash causation survey (No. DOT HS 812 115).

Songchitruksa, P., & Tarko, A. P. (2006). The extreme value theory approach to safety estimation. Accident Analysis & Prevention, 38(4), 811-822.

Stavrakaki, A. M., Tselentis, D. I., Barmpounakis, E., Vlahogianni, E. I., & Yannis, G. (2020). Estimating the necessary amount of driving data for assessing driving behavior. Sensors, 20(9), 2600.

Stephens, A. N., & Groeger, J. A. (2009). Situational specificity of trait influences on drivers' evaluations and driving behaviour. Transportation research part F: traffic psychology and behaviour, 12(1), 29-39.

Stipancic, J., Miranda-Moreno, L., & Saunier, N. (2018b). Vehicle manoeuvers as surrogate safety measures: Extracting data from the gps-enabled smartphones of regular drivers. Accident Analysis & Prevention, 115, 160-169.

Stipancic, J., Miranda-Moreno, L., Saunier, N., & Labbe, A. (2018a). Surrogate safety and network screening: Modelling crash frequency using GPS travel data and latent Gaussian Spatial Models. Accident Analysis & Prevention, 120, 174-187.

Stipancic, J., Miranda-Moreno, L., Saunier, N., & Labbe, A. (2019). Network screening for large urban road networks: using GPS data and surrogate measures to model crash frequency and severity. Accident Analysis & Prevention, 125, 290-301.

Stipancic, J., Racine, E. B., Labbe, A., Saunier, N., & Miranda-Moreno, L. (2021). Massive GNSS data for road safety analysis: Comparing crash models for several Canadian cities and data sources. Accident Analysis & Prevention, 159, 106232.

Strauss, J., Zangenehpour, S., Miranda-Moreno, L. F., & Saunier, N. (2017). Cyclist deceleration rate as surrogate safety measure in Montreal using smartphone GPS data. Accident Analysis & Prevention, 99, 287-296.

Strumbelj, E., & Kononenko, I. (2010). An efficient explanation of individual classifications using game theory. The Journal of Machine Learning Research, 11, 1-18.

Strumbelj, E., & Kononenko, I. (2014). Explaining prediction models and individual predictions with feature contributions. Knowledge and information systems, 41, 647-665.

Sullman, M. J. (2012). An observational study of driver distraction in England. Transportation research part F: traffic psychology and behaviour, 15(3), 272-278.

Tarko, A. P. (2018). Surrogate measures of safety. In Safe mobility: challenges, methodology and solutions (Vol. 11, pp. 383-405). Emerald Publishing Limited.

Tarko, A., Davis, G., Saunier, N., Sayed, T., & Washington, S. (2009). White paper: surrogate measures of safety. Committee on Safety Data Evaluation and Analysis (ANB20).

Tarlochan, F., Dun, S., Mohammed, S. O., Kharbeche, M., Soliman, A., & Gaben, B. (2022, November). Smartphone-based Vehicle Telematics For Naturalistic Driving Studies. In 2022 International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME) (pp. 1-5). IEEE.

Theofilatos, A., Yannis, G., Kopelias, P., & Papadimitriou, F. (2019). Impact of real-time traffic characteristics on crash occurrence: Preliminary results of the case of rare events. Accident Analysis & Prevention, 130, 151-159.

Thode, H. C. (2002). Testing for Normality; Marcel Dekker Inc.: New York, NY, USA.

Tselentis, D. (2018). Benchmarking driving efficiency using data science techniques applied on large-scale smartphone data (Doctoral dissertation, National Technical University of Athens).

Tselentis, D. I., Yannis, G., & Vlahogianni, E. I. (2017). Innovative motor insurance schemes: A review of current practices and emerging challenges. Accident Analysis & Prevention, 98, 139-148.

United Nations. (2022). Available online: https://www.undp.org/sustainable-development-goals (accessed on 18 December 2022).

Vapnik, V. (1999). The nature of statistical learning theory. Springer science & business media.

Wang, C., & Stamatiadis, N. (2014). Evaluation of a simulation-based surrogate safety metric. Accident Analysis & Prevention, 71, 82-92.

Wang, C., Xie, Y., Huang, H., & Liu, P. (2021). A review of surrogate safety measures and their applications in connected and automated vehicles safety modelling. Accident Analysis & Prevention, 157, 106157.

Wang, C., Xu, C., & Dai, Y. (2019). A crash prediction method based on bivariate extreme value theory and video-based vehicle trajectory data. Accident Analysis & Prevention, 123, 365-373.

Wang, X., Yang, J., Lee, C., Ji, Z., & You, S. (2016). Macro-level safety analysis of pedestrian crashes in Shanghai, China. Accident Analysis & Prevention, 96, 12-21.

Ward, M. D., & Gleditsch, K. S. (2018). Spatial regression models (Vol. 155). Sage Publications.

Washington, S., Karlaftis, M., Mannering, F., & Anastasopoulos, P. (2020). Statistical and econometric methods for transportation data analysis. Chapman and Hall/CRC.

World Health Organization. (2023). Global status report on road safety 2023. World Health Organization.

Wright, M. N., & Ziegler, A. (2015). ranger: A fast implementation of random forests for high dimensional data in C++ and R. arXiv preprint arXiv:1508.04409.

Wu, J., Xu, H., Zheng, Y., & Tian, Z. (2018). A novel method of vehicle-pedestrian near-crash identification with roadside LiDAR data. Accident Analysis & Prevention, 121, 238-249.

Xie, K., Yang, D., Ozbay, K., & Yang, H. (2019). Use of real-world connected vehicle data in identifying high-risk locations based on a new surrogate safety measure. Accident Analysis & Prevention, 125, 311-319.

Yang, D., Xie, K., Ozbay, K., & Yang, H. (2021). Fusing crash data and surrogate safety measures for safety assessment: Development of a structural equation model with conditional autoregressive spatial effect and random parameters. Accident Analysis & Prevention, 152, 105971.

Yang, D., Xie, K., Ozbay, K., Yang, H., & Budnick, N. (2019). Modelling of time-dependent safety performance using anonymized and aggregated smartphone-based dangerous driving event data. Accident Analysis & Prevention, 132, 105286.

Yannis, G., Laiou, A., Dragomanovits, A., Nikolaou, D., Folla, K., Michelaraki, E., Kallidoni, M., Apostoleris, K., Mavromatis, S., Georgiopoulos, S., Parissis, M. (2023). Development of the road safety strategic plan in Greece, 2021-2030. Transportation Research Procedia, 72, 256-262.

Yannis, G., Laiou, A., Vardaki, S., Papadimitriou, E., Dragomanovits, A., & Kanellaidis, G. (2011). Parameters affecting seat belt use in Greece. International journal of injury control and safety promotion, 18(3), 189-197.

Yannis, G., Nikolaou, D., Laiou, A., Stürmer, Y. A., Buttler, I., & Jankowska-Karpa, D. (2020). Vulnerable road users: Cross-cultural perspectives on performance and attitudes. IATSS research, 44(3), 220-229.

Yannis, G., Papadimitriou, E., Chaziris, A., & Broughton, J. (2014). Modelling road accident injury under-reporting in Europe. European transport research review, 6(4), 425-438.

Zeileis, A., Kleiber, C., & Jackman, S. (2008). Regression models for count data in R. Journal of statistical software, 27(8), 1-25.

Zhang, H., & Malczewski, J. (2019). Quality evaluation of volunteered geographic information: The case of OpenStreetMap. Crowdsourcing: Concepts, methodologies, tools, and applications, 1173-1201.

Zhao, X., Yang, H., Yao, Y., Qi, H., Guo, M., & Su, Y. (2022). Factors affecting traffic risks on bridge sections of freeways based on partial dependence plots. Physica A: Statistical Mechanics and its Applications, 598, 127343.

Zheng, L., & Sayed, T. (2019). Comparison of traffic conflict indicators for crash estimation using peak over threshold approach. Transportation research record, 2673(5), 493-502.

Zheng, L., Sayed, T., & Essa, M. (2019). Validating the bivariate extreme value modelling approach for road safety estimation with different traffic conflict indicators. Accident Analysis & Prevention, 123, 314-323.

Zheng, L., Sayed, T., & Mannering, F. (2021). Modeling traffic conflicts for use in road safety analysis: A review of analytic methods and future directions. Analytic methods in accident research, 29, 100142.

Ziakopoulos, A. (2021). Spatial analysis of harsh driving behavior events in urban networks using high-resolution smartphone and geometric data. Accident Analysis & Prevention, 157, 106189.

Ziakopoulos, A., & Yannis, G. (2020). A review of spatial approaches in road safety. Accident Analysis & Prevention, 135, 105323.

Ziakopoulos, A., Tselentis, D., Kontaxi, A., & Yannis, G. (2020). A critical overview of driver recording tools. Journal of safety research, 72, 203-212.

Ziakopoulos, A., Vlahogianni, E., Antoniou, C., & Yannis, G. (2022). Spatial predictions of harsh driving events using statistical and machine learning methods. Safety science, 150, 105722.

Appendix I

List of publications produced within the framework of this dissertation

Publications in scientific journals with peer review

- **pj4.** Nikolaou D., Ziakopoulos A., Kontaxi A., Theofilatos A., Yannis G., "Spatial analysis of telematics-based surrogate safety measures" (under review).
- **pj3.** Nikolaou D., Ziakopoulos A., Dragomanovits A., Roussou J., Yannis G., "Comparing Machine Learning Techniques for Predictions of Motorway Segment Crash Risk Level", Safety, Vol.9, Issue 2, 2023.
- **pj2.** Nikolaou D., Ziakopoulos A., Yannis G., "A Review of Surrogate Safety Measures Uses in Historical Crash Investigations", Sustainability, Vol.15, Issue 9, 2023.
- **pj1.** Nikolaou D., Dragomanovits A., Ziakopoulos A., Deliali A., Handanos I., Karadimas C., Kostoulas G., Frantzola E.K., Yannis G., "Exploiting Surrogate Safety Measures and Road Design Characteristics towards Crash Investigations in Motorway Segments", Infrastructures, Vol.8, Issue 3, 2023.

Publications in scientific conference proceedings (full papers with review)

- **pc3.** Nikolaou D., Kontaxi A., Ziakopoulos A., Yannis G., Fortsakis P., Frantzola E., Sigalos K., Kouridakis G., "Naturalistic Spatial Road Safety Analysis: The SmartMaps Project", Proceedings of the Transport Research Arena (TRA) Conference 2024, April 15-18, 2024, Dublin, Ireland.
- **pc2.** Nikolaou D., Kontaxi A., Ziakopoulos A., Yannis G., "Spatial analysis of telematics surrogate safety measures across road environments", Proceedings of the 11th International Congress on Transportation Research, Heraklion, Greece, 20-22 September 2023.
- **pc1.** Yannis G., Nikolaou D., Dragomanovits A., "Investigation of accident modelling data in Greece", Proceedings of the 8th Road Safety & Simulation International Conference, Athens, Greece, 8-10 June 2022.

Scientific awards for the publications produced within the framework of this dissertation

- **sa1.** 09/2023 Road Safety Award – Young Researcher Best Paper for the paper "Spatial analysis of telematics surrogate safety measures across road environments" at the 11th ICTR Congress in Heraklion, Greece.

List of publications in other research thematic areas

Publications in scientific journals with peer review

- **pj7.** Nikolaou D., Ntontis A., Michelaraki E., Ziakopoulos A., Yannis G., "Pedestrian safety attitudes and self-declared behaviour in Greece", IATSS Research, Vol.47, Issue 1, 2023, pp. 14-24.
- **pj6.** Nikolaou D., Typa D., Yannis G., "Investigation of traffic and safety behavior of pedestrians while talking on mobile phone", Advances in Transportation Studies, Special Issue, Vol. 3, 2022, pp. 73-82.
- **pj5.** Nikolaou D., Folla K., Yannis G., "Impact of socioeconomic and transport indicators on road safety during the crisis period in Europe", International Journal of Injury Control and Safety Promotion, Vol.28, Issue 4, 2021, pp.479-485.
- **pj4.** Ziakopoulos A., Nikolaou D., Yannis G., "Correlations of multiple rider behaviors with self-reported attitudes, perspectives on traffic rule strictness and social desirability", Transportation Research Part F: Traffic Psychology and Behaviour, Vol.80, 2021, pp.313-327.
- **pj3.** Yannis G., Nikolaou D., Laiou A., Achermann Stürmer Y., Buttler I., Jankowska-Karpa D., "Vulnerable road users: Cross-cultural perspectives on performance and attitudes", IATSS Research, Vol.44, Issue 3, 2020, pp. 220-229.
- **pj2.** Pires C., Torfs K., Areal A., Goldenbeld C., Vanlaar W., Granié M.A., Achermann Stürmer Y., Usami D.S., Kaiser S., Jankowska-Karpa D., Nikolaou D., Holte H., Kakinuma T., Trigoso J., Meesmann U., Van den Berghe, W, "Car drivers' road safety performance: A benchmark across 32 countries", IATSS Research, Vol.44, Issue 3, 2020, pp. 166-179.
- **pj1.** Ropaka M., Nikolaou D., Yannis G., "Investigation of traffic and safety behavior of pedestrians while texting or web-surfing", Traffic Injury Prevention, Vol.21, Issue 6, 2020, pp. 389-394.

Publications in scientific conference proceedings (full papers with review)

- **pc25.** Folla K., Kallidoni M., Nikolaou D., Yannis G., "Road Safety Key Performance Indicators in Greece", Proceedings of the 11th International Congress on Transportation Research, Heraklion, Greece, 20-22 September 2023.
- **pc24.** Merakou M., Nikolaou D., Folla K., Yannis G., "Analysis of distraction characteristics due to mobile phone use in Greece", Proceedings of the 11th International Congress on Transportation Research, Heraklion, Greece, 20-22 September 2023.
- **pc23.** Krouskos S., Nikolaou D., Folla K., Provatari E., Yannis G., "Analysis of speeding characteristics in Greece", Proceedings of the 11th International

Congress on Transportation Research, Heraklion, Greece, 20-22 September 2023.

- **pc22.** Ziakopoulos A., Michelaraki E., Nikolaou D., Folla K., Yannis G., "Association Rule Mining for Island and Mainland Road Crash Injuries in Greece", Proceedings of the Transport Research Arena (TRA) Conference 2022, November 14-17, 2022, Lisbon, Portugal.
- **pc21.**Georgakopoulos D., Nikolaou D., Roussou J., Yannis G., "The impact of mobility characteristics on public transport and road safety performance in selected European cities", Proceedings of the Transport Research Arena (TRA) Conference 2022, November 14-17, 2022, Lisbon, Portugal.
- **pc20.** Kallidoni M., Nikolaou D., Folla K., Yannis G., "EU countries' ranking in different road crash types", Proceedings of the Transport Research Arena (TRA) Conference 2022, November 14-17, 2022, Lisbon, Portugal.
- **pc19.** Yannis G., Folla K., Nikolaou D., Chaziris A., Kallidoni M., "Assessing Driver Safety Behaviour in Greece", Proceedings of the Transport Research Arena (TRA) Conference 2022, November 14-17, 2022, Lisbon, Portugal.
- **pc18.** Yannis G., Laiou A., Dragomanovits A., Nikolaou D., Michelaraki E., Folla K., Kallidoni M., Georgiopoulos S., Parissis M., "Effective road safety measures in Greece", Proceedings of the Transport Research Arena (TRA) Conference 2022, November 14-17, 2022, Lisbon, Portugal.
- **pc17.** Yannis G., Laiou A., Dragomanovits A., Nikolaou D., Folla K., Michelaraki E., Kallidoni M., Apostoleris K., Mavromatis S., Georgiopoulos S., Parissis M., "Development of the Road Safety Strategic Plan in Greece, 2021-2030", Proceedings of the Transport Research Arena (TRA) Conference 2022, November 14-17, 2022, Lisbon, Portugal.
- **pc16.** Yannis G., Dragomanovits A., Roussou J., Nikolaou D., "Development and implementation of a methodology for the economic appraisal of road infrastructure safety schemes", Proceedings of 6th International Symposium on Highway Geometric Design and Urban Street Symposium, Amsterdam, 26-29 June 2022.
- **pc15.** Typa D., Nikolaou D., Yannis G., "Investigation of traffic and safety behavior of pedestrians while talking on mobile phone", Proceedings of the 8th Road Safety & Simulation International Conference, Athens, Greece, 8-10 June 2022.
- **pc14.** Nikolaou D., Dragomanovits A., Efstathiadis S., Diaconu S., Yannis G., "Best practice for safe roads around schools" Proceedings of the 10th International Congress on Transportation Research, Rhodes, Greece, 2-3 September 2021.
- **pc13.** Michelaraki E., Nikolaou D., Yannis G., "Assessment of the evolution of road safety in Greece" Proceedings of the 10th International Congress on Transportation Research, Rhodes, Greece, 2-3 September 2021.
- **pc12.** Yannis G., Laiou A., Dragomanovits A., Nikolaou D., Folla K., Apostoleris K., Mavromatis S., Georgiopoulos S., Parisis M., "Development of the Road

Safety Strategic Plan in Greece, 2021-2030" Proceedings of the 10th International Congress on Transportation Research, Rhodes, Greece, 2-3 September 2021.

- **pc11.** Papantoniou G., Nikolaou D., Papantoniou P., "Investigation of road accidents' characteristics in the Dodecanese" Proceedings of the 10th International Congress on Transportation Research, Rhodes, Greece, 2-3 September 2021.
- **pc10.** Papantoniou P., Gonidi C., Papatzikou E., Chaziris A., Papadakos P., Nikolaou D., Folla K., Vlahogianni E., Yannis G., "Analysis of traffic and parking characteristics in the Municipality of Athens", Proceedings of the 10th International Congress on Transportation Research, Rhodes, Greece, 2-3 September 2021.
- **pc9.** Yannis G., Folla K., Nikolaou D., Dragomanovits A., Wang X., "Development of a Platform for Global Road Safety Data Analysis" Proceedings of the 8th Transport Research Arena TRA 2020, April 27-30, 2020, Helsinki, Finland.
- **pc8.** Nikolaou D., Goldenbeld C., Ziakopoulos A., Laiou A., Yannis G., "Road user safety attitudes towards driver fatigue", Proceedings of the 8th Transport Research Arena TRA 2020, April 27-30, 2020, Helsinki, Finland.
- **pc7.** Yannis G., Mavromatis S., Folla K., Laiou A., Nikolaou D., Zammataro S., Funk J., Theofilatos A., Welsh R., Talbot R., Fernandez E., Sogodel V., Wismans J., Kluppels L., Carnis L., Mignot D., "Identification of Road Safety Risk Factors in Africa", Proceedings of the 8th Transport Research Arena TRA 2020, April 27-30, 2020, Helsinki, Finland.
- **pc6.** Ropaka M., Nikolaou D., Yannis G., "Investigation of Traffic and Safety Behavior of Pedestrians Texting or Web-Surfing", Proceedings of the Transportation Research Board (TRB) 99th Annual Meeting, Washington, D.C., 12-16 January 2020.
- **pc5.** Yannis G., Dragomanovits A., Roussou J., Nikolaou D., "Economic Assessment of Road Infrastructure Safety Schemes in Greece Using Crash Prediction Methodology", Proceedings of the Transportation Research Board (TRB) 99th Annual Meeting, Washington, D.C., 12-16 January 2020.
- **pc4.** Nikolaou D., Folla K., Bellos E., Yannis G., "Tourism and Road Accidents in Greece", Proceedings of the 9th International Congress on Transportation Research, Athens, Greece, 24-25 October 2019.
- **pc3.** Yannis G., Mavromatis S., Laiou A., Folla K., Nikolaou D., "Thematic Fact Sheets on Road Safety Risk Factors in Africa – A Knowledge and Management Tool", Proceedings of the 9th International Congress on Transportation Research, Athens, Greece, 24-25 October 2019.
- **pc2.** Yannis G., Dragomanovits A., Roussou J., Nikolaou D., "Methodology for the economic assessment of road safety interventions", Proceedings of the 9th International Congress on Transportation Research, Athens, Greece, 24-25 October 2019.

- **pc1.** Nikolaou D., Folla K., Yannis G., “The effect of socioeconomics and transportation conditions on road safety in the European Union”, Proceedings of the 7th Panhellenic Road Safety Conference, Larissa, Greece, 11-12 October 2018.

Scientific awards for the publications in other research thematic areas

- **sa5.** 07/2023 Thomaideio Award (NTUA) for the paper “Investigation of traffic and safety behavior of pedestrians while talking on mobile phone”, published in Advances in Transportation Studies, Vol. 3 (Special Issue).
- **sa4.** 11/2022 Thomaideio Award (NTUA) for the paper “Correlations of multiple rider behaviors with self-reported attitudes, perspectives on traffic rule strictness and social desirability”, published in 2021 in Transportation Research Part F: Traffic Psychology and Behaviour, Vol.80.
- **sa3.** 06/2022 Best paper award for the paper titled “Development and implementation of a methodology for the economic appraisal of road infrastructure safety schemes” during the 6th International Symposium on Highway Geometric Design and Urban Street Symposium.
- **sa2.** 05/2022 Thomaideio Award (NTUA) for the paper “Vulnerable road users: Cross-cultural perspectives on performance and attitudes”, published in 2020 in IATSS Research, Vol.44, Issue 3.
- **sa1.** 11/2021 Thomaideio Award (NTUA) for the paper “Tourism and Road Accidents in Greece.” published in 2019 in the proceedings of the 9th International Congress on Transportation Research “Transport 4.0: The Smart Evolution”.