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.

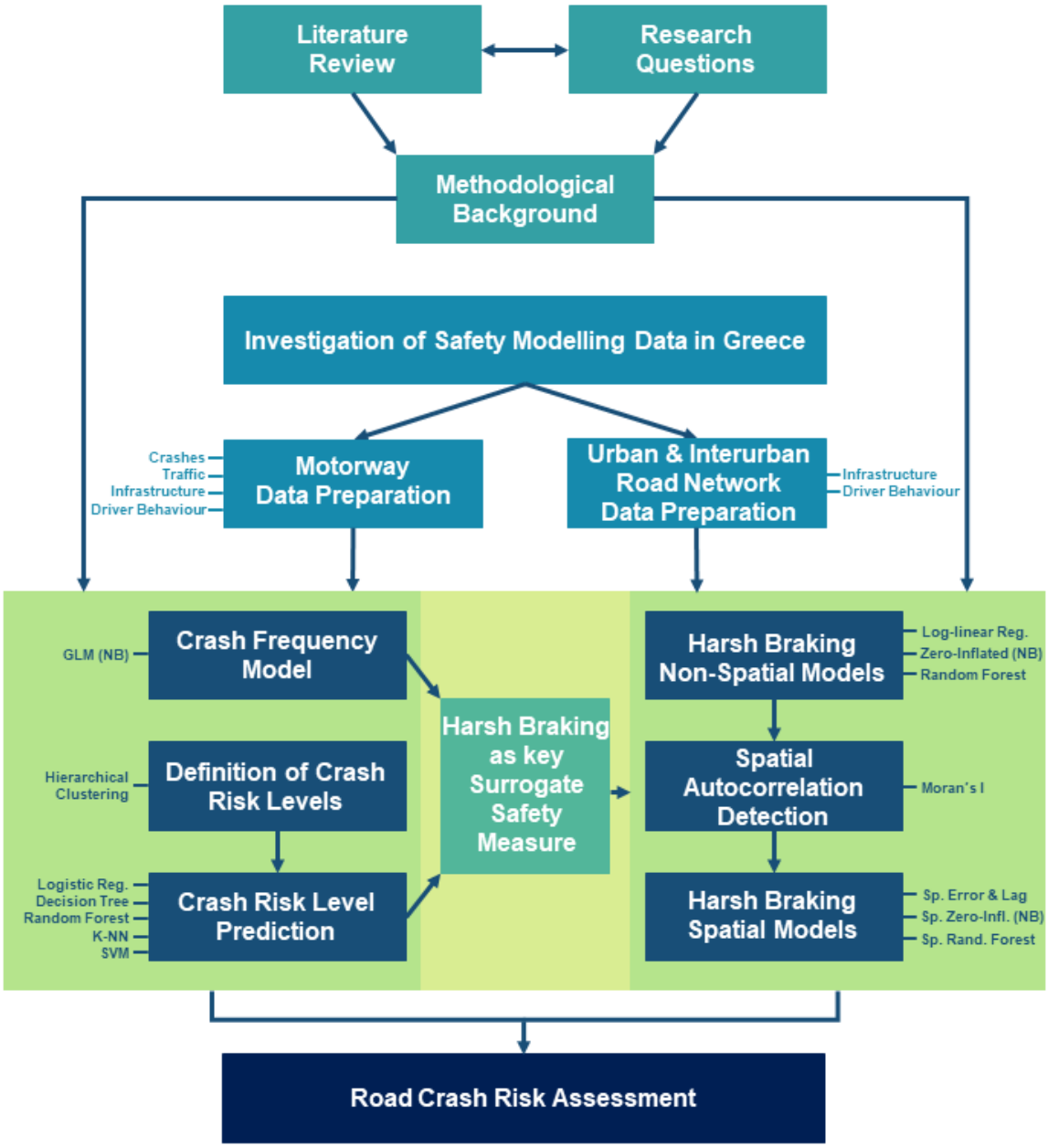

Figure I: Graphical representation of the overall methodological framework of the doctoral dissertation

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

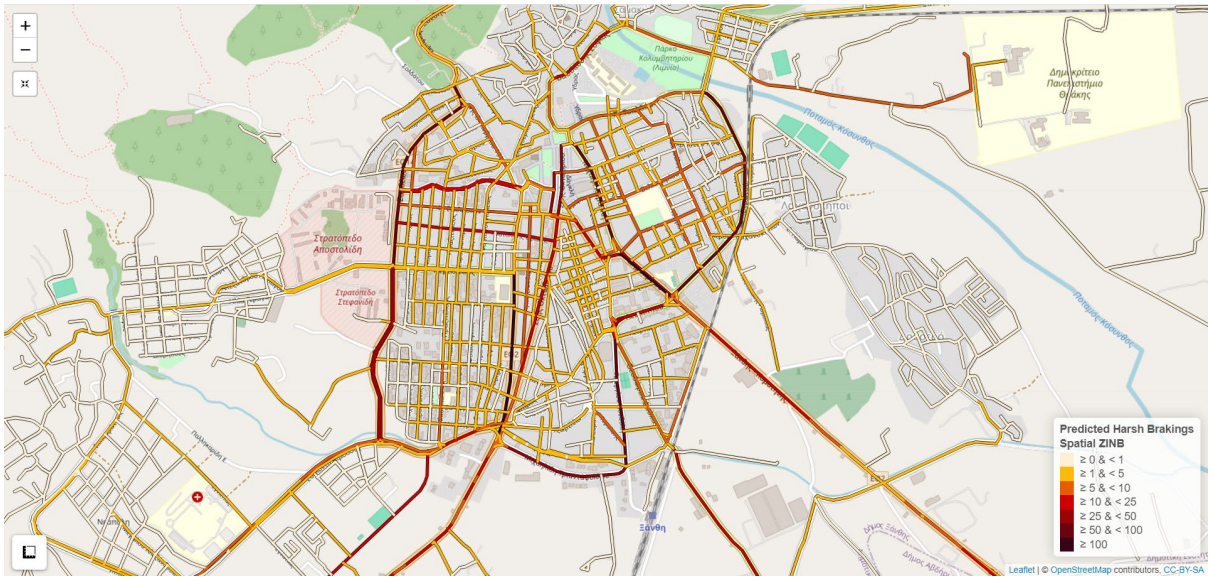

Figure III: Zoomed-in view of the SZINB results for the center of Xanthi

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**.