Extended Synopsis

Road safety is an ever-present issue for modern, motorized societies. Road crashes incur heavy human costs in the form of lives, incapacitations and injuries, as well as a number of additional costs such as direct property damage, disruption costs and service costs, among others. In order to mitigate the consequences of road crashes and to increase road safety levels, a critical tool is the detection of problematic locations, known as hotspots. As this problem involves the examination of entire study areas, dimensions and distances come to play an important role. **Spatial analyses** offer meaningful insights in the calculation of **event frequencies across areas** and for the respective hotspot detection. Traditionally, and due to the scarceness of crash data, spatial analyses were usually conducted at a high level (e.g. counties or municipalities). Rapid technological advancements in driving monitoring and acquisition of rich naturalistic driving data from smartphone sensors open new venues for more detailed and accurate research approaches. Spatial analysis can be conducted **using road segments as basis**, using the more abundant dependent variables of harsh events (namely harsh brakings and harsh accelerations) as proxies for hotspot detection, and utilizing the individual geometric and road network characteristic variables of each one as independent variables for model calibration.

In light of the aforementioned, the main objective of the present doctoral dissertation is the **spatial analysis of harsh event frequencies in road segments** using multi-parametric data, including (i) high resolution naturalistic driving and driver behavior data from smartphone sensors, (ii) microscopic road segment geometry and road network characteristic data from digital maps and (iii) high resolution traffic data.

An exhaustive literature review was conducted across three pillars, namely (i) Spatial approaches in road safety, (ii) Quantitative meta-regressions of exposure parameters used in spatial analyses in road safety and (iii) Overview of driver recording tools. From the review process, it was concluded that spatial analyses of harsh events on urban networks is a **novel, unexplored, and informative research direction**. Smartphone sensors can provide core trip data reliably and consistently, while offering additional information such as mobile use and speeding parameters. Such an approach was best served by naturalistic (and therefore reasonably uninfluenced) driving. The resulting big dataset is required to include extensive coverage of the study area for better calibration of the considered models. The execution of such research can be facilitated from readily available open-source rich data, which will allow the augmentation of high-resolution driver behavior data from smartphones with information of comparable quality.

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

1. How can smartphone data and map data be combined (map-matched) and examined in order to reach meaningful conclusions for road safety levels and to pinpoint possible hotspots in urban road environments?
2. How can harsh event frequencies be analyzed spatially in these environments, and which methods are appropriate for that purpose?
3. Is there spatial autocorrelation present in harsh event frequencies for road segments in urban road environments?
4. Which road geometry and network characteristics affect harsh event frequencies in urban road network environments? Are they the same for harsh brakings and harsh accelerations, and are their effects comparable? How transferable are the previous results in a different study area?
5. Do traffic and driver behavioral parameters have any statistical impact on harsh event frequencies? Are they the same per traffic state?

In order to answer these research questions, an elaborate **methodological framework** was devised, which is shown on Figure I.

The initial stage for spatial analyses involved the selection of **statistical tools** that would be useful and produce informative results. As part of the exploratory spatial analyses, global and local Moran's *I* coefficients, as well as merged and direction-based variograms were selected. Regarding statistical models, it was decided to utilize a balanced variety between classic functional (frequentist) methods, Bayesian stochastic methods and machine learning methods. Specifically, Geographically Weighted Poisson Regression (GWPR) models, Bayesian Conditional Autoregressive Prior (CAR) models and Extreme Gradient Boosting algorithms with random cross-validation (RCV XGBoost) and spatial cross-validation (SPCV XGBoost) were selected. As the dependent variables were frequency (count) variables, all analyses were conducted within a Poisson log-linear framework. The error metrics of (a) (Root) Mean Squared Error (RMSE/MSE), (b) Mean Absolute Error (or Deviation) (MAE/MAD) and (c) (Root) Mean Squared Log Error (RMSLE/MSLE) were adopted to evaluate model performance both for model fit and for predictions. A Custom Accuracy (CA) metric was devised as well.

The next stage involved the **definition of the necessary study areas**. However, a conundrum arose when integrating road user behavior and traffic input data: while they can be used as independent variables to calibrate statistical models, they cannot be meaningfully estimated for areas without data because they are snapshots of a particular instant. This limitation does not arise with geometric/infrastructure data which are fixed attributes. Therefore a critical decision was made for the analyses to be performed on **two parallel pillars**: (1) Prediction models were developed in an urban road network training area, with the intent to transfer them to a second urban road network testing area and assess their predictive performance and (2) Causal models including road user behavior and traffic input data to investigate additional underlying correlations in an effort to further understand the phenomena of harsh braking and harsh acceleration frequencies, and to explore whether there are noteworthy spatial correlations between segments regarding these phenomena. These models were created in an urban arterial study area, as traffic parameters are more clearly defined there.

Afterwards, **digital map data** from OpenStreetMap was extracted and processed, consisting mainly of nodes and ways of the examined road segments. The training urban network area was in Chalandri, Athens, and comprised 869 road segments. Similarly, the test urban network area was in Omonoia, Athens, and comprised 1,237 road segments. The study urban arterial area was a portion of Kifisias Avenue, Athens, and comprised 152 road segments. OSM segmentation is used, a practice that ensures homogeneous road segments that are split only when there is a reason to, such as a change of signage or lanes.

Based on the node coordinates as primary data, and also by augmenting OSM data with NASA's SRTM altitude data, several **road segment geometrical characteristics** were calculated: length, gradient, curvature and neighborhood complexity. In addition, information regarding the presence of traffic lights and pedestrian crossings was extracted in a binary format.

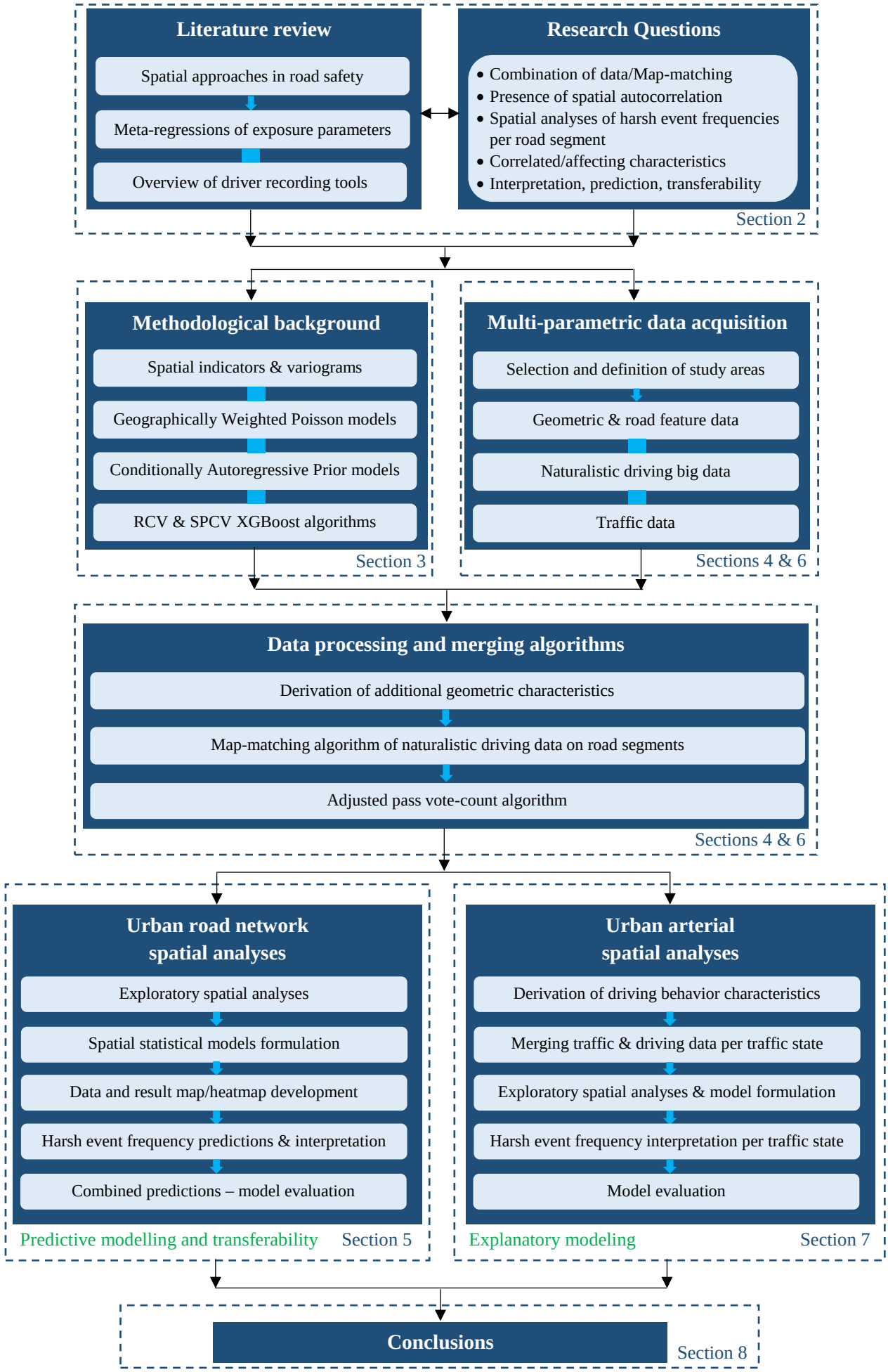

Figure I: Overall methodological framework of the doctoral dissertation

The **naturalistic trip data** in this dissertation was collected and provided by OSeven Telematics through an innovative smartphone application that seamlessly and non-intrusively records driving trips when users drive their vehicles normally. A wealth of naturalistic driving behavior metrics is collected through the use of smartphone sensors with no other equipment required.

Subsequently, a **novel purpose-made map-matching algorithm** was applied so as to match each trip-second of the naturalistic driving smartphone big dataset to the corresponding road segment. Each row of the resulting spatial data-frame represented a different road segment based on OSM segmentation, as per the demands of spatial analysis and the convention of this doctoral dissertation. In locations of several parallel segment axes with high density, such as Kifisias Avenue and its auxiliary parallel roads, another **custom vote-count algorithm** was implemented that compared the trip-seconds assigned to competing segments and ultimately assigned the portion of the trip to the segment with the majority of votes.

For the two urban network areas, the provided dataset corresponded to a period of two months; specifically during October and November 2019. In the training area of Chalandri, 3,294 trips were provided from 230 individual drivers during that period, resulting in 1,000,273 trip-seconds including **1,348 harsh brakings and 921 harsh accelerations** that were analyzed. In the test area of Omonoia, 2,615 trips were provided from 257 individual drivers during that period, resulting in 964,693 trip-seconds including **1,036 harsh brakings and 938 harsh accelerations** that were analyzed.

For urban arterial segments, the provided dataset corresponded to a period of three months, from September and November 2019. In that period, 8,756 trips were provided from 314 individual drivers resulting in 930,346 trip-seconds seconds including **1,543 harsh brakings and 1,033 harsh accelerations** that were analyzed. More importantly, naturalistic driving data were enhanced with traffic data from the **nearest spatio-temporally corresponding measurement location**. Traffic data was provided by the Traffic Management Centre of Athens and featured high resolution (90s) measurements to match the naturalistic driving dataset. All trip-seconds were then classified into three separate traffic flow states (i) free flow, (ii) synchronized flow and (iii) congested flow, based on limits defined from earlier research on Vasileos Konstantinou Avenue which is an extension of Kifisias Avenue to the south. The spatial data-frames were then **formulated separately for free flow and synchronized flow** (congested flow included very scarce harsh events), and the corresponding models were calibrated. Additional information based on the average speeding seconds and average mobile phone seconds of drivers was calculated and utilized in the models as well. All traffic and driver variables, which are non-fixed parameters, were calculated as updating averages per pass for each road segment. This essentially entailed their removal from being snapshots of an instant; their averages are treated as an infrastructure – road segment – characteristic.

With that step, the spatial data-frames were formulated and ready for spatial analyses. Numerous original and interesting results were obtained. In urban road networks, and based on global and local Moran's *I* coefficients, **there is spatial autocorrelation in harsh event frequencies** if only spatially correlated segments are considered. Based on direction based variograms, the average spatial autocorrelation lies within 190 m for harsh braking events and within 200 m for harsh acceleration events. After this distance spatial autocorrelation smoothens out. Furthermore, there is geographic anisotropy in the test urban network area – fluctuations of harsh event frequency semivariance along the North-South axis but not the East-West axis.

For **harsh brakings**, results showed that the exposure parameters of segment length and pass count increase their frequencies. Conversely, increases in gradient and neighborhood complexity reduce harsh

event frequencies. The effect of lane number is unclear and though significant, it is highly influenced by the spatial effects uniquely present in each road segment. This mostly applies to the effect of road type as well, though residential roads have consistently reduced harsh braking counts compared to primary roads. The presence of traffic lights and pedestrian crossings have marginally significant events – in other words, they are significant in one of the regression models and lowest in XGBoost gain. Curvature and road direction is not statistically significant for harsh event frequencies.

For **harsh accelerations**, results also showed that the exposure parameters of segment length and pass count increase their frequencies. Road segment curvature and the presence of traffic lights are positively correlated with harsh accelerations as well. Again, road type and lane number have an unclear effect, although secondary and tertiary roads showed are found as consistently correlated with increases in harsh accelerations compared to primary roads. The presence of pedestrian crossings has marginally significant events, while road direction was not a statistically significant variable for harsh acceleration frequency.

GWPR and CAR models shed more light to the **exact statistical impact** of variables through the more traditional variable coefficients and confidence/credible intervals. XGBoost does not feature traditional econometric variable significance, but can be used to verify that impact through information gain metrics. GWPR and CAR exhibit transferability issues to other areas. Their GLM counterparts can be used for harsh event prediction, however.

On the other hand, XGBoost can be **transferred seamlessly** to new areas. This is due to the fact that XGBoost does not incorporate spatial effects explicitly, but is inherently data-driven. SPCV XGBoost provided improved predictions compared to RCV XGBoost by allowing for spatial splits in the tree ensembles for both harsh brakings and harsh accelerations. Its performance indicates that ML methods are comparable to traditional methods, and not a panacea – although the transformed road segment spatial dataset was not as large as typically employed in ML.

CAR models can **fit on a specific study area extremely well** for harsh event frequencies with a Custom Accuracy (CA: accurate predictions with a ± 1 count tolerance) of more than 95% thanks to the combination of spatially structured and unstructured effects as well as Bayesian inference. In a way, spatial effects 'overfit' the data, but predictions are conducted without them.

Both for harsh brakings and harsh accelerations, the optimal predictive capabilities were obtained by prediction averaging of all four model types. This led to CAs of 87.55% for harsh brakings and 89% for harsh accelerations. There is a gain of more than 2% in CA compared to the next best individual performing models. The models mitigated the weaknesses and outliers of each other and led to a balanced predictive outcome for harsh brakings and harsh accelerations, with promising transferability.

Apart from the numerous statistical results, a large number of **maps and heatmaps** have been produced in the present dissertation, both from raw data and from statistical results. Indicatively, Figure II depicting the recorded harsh brakings in the test area segments and Figure III depicting the respective combined predictions for those segments (CA 87.55%) are shown indicatively below:

Individually, the best performing models regarding predictive capabilities are **different for harsh brakings and harsh accelerations**, as is the amount of improvement in model performance. Specifically, if CA is considered: SPCV XGBoost showed the best performance for harsh brakings (CA>85%), while frequentist and Bayesian GLMs were tied with SPCV XGBoost for harsh accelerations (CA>87%).

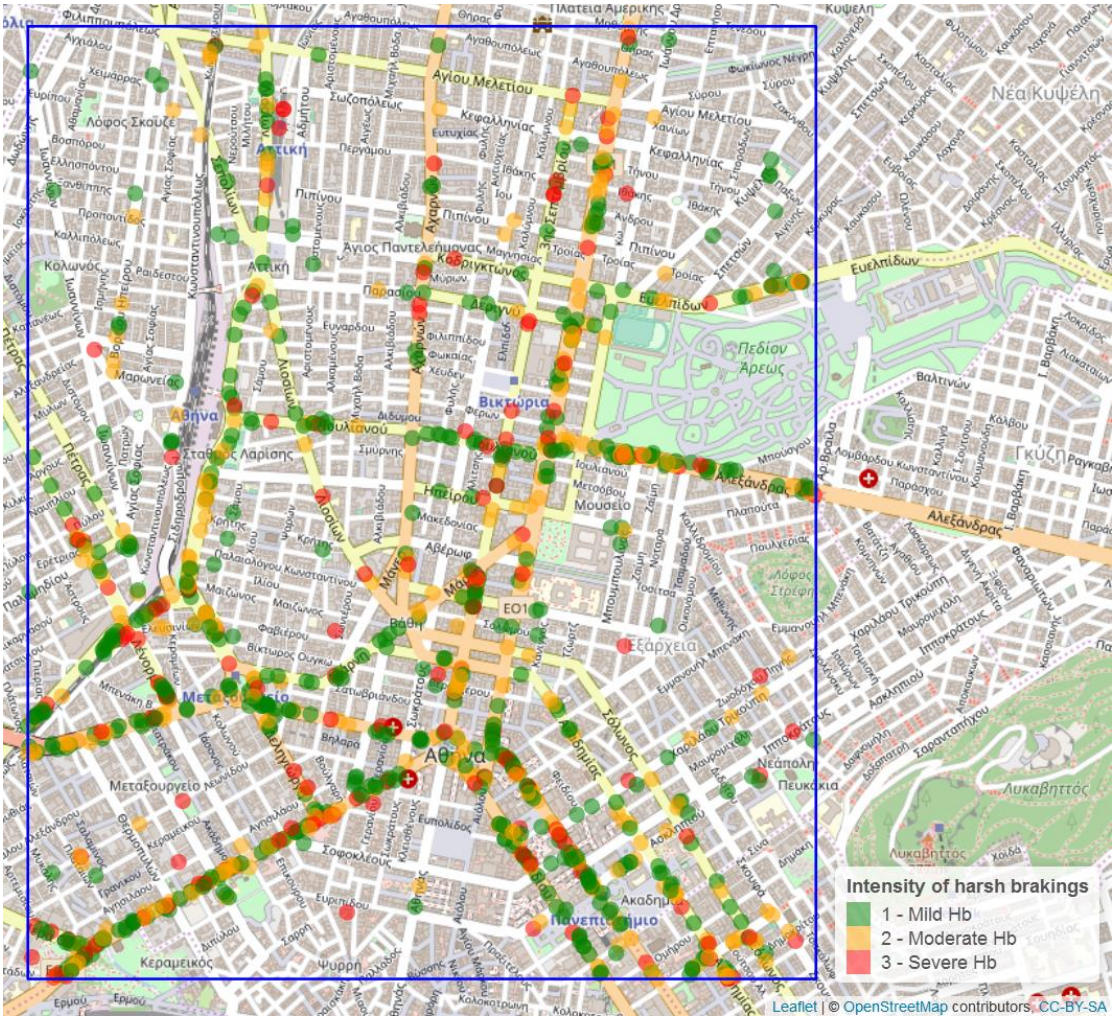

Figure II: Harsh braking events in Omonoia area

RMSE, RMSLE and MAE are **mathematically meaningful error metrics** when dealing with harsh event counts. Since their fluctuations differ based on the existence and distribution of more extreme values, all three are recommended when comparing model performance. The devised CA metric for frequencies augments the **capability assessment** for each model by providing a straightforward comprehensive percentage.

In urban arterial segments, from the initial spatial analyses it was determined that there is **large spatial autocorrelation in harsh braking and harsh acceleration** frequencies of certain segments towards the middle of the study area. This finding applies if only spatially correlated segments are considered, as suggested in the literature, and is based on global and local Moran's *I* coefficient values. These outcomes are in line with the findings for urban road networks.

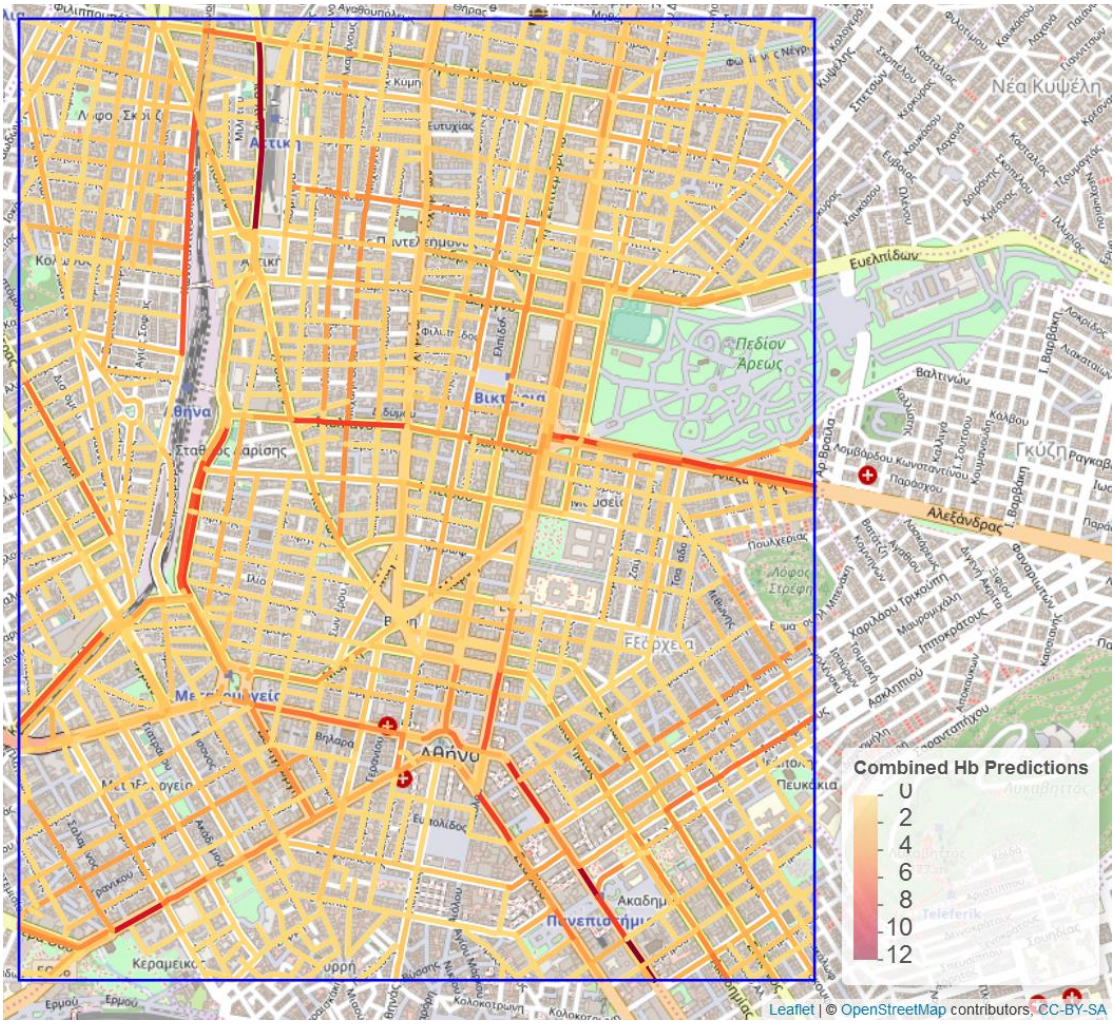

Figure III: Combined prediction heatmap of harsh braking frequencies in Omonoia area

Merged variograms show that the average spatial autocorrelation lies within 310 m for harsh braking events and within 320 m for harsh acceleration events. After this distance spatial autocorrelation smoothens out. **Variograms** for urban arterial segments appear to be **more volatile** compared to those of urban road networks. Moreover, there is spatial cyclicity observed in the axis for both harsh braking and harsh acceleration frequencies; in other words, there is some repetitiveness in the patterns of harsh event frequencies.

In **free flow conditions**, results indicated that the exposure parameters of segment length and pass count, as well as average mobile use seconds of drivers in road segments were all found to contribute positively to **harsh braking frequencies**. Regarding traffic parameters, speed difference between traffic and driver was found to be positively correlated with harsh braking frequencies, while the influence of the averaged standardized current traffic volume was found to be negative. The southbound segments of the study area were found to exhibit systematically fewer harsh brakings compared to the northbound ones. Lastly, average occupancy was found to exert a circumstantially positive influence and gradient was found to exert a circumstantially negative influence in harsh braking frequencies per road segment, depending on the employed method.

Respectively, for **harsh brakings in synchronized flow conditions**, results indicated that segment length, pass count and mobile use seconds all retain their positive contributions. Regarding traffic

parameters, average occupancy seems to assume a stronger role in influencing harsh brakings with a statistically significant positive correlation. The influence of traffic volume (standardized or hourly) was found to be circumstantially negative. The effects of curvature, gradient, number of lanes and road segment bearing weaken to be very circumstantial, depending on the employed method.

In **free flow conditions**, results indicated that segment length, pass count and mobile use seconds (with one exception) all have positive contributions for **harsh acceleration frequency**. The effect of average occupancy was found be consistently positive, while the variable of average speeding seconds of drivers per segments was found to have a marginally positive correlation as well. Average traffic speed was found to have a circumstantially negative influence, depending on the employed method. Geometric and road network characteristic variables were found to have very circumstantial effects.

Respectively, **for harsh accelerations in synchronized flow conditions**, results indicated that pass count and mobile use seconds all retain their positive contributions. For the first time in all arrays of analyses in this dissertation, segment length does not appear to significantly influence harsh acceleration frequency. Traffic volume (standardized or hourly) was found to be positively correlated with harsh accelerations as well. Conversely, an increased number of lanes was found to be negatively correlated with harsh accelerations in CAR models only.

Once again, based on performance error metrics and custom accuracy, it was found that **all three methods of GWPR, CAR and XGBoost** – with random or spatial cross-validation – **are valid and fruitful** methods for the analysis of harsh braking and harsh acceleration frequencies across road segments when employed within a Poisson-lognormal framework. Conducting predictions with the urban arterial dataset is not as meaningful as in urban road networks, however. This is due to the inclusion of traffic and road behavior variables which are not readily available in any location and would require forecasting estimations themselves.

A noteworthy observation is that the **inclusion of traffic and driver behavior variables** in the models weakens the correlations obtained from geometric and road characteristic variables, **substituting them** in a way. Furthermore, it was once again confirmed that harsh accelerations and harsh brakings are two different road safety phenomena. Their frequencies are correlated with certain common variables, albeit with different magnitudes, and also some entirely different parameters.

The **linearity of Kifisias Avenue has led to a more homogenous study area**, with less uncertainty for the acquisition of traffic variables and for the compilation of the urban arterial segment spatial dataset. At the same time, it is possible that this linearity also causes some loss of information or different model performance. Specifically, it was not possible to create direction-based variograms, and GWPR models suffered reductions in their capabilities to adapt to the data more accurately.

Bayesian CAR and XGBoost models did not appear to be affected in the same manner from the study area linearity. In most cases, XGBoost fitted the dataset better, drawing informative gains from more independent variables, especially geometric and road network characteristics. Learning rate (ETA) appeared as the most important hyperparameter during the tuning phase. For SPCV XGBoost, gamma – which governs the minimum loss reduction that can justify making a partition on a tree – was found to affect performance as well.

In summary, the present doctoral dissertation offers **significant innovative contributions** in the field of road safety and traffic behavior analysis:

1. A **novel methodological research framework** was conceived and implemented in order to conduct road safety spatial analyses of harsh driving event frequencies using high resolution multi-parametric data in road segments, providing highly detailed knowledge for hotspot identification.

2. To augment and realize the envisioned framework, a number of **purpose-made big data algorithms** were devised and implemented in intermediate steps, performing critical functions necessary for the spatial analyses, such as derivation of additional characteristics, data merging/processing and map-matching.

3. The methodology was applied in **innovative types of spatial analyses for urban road networks:** (i) spatial analyses of harsh events were conducted at the road segment level and (ii) results were used for successful prediction of event frequencies in a different urban network test area.

4. Additionally, an array of analyses with **additional depth** was conducted in **urban arterial segments**, which were spatially analyzed separately for the traffic states of free flow and synchronized flow.

5. From the detailed microscopic investigations of the present dissertation, **original insights and statistical correlations** between the frequencies of harsh braking and harsh acceleration events per segment and geometrical, road network, traffic and driver behavior variables were revealed.

The availability of multi-parametric high resolution data – and the relative abundance of harsh driving events compared to road crashes – served as impetus to explore the venue of **conducting spatial analysis of harsh events to the much more detailed, microscopic road segment level**, as opposed to the more traditional macroscopic areal analysis (for instance on the county or municipality/district levels). The investigation of harsh event frequencies spatially in general, and in road segments in particular, outlined a completely unexplored research area.

From a scientific standpoint, an added benefit of the adopted approach is the **circumvention of the boundary problem and the modifiable areal unit problem** (MAUP). These problems are ever-present in spatial analyses. The presence of MAUP in particular was confirmed by the meta-regression of Vehicle-Miles Travelled in the quantitative part of the conducted literature review. By modulating the road safety study areas each time, there is no ambiguity on how to treat an event which occurs on the border of a study area, once its respective segment is determined. Furthermore, the modulation that road segments provide standardizes the process of selecting units for analysis, removing MAUP uncertainties for future endeavors.

The inception and creation of the several purpose-made algorithms that were implemented in this doctoral dissertation merits specific mention. The algorithms were devised and implemented in intermediate steps, **performing critical functions** such as derivation of additional geometrical characteristics, data merging (in the form of fusion and aggregation) and map-matching. As such, they provided the means for realizing the envisioned innovative framework and prepared the spatial data-frames comprising of road segments

that were analyzed afterwards. They enabled the **seamless transferability of the entire methodological and data processing framework** followed in the present doctoral dissertation.

Specifically, the algorithm for the derivation of **additional geometric characteristics** draws information from the digital nodes that define road segments (or ways in OpenStreetMap). From the node coordinates, segment length, gradient, curvature and neighborhood complexity are calculated. The iterative nature of the algorithm ensures **its functionality in all segments** regardless of total node number, road type or segment location.

Afterwards, a **map-matching algorithm** was implemented in order to match the naturalistic driving data to the road segments of the study areas. To that end, for each trip-second the nearest road segment, termed Minimum-Distance Way (MDW), was determined using a composite two-step calculation of point-to-point and point-to-polyline distances. Moreover, the algorithm included moving-window approaches that reduced dimensions for the comparison matrices, thus reducing computational times. The adoption of this approach enabled **hands-on implementation** of the map-matching process with direct control over the outcomes, without having to rely on third party services which are unknown 'black box' processes that also require processing fees.

As a necessary subroutine complementary to the map-matching algorithm, an adjusted pass vote-count algorithm was devised. This was an essential subroutine in order to **mitigate GPS uncertainties**, through an advanced vote-count algorithm that assigned the trip to the road segment winning the majority of matched instances. The use of the subroutine proved critical in locations of several parallel segment axes with high density, such as Kifisias Avenue and its auxiliary parallel roads, **increasing the overall robustness of the process**.

The implementation of a final custom algorithm was required for urban arterial analyses in order to **enhance the naturalistic driving dataset with traffic data** prior to map-matching. This algorithm entailed the separation of segments and measurement locations per direction (northbound, southbound) and the determination of the measurement with the **minimum spatio-temporal distance** of each trip-second between the two very large naturalistic data and traffic measurement datasets.

The importance of examining the spatial autocorrelation of harsh events (through global and local Moran's *I* indicators) only in relation with correlated segments **confirmed** both the overall suggested good practices but also the road safety practices followed when analyzing crashes. Furthermore, for the first time **distances measuring the influencing range of spatial autocorrelation** of harsh brakings and harsh accelerations were calculated using variograms, which also determined that these distances differ per road type.

Furthermore, the wealth of high-resolution multi-parametric data and the robustness of the data processing and merging phases permitted the execution of **innovative types of spatial analyses**. It is the first time that harsh driving events are analyzed on the road segment level for urban road networks. The present dissertation managed to **overcome the typical issues of data scarcity** for urban road networks, which are heavily understudied areas in road safety.

An equally important innovation, to the knowledge of the author, is that spatial data-frames and spatial approaches are used to **conduct road safety predictions in a different urban network test area, which also showed a high rate of success**. This constitutes a solid basis to claim high transferability of prediction results in similar areas. In addition to the previous, it is the **first time that XGBoost**

algorithms are used for spatial analyses in road safety. XGBoost proved to be a very potent and overall promising analysis method. The exploration of random cross-validation and **spatial cross-validation**, which is a very recent concept, provides further depth to the results of the algorithm.

Moreover, the results of the urban road network analysis confirm that a utility balance exists between functional (frequentist) methods (GWPR), Bayesian stochastic methods (CAR) and machine learning methods (XGBoost). These methods created models which fit the data differently, and they predicted peak frequencies for different segments. However, their **combination through prediction averaging yielded more accurate results** compared to individual models, as the outliers were mitigated and the correct predictions were enhanced.

For urban arterial segments, it was revealed that **different variables** are significantly correlated with harsh event occurrence **per traffic state.** To the knowledge of the author, this is one of the very few research endeavors that **captured the traffic conditions at the instance** of the examined phenomenon, and the **only one for harsh events.** Variables such as speed difference of traffic and individual driver become much more meaningful for the interpretation of harsh event frequencies, even if they are aggregated per road segment. Overall, the complex non-linear manner in which traffic parameters impact harsh event frequencies was revealed by the present research.

As an overall remark for the numerous conducted analyses, most geometrical, road network, traffic and driver behavior variables were found as statistically significant at least once. These results **showcase the inherent differences** of harsh braking and harsh acceleration phenomena, as the respective frequencies are correlated with **consistently different variables.** What is more, they support holistic approaches for road safety that include **multi-parametric data**, in an effort to capture most sides of the road environment and its users in statistical models.

The creation of **comprehensive road safety maps and heatmaps** for harsh events offers a unique tool to road management authorities, stakeholders and road users that depicts complex data and model predictions in a straightforward manner that is easy to follow, to communicate and to integrate in any working environment or personal decision. In the produced maps, the multi-layered effort of this dissertation is instilled and disseminated from the scientific to the public domain.

One final niche innovation of the present research is the inception and implementation of the **unique model performance metric** of Custom Accuracy. Custom Accuracy offered a useful way to measure the accuracy of predictions for count models that borrows both from classification metrics (such as the confusion matrix) and from regression metrics (such as Mean Absolute Percent Error). By measuring the percentage of correct predictions with a ±1 tolerance, this metric is intuitive and readily comprehensible.