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

Spatial Analysis of Road Safety and Traffic Behaviour using High Resolution Multi-parametric Data

Apostolos Ziakopoulos

Doctoral Dissertation

Supervising Committee:
George Yannis, Professor NTUA
Constantinos Antoniou, Professor TUM
Eleni Vlahogianni, Associate Professor NTUA

Athens, July 2020

This research is co-financed by Greece and the European Union (European Social Fund- ESF) through the Operational Programme «Human Resources Development, Education and Lifelong Learning» in the context of the project "Strengthening Human Resources Research Potential via Doctorate Research" (MIS-5000432), implemented by the State Scholarships Foundation (IKY).

Ευρωπαϊκή Ένωση
European Social Fund

Operational Programme
Human Resources Development,
Education and Lifelong Learning

Co-financed by Greece and the European Union

ΕΣΠΑ
2014-2020
ανάπτυξη - εργασία - αλληλεγγύη

Contact details

For any inquiries about the contents of this PhD dissertation, please contact the author using any of the following contact details:

Apostolos Ziakopoulos

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: apziak@central.ntua.gr
Personal e-mail: apziak@windowslive.com

Phone (NTUA): (+30) 2107721575
Phone (mobile): (+30) 6945362677

Researcher website: http://www.nrso.ntua.gr/apziak/

Any errors in the text are the sole responsibility of the author.

Acknowledgements

As anyone who has ever undertaken it will attest, the inception, conduct and completion of a doctoral dissertation resembles a long, arduous journey towards a mountaintop. It includes lengthy, winding roads that often lead to dead ends, and force you to backtrack, and a lot of steep uphill climbing with few and scarce downhill breaks. Determination, patience, discipline and hard work are constantly required, and at times you are brought to your limits. Nonetheless, my journey was made possible – and also easier – by certain people, all of whom I would like to thank deeply.

Firstly, I would like to thank my supervisor, Professor George Yannis, who has been a constant source of inspiration and guidance ever since my undergraduate years. Our cooperation has always been excellent, and I deeply treasure his advice, both for scientific matters and for life in general. He constantly endeavors to reveal the best in me, provides immense support and always believes in my capabilities, even more than myself at times. Moreover, he created and leads the superb National Road Safety Observatory (NRSO) research team, in a truly enjoyable working environment at NTUA.

I would also like to sincerely thank my two co-supervisors: Professor Constantinos Antoniou offered valuable contributions on the conceptual direction of the dissertation and on appropriate analytical methods to implement. Associate Professor Eleni Vlahogianni provided extremely keen insights for several critical methodological steps in this research, and greatly enhanced my analytical approach and overall scope with her guidance.

In addition, I would like to express my gratitude for the recommendations and constructive input which I received from the remaining members of the examination committee: Professor John Golias, Professor Andreas Loizos, Associate Professor Nikolaos Geroliminis and Assistant Professor Eleonora Papadimitriou. Their comments enabled the creation of a more thorough dissertation and increased its scientific quality overall. I feel the need to explicitly thank Dr. Eleonora Papadimitriou also for her mentoring and her great lessons in conducting research and managing people, projects and data during my first steps at NTUA.

I would like to express my gratitude to OSeven Telematics and its team, who have provided the naturalistic driving data that served as the core of the undertaken research. Their support and aid in several technical and implementation issues was critical for the transition from concept to results. OSeven has a long lasting multilevel cooperation with NTUA, which I am grateful to be part of. Dr. Petros Fortsakis, Co-Founder and Chief Research Officer, deserves specific acknowledgement for his personal interest in the present dissertation, becoming enthusiastically involved from start to completion.

Equally, I would like to thank the people of the Traffic Management Centre of Athens, who provided the traffic data used in this dissertation. Specifically, I would like to thank the TMC Supervisor, Mr. Dimitrios Kefallinos, for making this research possible by providing access to the data. I would also like to particularly thank my good friend Mr. Antonis Chaziris, Traffic Engineer at the TMC, for his advice during the data extraction process and his availability and eagerness to help with all my questions whenever they arose.

It feels important to acknowledge the contribution of the several open-source platforms that were utilized during the process of conducting this dissertation. The OpenStreetMap community provides and maintains the worldwide open-source maps that served as a high quality basis for my spatial analyses. NASA's freely provided SRTM data greatly complemented the digital maps. The R core team and individual package authors provide a mighty open-source arsenal with an ever-growing wealth of tools for data science, statistical modelling and production of graphics, and the respective documentation at CRAN. Moreover, I received countless hours of support and helpful tips that each took me one step further from platforms such as Stack Overflow, Stack Exchange, Udemy, YouTube and other online sources. I thank all these eponymous and anonymous heroes that offered them and give them a nod of respect for the time and effort they spend to aid others, mostly without compensation.

Furthermore, I would like to thank the Greek State Scholarships Foundation (IKY) that provided the scholarship funding this doctoral dissertation. Apart from financial support, the scholarship provided an added incentive to focus on this work and the pace to stay on schedule.

At this point, I would like to deeply thank my great friend Dr. Athanasios (Akis) Theofilatos. Ever since our first cooperation, during my Diploma thesis, you infected me with a passion for research, statistics and the constant pursuit of learning and self-improvement. You have been a mentor and gave the best advice at hard times. Akare, thank you.

Likewise, I would also like to truly thank all my great friends and colleagues from the NRSO and other Department of the Transportation Planning and Engineering research groups. Their company, encouragement, active help or even mere presence at times was an invaluable support which countered the weight of undertaking a PhD. Alexandra, Tassos, Panagiotis P., Dimos, Christos, Panagiotis K., Dimitris T., Katerina, Foteini, Areti, Manos B., Manos K., Elena, Panagiotis F., Dimitris N., Julia, Vassilis, Kostas, Eva, Virginia, Christina, Maria, Marios, and especially Armira, who eased a lot of my working burdens in order to make the finalization of this dissertation possible, thank you all so much. You made time spent at the office a joy.

My friends outside of work were also there for me in a similar manner, in good times and in bad ones, and likewise have my endless gratitude. Christos, George, Alexandros T., Thanos, Avgoustina, Elena, Thrasyvoulos, John, Petros, Georgina, Antonis, Antonis, Alexandros K., Sotiris, Vangelis, whether chatting over a cup of coffee – or a mug of beer –, head-banging at a metal concert, or exploring mystical worlds at a gaming table, I am glad it's with you guys. Thank you all so much. Here’s to many good memories and here's to many more yet to come.

Words are too small to express my gratitude towards my family: my parents, Nikos and Nikoletta, and my sister, Ellie. You provide me with immense love, support and encouragement and always believe in me, no matter the task. I owe you all good traits in the person I have come to be. This would have been impossible without you. Thank you, from my heart of hearts.

And now that this journey is concluded, I turn my eyes towards the very special one who has been there every step of the way, with love, faith, patience, understanding. You are my muse. Elena, thank you, for everything.

I got no reason, to lie to you
What's in the cards, that's what I do
I was born a-running, and laughing out loud
With my feet on the ground
And my head in the clouds

You better run
Baby, you better run
I got a blade like lightning
Silver bullets in my gun

I'm short and I'm tall, I'm black and I'm white
Sometimes I'd be wrong, sometimes I'd be right
I'm iron, I'm steel, and I'm bad to the bone
You come looking for trouble, honey
Don't you come alone

You better run
Baby, you better run
I got a blade like lightning
Silver bullets in my gun

I've seen them come, and I've seen them go
I've seen things and been people, that nobody knows
I'm talking in pictures, and I'm painting them black
I'm telling you, fish-face,
You ain't never coming back

You better run
Baby, you better run
– Oh yeah –
I got a blade like lightning
Silver bullets in my gun

– **Motörhead**, You better run

Table of Contents

Table of Tables

Table of Figures

List of Acronyms

A(A)DT	Average (Annual) Daily Traffic
AIC(c)	Akaike Information Criterion (corrected)
API	Application Program Interface
B-V	Bicycle-Vehicle crashes
CA	Custom Accuracy
CAR	Conditionally Auto-Regressive
CPM	Crash Prediction Model
DIC	Deviance Information Criterion
GLM	Generalized Linear Model
GPS	Global Positioning System
GWGLM	Geographically Weighted Generalized Linear Model
GW(P)R	Geographically Weighted (Poisson) Regression
HA	Harsh Acceleration
HB	Harsh Braking
IVDR	In-Vehicle Data Recorders
kNN	*k* Nearest-Neighbors
LMPL	Log Marginal Predictive Likelihood
MAE/MAD	Mean Absolute Error/Deviation
MAPE	Mean Absolute Percent Error
MAUP	Modifiable Areal Unit Problem
MC	Motorcycle Crashes
MCMC	Markov Chain Monte Carlo
MDI	Mean Decrease in Impurity
MDW	Minimum-Distance Way
ML	Machine Learning
MLE	Maximum Likelihood Estimation
NTUA	National Technical University of Athens
OBD	On-Board Diagnostics
OSM	OpenStreetMap
P-V	Pedestrian-Vehicle crashes
RCV	Random Cross-Validation
RF	Random Forest
(R)MSE	(Root) Mean Squared Error
(R)MSLE	(Root) Mean Squared Logarithmic Error
SGW(P)R	Semi-parametric Geographically Weighted (Poisson) Regression
SPCV	Spatial Cross-Validation
SRTM	Shuttle Radar Topography Mission
SVM	Support Vector Machine
TC	Total Crashes
TMC	Traffic Management Center
V-V	Vehicle-Vehicle crashes
VMT/VKT	Vehicle Miles/Kilometers Travelled
WAIC	Watanabe's modified Akaike Information Criterion
WGS84	World Geodetic System 1984
XGBoost	EXtreme Gradient Boosting

Abstract

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. Naturalistic driving data were collected and processed with purpose-made spatial processing algorithms, performing critical functions such as derivation of additional geometrical characteristics, data merging and map-matching. The resulting spatial data-frames were then analyzed and modelled on a road segment basis. Moran's I coefficients, as well as merged and directional variograms were calculated. Spatial analyses were performed on two parallel pillars: (i) 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 (ii) Causal models including road user behavior and traffic input data were calibrated in an urban arterial study area per traffic state, in order to investigate additional underlying correlations in an effort to further understand the phenomena of harsh braking and harsh acceleration frequencies. 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 implemented.

From the spatial analyses, numerous informative results were obtained. Spatial autocorrelation was identified in both harsh braking and harsh acceleration frequencies, and its range of influence was determined for each study area. In urban networks, certain geometrical characteristics were found to affect harsh braking frequencies per road segment: Segment length is positively correlated with harsh brakings, while gradient and neighborhood complexity are negatively correlated with them. Different geometrical characteristics were found to affect harsh acceleration frequencies per road segment: Segment length, curvature and the presence of traffic lights are positively correlated with harsh accelerations. For both harsh event types, pass count increased frequencies of both types of harsh events, while lane number and road type have more unclear circumstantial effects, depending on the utilized models. Furthermore, successful spatial predictions were conducted by averaging the results of all four methods, achieving accuracy of 87% for harsh brakings and 89% for harsh accelerations.

In urban arterial segments, segment length and pass count were consistently positively correlated with harsh event occurrence overall. In addition, it was determined that different variables are significantly correlated with harsh event occurrence per traffic state: For harsh brakings in free flow conditions, speed difference between traffic and driver was found to exert a positive influence, while the influence of the averaged standardized current traffic volume was found to be negative. In synchronized flow conditions, average occupancy assumes a statistically significant positive correlation for harsh braking frequencies, while the influence of traffic volume was found to be circumstantially negative. For harsh accelerations in free flow conditions, the influence of average occupancy was found be consistently positive, as was the average mobile use seconds of drivers. In synchronized flow conditions, traffic volume was found to be positively correlated with harsh accelerations as well. In both traffic states, geometric and road network characteristic variables were found to have very circumstantial effects.

Περίληψη [Abstract in Greek]

Ο κύριος στόχος της παρούσας διδακτορικής διατριβής είναι η χωρική ανάλυση συχνοτήτων απότομων οδηγικών συμβάντων σε οδικά τμήματα με χρήση πολυπαραμετρικών δεδομένων, ήτοι (i) δεδομένα υψηλής ανάλυσης οδηγικής συμπεριφοράς υπό πραγματικές συνθήκες από αισθητήρες έξυπνων κινητών τηλεφώνων, (ii) γεωμετρία οδικών τμημάτων και χαρακτηριστικά οδικού δικτύου από ψηφιακούς χάρτες και (iii) δεδομένα κυκλοφορίας υψηλής ανάλυσης. Για αυτό το σκοπό, συλλέχθηκαν δεδομένα οδήγησης υπό πραγματικές συνθήκες μέσω μιας καινοτόμου εφαρμογής έξυπνων κινητών τηλεφώνων, δεδομένα από λεπτομερείς ψηφιακούς χάρτες καθώς και δεδομένα κυκλοφορίας. Τα δεδομένα υποβλήθηκαν σε επεξεργασία μέσω εξειδικευμένων χωρικών αλγορίθμων οι οποίοι εκέλεσαν κρίσιμες λειτουργίες όπως ο υπολογισμός πρόσθετων γεωμετρικών χαρακτηριστικών, η συγχώνευση βάσεων δεδομένων και η αντιστοίχιση οδηγικών και κυκλοφοριακών δεδομένων σε οδικά τμήματα. Προέκυψαν πλούσιες βάσεις χωρικών δεδομένων με βάση τις οποίες υπολογίστηκαν ολικοί και τοπικοί συντελεστές I του Moran και βαριογράμματα (variograms) συγχωνευμένα και ανά κατεύθυνση. Πραγματοποιήθηκαν χωρικές αναλύσεις ανά οδικό τμήμα σε δύο παράλληλους άξονες: (i) Ανάπτυξη μοντέλων πρόβλεψης απότομων συμβάντων σε περιοχή δοκιμής αστικού οδικού δικτύου, με σκοπό την μεταφορά τους σε περιοχή ελέγχου και την αξιολόγηση της προβλεπτικής τους ικανότητας και (ii) Ανάπτυξη μοντέλων εμβαθυμένης επεξήγησης απότομων συμβάντων, με συμπερίληψη της οδηγικής συμπεριφοράς και της κυκλοφορίας, τα οποία αναπτύχθηκαν σε περιοχή αστικής λεωφόρου ανά κατάσταση κυκλοφορίας. Σκοπός ήταν να διερευνηθούν επιπλέον υποκείμενες στατιστικές συσχετίσεις για την περαιτέρω κατανόηση των φαινομένων των απότομων επιταχύνσεων και επιβραδύνσεων. Συγκεκριμένα, αναπτύχθηκαν μοντέλα Γεωγραφικά Σταθμισμένης Παλινδρόμησης Poisson (Geographically Weighted Poisson Regression – GWPR), μοντέλα Μπευζιανής Υπό Όρους Αυτοπαλινδρόμησης (Bayesian Conditional Autoregressive Prior – CAR), καθώς και αλγόριθμοι Ραγδαίας Βελτιστοποίησης Συναρτήσεων Απωλειών με τυχαία επικύρωση (Random Cross-validation Extreme Gradient Boosting – RCV-XGBoost) και με χωρική επικύρωση (Spatial Cross-validation Extreme Gradient Boosting – SPCV-XGBoost).

Από τις χωρικές αναλύσεις προέκυψαν πολυάριθμα ενδιαφέροντα αποτελέσματα: Εντοπίστηκε χωρική αυτοσυσχέτιση στις συχνότητες απότομων επιβραδύνσεων και απότομων επιταχύνσεων ανά οδικό τμήμα, και υπολογίστηκε το εύρος επιρροής κατά περίπτωση για κάθε περιοχή μελέτης. Στα αστικά δίκτυα, προέκυψε ότι ορισμένα γεωμετρικά χαρακτηριστικά επηρεάζουν τις συχνότητες απότομων επιβραδύνσεων ανά οδικό τμήμα: το μήκος του τμήματος παρουσιάζει θετική συσχέτιση, ενώ η κλίση και η πολυπλοκότητα της γειτονιάς παρουσιάζουν αρνητική συσχέτιση. Διαφορετικά γεωμετρικά χαρακτηριστικά επηρεάζουν τις συχνότητες απότομων επιταχύνσεων ανά οδικό τμήμα: το μήκος του οδικού τμήματος, η καμπυλότητα και η παρουσία σηματοδότησης παρουσιάζουν θετική συσχέτιση. Επίσης ποσοτικοποιήθηκε η θετική συσχέτιση του αριθμού διελεύσεων για τους δύο τύπους απότομων συμβάντων, ενώ πιο αδύναμες συσχετίσεις παρουσιάστηκαν για τον αριθμό λωρίδων και τον τύπο οδού, ανάλογα με τα χρησιμοποιούμενα μοντέλα. Επιπλέον, πραγματοποιήθηκαν επιτυχώς χωρικές προβλέψεις μεσοσταθμίζοντας τα αποτελέσματα των τεσσάρων μοντέλων και επιτυγχάνοντας ακρίβεια 87% για τις απότομες επιβραδύνσεις και 89% για τις απότομες επιταχύνσεις.

Στα οδικά τμήματα αστικών λεωφόρων, το μήκος των τμημάτων και ο αριθμός των διελεύσεων παρουσίασαν σταθερά θετικές συσχετίσεις με τη συχνότητα απότομων συμβάντων. Επιπλέον, καθορίστηκε ότι διαφορετικές μεταβλητές συσχετίζονται σημαντικά με την εμφάνιση απότομων συμβάντων ανά κατάσταση κυκλοφορίας: Η διαφορά ταχύτητας μεταξύ κυκλοφορίας και οδηγού προέκυψε ότι ασκεί θετική επιρροή στις απότομες επιβραδύνσεις, σε συνθήκες ελεύθερης ροής, ενώ η επιρροή του μέσου όρου τρέχοντος κυκλοφοριακού φόρτου διαπιστώθηκε ότι είναι αρνητική. Σε συνθήκες συγχρονισμένης ροής, η μέση κατάληψη είναι θετικά συσχετισμένη με τις απότομες επιβραδύνσεις, ενώ η επίδραση του κυκλοφοριακού φόρτου βρέθηκε ότι ήταν περιστασιακά αρνητική. Η επιρροή της μέσης κατάληψης βρέθηκε σταθερά θετική στις απότομες επιταχύνσεις σε συνθήκες ελεύθερης ροής, όπως και ο μέσος όρος των δευτερολέπτων χρήσης κινητών τηλεφώνων από τους οδηγούς. Σε συνθήκες συγχρονισμένης ροής, ο κυκλοφοριακός φόρτος προέκυψε ότι συσχετίζεται θετικά τη συχνότητα απότομων επιταχύνσεων. Και στις δύο καταστάσεις κυκλοφορίας, οι γεωμετρικές μεταβλητές και τα χαρακτηριστικά του οδικού δικτύου βρέθηκε ότι έχουν πολύ περιστασιακές και ασθενέστερες συσχετίσεις.

Το έργο συγχρηματοδοτείται από την Ελλάδα και την Ευρωπαϊκή Ένωση (Ευρωπαϊκό Κοινωνικό Ταμείο) μέσω του Επιχειρησιακού Προγράμματος «Ανάπτυξη Ανθρώπινου Δυναμικού, Εκπαίδευση και Διά Βίου Μάθηση», στο πλαίσιο της Πράξης «Ενίσχυση του ανθρώπινου ερευνητικού δυναμικού μέσω της υλοποίησης διδακτορικής έρευνας» (MIS-5000432), που υλοποιεί το Ίδρυμα Κρατικών Υποτροφιών (IKY).

Ευρωπαϊκή Ένωση
Ευρωπαϊκό Κοινωνικό Ταμείο

Επιχειρησιακό Πρόγραμμα
Ανάπτυξη Ανθρώπινου Δυναμικού,
Εκπαίδευση και Διά Βίου Μάθηση
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

ΕΣΠΑ
2014-2020
ανάπτυξη - εργασία - αλληλεγγύη

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.

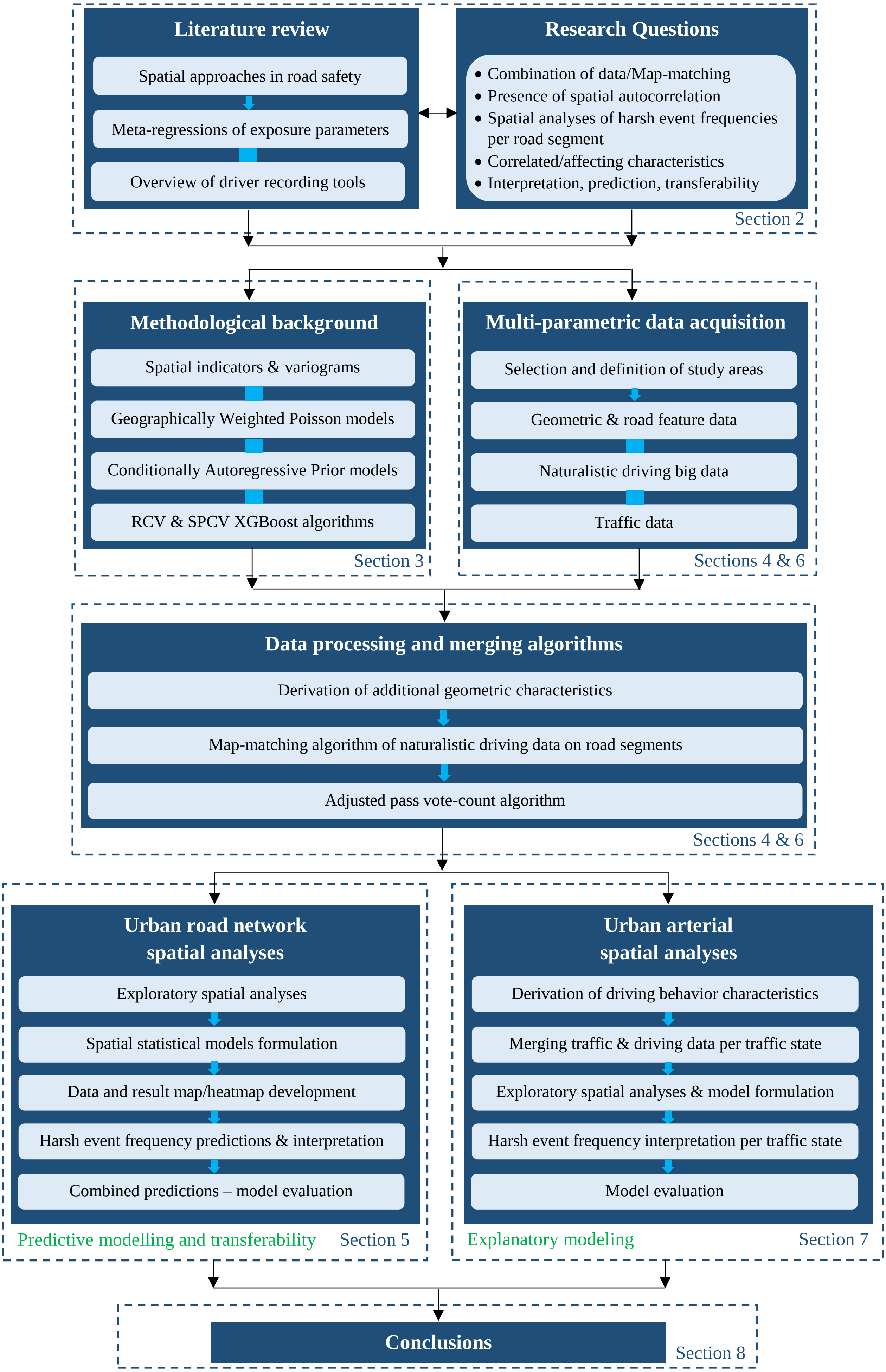

Literature review
Spatial approaches in road safety
Meta-regressions of exposure parameters
Overview of driver recording tools
Research Questions
Combination of data/Map-matching
Presence of spatial autocorrelation
Spatial analyses of harsh event frequencies per road segment
Correlated/affecting characteristics
Interpretation, prediction, transferability
Section 2
Methodological background
Spatial indicators & variograms
Geographically Weighted Poisson models
Conditionally Autoregressive Prior models
RCV & SPCV XGBoost algorithms
Section 3
Multi-parametric data acquisition
Selection and definition of study areas
Geometric & road feature data
Naturalistic driving big data
Traffic data
Sections 4 & 6
Data processing and merging algorithms
Derivation of additional geometric characteristics
Map-matching algorithm of naturalistic driving data on road segments
Adjusted pass vote-count algorithm
Sections 4 & 6
Urban road network spatial analyses
Exploratory spatial analyses
Spatial statistical models formulation
Data and result map/heatmap development
Harsh event frequency predictions & interpretation
Combined predictions – model evaluation
Predictive modelling and transferability
Section 5
Urban arterial spatial analyses
Derivation of driving behavior characteristics
Merging traffic & driving data per traffic state
Exploratory spatial analyses & model formulation
Harsh event frequency interpretation per traffic state
Model evaluation
Explanatory modeling
Section 7
Conclusions
Section 8

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%).

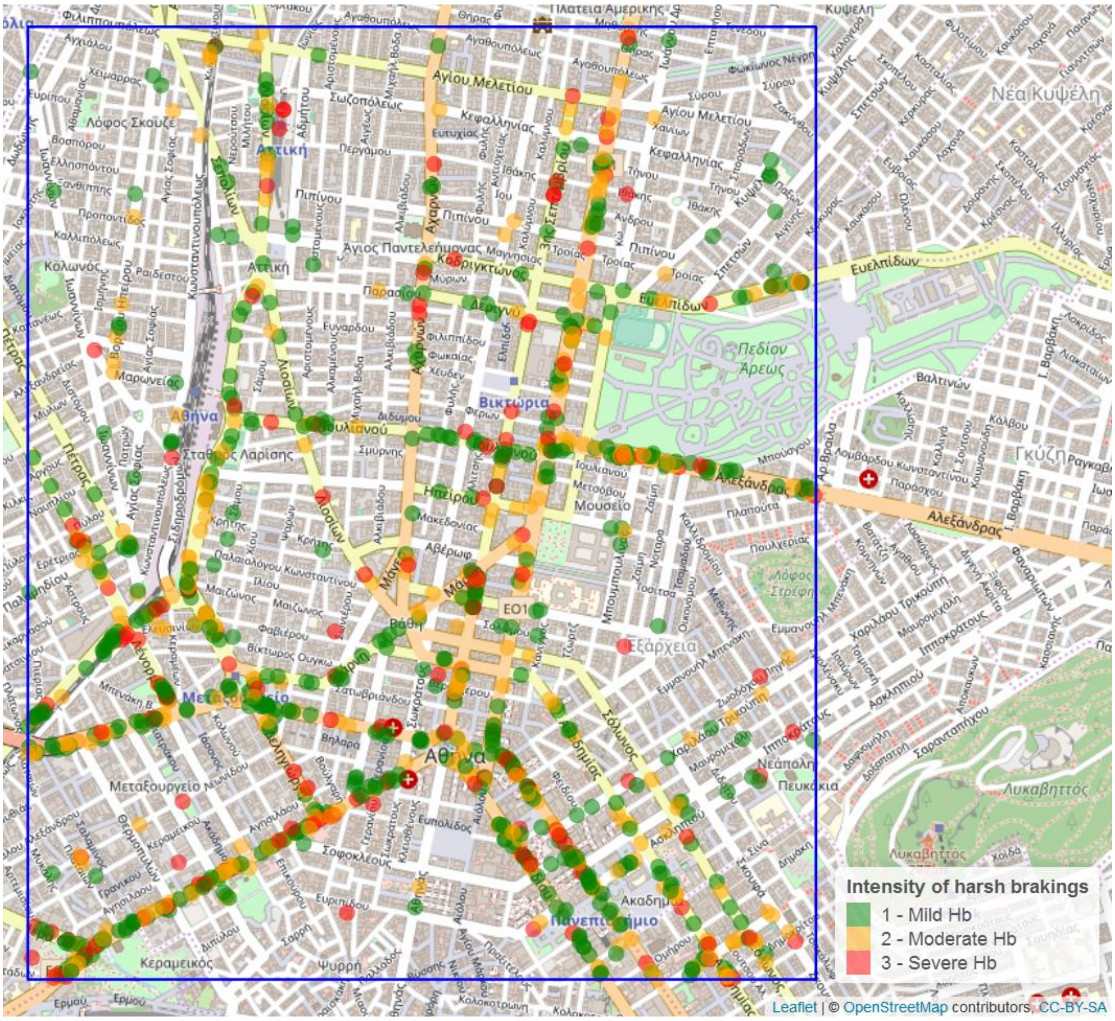

Intensity of harsh brakings
🟩 1 - Mild Hb
🟨 2 - Moderate Hb
🟥 3 - Severe Hb
Leaflet | © OpenStreetMap contributors, CC-BY-SA

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.

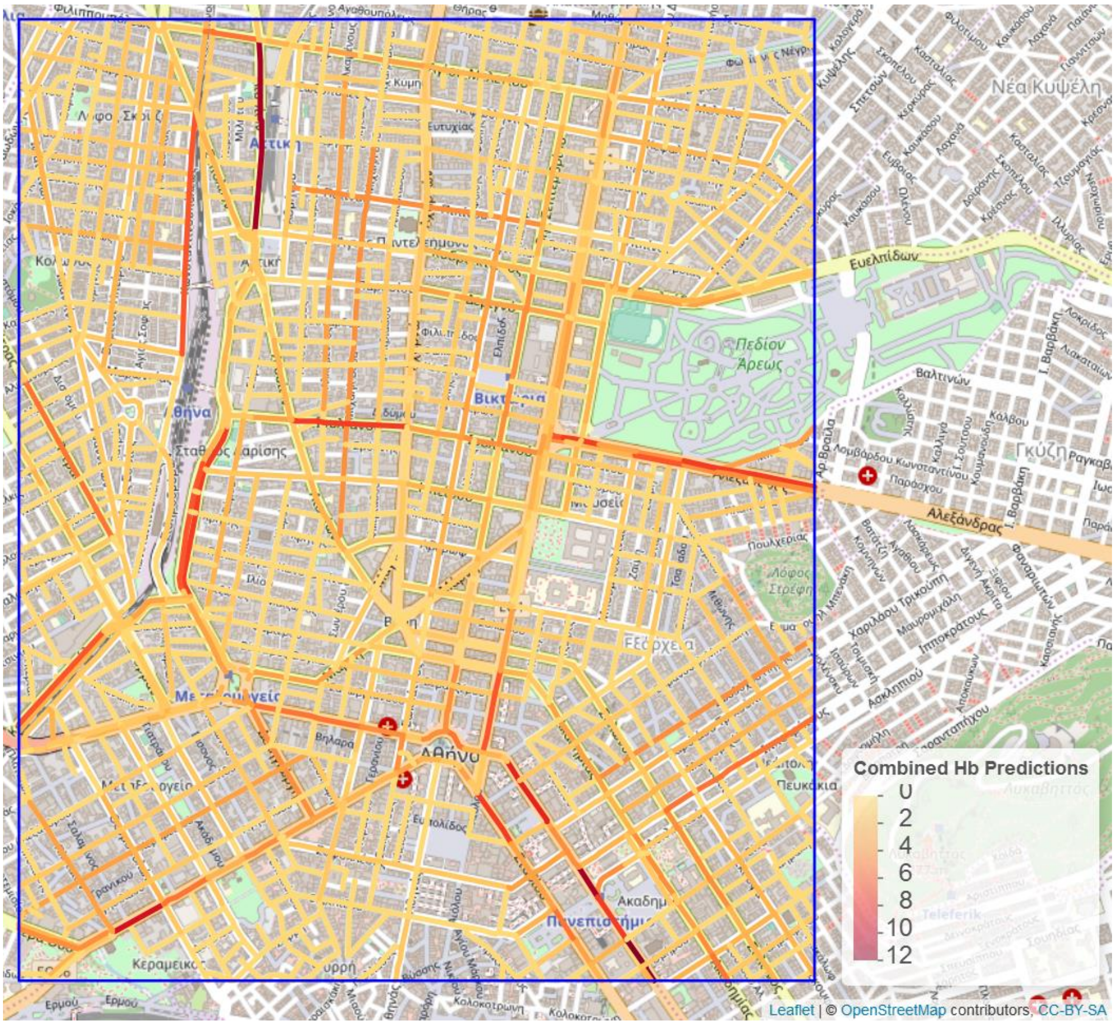


Figure III: Combined prediction heatmap of harsh braking frequencies in Omonoia area

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.

Σύνοψη [Extended Synopsis in Greek]

Η οδική ασφάλεια αποτελεί ένα μόνιμο ζήτημα στις σύγχρονες κοινωνίες οι οποίες διαθέτουν μεγάλο αριθμό οχημάτων. Τα οδικά ατυχήματα προξενούν δυσβάσταχτα ανθρώπινα κόστη, μέσω τραυμάτων, αναπηριών και τραυματισμών, καθώς και μια σειρά πρόσθετων δαπανών και συνεπειών, όπως άμεσες υλικές και περιουσιακές ζημίες, κόστη διακοπής κυκλοφορίας, καθώς και υπηρεσιακά και διαχειριστικά κόστη, μεταξύ άλλων.

Προκειμένου να μετριαστούν οι συνέπειες των τροχαίων ατυχημάτων και να αυξηθούν τα επίπεδα οδικής ασφάλειας, η ανίχνευση επικίνδυνων θέσεων (hotspots) αποτελεί ένα κρίσιμο εργαλείο. Καθώς αυτός ο τύπος προβλήματος περιλαμβάνει την εξέταση ολόκληρων περιοχών μελέτης, οι διαστάσεις και οι αποστάσεις διαδραματίζουν σημαντικό ρόλο. Οι χωρικές αναλύσεις προσφέρουν σημαντικές δυνατότητες για τον υπολογισμό των συχνοτήτων συμβάντων σε διάφορες περιοχές και για την αντίστοιχη ανίχνευση των επικίνδυνων θέσεων. Κατά το παρελθόν, λόγω της έλλειψης λεπτομερών δεδομένων ατυχημάτων σε κάθε θέση, οι χωρικές αναλύσεις διεξαγόταν συνήθως σε μεγάλη κλίμακα παγκοσμίως (περιοχές αντίστοιχες με νομούς ή δήμους). Επί του παρόντος, οι ραγδαίες τεχνολογικές εξελίξεις στην παρακολούθηση της οδηγικής συμπεριφοράς επιτρέπουν την απόκτηση πλούσιων δεδομένων οδήγησης υπό πραγματικές συνθήκες από αισθητήρες έξυπνων κινητών τηλεφώνων (smartphones) και αποκαλύπτουν νέες δυνατότητες για πιο λεπτομερείς και ακριβείς ερευνητικές προσεγγίσεις. Υπάρχει πλέον η δυνατότητα εκτέλεσης χωρικών αναλύσεων με βάση τα μεμονωμένα οδικά τμήματα, χρησιμοποιώντας ως ανεξάρτητες μεταβλητές τα χαρακτηριστικά γεωμετρίας και οδικού δικτύου σε κάθε οδικό τμήμα. Ως εξαρτημένες μεταβλητές χρησιμοποιούνται οι συχνότητες απότομων συμβάντων κατά την οδήγηση (συγκεκριμένα απότομες επιβραδύνσεις και επιταχύνσεις). Οι συγκεκριμένες μεταβλητές είναι πολυπληθέστερες σε σύγκριση με τα οδικά ατυχήματα, και δύναται να λειτουργήσουν ως διαμεσολαβητές για τον εντοπισμό επικίνδυνων οδικών τμημάτων.

Με βάση τα προαναφερθέντα, ο κύριος στόχος της παρούσας διδακτορικής διατριβής είναι η χωρική ανάλυση των συχνοτήτων απότομων οδηγικών συμβάντων ανά οδικό τμήμα με την αξιοποίηση πολυπαραμετρικών δεδομένων, συμπεριλαμβανομένων (i) δεδομένων υψηλής ανάλυσης οδήγησης υπό πραγματικές συνθήκες από αισθητήρες smartphones (ii) γεωμετρίας οδικών τμημάτων και χαρακτηριστικών οδικού δικτύου από ψηφιακούς χάρτες και (iii) δεδομένων κυκλοφορίας υψηλής ανάλυσης.

Για το σκοπό αυτό, διεξήχθη εκτενής βιβλιογραφική ανασκόπηση σε τρεις πυλώνες, συγκεκριμένα: (i) Χωρικές προσεγγίσεις στην οδική ασφάλεια, (ii) Ποσοτικές μετα-παλινδρομήσεις παραμέτρων έκθεσης που χρησιμοποιούνται σε χωρικές αναλύσεις στην οδική ασφάλεια και (iii) Επισκόπηση των εργαλείων καταγραφής οδηγικής συμπεριφοράς. Από την ανασκόπηση, προέκυψε ότι οι χωρικές αναλύσεις αστικών δικτύων είναι ένα καινοτόμο, ανεξερεύνητο και υποσχόμενο ερευνητικό πεδίο. Οι αισθητήρες των smartphones μπορούν να παρέχουν αξιόπιστα δεδομένα οδηγικών διαδρομών, καθώς και επιπλέον δεδομένα περί της χρήσης τηλεφώνου και παραμέτρους ταχύτητας κατά την οδήγηση. Για την εξερεύνηση του συγκεκριμένου ερευνητικού πεδίου, προτιμώνται δεδομένα οδήγησης υπό πραγματικές συνθήκες, τα οποία δέχονται τις λιγότερες επιρροές. Δεδομένα οδήγησης μεγάλης κλίμακας (big data) απαιτούνται για την εκτεταμένη κάλυψη της περιοχής μελέτης και την πληρέστερη ανάπτυξη των αντίστοιχων χωρικών μοντέλων.

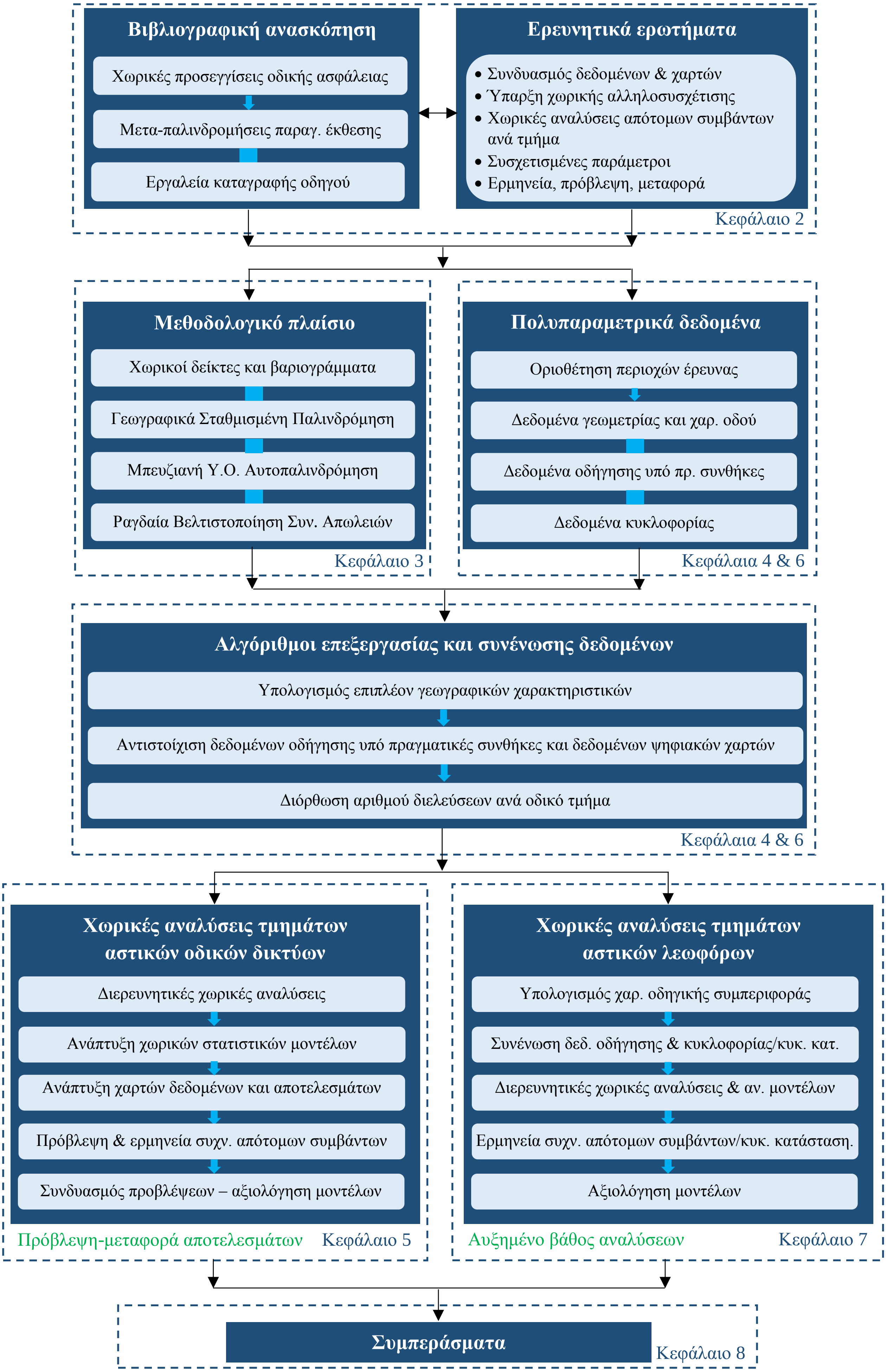
Ακολούθως, διατυπώθηκαν τα εξής ερευνητικά ερωτήματα:

1. Πώς μπορούν να συνδυαστούν και να εξεταστούν τα δεδομένα από smartphones και τα δεδομένα ψηφιακών χαρτών (αντιστοίχιση χαρτών – map-matching) προκειμένου να παραχθούν ουσιαστικά συμπεράσματα για τα επίπεδα οδικής ασφάλειας και να εντοπίσουν πιθανές επικίνδυνες θέσεις σε αστικά οδικά περιβάλλοντα;
2. Πώς μπορούν να αναλυθούν χωρικά οι συχνότητες απότομων οδηγικών συμβάντων σε αυτά τα περιβάλλοντα και ποιες μέθοδοι είναι κατάλληλες για αυτό το σκοπό;
3. Υπάρχει χωρική αυτοσυσχέτιση σε συχνότητες απότομων συμβάντων ανά οδικό τμήμα σε αστικά οδικά περιβάλλοντα;
4. Ποια χαρακτηριστικά γεωμετρίας και οδικού δικτύου επηρεάζουν τις συχνότητες απότομων συμβάντων σε αστικά οδικά περιβάλλοντα; Είναι τα ίδια για απότομες επιβραδύνσεις και απότομες επιταχύνσεις, και είναι συγκρίσιμα τα αποτελέσματά τους; Πόσο μεταβιβάσιμα είναι τα προηγούμενα αποτελέσματα σε διαφορετική περιοχή μελέτης;
5. Έχουν τα χαρακτηριστικά της κυκλοφορίας και της συμπεριφοράς του οδηγού στατιστικά σημαντικές επιρροές στις συχνότητες απότομων συμβάντων; Είναι τα ίδια χαρακτηριστικά και οι ίδιες επιρροές ανά κατάσταση κυκλοφορίας;

Προκειμένου να απαντηθούν αυτά τα ερευνητικά ερωτήματα, επινοήθηκε ένα σύνθετο μεθοδολογικό πλαίσιο, το οποίο φαίνεται στο Σχήμα I.

Το αρχικό στάδιο των χωρικών αναλύσεων περιελάμβανε την επιλογή στατιστικών εργαλείων τα οποία θα ήταν χρήσιμα και θα παρήγαγαν αξιόλογα αποτελέσματα. Ως αρχικό διερευνητικό μέρος των χωρικών αναλύσεων, επιλέχθηκαν οι ολικοί και τοπικοί συντελεστές I του Moran και τα βαριογράμματα (variograms) συγχωνευμένα και ανά κατεύθυνση. Όσον αφορά τα στατιστικά μοντέλα, αποφασίστηκε να χρησιμοποιηθεί ένα ισορροπημένο μίγμα μεταξύ κλασικών συναρτησιακών μεθόδων, Μπευζιανών μεθόδων και Μηχανικής Μάθησης. Συγκεκριμένα, αναπτύχθηκαν μοντέλα Γεωγραφικά Σταθμισμένης Παλινδρόμησης Poisson (Geographically Weighted Poisson Regression – GWPR), μοντέλα Μπευζιανής Υπό Όρους Αυτοπαλινδρόμησης (Bayesian Conditional Autoregressive Prior – CAR), καθώς και αλγόριθμοι Ραγδαίας Βελτιστοποίησης Συναρτήσεων Απωλειών με τυχαία επικύρωση (Random Cross-validation Extreme Gradient Boosting – RCV-XGBoost) και με χωρική επικύρωση (Spatial Cross-validation Extreme Gradient Boosting – SPCV-XGBoost). Δεδομένου ότι οι εξαρτημένες μεταβλητές ήταν μεταβλητές συχνότητας (δεδομένα φυσικών αριθμών), όλες οι αναλύσεις διεξήχθησαν μέσα σε ένα λογαριθμικό-Poisson πλαίσιο. Οι δείκτες σφαλμάτων που χρησιμοποιήθηκαν για την αξιολόγηση της απόδοσης των μοντέλων τόσο κατά την προσαρμογή όσο και κατά τις προβλέψεις ήταν οι (α) (ρίζα) μέσου τετραγώνου σφάλματος (RMSE / MSE), (β) μέσο απόλυτο σφάλμα (ή απόκλιση) (MAE / MAD) και (γ) (ρίζα) μέσου τετραγώνου λογαριθμικού σφάλματος (RMSLE / MSLE). Επινοήθηκε επίσης ένας δείκτης προσαρμοσμένης ακρίβειας (Custom Accuracy – CA).

Ο καθορισμός των απαραίτητων περιοχών μελέτης αποτέλεσε το επόμενο στάδιο. Ωστόσο, προέκυψε ένα δίλημμα κατά την προσπάθεια ενσωμάτωσης των δεδομένων κυκλοφορίας και συμπεριφοράς των οδηγών. Παρότι μπορούσαν να χρησιμοποιηθούν ως ανεξάρτητες μεταβλητές για την ανάπτυξη στατιστικών μοντέλων, δεν μπορούσαν να εκτιμηθούν ουσιαστικά για περιοχές χωρίς δεδομένα, επειδή αποτελούν στιγμιότυπα μιας συγκεκριμένης κυκλοφοριακής χρονικής στιγμής.


Βιβλιογραφική ανασκόπηση
Χωρικές προσεγγίσεις οδικής ασφάλειας
Μετα-παλινδρομήσεις παραγ. έκθεσης
Εργαλεία καταγραφής οδηγού
Ερευνητικά ερωτήματα
• Συνδυασμός δεδομένων & χαρτών
• Ύπαρξη χωρικής αλληλοσυσχέτισης
• Χωρικές αναλύσεις απότομων συμβάντων ανά τμήμα
• Συσχετισμένες παράμετροι
• Ερμηνεία, πρόβλεψη, μεταφορά
Κεφάλαιο 2
Μεθοδολογικό πλαίσιο
Χωρικοί δείκτες και βαριογράμματα
Γεωγραφικά Σταθμισμένη Παλινδρόμηση
Μπευζιανή Υ.Ο. Αυτοπαλινδρόμηση
Ραγδαία Βελτιστοποίηση Συν. Απωλειών
Κεφάλαιο 3
Πολυπαραμετρικά δεδομένα
Οριοθέτηση περιοχών έρευνας
Δεδομένα γεωμετρίας και χαρ. οδού
Δεδομένα οδήγησης υπό πρ. συνθήκες
Δεδομένα κυκλοφορίας
Κεφάλαια 4 & 6
Αλγόριθμοι επεξεργασίας και συνένωσης δεδομένων
Υπολογισμός επιπλέον γεωγραφικών χαρακτηριστικών
Αντιστοίχιση δεδομένων οδήγησης υπό πραγματικές συνθήκες και δεδομένων ψηφιακών χαρτών
Διόρθωση αριθμού διελεύσεων ανά οδικό τμήμα
Κεφάλαια 4 & 6
Χωρικές αναλύσεις τμημάτων αστικών οδικών δικτύων
Διερευνητικές χωρικές αναλύσεις
Ανάπτυξη χωρικών στατιστικών μοντέλων
Ανάπτυξη χαρτών δεδομένων και αποτελεσμάτων
Πρόβλεψη & ερμηνεία συχν. απότομων συμβάντων
Συνδυασμός προβλέψεων – αξιολόγηση μοντέλων
Πρόβλεψη-μεταφορά αποτελεσμάτων
Κεφάλαιο 5
Χωρικές αναλύσεις τμημάτων αστικών λεωφόρων
Υπολογισμός χαρ. οδηγικής συμπεριφοράς
Συνένωση δεδ. οδήγησης & κυκλοφορίας/κυκ. κατ.
Διερευνητικές χωρικές αναλύσεις & αν. μοντέλων
Ερμηνεία συχν. απότομων συμβάντων/κυκ. κατάσταση.
Αξιολόγηση μοντέλων
Αυξημένο βάθος αναλύσεων
Κεφάλαιο 7
Συμπεράσματα
Κεφάλαιο 8

Σχήμα I: Γενικό μεθοδολογικό πλαίσιο της διδακτορικής διατριβής

Παράλληλα, αυτός ο περιορισμός δεν προκύπτει για τα χαρακτηριστικά γεωμετρίας και οδικού δικτύου που αποτελούν σταθερά χαρακτηριστικά της υποδομής. Ως εκ τούτου, ελήφθη μια κρίσιμη απόφαση: Οι αναλύσεις πραγματοποιήθηκαν σε δύο παράλληλους πυλώνες: (i) Ανάπτυξη μοντέλων πρόβλεψης απότομων συμβάντων σε περιοχή δοκιμής αστικού οδικού δικτύου, με σκοπό την μεταφορά τους σε περιοχή ελέγχου και την αξιολόγηση της προβλεπτικής τους ικανότητας και (ii) Ανάπτυξη μοντέλων εμβαθυμένης επεξήγησης απότομων συμβάντων, με συμπερίληψη της οδηγικής συμπεριφοράς και της κυκλοφορίας, τα οποία αναπτύχθηκαν σε περιοχή αστικής λεωφόρου ανά κατάσταση κυκλοφορίας, λόγω του μονοσήμαντου καθορισμού των δεδομένων κυκλοφορίας σε αυτό το περιβάλλον. Σκοπός αυτής της ανάλυσης ήταν να διερευνηθούν επιπλέον υποκείμενες στατιστικές συσχετίσεις για την περαιτέρω κατανόηση των φαινομένων των απότομων επιταχύνσεων και επιβραδύνσεων.

Στη συνέχεια, εξήχθησαν και επεξεργάστηκαν ψηφιακά δεδομένα χαρτών από την πλατφόρμα OpenStreetMap (OSM), αποτελούμενα από κόμβους και τμήματα των οδών που εξετάστηκαν. Η περιοχή δοκιμής αστικού οδικού δικτύου βρίσκεται στο Χαλάνδρι της Αθήνας και περιλαμβάνει 869 οδικά τμήματα. Παρομοίως, η περιοχή ελέγχου αστικού οδικού δικτύου ήταν στην Ομόνοια της Αθήνας και περιλαμβάνει 1.237 οδικά τμήματα. Η περιοχή έρευνας αστικής λεωφόρου είναι τμήμα της Λεωφόρου Κηφισίας στην Αθήνα και περιλαμβάνει 152 οδικά τμήματα. Επιπροσθέτως, χρησιμοποιείται η κατάτμηση OSM, μια πρακτική που εξασφαλίζει ομοιογενή τμήματα δρόμων που χωρίζονται μόνο όταν συντρέχει συγκοινωνιακός λόγος, όπως αλλαγή σήμανσης ή αριθμού λωρίδων.

Με βάση τις συντεταγμένες κόμβων ως πρωτογενή δεδομένα, καθώς επίσης και με την ενίσχυση των δεδομένων OSM με τοπογραφικά δεδομένα υψόμετρου από το SRTM της NASA, υπολογίστηκαν τα γεωμετρικά χαρακτηριστικά των οδικών τμημάτων: μήκος, κλίση, καμπυλότητα και πολυπλοκότητα γειτονιάς. Επιπλέον, πληροφορίες σχετικές με την παρουσία σηματοδοτών και διαβάσεων πεζών αντλήθηκαν σε δυαδική μορφή.

Τα δεδομένα μεγάλης κλίμακας οδήγησης υπό πραγματικές συνθήκες που αξιοποιήθηκαν σε αυτή τη διατριβή συλλέχθηκαν και παρασχέθηκαν από την OSeven Telematics μέσω μιας καινοτόμου εφαρμογής για smartphones που καταγράφει αδιάκοπα και χωρίς παρεμβολές τις διαδρομές καθώς οι χρήστες οδηγούν τα οχήματά τους κανονικά. Ένας μεγάλος αριθμός δεικτών συμπεριφοράς συλλέγεται μέσω της χρήσης αισθητήρων smartphone χωρίς να απαιτείται άλλος εξοπλισμός.

Ακολούθως, εφαρμόστηκε ένας πρότυπος αλγόριθμος αντιστοίχισης χαρτών ειδικά σχεδιασμένος έτσι ώστε να αντιστοιχίζεται κάθε δευτερόλεπτο διαδρομής από τα δεδομένα μεγάλης κλίμακας οδήγησης υπό πραγματικές συνθήκες με το αντίστοιχο οδικό τμήμα. Κάθε σειρά του προκύπτοντος αρχείου χωρικών δεδομένων αντιπροσώπευε ένα διαφορετικό οδικό τμήμα δρόμου με βάση την κατάτμηση OSM, σύμφωνα με τις απαιτήσεις χωρικής ανάλυσης και την προσέγγιση της διδακτορικής διατριβής. Σε τοποθεσίες πολλών παράλληλων αξόνων οδικών τμημάτων με υψηλή πυκνότητα, όπως η Λεωφόρος Κηφισίας και οι βοηθητικές παράλληλες οδοί της, εφαρμόστηκε ένας επιπλέον πρότυπος αλγόριθμος καταμέτρησης ψήφων που συνέκρινε τα δευτερόλεπτα διαδρομής που αντιστοιχούσαν σε ανταγωνιστικά οδικά τμήματα και τελικά ανέθετε τη διαδρομή στο οδικό τμήμα με την πλειοψηφία των ψήφων.

Για τις δύο περιοχές αστικών οδικών δικτύων, παρείχθησαν δεδομένα διαδρομών εντός χρονικού διαστήματος διάρκειας δύο μηνών, συγκεκριμένα τον Οκτώβριο και τον Νοέμβριο του 2019. Στην περιοχή δοκιμών του Χαλανδρίου, αποκτήθηκαν 3.294 διαδρομές από 230 διαφορετικούς οδηγούς κατά τη διάρκεια αυτής της περιόδου, με αποτέλεσμα να παραχθούν 1.000.273 δευτερόλεπτα διαδρομών που περιείχαν 1.348 απότομες επιβραδύνσεις και 921 απότομες επιταχύνσεις. Στην περιοχή ελέγχου της Ομόνοιας, πραγματοποιήθηκαν 2.615 διαδρομές από 257 διαφορετικούς οδηγούς κατά τη διάρκεια αυτής της περιόδου, με αποτέλεσμα να παραχθούν 964.693 δευτερόλεπτα διαδρομής που περιείχαν 1.036 απότομες επιβραδύνσεις και 938 απότομες επιταχύνσεις.

Για τα τμήματα αστικών λεωφόρων, παρείχθησαν δεδομένα διαδρομών εντός χρονικού διαστήματος διάρκειας τριών μηνών, συγκεκριμένα από τον Σεπτέμβριο έως τον Νοέμβριο του 2019. Κατά την περίοδο αυτή, αποκτήθηκαν δεδομένα 8.756 διαδρομών από 314 διαφορετικούς οδηγούς με αποτέλεσμα να παραχθούν 930.346 δευτερόλεπτα διαδρομών τα οποία περιείχαν 1.543 απότομες επιβραδύνσεις και 1.033 απότομες επιταχύνσεις. Τα δεδομένα οδήγησης ενισχύθηκαν με δεδομένα κυκλοφορίας από την πλησιέστερη χωροχρονικά αντίστοιχη θέση μέτρησης. Τα δεδομένα κυκλοφορίας παρασχέθηκαν από το Κέντρο Διαχείρισης Κυκλοφορίας της Αθήνας και περιελάμβαναν μετρήσεις υψηλής ανάλυσης (90s) ώστε να ταιριάζουν με τα δεδομένα οδήγησης υπό πραγματικές συνθήκες. Όλα τα δευτερόλεπτα διαδρομών στη συνέχεια ταξινομήθηκαν σε τρεις ξεχωριστές καταστάσεις ροής κυκλοφορίας (i) ελεύθερη ροή, (ii) συγχρονισμένη ροή και (iii) ροή υπό συμφόρηση, με βάση όρια που είχαν καθοριστεί από προηγούμενη έρευνα στη Λεωφόρο Βασιλέως Κωνσταντίνου, η οποία αποτελεί νότια επέκταση της Λεωφόρου Κηφισίας. Τα αρχεία χωρικών δεδομένων στη συνέχεια διαμορφώθηκαν χωριστά για ελεύθερη ροή και συγχρονισμένη ροή (η κορεσμένη ροή περιελάμβανε ελάχιστα απότομα συμβάντα) και αναπτύχθηκαν τα αντίστοιχα μοντέλα.

Πρόσθετα χαρακτηριστικά υπολογίστηκαν βάσει των μέσων δευτερολέπτων υπέρβασης ταχύτητας και των μέσων δευτερολέπτων χρήσης κινητού τηλεφώνου από τους οδηγούς. Τα χαρακτηριστικά εισήχθησαν στα μοντέλα ως επιπλέον ανεξάρτητες μεταβλητές. Όλες οι μεταβλητές κίνησης και οδήγησης, οι οποίες είναι μη σταθερές παράμετροι, υπολογίστηκαν ως μέσοι όροι οι οποίοι ενημερωνόταν ανά διέλευση από το αντίστοιχο οδικό τμήμα. Αυτή η διαδικασία ουσιαστικά συνεπαγόταν τη μετατροπή τους από στιγμιότυπα και την μεταχείριση αυτών των μέσων όρων ως χαρακτηριστικά οδικών τμημάτων – άρα υποδομής, βήμα απαραίτητο για τις χωρικές αναλύσεις.

Με αυτό το βήμα, διαμορφώθηκαν τα αρχεία χωρικών δεδομένων και εκτελέστηκαν οι χωρικές αναλύσεις, από τις οποίες προέκυψαν διάφορα πρωτότυπα και ενδιαφέροντα αποτελέσματα. Στα αστικά οδικά δίκτυα εντοπίστηκε χωρική αυτοσυσχέτιση με βάση τους ολικούς και τοπικούς συντελεστές *I* του Moran, εάν λαμβάνονται υπόψη μόνο τα χωρικά συσχετισμένα οδικά τμήματα. Με βάση τα βαρογραφήματα ανά κατεύθυνση, η μέση χωρική αυτοσυσχέτιση έχει απόσταση επιρροής τα 190 m για τις απότομες επιβραδύνσεις και στα 200 m για τις απότομες επιταχύνσεις, Πέρα από αυτή την απόσταση, η χωρική αυτοσυσχέτιση εξομαλύνεται. Επιπλέον, υπάρχει γεωγραφική ανισοτροπία στην περιοχή δοκιμής: διακυμάνσεις της ημι-διακύμανσης της συχνότητας απότομων συμβάντων εντοπίζονται κατά μήκος του άξονα Βορρά-Νότου αλλά όχι του άξονα Ανατολής-Δύσης.

Για τις απότομες επιβραδύνσεις, τα αποτελέσματα έδειξαν ότι οι παράμετροι έκθεσης του μήκους τμήματος και του αριθμού περάσματος αυξάνουν τις συχνότητες εμφάνισής τους ανά οδικό τμήμα. Αντίθετα, αυξήσεις στην κλίση και την πολυπλοκότητα της γειτονιάς μειώνουν τις συχνότητες απότομων επιβραδύνσεων. Η επίδραση του αριθμού λωρίδων δεν είναι μονοσήμαντη (αν και σημαντική) και επηρεάζεται σε μεγάλο βαθμό από τους χωρικούς όρους και τις αντίστοιχες τοπικές συσχετίσεις που εμφανίζονται τοπικά σε κάθε οδικό τμήμα. Αυτό ισχύει σε μεγάλο βαθμό και για την επίδραση του τύπου οδού, αν και οι οικιστικές οδοί μειώνουν σταθερά τη συχνότητα των απότομων επιβραδύνσεων σε σύγκριση με τις κεντρικές (πρωτεύουσες) οδούς. Η παρουσία σηματοδοτών και διαβάσεων πεζών έχει οριακά σημαντική επιρροή – συγκεκριμένα, είναι στατιστικά σημαντικά σε ένα από τα μοντέλα παλινδρόμησης και εμφανίζεται χαμηλά στην κατάταξη πληροφορίας στους αλγορίθμους XGBoost. Η καμπυλότητα και η κατεύθυνση της οδού δεν είναι στατιστικά σημαντικές μεταβλητές για την εξήγηση της συχνότητας των απότομων επιβραδύνσεων.

Τα αποτελέσματα έδειξαν ότι οι παράμετροι έκθεσης του μήκους τμήματος και του αριθμού περάσματος αυξάνουν επίσης τις συχνότητες των απότομων επιταχύνσεων. Η καμπυλότητα του οδικού τμήματος και η παρουσία σηματοδοτών συσχετίζονται θετικά με απότομες επιταχύνσεις. Ο τύπος οδού και ο αριθμός λωρίδων έχουν μη μονοσήμαντη επίδραση, αν και οι δευτερεύουσες και τριτεύουσες οδοί είναι σταθερά συσχετισμένες με αυξημένες απότομες επιταχύνσεις σε σύγκριση με τις κεντρικές οδούς. Η παρουσία διαβάσεων πεζών έχει οριακά σημαντική επιρροή, ενώ η κατεύθυνση της οδού δεν ήταν στατιστικά σημαντική μεταβλητή με τη συχνότητα απότομων επιταχύνσεων.

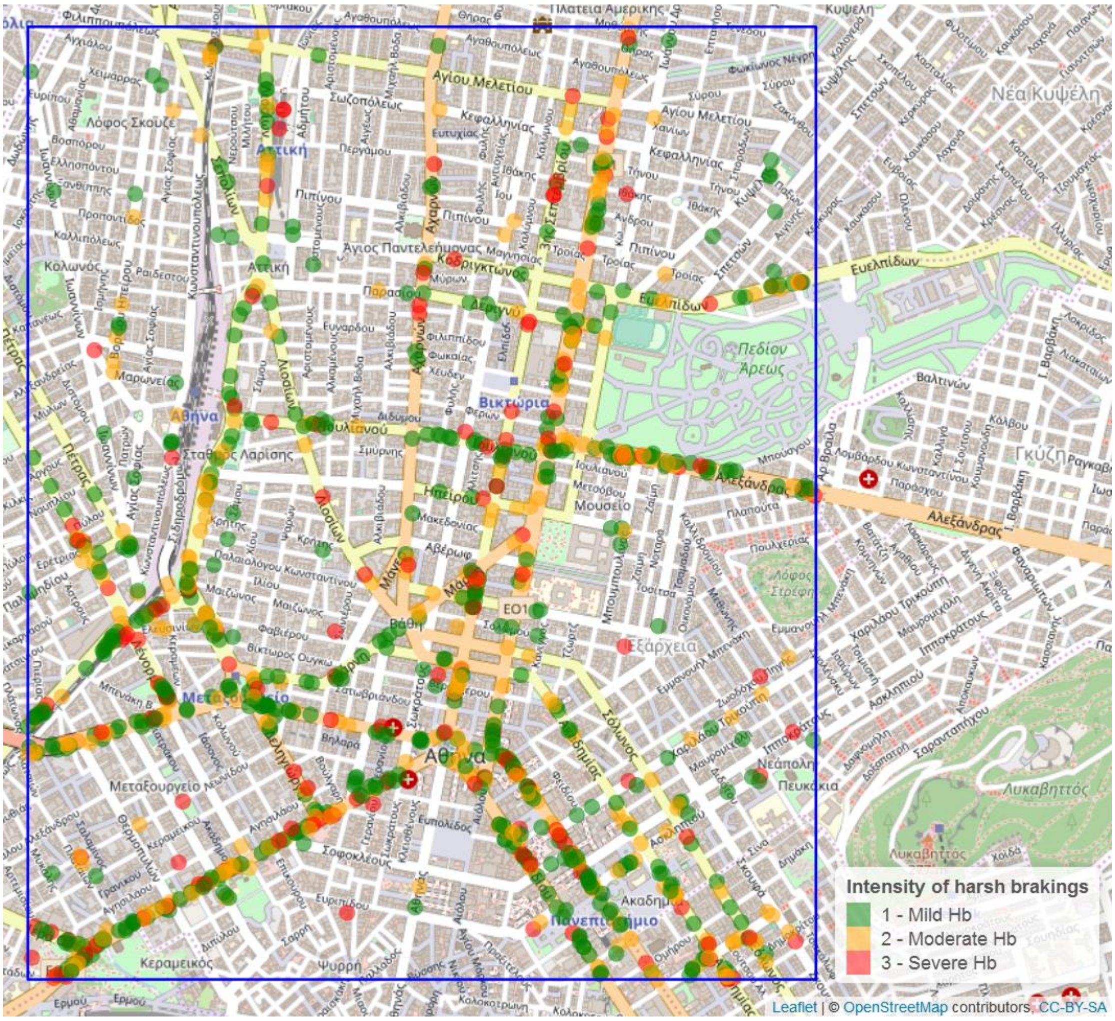
Τα μοντέλα GWPR και CAR παρέχουν περισσότερες πληροφορίες ποσοτικοποιώντας τις ακριβείς στατιστικές επιδράσεις των ανεξάρτητων μεταβλητών μέσω των ευρέως χρησιμοποιούμενων συντελεστών συσχέτισης και των διαστημάτων εμπιστοσύνης/ αξιοπιστίας. Υπό αυτή την έννοια, οι αλγόριθμοι XGBoost μπορούν να χρησιμοποιηθούν μόνο για την επαλήθευση αυτών των επιδράσεων μέσω της μέτρησης και κατάταξης της απόκτησης πληροφοριών από κάθε μεταβλητή. Παρόλα αυτά, τα μοντέλα GWPR και CAR παρουσιάζουν προβλήματα μεταφοράς σε άλλες περιοχές. Ωστόσο, τα αντίστοιχα GLM μπορούν να χρησιμοποιηθούν για πρόβλεψη απότομων συμβάντων σε περιοχές ελέγχου εκτός της περιοχής δοκιμών.

Αντιθέτως, οι αλγόριθμοι XGBoost μπορούν να μεταφερθούν απρόσκοπτα σε νέες περιοχές. Αυτό οφείλεται στο γεγονός ότι το XGBoost δεν ενσωματώνει ρητά τοπικές-χωρικές παραμέτρους, αλλά εξάγει συμπεράσματα βάσει των δεδομένων. Το SPCV XGBoost παρέχει βελτιωμένες προβλέψεις σε σύγκριση με το RCV XGBoost διότι επιτρέπει τις χωρικές ομαδοποιήσεις δεδομένων στα δένδρα αποφάσεων τόσο για απότομες επιβραδύνσεις όσο και για απότομες επιταχύνσεις. Η απόδοση των αλγορίθμων XGBoost δείχνει ότι οι μέθοδοι μηχανικής μάθησης είναι συγκρίσιμες με τις παραδοσιακές μεθόδους και όχι πανάκεια – παρόλο που το τελικό αρχείου χωρικών δεδομένων δεν είχε το τυπικό πλήθος σειρών που χρησιμοποιείται σε προβλήματα μηχανικής μάθησης.

Τα μοντέλα CAR εφαρμόζουν σε μεγάλο βαθμό στα δεδομένα μιας συγκεκριμένης περιοχής μελέτης σε πολύ υψηλό βαθμό για τις συχνότητες απότομων συμβάντων (CA>95%) χάρη στο συνδυασμό χωρικά δομημένων και μη δομημένων τοπικών παραμέτρων, και της Μπευζιανής ανανέωσης του μοντέλου με βάση νεότερες πληροφορίες (Bayesian inference). Κατά μια έννοια, οι τοπικές παράμετροι οδηγούν σε πλήρη ταύτιση μοντέλου και δεδομένων, παρόλα αυτά οι προβλέψεις πραγματοποιούνται εν τέλει χωρίς αυτές.

Τόσο για απότομες επιβραδύνσεις όσο και για απότομες επιταχύνσεις, οι βέλτιστες προβλέψεις αποκτήθηκαν με τη μεσοστάθμιση (μέσο όρο) των προβλέψεων από το σύνολο των τεσσάρων τύπων μοντέλων. Επιτεύχθηκε ακρίβεια CA 87,55% για απότομες επιβραδύνσεις και 89% για απότομες επιταχύνσεις. Σε σύγκριση με τα καλύτερα μεμονωμένα μοντέλα, υπάρχει αύξηση ακρίβειας άνω του 2% όταν οι προβλέψεις μεσοσταθμίζονται. Η μεσοστάθμιση των μοντέλων άμβλυνε τις μεμονωμένες αδυναμίες και τα ακρότατα στατιστικά σημεία και οδήγησε σε ένα ισορροπημένο προγνωστικό εργαλείο για απότομες επιβραδύνσεις και για απότομες επιταχύνσεις, με πολλά υποσχόμενες δυνατότητες μεταφοράς σε άλλες περιοχές.

Εκτός από τα πολυάριθμα στατιστικά αποτελέσματα, ένας μεγάλος αριθμός χαρτών σημειακής απεικόνισης και χαρτών θερμότητας (heatmaps) έχουν παραχθεί στην παρούσα διατριβή, τόσο από τα αρχικά δεδομένα όσο και από στατιστικά αποτελέσματα. Ενδεικτικά, το Σχήμα II απεικονίζει τις καταγεγραμμένες απότομες επιβραδύνσεις στα οδικά τμήματα της περιοχής ελέγχου και το Σχήμα III απεικονίζει τις αντίστοιχες συνδυασμένες προβλέψεις για αυτά τα τμήματα (CA 87,55%):

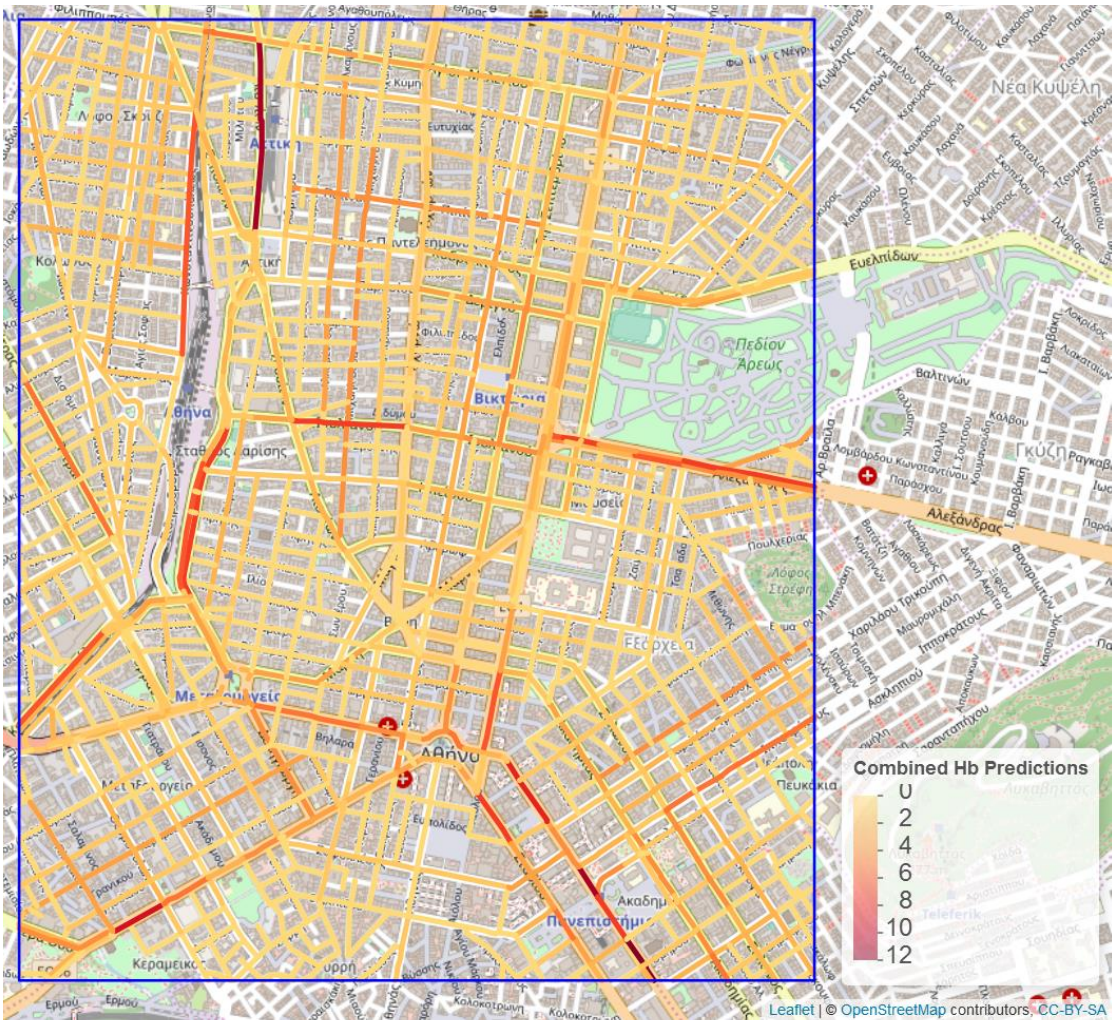

Intensity of harsh brakings
🟩 1 - Mild Hb
🟨 2 - Moderate Hb
🟥 3 - Severe Hb
Leaflet | © OpenStreetMap contributors, CC-BY-SA

Σχήμα ΙΙ: Απότομες επιβραδύνσεις στην περιοχή της Ομόνοιας

Μεμονωμένα, τα μοντέλα με τις καλύτερες αποδόσεις σχετικά με τις προγνωστικές δυνατότητές τους είναι διαφορετικά για τις απότομες επιβραδύνσεις και για τις απότομες επιταχύνσεις, όπως και η ίδια η απόδοση τους. Συγκεκριμένα, εάν ληφθεί υπόψη η ακρίβεια CA: ο αλγόριθμος SPCV XGBoost παρουσίασε την καλύτερη απόδοση για τις απότομες επιβραδύνσεις (CA>85%), ενώ οι κλασικές συναρτησιακές μέθοδοι και οι Μπευζιανές μέθοδοι ισοβάθμησαν με το SPCV XGBoost για τις απότομες επιταχύνσεις (CA>87%).

Οι δείκτες RMSE, RMSLE και MAE είναι μαθηματικά κατάλληλοι δείκτες σφαλμάτων για την μοντελοποίηση συχνοτήτων απότομων συμβάντων. Δεδομένου ότι οι διακυμάνσεις τους διαφέρουν ανάλογα με την ύπαρξη και την κατανομή στατιστικά ακραίων τιμών, συνιστώνται και οι τρεις κατά τη σύγκριση της απόδοσης των μοντέλων. Η επινόηση της προσαρμοσμένης ακρίβειας για δεδομένα συχνοτήτων αυξάνει την ικανότητα αξιολόγησης κάθε μοντέλου παρέχοντας ένα απλό και εύκολα κατανοητό ποσοστό ακρίβειας.

Σε τμήματα αστικών λεωφόρων, υπάρχει χωρική αυτοσυσχέτιση με βάση τους ολικούς και τοπικούς συντελεστές *I* του Moran, εάν λαμβάνονται υπόψη μόνο τα χωρικά συσχετισμένα οδικά τμήματα, όπως προτείνεται στη βιβλιογραφία. Μεγάλη χωρική αυτοσυσχέτιση εντοπίζεται στο μέσον (σε κεντρικά τμήματα) της λεωφόρου υπό εξέταση. Αυτά τα αποτελέσματα είναι σύμφωνα με τα ευρήματα για τα αστικά οδικά δίκτυα.



Σχήμα ΙΙΙ: Χάρτης συνδυασμένων προβλέψεων στην περιοχή της Ομόνοιας

Από τα συγχωνευμένα βαριογραφήματα προκύπτει ότι η μέση χωρική αυτοσυσχέτιση βρίσκεται σε απόσταση 310 m για τις απότομες επιβραδύνσεις και εντός 320 m για τις απότομες επιταχύνσεις. Πέρα από αυτή την απόσταση, η χωρική αυτοσυσχέτιση εξομαλύνεται. Τα βαριογραφήματα για οδικά τμήματα αστικών λεωφόρων φαίνεται να είναι πιο ασταθή σε σύγκριση με εκείνα των αστικών οδικών δικτύων. Επιπλέον, παρατηρείται χωρική κυκλικότητα στον άξονα τόσο για απότομες επιβραδύνσεις όσο και για απότομες επιταχύνσεις. Με άλλα λόγια, παρατηρείται κάποια επανάληψη στα μοτίβα (patterns) των συχνοτήτων απότομων συμβάντων.

Υπό συνθήκες ελεύθερης ροής, τα αποτελέσματα έδειξαν ότι οι παράμετροι έκθεσης κινδύνου του μήκους οδικού τμήματος και του αριθμού διελεύσεων, καθώς και τα μέσα δευτερόλεπτα χρήσης κινητού τηλεφώνου από τους οδηγούς είναι παράμετροι θετικά συσχετισμένες με τις συχνότητες απότομων επιβραδύνσεων. Όσον αφορά τις παραμέτρους κυκλοφορίας, η διαφορά ταχύτητας μεταξύ κυκλοφορίας και οδηγού συσχετίζεται θετικά με τις συχνότητες απότομων επιβραδύνσεων, ενώ η επιρροή του μέσου όρου τρέχοντος όγκου κυκλοφορίας προέκυψε ότι είναι αρνητική. Τα τμήματα της περιοχής μελέτης με κατεύθυνση προς το Νότο βρέθηκε ότι εμφανίζουν συστηματικά λιγότερες απότομες επιβραδύνσεις σε σύγκριση με αυτά που κατευθύνονται προς Βορρά. Τέλος, διαπιστώθηκε ότι η μέση κατάληψη ασκεί μια περιστασιακά θετική επιρροή και ότι η κλίση ασκεί μια περιστασιακά αρνητική επιρροή στις συχνότητες απότομων επιβραδύνσεων ανά οδικό τμήμα, ανάλογα με τη χρησιμοποιούμενη μέθοδο.

Αντίστοιχα, καθορίστηκε ότι οι παράμετροι έκθεσης κινδύνου του μήκους οδικού τμήματος και του αριθμού διελεύσεων, καθώς και τα μέσα δευτερόλεπτα χρήσης κινητού τηλεφώνου από τους οδηγούς διατηρούν τη θετική επιρροή τους στις συχνότητες απότομων επιβραδύνσεων υπό συνθήκες συγχρονισμένης ροής. Όσον αφορά τις παραμέτρους κυκλοφορίας, η μέση κατάληψη φαίνεται ότι αναλαμβάνει ενισχυμένη επιρροή με στατιστικά σημαντική θετική συσχέτιση με την εξεταζόμενη συχνότητα. Η επιρροή του κυκλοφοριακού φόρτου (ωριαίου ή τρέχοντος) εμφανίζεται ως αρνητική κατά περίπτωση. Οι επιρροές της καμπυλότητας, της κλίσης, του αριθμού λωρίδων και της κατεύθυνσης των οδικών τμημάτων εξασθενούν και είναι πολύ περιστασιακές, αναλόγως με τη χρησιμοποιούμενη μέθοδο.

Υπό συνθήκες ελεύθερης ροής, τα αποτελέσματα έδειξαν ότι οι παράμετροι έκθεσης κινδύνου του μήκους οδικού τμήματος και του αριθμού διελεύσεων, καθώς και τα μέσα δευτερόλεπτα χρήσης κινητού τηλεφώνου από τους οδηγούς (με μια εξαίρεση) είναι παράμετροι θετικά συσχετισμένες με τις συχνότητες απότομων επιταχύνσεων. Αναφορικά με τις παραμέτρους κυκλοφορίας, η επιρροή της μέσης κατάληψης εμφανίζεται ως σταθερά θετική, ενώ τα μέσα δευτερόλεπτα υπέρβασης του ορίου ταχύτητας από τους οδηγούς προέκυψε ότι έχουν οριακά θετική συσχέτιση. Η μέση ταχύτητα κίνησης εμφανίζει μια περιστασιακά αρνητική επιρροή, αναλόγως με τη χρησιμοποιούμενη μέθοδο. Οι μεταβλητές γεωμετρίας και χαρακτηριστικών οδικού δικτύου βρέθηκαν ότι είναι πολύ περιστασιακά συσχετισμένες με τη συχνότητα απότομων επιταχύνσεων.

Αντίστοιχα, διαπιστώθηκε ότι ο αριθμός διελεύσεων και τα μέσα δευτερόλεπτα χρήσης κινητού τηλεφώνου από τους οδηγούς διατηρούν τη θετική επιρροή τους στις συχνότητες απότομων επιταχύνσεων υπό συνθήκες συγχρονισμένης ροής. Για πρώτη φορά σε όλες τις χωρικές αναλύσεις, το μήκος οδικού τμήματος δεν εμφανίζεται να επηρεάζει τις συχνότητες απότομων επιταχύνσεων ανά οδικό τμήμα. Ο κυκλοφοριακός φόρτος (ωριαίος ή τρέχων) βρέθηκε θετικά συσχετισμένος με τις απότομες επιταχύνσεις. Αντιθέτως, ο αριθμός λωρίδων προέκυψε αρνητικά συσχετισμένος με τις απότομες επιταχύνσεις μόνο στα μοντέλα CAR.

Για ακόμα μια φορά, διαπιστώθηκε ότι και οι τρεις μέθοδοι, GWPR, CAR και XGBoost - με τυχαία ή χωρική εγκάρσια επικύρωση – είναι κατάλληλες και δόκιμες μέθοδοι για την ανάλυση των συχνοτήτων απότομων επιβραδύνσεων και επιταχύσεων σε ένα μέσα σε ένα λογαριθμικό-Poisson πλαίσιο. Παρόλα αυτά, η διεξαγωγή προβλέψεων με τη βάση δεδομένων τμημάτων αστικών λεωφόρων δεν είναι τόσο δόκιμη όσο με την αντίστοιχη βάση αστικών οδικών δικτύων. Ο λόγος είναι η συμπερίληψη μεταβλητών κυκλοφορίας και συμπεριφοράς οδηγών οι οποίες δεν είναι γνωστές και διαθέσιμες εκ των προτέρων σε κάθε τοποθεσία. Αντιθέτως, θα απαιτούνταν επιπλέον μοντέλα εκτιμήσεων-προβλέψεων για αυτές τις μεταβλητές.

Μια αξιοσημείωτη γενική παρατήρηση είναι ότι η συμπερίληψη μεταβλητών κυκλοφορίας και συμπεριφοράς οδηγών στα μοντέλα αποδυναμώνει τις συσχετίσεις που προκύπτουν με τις μεταβλητές γεωμετρίας και χαρακτηριστικών οδικού δικτύου, αντικαθιστώντας τις κατά κάποιον τρόπο. Επιπλέον, επιβεβαιώνεται ότι οι απότομες επιβραδύνσεις και οι απότομες επιταχύνσεις είναι δύο διαφορετικά φαινόμενα οδικής ασφάλειας. Οι συχνότητές τους συσχετίζονται με ορισμένες κοινές μεταβλητές, αν και με διαφορετικούς συντελεστές, αλλά επίσης και με ορισμένες εντελώς διαφορετικές μεταβλητές.

Η γραμμικότητα της λεωφόρου Κηφισίας οδήγησε σε μια πιο ομοιογενή περιοχή μελέτης, με λιγότερη αβεβαιότητα κατά τον υπολογισμό των δεδομένων κυκλοφορίας ανά δευτερόλεπτο διαδρομής και τη σύνθεση του αντίστοιχου αρχείου χωρικών δεδομένων. Ταυτόχρονα, είναι πιθανό ότι αυτή η γραμμικότητα προκαλεί επίσης απώλεια πληροφοριών ή διαφορετική απόδοση ορισμένων μοντέλων.

Συγκεκριμένα, δεν ήταν δυνατή η δημιουργία βαριογραφημάτων βάσει κατεύθυνσης και τα μοντέλα GWPR προσαρμόστηκαν στα δεδομένα με μειωμένη ακρίβεια.

Τα μοντέλα Bayesian CAR και XGBoost δεν φάνηκαν να επηρεάζονται με τον ίδιο τρόπο από τη γραμμικότητα της περιοχής μελέτης αστικής λεωφόρου. Στις περισσότερες περιπτώσεις, το XGBoost ταιριάζει καλύτερα στο σύνολο δεδομένων, αντλώντας πληροφορίες από μεγαλύτερο αριθμό ανεξάρτητων μεταβλητών, και ιδιαίτερα από μεταβλητές γεωμετρίας και χαρακτηριστικών οδικού δικτύου. Ο ρυθμός εκμάθησης (learning rate – ETA) εμφανίστηκε ως η πιο σημαντική υπερπαραμέτρος κατά τη φάση επιλογής υπερπαραμέτρων του XGBoost. Για το SPCV XGBoost, η υπερπαράμετρος γάμμα (gamma) – η οποία διέπει την ελάχιστη μείωση της συνάρτησης απώλειας που μπορεί να δικαιολογήσει τον διαχωρισμό σε ένα δένδρο – βρέθηκε να επηρεάζει επίσης την απόδοση του αλγορίθμου.

Συνοψίζοντας, η παρούσα διδακτορική διατριβή προσφέρει σημαντικές καινοτομίες στον επιστημονικό τομέα της οδικής ασφάλειας και της ανάλυσης συμπεριφοράς της κυκλοφορίας:

1. Ένα νέο μεθοδολογικό ερευνητικό πλαίσιο σχεδιάστηκε και εφαρμόστηκε με σκοπό τη διεξαγωγή χωρικών αναλύσεων οδικής ασφάλειας των συχνοτήτων απότομων οδηγικών συμβάντων με χρήση πολυπαραμετρικών δεδομένων υψηλής ανάλυσης ανά οδικό τμήμα, παρέχοντας πολύ λεπτομερείς γνώσεις για την διαδικασία εντοπισμού επικίνδυνων θέσεων.
2. Για την υλοποίηση και ενίσχυση του ερευνητικού πλαισίου, επινοήθηκαν και εφαρμόστηκαν αλγόριθμοι δεδομένων μεγάλης κλίμακας, με σκοπό την εκτέλεση κρίσιμων λειτουργιών απαραίτητων για τις χωρικές αναλύσεις, όπως ο υπολογισμός πρόσθετων γεωμετρικών χαρακτηριστικών, η επεξεργασία και συνένωση δεδομένων και η αντιστοίχιση χαρτών (map-matching).
3. Η μεθοδολογία εφαρμόστηκε σε καινοτόμους τύπους χωρικών αναλύσεων για αστικά οδικά δίκτυα: Για πρώτη φορά (i) χωρικές αναλύσεις απότομων συμβάντων πραγματοποιήθηκαν σε επίπεδο οδικού τμήματος και (ii) χρησιμοποιήθηκαν αποτελέσματα χωρικών αναλύσεων για την επιτυχή πρόβλεψη συχνοτήτων απότομων συμβάντων σε διαφορετική περιοχή ελέγχου αστικού δικτύου.
4. Πραγματοποιήθηκε επίσης μια σειρά αναλύσεων με επιπλέον χαρακτηριστικά σε οδικά τμήματα αστικών λεωφόρων, οι οποίες αναλύθηκαν χωρικά για τις καταστάσεις κυκλοφορίας ελεύθερης ροής και συγχρονισμένης ροής.
5. Από τις λεπτομερείς μικροσκοπικές έρευνες της παρούσας διατριβής, προέκυψαν πρωτότυπες πληροφορίες και στατιστικές συσχετίσεις μεταξύ των συχνοτήτων απότομων επιβραδύνσεων και επιταχύνσεων ανά οδικό τμήμα, και μεταβλητών γεωμετρίας, χαρακτηριστικών οδικού δικτύου, κυκλοφορίας και συμπεριφοράς οδηγών.

Η διαθεσιμότητα πολυπαραμετρικών δεδομένων υψηλής ανάλυσης – και η σχετική αφθονία των απότομων οδηγικών συμβάντων σε σύγκριση με τα οδικά ατυχήματα – χρησίμευσε ως έναυσμα προς την εξερεύνηση της διεξαγωγής χωρικών αναλύσεων απότομων συμβάντων στο πιο λεπτομερές, μικροσκοπικό επίπεδο οδικού τμήματος, σε αντίθεση με την πιο παραδοσιακή μακροσκοπική ανάλυση περιοχών (για παράδειγμα σε επίπεδο περιφερειών, νομών ή δήμων). Η χωρική διερεύνηση των απότομων συμβάντων εν γένει, αλλά και συγκεκριμένα ανά οδικά τμήματα, αποτελεί έναν πλήρως ανεξερεύνητο ερευνητικό χώρο.

Από επιστημονικής άποψης, ένα πρόσθετο πλεονέκτημα της υιοθετούμενης προσέγγισης είναι η παράκαμψη του οριακού προβλήματος (border problem) και του προβλήματος τροποποιημένων επιφανειών (modifiable areal unit problem – MAUP). Τα συγκεκριμένα προβλήματα είναι πάντα παρόντα στις χωρικές αναλύσεις, και η παρουσία του MAUP επιβεβαιώθηκε από τη μετα-παλινδρόμηση της μεταβλητής της διανυθείσας απόστασης στο ποσοτικό μέρος της βιβλιογραφικής επισκόπησης της διατριβής. Με την διακριτοποίηση και κατάτμηση των περιοχών έρευνας οδικής ασφάλειας σε οδικά τμήματα κάθε φορά, δημιουργείται ένας μονοσήμαντος τρόπος αντιμετώπισης των συμβάντων που πραγματοποιούνται στα όρια μιας περιοχής μελέτης, μόλις καθοριστεί ότι το αντίστοιχο τμήμα ανήκει (ή δεν ανήκει) στην περιοχή. Επιπλέον, τυποποιείται η διαδικασία επιλογής μονάδων για ανάλυση μέσω των οδικών τμημάτων, αφαιρώντας τις αβεβαιότητες από το MAUP σε μελλοντικές αναλύσεις.

Η επινόηση και δημιουργία των διαφόρων αλγορίθμων που έχουν εφαρμοστεί σε αυτή τη διδακτορική διατριβή επιδέχονται σχολιασμού. Οι αλγόριθμοι επινοήθηκαν και εφαρμόστηκαν σε ενδιάμεσα στάδια, εκτελώντας κρίσιμες λειτουργίες, όπως ο υπολογισμός πρόσθετων γεωμετρικών χαρακτηριστικών, η συγχώνευση δεδομένων και η αντιστοίχιση χαρτών. Ως εκ τούτου, παρείχαν το μέσον για την υλοποίηση και εφαρμογή του καινοτόμου μεθοδολογικού πλαισίου και προετοίμασαν τα χωρικά πλαίσια δεδομένων ανά οδικό τμήμα για τις αναλύσεις που ακολούθησαν. Οι αλγόριθμοι επιτρέπουν την απρόσκοπτη μεταφορά αυτούσιου του πλαισίου μεθοδολογίας και επεξεργασίας δεδομένων που εφαρμόστηκε στην παρούσα διδακτορική διατριβή σε άλλες περιπτώσεις.

Συγκεκριμένα, ο αλγόριθμος για τον υπολογισμό πρόσθετων γεωμετρικών χαρακτηριστικών αντλεί πληροφορίες από τους ψηφιακούς κόμβους που ορίζουν οδικά τμήματα (ή ways στο OpenStreetMap). Από τις συντεταγμένες κόμβων, υπολογίζεται το μήκος τμήματος, η κλίση, η καμπυλότητα και η πολυπλοκότητα της γειτονιάς. Η επαναληπτική φύση του αλγορίθμου διασφαλίζει τη λειτουργικότητά του σε όλα τα τμήματα ανεξάρτητα από τον συνολικό αριθμό κόμβου, τον τύπο του δρόμου ή την τοποθεσία του τμήματος.

Στη συνέχεια, ένας αλγόριθμος αντιστοίχισης χαρτών (map-matching) επινοήθηκε ώστε να ταιριάξει τα δεδομένα οδήγησης υπό πραγματικές συνθήκες με τα οδικά τμήματα των περιοχών μελέτης. Για αυτό το σκοπό, για κάθε δευτερόλεπτο διαδρομής, προσδιορίστηκε το πλησιέστερο οδικό τμήμα (Minimum-Distance-Way – MDW), χρησιμοποιώντας έναν σύνθετο υπολογισμό σε δύο στάδια: ελάχιστη απόσταση από σημείο προς σημείο και από σημείο προς καμπύλη. Επιπλέον, ο αλγόριθμος περιελάμβανε προσεγγίσεις κινούμενων παραθύρων που μείωναν τις διαστάσεις για τους πίνακες σύγκρισης αποστάσεων, ελαττώνοντας έτσι τους υπολογιστικούς χρόνους. Αυτή η προσέγγιση επέτρεψε την πρακτική εφαρμογή της διαδικασίας αντιστοίχισης χαρτών με άμεσο έλεγχο των αποτελεσμάτων, χωρίς να χρειάζεται η εξάρτηση από τρίτες υπηρεσίες, οι οποίες είναι άγνωστες διαδικασίες «μαύρου κουτιού» και συχνά χρεώνουν κόστη επεξεργασίας.

Ως απαραίτητη συμπληρωματική υπορουτίνα στον αλγόριθμο αντιστοίχισης χαρτών, επινοήθηκε ένας αλγόριθμος διόρθωσης αριθμού διελεύσεων με βάση την καταμέτρηση ψήφων. Η συγκεκριμένη υπορουτίνα ουσιαστικά εξυπηρετούσε την ελάττωση των αβεβαιοτήτων των αισθητήρων GPS, μέσω ενός προηγμένου αλγορίθμου καταμέτρησης ψήφων που ανέθετε κάθε διέλευση στο οδικό τμήμα το οποίο θα λάμβανε την πλειοψηφία των περιπτώσεων αντιστοίχησης χαρτών. Η χρήση της υπορουτίνας αποδείχθηκε κρίσιμη σε τοποθεσίες πολλών παράλληλων αξόνων οδικών τμημάτων με υψηλή πυκνότητα, όπως η Λεωφόρος Κηφισίας και οι βοηθητικές παράλληλες οδοί της, ενώ αύξησε την συνολική ακρίβεια της διαδικασίας.

Η εφαρμογή ενός επιπλέον προσαρμοσμένου αλγορίθμου ήταν απαραίτητη για τις αναλύσεις τμημάτων αστικών λεωφόρων, προκειμένου να ενισχυθούν τα δεδομένα οδήγησης υπό πραγματικές συνθήκες με τα δεδομένα κυκλοφορίας πριν από την αντιστοίχιση χαρτών. Αυτός ο αλγόριθμος περιελάμβανε το διαχωρισμό τμημάτων και θέσεων μέτρησης ανά κατεύθυνση (προς Βορρά ή Νότο) και τον προσδιορισμό της μέτρησης με την ελάχιστη χωροχρονική απόσταση κάθε δευτερολέπτου διαδρομής μεταξύ των δύο πολύ μεγάλων βάσεων δεδομένων οδήγησης και κυκλοφορίας.

Η σημασία της εξέτασης της χωρικής αυτοσυσχέτισης απότομων συμβάντων (μέσω των ολικών και τοπικών δεικτών *I* του Moran) αποκλειστικά με βάση τα χωρικά συσχετισμένα τμήματα επιβεβαίωσε τόσο τις προτεινόμενες καλές πρακτικές χωρικών αναλύσεων όσο και τις πρακτικές οδικής ασφάλειας που εφαρμόζονται κατά την ανάλυση των ατυχημάτων. Υπολογίστηκαν για πρώτη φορά αποστάσεις επιρροής της χωρικής αυτοσυσχέτισης των απότομων επιβραδύνσεων και επιταχύνσεων με χρήση βαριογραφημάτων, τα οποία επίσης καθόρισαν ότι αυτές οι αποστάσεις διαφέρουν ανά τύπο οδού.

Επιπλέον, ο πλούτος των πολυπαραμετρικών δεδομένων υψηλής ανάλυσης και η εγκυρότητα των φάσεων επεξεργασίας και συγχώνευσης δεδομένων επέτρεψαν την εκτέλεση καινοτόμων χωρικών αναλύσεων. Είναι η πρώτη φορά που αναλύονται απότομα οδηγικά συμβάντα σε επίπεδο οδικού τμήματος για αστικά οδικά δίκτυα. Η παρούσα διατριβή κατάφερε να ξεπεράσει τα τυπικά ζητήματα της έλλειψης δεδομένων για αστικά οδικά δίκτυα, τα οποία είναι σε μεγάλο βαθμό υπομελετημένες περιοχές οδικής ασφάλειας.

Είναι εξίσου σημαντικό ότι για πρώτη φορά, εξ όσων γνωρίζει ο συγγραφέας, χρησιμοποιούνται αρχεία χωρικών δεδομένων και χωρικές προσεγγίσεις για τη διεξαγωγή προβλέψεων οδικής ασφάλειας σε διαφορετική περιοχή αστικού δικτύου, με την επίτευξη υψηλού ποσοστού επιτυχίας. Αυτό το εύρημα αποτελεί μια ισχυρή ένδειξη της μεγάλης δυνατότητας μεταφοράς αποτελεσμάτων πρόβλεψης σε παρόμοιες περιοχές. Εκτός των προηγουμένων, είναι η πρώτη φορά που οι αλγόριθμοι XGBoost χρησιμοποιούνται για χωρικές αναλύσεις στην οδική ασφάλεια. Το XGBoost αποδείχθηκε μια πολύ ισχυρή και πολλά υποσχόμενη μέθοδος ανάλυσης. Η διερεύνηση της τυχαίας επικύρωσης και της χωρικής επικύρωσης, η οποία είναι μια πολύ πρόσφατη ιδέα, παρέχει περαιτέρω εμβάθυνση στα αποτελέσματα του αλγορίθμου.

Επιπλέον, τα αποτελέσματα της ανάλυσης αστικών οδικών δικτύων επιβεβαιώνουν ότι υπάρχει μια ισορροπία στη χρηστικότητα μεταξύ κλασικών συναρτησιακών μεθόδων (GWPR), Μπευζιανών μεθόδων (CAR) και Μηχανικής Μάθησης (XGBoost). Αυτές οι μέθοδοι δημιούργησαν μοντέλα που ταιριάζουν διαφορετικά στα δεδομένα και προέβλεψαν μεγαλύτερες συχνότητες για διαφορετικά τμήματα κατά περίπτωση. Ωστόσο, ο συνδυασμός τους μέσω του μέσου όρου προβλέψεων απέδωσε ακριβέστερα αποτελέσματα σε σύγκριση με τα μεμονωμένα μοντέλα, καθώς οι ακραίες τιμές μειώθηκαν και οι σωστές προβλέψεις βελτιώθηκαν.

Αναφορικά με τα τμήματα αστικών λεωφόρων, διαπιστώθηκε ότι διαφορετικές μεταβλητές σχετίζονται σημαντικά με τις συχνότητες απότομων συμβάντων ανά κατάσταση κυκλοφορίας. Η παρούσα διατριβή είναι μία από τις ελάχιστες ερευνητικές προσπάθειες που συμπεριλαμβάνουν τις συνθήκες κυκλοφορίας κατά τη στιγμή του εξεταζόμενου φαινομένου και η μόνη για τα απότομα συμβάντα. Μεταβλητές όπως η διαφορά ταχύτητας μεταξύ κυκλοφορίας και μεμονωμένου οδηγού καθίστανται πολύ σημαντικές για την ερμηνεία των συχνοτήτων απότομων συμβάντων, ακόμη και αν τα δεδομένα συγκεντρώνονται ανά οδικό τμήμα. Συνολικά, ο περίπλοκος μη γραμμικός τρόπος με τον οποίο οι παράμετροι κυκλοφορίας επηρεάζουν τις συχνότητες απότομων συμβάντων αποκαλύφθηκε από την παρούσα έρευνα.

Ως συνολική παρατήρηση από τις πολυάριθμες αναλύσεις που πραγματοποιήθηκαν, οι περισσότερες μεταβλητές γεωμετρίας, χαρακτηριστικών οδικού δικτύου, κυκλοφορίας και συμπεριφοράς οδηγού βρέθηκαν ως στατιστικά σημαντικές τουλάχιστον μία φορά. Αυτά τα αποτελέσματα καταδεικνύουν τις εγγενείς διαφορές των φαινομένων απότομης επιβράδυνσης και απότομης επιτάχυνσης, καθώς οι αντίστοιχες συχνότητες συσχετίζονται με σταθερά διαφορετικές μεταβλητές. Επιπλέον, τα αποτελέσματα υποστηρίζουν ολιστικές προσεγγίσεις για την οδική ασφάλεια που περιλαμβάνουν πολυπαραμετρικά δεδομένα, σε μια προσπάθεια να αποτυπώνονται όσο το δυνατόν περισσότερες πλευρές του οδικού περιβάλλοντος και των χρηστών του σε στατιστικά μοντέλα.

Η δημιουργία ολοκληρωμένων χαρτών σημειακών δεδομένων και θερμικών χαρτών οδικής ασφάλειας με βάση τα απότομα οδηγικά συμβάντα προσφέρει ένα μοναδικό εργαλείο στις διαχειριστικές αρχές των οδών, στους χρήστες των οδών και στους λοιπούς ενδιαφερόμενους. Οι χάρτες απεικονίζουν πολύπλοκα δεδομένα και προβλέψεις προηγμένων μοντέλων με έναν εύκολο και κατανοητό τρόπο, ο οποίος μπορεί να μεταφερθεί και να ενσωματωθεί σε οποιοδήποτε οδικό εργασιακό περιβάλλον ή σε προσωπικές αποφάσεις των χρηστών. Μέσω των παραγόμενων χαρτών, η πολυεπίπεδη προσπάθεια αυτής της διατριβής ενσταλάσσεται και μεταλαμπαδεύεται από τον επιστημονικό στο δημόσιο τομέα.

Μία τελευταία εξειδικευμένη καινοτομία της παρούσας έρευνας είναι η εφεύρεση και εφαρμογή του πρότυπου δείκτη απόδοσης μοντέλων που είναι η προσαρμοσμένη ακρίβεια (CA). Η προσαρμοσμένη ακρίβεια προσφέρει έναν άμεσο τρόπο μέτρησης της ακρίβειας των προβλέψεων ο οποίος αντλεί τη λειτουργία του τόσο από δείκτες μοντέλων ταξινόμησης (όπως ο πίνακας ακρίβειας – confusion matrix) όσο και από δείκτες μοντέλων παλινδρόμησης (όπως το μέσο απόλυτο ποσοστό σφάλματος) ο οποίος είναι διαισθητικός και εύκολα κατανοητός.

1 Introduction

1.1 Road safety overview

1.1.1 Global and national road safety state

Ever since the industrial revolution and rapid spread of motor vehicles, road crashes have become an integral part of road transport systems. Every year, road crashes, also known as road accidents, continue to incur heavy costs to societies. The most critical part, human costs, include fatalities, permanent incapacitations, severe and slight injuries and similar physical or psychological trauma inflicted on the involved individuals and to individuals related to them. Moreover, considerable material costs are incurred in various forms, such as direct property damage, traffic flow disruptions and delays, hospitalization treatment and rehabilitation services, lost production values from recovery time, police and fire brigade costs, insurance costs, court costs and administrative costs.

Indicatively, 1,350,000 people are killed every year in road crashes worldwide, as shown in Figure 1-1. Perhaps the most alarming statistic to convey the magnitude of the problem is that road crashes are steadily the leading fatality cause for individuals aged 5 to 29 years old, namely children and young adults (WHO, 2018).

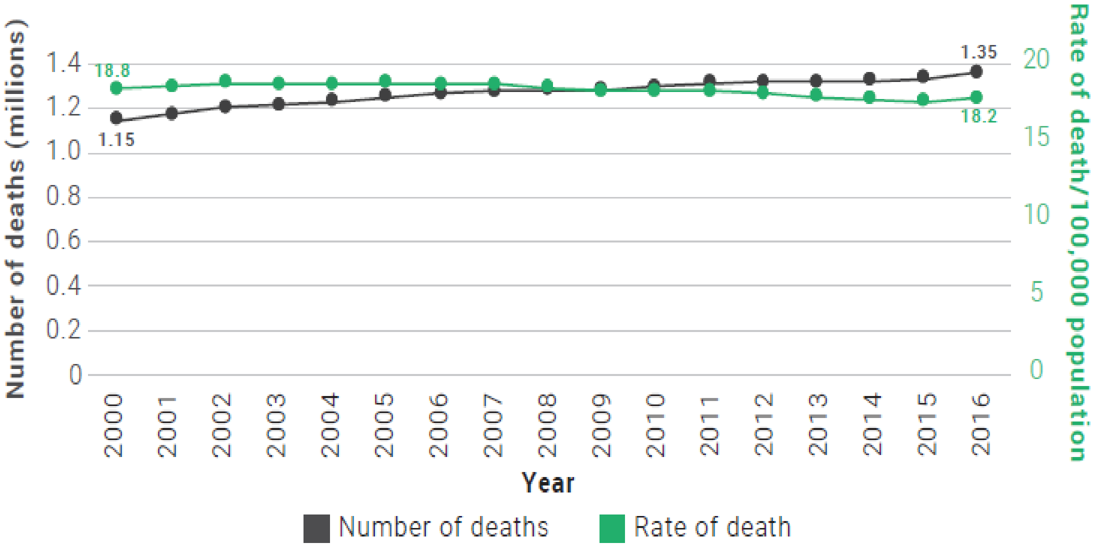


Figure 1-1: Number and rate of road traffic fatalities per 100,000 population during 2000–2016 [Source: WHO, 2018]

In Greece, the native country of the author and National Technical University of Athens, 14,002 people were involved in 10,848 road crashes during 2017; of those, 731 were fatalities. These numbers present a road crash reduction of 4.2% from the previous year 2016, as shown in Figure 1-2 (Hellenic Statistical Authority, 2019) – however, road crashes and casualties continue to be unacceptably high.

In order to mitigate the occurrence and consequences of road crashes, the science of road safety has been developed since decades, branching off of the more encompassing sciences of road design, civil engineering and transportation engineering. Road safety focuses on examining the three known pillars related to crash cause and prevention, namely (i) road infrastructure, (ii) vehicles and (iii) road users.

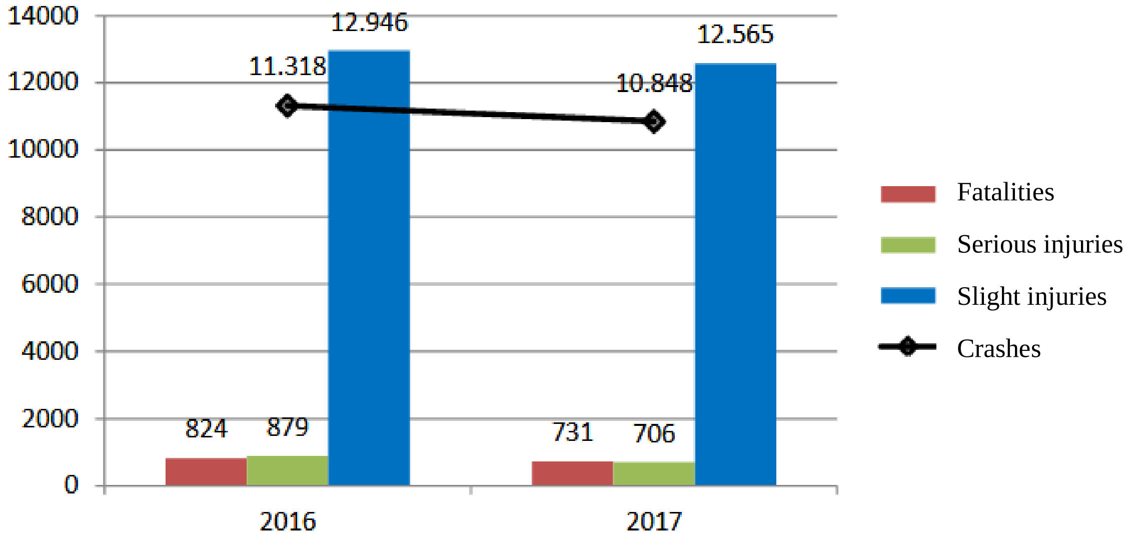


Figure 1-2: Numbers of road crashes and involved individuals in Greece during 2016–2017 [Source: Hellenic Statistical Authority, 2019]

The origins of road safety are based on civil engineering, and as such, a considerable number of road infrastructure measures have been investigated and applied by road safety experts during previous decades, such as road surface, superelevation, median barrier and lane treatments, traffic signal installations, and many more (e.g. Papadimitriou et al., 2019a). Naturally, the implementation of road safety measures has its own limitations. Low-income countries have approximately 3 times higher fatality rates compared to high-income countries (WHO, 2018), a testament to the fact that road safety improvements are not free but require dedicated investments. Furthermore, road safety measures have different costs and effectiveness, (Daniels et al., 2019), thus the allocation of limited administrative funds requires decisions that are informed from scientific knowledge to achieve the maximum possible benefits.

Simultaneously, considerable progress was made in mechanical engineering and vehicle design. A number of technological advancements and innovative systems have been integrated in vehicles, such as more resilient crash designs, airbags, rollover protection systems, electronic stability control, autonomous emergency braking, emergency brake assist, anti-lock braking system and many more (e.g. Winner et al, 2016).

The final pillar, road users, has been notoriously difficult to improve. This pillar encompasses what is known as the human factor, which is related to errors, poor judgement, lack of knowledge or experience and overall unpredictable behavior from drivers and other road users. In relevant studies conducted by the U.S. National Highway Traffic Safety Administration (NHTSA), it was found that the critical reason of crash occurrence lied with the drivers in 94% of cases (±2.2%) in a sample of more than 2,000,000 drivers (Singh, 2015; 2018), and human factor causes in that range are commonly accepted by road safety researchers as the norm.

Revisiting Figure 1-1, it can be seen that road safety progress, as well as additional, unobserved factors, have caused the number of road fatalities to plateau during recent years – if the rate of population growth is accounted for. The goal of halving road fatalities and injuries from crashes globally by 2020, which was set by the General Assembly of the United Nations, will unfortunately not be realized, as insufficient progress has been made (United Nations, 2015).

With a magnitude that is accepted as approximately 95%, it can be surmised that the human factor is still not adequately addressed in road safety, and continues to manifest as critical reason in crash occurrence. While drivers and other road users have been under heavy scrutiny and the target of many studies, such as the famous driver distraction studies (e.g. Young et al., 2007), the means to translate the acquired knowledge to crash reduction seem to be yet undiscovered. As overall "blanket" approaches have been applied across countries and large road networks, it becomes apparent that more focus is required in determining precise problematic spots warranting intervention pertinent to human factors, such as driver behavior.

1.1.2 Harsh events of driving behavior

One interesting and underused metric that can be used to investigate driver behavior is harsh events of drivers. The term harsh event refers to instances of any rapid and abrupt acceleration and deceleration of a vehicle by its driver. Harsh events are usually detected when acceleration sensors exceed certain thresholds that are predetermined from researchers, or more dynamically via machine learning methods, and can be treated as point-type data, similar to crashes. Harsh events are ultimately driver behavior metrics, but there is great potential in their analysis: they are much more frequent than crashes, thus providing richer data for many driver environments, they are a proactive road safety measure proxy, meaning that research can be conducted in a naturalistic setting without any crashes occurring, and they appear to be adequately representative of crash occurrence probability (Tselentis et al., 2017).

Harsh events have been adopted as a parameter for measurement of road safety in the past, as they are strongly correlated with reduced spatial and temporal headways (unsafe distance) from neighboring vehicles, near misses with road users or stationary objects, and also include additional behavioral parameters such as lack of concentration or experience. Harsh events have been determined as closely linked with driving risk (Tselentis et al., 2017), while research has also documented harsh driving behavior as critical for driving risk assessment (Bonsall et al., 2005; Gündüz et al., 2018). Harsh accelerations and decelerations, and their correlations with crash risk, have been investigated by the insurance industry as well (Paefgen et al., 2014).

However, to the experience of the author, studies focusing on factors influencing harsh event occurrence and similar characteristics are very scarce, and significantly outnumbered by studies analyzing crashes, indicating significant research gaps in this field. The opportunities that harsh event analysis offers are considerable, provided that a proper data collection scheme has been set up, such as data collection via smartphone sensors.

1.1.3 Spatial analysis in road safety

A standing problem in road safety, which is explored both in research and in practice, is the determination of problematic spots regarding road safety, also known as hotspots or blackspots. Obviously, not all network locations operate on the same road safety levels. In any given road network, hotspots are locations where road safety events, typically crashes, are much more frequent and severe. Hotspot detection may come from observation, in any case, it must be verified through robust statistical analysis to rule out any possible randomness of events.

The reader is invited to consider their personal experience as a road user, when either driving or otherwise observing road environments. Specifically, one can picture the following scenes:

- A motorcycle rushes to cross a signalized intersection before orange turns to red
- A car brakes abruptly after detecting an inadequately visible stop sign
- A truck accelerates harshly in a section of road before a steep slope
- A van slows down noticeably to merge with traffic at tight curve

One or more of these paradigms may then appear familiar, and well within ordinary observation. It stands to reason to surmise that the areas described above are candidate harsh event hotspots, and, by proxy, candidate hotspots with increased crash probability. The hotspots may be of any reasonable unit, such as a location (a small surface of 20 m radius), a road segment or a slightly larger area (e.g. a fraction of a municipality).

In practice, frequent road users may know and anticipate certain problematic spots on the network, whether they are drivers or pedestrians. For instance, if the van driver of the last example regularly passes from the particular tight curve, they may not accelerate as they would have to abruptly decelerate immediately afterwards. New or inexperienced road users, however, do not possess such knowledge, and are possibly involved in more harsh events. Furthermore, relying on road user anticipation is not within the state-of-the-art approaches that promote more holistic, 'safe system' designs. This fact alone proves the necessity of highly accurate hotspot detection.

There is a critical relationship that is implied in the previous situations: the road safety level of a location can be affected not only by its own characteristics, but in addition, by the characteristics of the surrounding/neighboring locations as well. If a small road segment feeds into a larger one without proper provision, for instance, a hotspot may be created in the larger segment due to the presence of the smaller one.

In order to tackle the previous problems, the most appropriate tool to utilize is spatial analysis. Using spatial analysis, in addition to the influence of local features of an area such as a road segment, it is possible to examine the influence of the characteristics of neighboring areas as well. In other words, the relative position of the studied areas comes into importance and plays a role in all spatial analyses techniques. Moreover, spatial analyses provide inherent comparative advantages compared with ordinary statistical or econometric methods. Examples are the inclusion of spatial effects in models from unobserved or unknown parameters, the consideration of location for the estimation of the influence of parameters and the intuitive presentation of results on maps (e.g. Loo & Anderson, 2015).

1.1.4 Objective of the dissertation

Taking the previous into account, 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.

This combination of data, study areas (urban network and urban arterial segments) and methodological tools outlines a very promising – and previously unexplored – research field. The exploitation of high resolution big data via smartphone sensors provides an environment which is rich in information and has adequate network coverage as basis for analysis. Data analysis through powerful state-of-the-art spatial models is expected to highlight the extent of the influence of several factors contributing to the occurrence of harsh events, namely harsh braking and harsh acceleration frequencies. Additionally, the effect of segment locations is taken into account for each road segment in several parts of the analysis. The investigation of urban road networks in particular is of high importance, as these areas are relatively unexplored in road safety. Given the abundant risk exposure of urban road users regarding harsh events, fruitful results are expected. These results in turn are expected to lead to knowledge which will be useful for reducing harsh event occurrence, and thus crash occurrence, and increasing overall road safety levels.

To achieve the aforementioned objective, the dissertation utilized three analytical tools:

1. Geographically Weighted Regression models, an econometric/functional method
2. Conditional Autoregressive Prior Bayesian models, a Bayesian statistical method
3. XGBoost algorithms, a machine learning method (i) with aspatial random cross-validation and (ii) with spatial cross-validation

In addition, to accomplish the aims of the dissertation, several methodological innovations are devised as answers to the respective challenges:

1. Derivation of secondary geometrical characteristics from extracted digital map data
2. Assignment of naturalistic driving trips and harsh events to segments with a purpose-made map-matching algorithm
3. Integration of traffic characteristics in urban arterial environments
4. Evaluation of model performance prediction capabilities and employment of the respective metrics

1.2 Methodology of the dissertation

In order to achieve the scientific objective set for the present doctoral dissertation, a series of subsequent methodological steps were undertaken.

I – Literature review

Initially, an exhaustive literature review was conducted covering spatial analyses in road safety on various areal units, as well as the examination of famous related problems such as the boundary problem and modifiable areal unit problem. Meta-regression techniques were then applied on exposure variables used in spatial analyses to obtain quantitative results. The various tools employed for the recording of driver behavior were also examined and assessed on their comparative advantages and disadvantages. An additional exploration of the literature was conducted based on the merits of harsh event analysis. From the critical synthesis of the findings of this extensive process, the research questions of the dissertation were formulated.

II – Methodological background

With the problem under consideration as well as the scientific literature in mind, a methodological investigation was also conducted. Several spatial exploratory tools are presented, such as global and local Moran's *I* coefficients and merged and directional variograms. The underlying theory for statistical models and algorithms appropriate for spatial analysis – but also novel ones – was subsequently investigated. Specifically, functional (frequentist) methods (GWPR), Bayesian stochastic methods (CAR) and machine learning methods (XGBoost) were all explored regarding their feasibility.

III – Multi-parametric data acquisition

A series of different data sources were investigated on their capability to provide large-scale data that could be seamlessly integrated in order to capture more aspects of the road environment and thus study the phenomena of harsh event frequencies more spherically. Appropriate study areas (two urban networks and one urban arterial) and a fixed study period were defined. Afterwards, high-resolution naturalistic driving big data collected using smartphone sensors via the OSeven platform were obtained for these areas. These data were augmented by geometric data from OpenStreetMap and NASA SRTM topography data. Additionally for urban arterial segments, Traffic Management Center data regarding traffic conditions were obtained for the specific areas and study periods.

IV – Data processing and merging algorithms

Data from all sources were then combined using complex algorithms developed in R-studio for the purposes of this dissertation. The tasks undertaken by the algorithms include data cleaning, derivation of additional geometrical characteristics from the existing geometry, parameter map-matching to each road segment and selective traffic parameter integration. Thus a number of variables for the examined road segments that included geometry and similar fixed attributes, road user behavior and, for urban arterials, traffic parameters was obtained. These datasets were then ready for descriptive and spatial analysis.

V – Urban road network spatial analyses

Exploratory spatial analyses were conducted by calculating global and local Moran's I indicators as well as variograms for harsh braking and harsh acceleration frequencies. The three diverse analytical tools that were determined from the extensive review processed were applied in the training and test datasets for urban networks. By employing count-based Geographically Weighted Regression, Conditional Autoregressive Prior Bayesian Regression and XGBoost algorithms with random and spatial cross-validation, several statistical models were developed separately for harsh braking and harsh acceleration frequencies. These models describe a wealth of statistical relationships between the frequencies and the independent variables. Furthermore, by applying the models on the test area dataset, information was obtained on the performance and predictive capabilities of each model. Combined final predictions were made from the four employed methods for the test area that were more accurate than those of individual models.

VI – Urban arterial spatial analyses

To further explain harsh braking and harsh acceleration frequencies, it was decided to eschew predictive power in favor of including more non-fixed variables in the dataset, such as traffic and driver behavior variables. A section of a directionally separated urban arterial was selected as the study area, due to the straightforward nature of traffic measurement locations and the availability of traffic data therein. Furthermore, the classification of trip-seconds in each traffic state and the subsequent examination and modelling of each traffic state separately was chosen as the most appropriate approach. Therefore, a similar process is applied to the urban arterial dataset, with added traffic and driver behavior variables for each traffic state. The same analytical methods were then applied and additional spatial relationships were obtained with the inclusion of variables related to traffic and driver behavior. A number of new informative statistical relationships was discovered.

VII – Conclusions

As a final step, the findings of all the previous processes are presented in a compact format and evaluated. They serve as basis for drawing useful conclusions, as well as for presenting the several innovative contributions of the present research. Challenges, limitations and future research directions are also discussed.

The methodological steps that were followed in this doctoral dissertation are also presented visually on Figure 1-3:

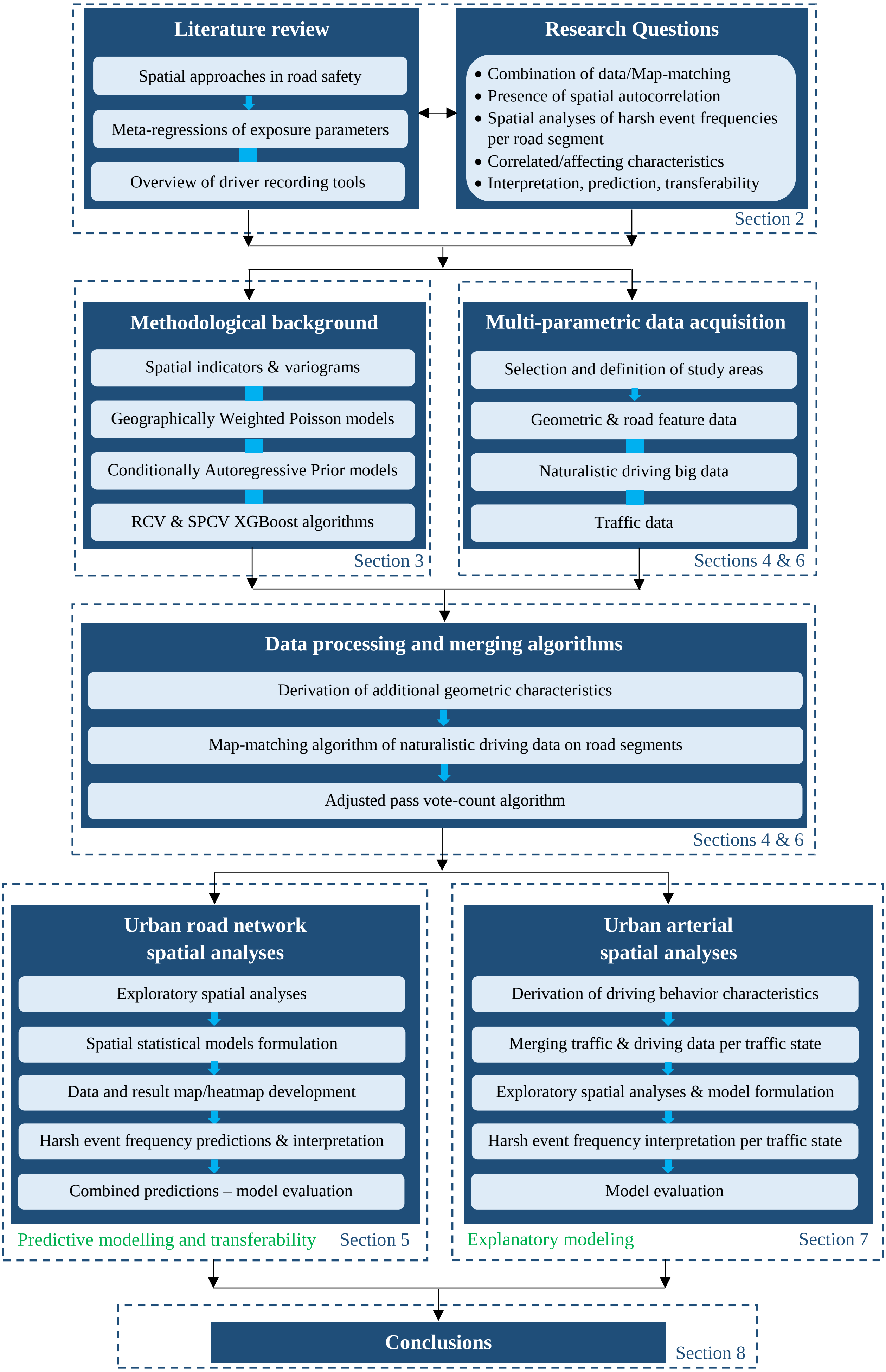

Literature review
Spatial approaches in road safety
Meta-regressions of exposure parameters
Overview of driver recording tools
Research Questions
• Combination of data/Map-matching
• Presence of spatial autocorrelation
• Spatial analyses of harsh event frequencies per road segment
• Correlated/affecting characteristics
• Interpretation, prediction, transferability
Section 2
Methodological background
Spatial indicators & variograms
Geographically Weighted Poisson models
Conditionally Autoregressive Prior models
RCV & SPCV XGBoost algorithms
Section 3
Multi-parametric data acquisition
Selection and definition of study areas
Geometric & road feature data
Naturalistic driving big data
Traffic data
Sections 4 & 6
Data processing and merging algorithms
Derivation of additional geometric characteristics
Map-matching algorithm of naturalistic driving data on road segments
Adjusted pass vote-count algorithm
Sections 4 & 6
Urban road network spatial analyses
Exploratory spatial analyses
Spatial statistical models formulation
Data and result map/heatmap development
Harsh event frequency predictions & interpretation
Combined predictions – model evaluation
Predictive modelling and transferability
Section 5
Urban arterial spatial analyses
Derivation of driving behavior characteristics
Merging traffic & driving data per traffic state
Exploratory spatial analyses & model formulation
Harsh event frequency interpretation per traffic state
Model evaluation
Explanatory modeling
Section 7
Conclusions
Section 8

Figure 1-3: Overall methodological framework of the doctoral dissertation

1.3 Structure of the dissertation

Section 1 serves as the introduction of both the research areas and the respective scientific problems this dissertation aims to explore, and the methodology and processes it employs in endeavoring to tackle them. Starting with the overall description of the current – at the time of writing – state of road safety and overall macroscopic crash statistics, the three pillars of road safety risk factors are explained. Amongst road infrastructure elements, vehicle factors and human factors, human factors are designated as the overwhelming cause of crashes. A specific aspect of human factors, harsh events of driver behavior, is showcased as but understudied field, and spatial analysis of harsh event occurrence is determined as an interesting and promising research direction. The objectives and methodological structure are presented subsequently, followed by the structure and main contributions of the doctoral dissertation.

Section 2 presents the literature review of a wide body of scientific research pertinent to the topic and methodology of the dissertation. The initial sub-sections concern the applications of spatial analyses in road safety on various areal units, as well as the examination of famous problems such as the boundary and modifiable areal unit problems. Some applications on related fields are also mentioned. Afterwards, meta-regression techniques are applied to selected road safety exposure parameters, thus providing a quantitative review of study characteristics on exposure parameters. An overview of available driver recording tools is then provided, followed by a note on the merits of harsh event analysis. The literature review is concluded with the critical synthesis of results and the resulting research questions that this dissertation is called to answer.

Section 3 describes the methodological approach of this doctoral dissertation. Initially, the overall framework is outlined, and the necessity of the examination of both urban networks and urban arterials is explained. The theoretical background of the various statistical methods for analyzing harsh event frequencies is then provided. These methods include exploratory analytical tools, such as autocorrelation indicators and variograms, statistical techniques, namely Geographically Weighted Poisson Regression and Bayesian Conditional Autoregressive Prior Regression, and machine learning methods, such as XGBoost. Afterwards the description of the various data sources, along with the methodological steps followed in the dissertation is provided in detail. The data collection and pre-processing phases, as well as the derivation of additional geometric and driver behavior characteristics from the data and the map-matching method are all highlighted. The description of the respective algorithmic processes are then explained in parallel.

Section 4 outlines the results of data collection and processing for urban networks, including the initial study area examination, the derivation of geometrical characteristics, the exploration of the large-scale naturalistic driving data provided from smartphones and the results of the map-matching algorithm. Extensive descriptive statistics are provided for both urban network study areas, namely the training area (Chalandri) and the test area (Omonoia). Sample spatial data frames for each area are also provided.

Section 5 showcases the exploratory and spatial statistical analyses conducted for urban networks. Global and local Moran's I indicators are calculated for harsh braking and harsh acceleration frequencies, followed by variograms for these variables. Afterwards, the results of Geographically Weighted Poisson Regression, Conditional Autoregressive Prior Regression, and XGBoost with random and spatial cross-validation are presented, compared and discussed. For urban networks, emphasis is given to the predictive capability and transferability of model results. All methods are applied on the training area dataset, and then full predictions are conducted on the test area dataset.

Sections 6 and 7 largely mirror Sections 4 and 5, albeit for urban arterial segments. In Section 6, data collection and processing results are presented along with the spatial data frame for urban arterial segments. In Section 7, the exploratory and spatial statistical analyses conducted for urban arterial segments are showcased for each traffic state. An important difference is that urban arterial segments are exploratory and are thus not used for prediction by application of the models in a different dataset due to lack of fixed traffic variables. Comparison and discussion of results is conducted again as well.

Section 8 contains the conclusions of the present doctoral dissertation. The main findings are succinctly summarized regarding harsh braking and harsh acceleration frequencies for both types of study areas. Subsequently, the main contributions and innovations of the present research are presented. Present challenges and limitations as well as future research directions are also discussed. The full list of bibliographical references concludes the dissertation in Section 9.

1.4 Contribution of the dissertation

The present doctoral dissertation is an endeavor to widen the range of available road safety and traffic behavior knowledge overall, and to offer new insights regarding harsh event and traffic behavior spatial analysis and hotspot detection and prediction in particular. To achieve the maximum utility and transferability of results, multi-parametric big datasets are utilized and analyzed, describing several aspects of the road environment: (i) high resolution naturalistic driving and driver behavior data, (ii) microscopic road segment geometry and road network characteristic data from digital maps and (iii) high resolution traffic data.

The findings of this research are multi-faceted, and are expected to contribute to the broader field of road safety significantly: The spatial analysis of harsh events, which constitute pro-active road safety indicators, is a new, unexplored research direction that shows considerable promise. Furthermore, the microscopic approach of segment-based analyses, with the utilization of individual segment characteristics, allows for in-depth and precise research approaches.

To obtain the desired outcomes, this dissertation employs various purpose-made big data algorithms, which were devised and implemented in intermediate data-processing steps, performing critical functions necessary for the spatial analyses, such as derivation of additional characteristics, data merging & processing and map-matching.

Furthermore, various advanced spatial statistical models are innovatively utilized for segment-based harsh event spatial analyses. A balanced variety between classic functional (frequentist) methods, Bayesian stochastic methods and machine learning methods was chosen. 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.

The practical value and applicability of the present research is significant. Firstly, road environment aspects which are correlated with harsh braking and harsh acceleration occurrence are determined, and their effects are quantified. Based on that knowledge, effective respective countermeasures can be implemented in order to increase road safety levels in urban network and urban arterial areas. Secondly, innovative methodological tools and concepts for data merging and spatial predictions of harsh events are provided that be used to tackle similar problems.

A critical outcome of spatial analysis is the creation of maps and heatmaps based on both initial data and processed results. The creation of the corresponding maps showcasing important information, such as hotspot locations, offers a highly comprehensive and informative tool to individual road users, companies involved in road transport, road authorities and other road safety stakeholders. They offer a means to effectively increase road safety levels, which can also be adapted for any similar purpose.

2 Literature Review

The purpose of this section is to review and evaluate studies pertinent to the topic of this doctoral dissertation. To that end, three major topics are examined in three respective and distinct sections:

1. Spatial approaches in road safety. In this section the characteristics, findings and approaches of scientific literature relevant to spatial analysis are presented.
2. Meta-regressions of exposure parameters of spatial analyses: a selected number of the previously reviewed studies are meta-analyzed to obtain quantitative estimates that several study characteristics impose on the values of their coefficients. This section essentially complements the previous one by providing a quantitative review of study characteristics on exposure parameters.
3. Overview of driver recording tools. In this section, the large array of tools available to road safety researchers are discussed and evaluated comparatively.

These qualitative and quantitative review sections are followed by a note on the merits of harsh event analysis. Subsequently, a critical synthesis of the findings of the literature, as well as the formulation of the critical research questions that the present doctoral research endeavors to answer are provided.

2.1 Spatial approaches in road safety

2.1.1 Introduction

Road safety has been a major issue in contemporary societies, with road crashes incurring major human and material costs annually worldwide. Traffic and road safety practices have been implemented to save lives by halting the increase of road traffic fatalities against an ever-rising population (WHO, 2015), though it appears that the global target of halving road traffic deaths by 2020 will not be met (WHO, 2018).

The still occurring and plateauing crash casualties suggest a lot of untapped potential and margins for safety improvements that can be exploited if the occurrence of crashes can be predicted more accurately. Road safety scientists have invested considerable efforts in studying the impacts of several risk factors (e.g. Theofilatos & Yannis, 2014; Papadimitriou et al., 2019a) and road safety measures (e.g. Elvik et al., 2009) and have developed or adopted a number of mathematical methodologies to approach crash prediction problems (e.g. Lord & Mannering, 2010) or road safety site prioritization problems (e.g. Lee & Abdel-Aty, 2018).

Since road transport involves distances by nature, it stands to reason that spatial analyses would be considered by researchers. Spatial analyses in road safety typically involve the examination of crashes while taking their absolute or relative locations into account. Crashes face the typical issues of all point data: spatial dependence and spatial heterogeneity.

In simple terms, spatial dependence essentially refers to events at a location being highly influenced by events at neighboring locations. It is usually measured via spatial autocorrelation metrics. In turn, autocorrelation refers to the influence of variable values of given points on variable values of adjacent points (spatially or temporally). Spatial heterogeneity occurs in the modelled relationships as the coefficients between random parameters and observed events are not fixed spatially.

Therefore, researchers have discovered several caveats and merits in conducting spatial analysis. Road crashes are subject to both spatial and temporal variations (Loo & Anderson, 2015), intuitively suggesting spatial analyses as informative. By accounting for spatial dependence and heterogeneity in the estimates, spatial analyses describe how regions affect and are affected by the road safety attributes of their neighbors, and how the influence of explanatory parameters varies across space as well.

As a more specific example, when considering spatial correlation in crash models, estimates are effectively "pooling strength" from neighboring locations, thus improving the produced estimations (Aguero-Valverde & Jovanis, 2008). Road crashes are a complex phenomenon, and their analysis requires assumptions and merging of the examined parameters for a feasible approach, which unavoidably leads to some degree of loss of information or even misrepresentation of the actual conditions (Xu & Huang, 2015). Spatial analyses can counterbalance this loss by providing predictions of counts of crashes (and of similar incidents, such as near-misses) that vary across different units of analyses, thus capturing all the unobserved trends and particularities of each area. Thus not only is better theoretical understanding provided for crash occurrence across space, but the identification of high-risk sites (known as hotspots) becomes more accurate (El-Basyouny & Sayed, 2009; Aguero-Valverde, 2014).

After decades of research, the topic of spatial analysis of traffic crashes covers a wide range, including mapping and visualization of crash counts, identifying clustering patterns of traffic collisions, and use of spatial models to investigate the effects of contributory factors and recommend targeted countermeasures. The mathematic particulars of spatial analyses have been examined in several published studies, for instance in Bivand et al. (2009) for Global and Local Moran's I and in Ver Hoef et al. (2018) for conditional autoregressive priors (CAR) models or simultaneous autoregressive priors (SAR) models. The reader is also referred to Yao et al. (2016), for a review of major advancements of spatial crash analysis using applied GIS tools. The examined research there starts from a significantly older period (in 1976) and includes topics that fall out of the scope of the present research, such as visualizing and mapping of events.

The aim of the present section is to provide a review of the scientific literature regarding spatial approaches and spatial analyses in road safety. This passage constitutes is an endeavor to investigate how road safety researchers handle the dimension of space in its various aspects in their studies, whether that regards modelling of spatial events, selecting the scale of areal units or proximity structures, tackling boundary problems or other specific issues (such as vulnerable road users – VRUs). In order to achieve the aim of the current research, published scientific studies (in English) are critically examined. The selected studies were intended to be representative of a wide array of countries and adopted methodologies, in order to provide a well-rounded summary of the state-of-the art in road safety spatial analyses. Emphasis was given to more recent studies, with some seminal endeavors being included as well for completeness.

The main focus of the current study is on study characteristics, modelling approaches and methodological issues. It should be noted that this research only includes studies that conducted explicit and dedicated spatial or spatio-temporal analyses, as opposed to studies that examine different areas for purposes of cross-sectional or case-control studies (and as such do not examine the spatial aspect of road safety incidents). The second category of studies has its own merits and has been extensively implemented in road safety research, but falls out of the scope of this review.

This section of the literature review is organized as follows. Section 2.1.2 includes an examination of the different spatial units of analyses, together with famous boundary and zonal problems, as well as the issues of proximity structures. Section 2.1.3 outlines various modelling approaches, while Section 2.1.4 discusses issues in spatial analyses of VRUs. Finally, a discussion of overall findings from the review process and future research directions on this topic are provided in Section 2.1.5.

2.1.2 Examination of spatial units

Spatial analyses in road safety fundamentally involve the examination of road safety indicators (crash counts or rates, injury severity rates etc.) across spatial units of analyses. The manner in which researchers select and define these spatial units directly influences the scope of the study, as well as the interpretability of results, while this can apply to data preparation as well (Imprialou et al., 2016). There is a structural difference, for instance, in examining spatial distribution of road safety indicators in consequent road segments that feed traffic flow seamlessly into each other compared to examining junction clusters with several inflows and outflows for the distributions of the same indicators.

Different spatial units are discussed in the following section, and study characteristics for each spatial unit level are summarized on Tables 2-1 to 2-4. It was decided to include study characteristics initially considered by researchers on the Tables of this review, even if they were not found significant in the respective final models, to better showcase the scope of each research. The examined crash categories are denoted with the following acronyms with respect to the involved road users: Total Crashes (TC), Motorcycle crashes (MC), Single Vehicle crashes (V), Vehicle-vehicle crashes (V-V), Bicycle-vehicle crashes (B-V) and Pedestrian-vehicle crashes (P-V). When crash category details are not given about the examined crashes in a study, they are noted as TC. Additional details, such as the analysis of a specific crash type are noted as well.

2.1.2.1 Road segment and intersection approaches

Initial approaches of spatial analyses involved the more intuitive examination of road safety indicators across singular or multiple road sections, such as straight road segments and intersections. Earlier approaches involve the depiction and analysis of spatial distribution of crashes on (state) highways, in an attempt to perceive visual patterns of heightened concentration and possible correlation with touristic areas (Page & Meyer, 1996), albeit with a small sample. Furthermore, examination of the impact of the length of segments on crash counts and density which were found to follow Poisson distribution in the smaller segment scales growing from more intermediate distributions to normal distributions as segments increased, as shown by a study by Thomas (1996) that also first touched on the modifiable areal unit problem in road safety (discussed in Section 2.1.2.6).

It has been determined that local environment and road infrastructure are critical factors of crash occurrence (Flahaut, 2004; Wang et al., 2016a). A traditional division when examining straight road segments is road type; highways with divided traffic directions display different road safety mechanisms than undivided two-lane arterials and for decades have been analyzed separately, a practice that is continued in segment-based spatial analyses.

The environment of road segments has been traditionally examined separately in the literature, with researchers distinguishing between urban and rural segments and often producing comparative analyses between different types of segments. A spatial analysis by Flahaut (2004) determined 2-lane configurations as the most unsafe configuration for rural roads. For urban roads, it has been found that increases in the number of crosswalks and the densities of unsignalized intersections both increase crash occurrence (Barua et al., 2014). Furthermore, local and non-local drivers are found to cluster along road segments, and segments with adverse safety interactions between these two groups are estimated to transfer these effects spatially to neighboring segments (Wang et al., 2016a).

In spatial analyses, researchers examine intersections either in groups (Guo et al., 2010; El-Basyouny & Sayed, 2011) or in aggregation (Miaou & Lord, 2003; Wang & Abdel-Aty, 2006). Intersection geometry, location and traffic parameters are important within the context of spatial analyses. The size of intersection, the traffic conditions by turning movement, and the coordination of signal phase have significant impacts on the number of crashes at intersections (Guo et al., 2010). Xie et al. (2013) have shown intersections on segments with lower mean speeds were associated with fewer crashes than those with higher speeds, and that intersections on two-way roads, under elevated roads, and in close proximity to each other, tended to have higher crash frequencies as well. A seminal result of a study by Abdel-Aty & Wang (2006) shows that overall, three-legged intersections tend to exhibit lower crash rates than four-legged intersections, and that they exhibit different road safety mechanisms. Furthermore, effectiveness of implemented road safety treatments can vary between locations when considering injury severity levels (El-Basyouny & Sayed, 2011).

When proximal segments are considered, with the layout of a simple road network, it is important to note that there are spatial correlations between intersections and their adjacent segments, which have been found to be significant in the literature (Abdel-Aty and Wang, 2006; Quddus, 2008; Aguero-Valverde &

Jovanis, 2010; Dong et al., 2014; Dong et al., 2015; Wang & Huang, 2016). Spatial correlation is also found in crashes of intersections along the same corridor, due to similar traffic flow patterns, presence of traffic signals and geographic characteristics (Guo et al., 2010), an issue which ought to be properly addressed with proper modelling tools (Xie et al., 2014). Additionally, several studies have integrated corridor-level characteristics into segment-level or intersection-level analysis in an effort to capture factors explaining heterogeneity (Abdel-Aty and Wang, 2006; Guo et al., 2010; Xie et al., 2014).

A different effort was made by Zeng & Huang (2014), who endeavored to model crash counts on road segments and intersections simultaneously. They used Bayesian spatial joint models to account for spatial correlations between adjacent road segments and intersections that were found to be more accurate than simple Poisson and negative binomial models. The joint model integrated junctions and segments to the basic link function. An indicator variable which denoted whether a segment or intersection was examined was utilized. The authors highlight that the spatial correlations between intersections and their connected segments were more significant than those found between intersections or between segments only, presumably due to common unobserved parameters such as speed. The approach of joint simultaneous modelling of intersections and segments was further advanced by Alarifi et al. (2017) who developed four multi-level Bayesian joint models for that purpose. Specifically, the reasoning was to complement the intersection/segment examination by including corridor-level characteristics in the models. Because corridor characteristics vary along their length, random forest models were used to divide corridors into-sub corridors of fixed-value characteristics. Ultimately there were statistically significant variables at the segment level, at the intersection level and at the corridor/sub-corridor level; the importance of median opening density for crash occurrence was underlined from the results. However, spatial autocorrelation of adjacent road entities was not examined in that study. Moreover, Alarifi et al. (2018) (discussed in Section 2.7) also conducted analyses including intersection-, road segment- and corridor-level parameters, in an attempt to explore that research question.

Reviewed studies that primarily focus on spatial analyses at the individual road segment/intersection level are shown on Table 2-1.

Table 2-1: Studies with road safety spatial analyses primarily on the individual road segment/intersection level

Study Characteristics				Dependent variables				Independent variables – parameters													Spatial aggregation approach			
								Traffic				Road user		Road environment										
Author(s)	Year	Country of study	Crash type analyzed	Crash count/frequency	Crash rate	Injury Severity	Casualty rate	Speed	Traffic volume	Vehicle distance traveled	Number of Trips - OD	Road user/ Population age	Modal distinction	Speed Limit	Curvature	Gradient	Lane width	Lane number	Intersection nr./density	Roadway length	Regional level	Zonal level	Link/ segment/ intersection level	Analysis - Modelling approach
Abdel-Aty & Wang	2006	United States	TC	●				●	●					●			●	●	○	●			Intersections	Negative Binomial Regression with and without Generalized estimating equations | Cluster analysis
Aguero-Valverde	2014	United States	TC	●					●					●			●	○		●			Rural road segments	Full Bayes hierarchical Poisson model (1) with normal priors for spatial random effects | (2) with CAR priors for spatial random effects | (3) with a joint distribution
Aguero-Valverde & Jovanis	2010	United States	TC	●					●					●	●	●	●	○		●			Rural & Urban road segments	Full Bayes hierarchical Poisson model with CAR priors for spatial random effects
Aguero-Valverde & Jovanis	2008	United States	TC	●					●					●			●	○		●			Rural road segments	Bayesian Multivariate Poisson Lognormal Regression | Bayesian random effects models
Aguero-Valverde et al.	2016	United States	TC (6 Crash types)	●					●											●			Rural road segments	Full Bayes Poisson Regressions (Univariate, Univariate Spatial, Multivariate, Multivariate Spatial)
Alarifi et al.	2018	United States	TC	●					●				●	●				●	●	●			Intersections | Road segments	13 Bayesian hierarchical Poisson-lognormal joint spatial models with adjacency-based, adjacency-route, distance-order, and distance-based spatial weight features
Alarifi et al.	2017	United States	TC	●					●				●	●				●	●	●			Intersections | Road segments	Multilevel Poisson-lognormal joint model (1,2) with corridor and sub-corridor random effects (3,4) with corridor and sub-corridor random parameters
Barua et al.	2016	Canada	TC	●		○			●									●	●	●			Urban road segments	Full Bayesian Poisson lognormal multivariate random parameters models (1) with heterogenous effects (2) with CAR priors for spatial heterogeneity (3) with both
Barua et al.	2014	Canada	TC	●		○			●									●	●	●			Urban road segments	Full Bayesian Poisson lognormal univariate and multivariate random parameters models (1) with heterogenous effects (2) with CAR priors for spatial heterogeneity (3) with both
Chiou et al.	2014	Taiwan	TC	●		●			●				●		●	●		●	○	●			Highway segments	Multinomial-generalized Poisson with error-components (spatial error and spatial exogenous)
Effati et al.	2015	Iran	TC			●						●				●	●		●	●			Highway segments	Support Vector Machine Algorithms (SVMs) | Coactive neuro-fuzzy inference system
El-Basyouny & Sayed	2011	Canada	TC	●		○			●										○				Intersections	Univariate and Multivariate Poisson Lognormal Regressions | Full Bayes estimations
El-Basyouny & Sayed	2009	Canada	TC	●					●									●	●	●			Urban road segments	Full Bayesian Multivariate Poisson Lognormal with and without CAR Prior | Full Bayesian Multiple Membership model | Full Bayesian Extended Multiple Membership model

Study Characteristics				Dependent variables				Independent variables – parameters													Spatial aggregation approach			
								Traffic				Road user		Road environment										
Author(s)	Year	Country of study	Crash type analyzed	Crash count/frequency	Crash rate	Injury Severity	Casualty rate	Speed	Traffic volume	Vehicle distance traveled	Number of Trips - OD	Road user/ Population age	Modal distinction	Speed Limit	Curvature	Gradient	Lane width	Lane number	Intersection nr./density	Roadway length	Regional level	Zonal level	Link/ segment/ intersection level	Analysis - Modelling approach
Guo et al.	2010	United States	TC	●					○					●				○		○			Intersections	Fixed effects Bayesian Poisson Regression | Fixed and Mixed effects Bayesian Negative Binomial Regression | Spatial CAR Prior extended Poisson/Negative Binomial models
Huang et al.	2017	China	TC | V/V-V P-V | B-V	●					●			●	○	●					○				Intersections	Poisson Regression (Univariate, Multivariate Lognormal & Spatial random effects models)
Huang et al.	2016	United States	TC	●					●	●	●	●		●				●	●	●		TAZ	Intersections | Road segments	Bayesian spatial model with CAR prior (macroscopic) | Bayesian spatial joint models with CAR prior (microscopic)
Flahaut	2004	Belgium	TC	●		○			●					●		○		●	○	○			Rural & Highway segments	Logistic regression with and without spatial autocorrelation
Liu et al.	2017	United States	TC	●					●				●	●						●			Highway segments	Geographically Weighted Negative Binomial Regression | Negative Binomial Regression
Ma et al.	2017	United States	TC	●		○		●		●						●		●					Highway segments	Hierarchical Bayesian random parameters models (structured and unstructured spatio-temporal effects)
Miaou & Lord	2003	Canada	TC		●				●					○					○				Intersections	Full Bayes | Empirical Bayes
Miaou & Song	2005	Canada | United States	TC	●	●	●			●	●				○				○		●			Intersections | Rural segments	Multivariate spatial Bayesian generalized linear mixed models with and without CAR Prior
Mitra	2009	United States	TC	●		●			●														Intersections	Hierarchical Full Bayes Jointly specified spatial model | Negative Binomial Regression | Local Moran's I
Mountrakis & Gunson	2009	United States	V-A	●																○			Rural segments	Spatial, Temporal & Spatiotemporal kernel estimation | Ripley's K-function
Page & Meyer	1996	New Zealand	TC	●		○														○	National Parks		Highway segments	Percentage descriptive statistics
Thomas	1996	Belgium	TC	●		○			○											●			Highway segments	Univariate and bivariate descriptive statistics, chi^2 and W tests
Wang & Abdel-Aty	2006	United States	V-V (rear-end only)	●					●					●				●	○				Intersections	Generalized Estimating Equations with Negative Binomial link function
Wang & Huang	2016	United States	TC	●					●		●			●				●	●	●		TAZ	Intersections | Urban segments	Bayesian hierarchical joint Poisson Regression | Bayesian joint Poisson Regression | Negative Binomial Regression
Wang et al. (a)	2016	United States	TC	●		●			●					●	●	●			●	●			Highway segments	Multivariate Poisson Lognormal regression with CAR Prior
Wang et al.	2009	United Kingdom	TC	●		○		●	●						●	●		●		●			Highway segments	Bayesian Multivariate Poisson Lognormal | Negative Binomial Regression | Poisson Models with CAR priors (with first/second order neighbors)

Study Characteristics				Dependent variables				Independent variables – parameters													Spatial aggregation approach			Analysis - Modelling approach
								Traffic				Road user		Road environment										
Author(s)	Year	Country of study	Crash type analyzed	Crash count/frequency	Crash rate	Injury Severity	Casualty rate	Speed	Traffic volume	Vehicle distance traveled	Number of Trips - OD	Road user/ Population age	Modal distinction	Speed Limit	Curvature	Gradient	Lane width	Lane number	Intersection nr./density	Roadway length	Regional level	Zonal level	Link/ segment/ intersection level	
Wen et al.	2019	China	TC	●						●					●	●							Highway segments	(1) Poisson Lognormal regression with CAR Prior | (2) Poisson Lognormal regression with spillover effects | (3) Hybrid of (1) and (2)
Xie et al.	2014	China	TC	●				●	●									●	○	●			Intersections | Urban segments	Bayesian Negative Binomial regression (basic, random effect, random parameter, hierarchical, hierarchical CAR)
Xie et al.	2013	China	TC	●				●	●									●	○	●			Intersections | Urban segments	Bayesian Negative Binomial regression (basic, random parameter, hierarchical)
Zeng & Huang	2014	United States	TC	●				●						●				●	●	●			Intersections | Urban segments	Poisson Regression | Negative Binomial Regression | Bayesian spatial model with CAR prior | Bayesian spatial joint models with CAR prior

● Considered in the study design, ○ considered in the study process as filter/defining characteristic

2.1.2.2 Zonal approaches

A number of zonal units have been adopted by researchers, from smaller to larger ones. Their boundaries can be census-based, administrative-based or traffic-based, and are dependent on the country or environment of study. Studies in the UK might utilize enumeration districts, namely areas averaging circa 200 households (Noland & Quddus, 2005) or census wards, which include about 2000 households (Noland & Quddus, 2004; Quddus, 2008). Similarly, studies from other countries have used locally available spatial units, such as the Australian ABS structure units (Statistical areas 1,2 (SA1,2), state electoral divisions (SED)) used by Amoh-Gyimah et al. (2017).

Many studies originate from the US and have utilized units that are used there: Census Blocks (CBs) are the smallest unit, averaging 85 people and are expanded to Census Block Groups (CBGs), averaging 39 blocks with about 1500 people (Lee et al., 2017a). CBGs have been utilized by road safety researchers to some extent (Levine et al., 1995; Abdel-Aty et al., 2013).

Traffic Analysis Zones (TAZs) are created primarily in the US with the explicit purpose of collecting trip and traffic statistics and data, though they have been implemented in other countries as well (Ng et al., 2002; Gomes et al., 2017). From traditional zonal approaches, TAZs are the only traffic-related zone system (Lee et al., 2017a), which might explain their popularity for utilization in spatial analyses (e.g. Ng et al., 2002; Hadayeghi et al., 2003; Ladrón de Guevara et al., 2004; Lovegrove & Sayed, 2006; Lovegrove & Sayed, 2007; Hadayeghi et al., 2010; Naderan & Shahi, 2010; Abdel-Aty et al., 2011; Abdel-Aty et al., 2013; Dong et al., 2014; Lee et al., 2014b; Dong et al., 2015; Lee et al., 2015a; Xu & Huang, 2015; Dong et al., 2016; Nashad et al., 2016; Xu et al., 2017a, 2017b; Bao et al., 2017; Gomes et al., 2017). TAZs can be also expanded for road safety assessment purposes by aggregating TAZs groups with similar crash rates, thus creating Traffic Safety Analysis Zones (TSAZs), (Lee et al., 2014b; Abdel-Aty et al., 2016).

Census Tracts (CTs, or census output areas) are larger units containing about 4000 people of comparable socio-economic statuses in the US (or about 2500 people in the UK). They too have been adequately explored in road safety spatial analyses in the literature (e.g. LaScala et al., 2000; Loukaitou-Sideris et al., 2007; Delmelle & Thill, 2008; Wier et al., 2009; Cottrill & Thakuriah, 2010; Ukkusuri et al., 2011; Narayanamoorthy et al., 2013).

Similar to TAZs, Traffic Analysis Districts (TADs) are newly created, larger geographic traffic related units used for transport analyses. A few recent studies have utilized TADs as basis for analysis (e.g. Abdel-Aty et al., 2016, Cai et al., 2017b; Lee et al., 2017a). Other zonal areas have been used as well by exploiting existing utility systems, such as postal-ZIP codes (e.g. Lee et al., 2014a; Bao et al., 2018) and urban/rural areas defined by healthcare authorities (e.g. MacNab, 2004; Bu et al., 2018).

Reviewed studies that primarily focus on spatial analyses at zonal levels are shown on Table 2-2.

Table 2-2: Studies with road safety spatial analyses primarily on the zonal level

Study Characteristics				Dependent variables				Independent variables – parameters																Spatial aggregation approach			Analysis - Modelling approach
								Traffic				Road environment						Demographic			Socio-economic		Land Use				
Author(s)	Year	Country of study	Crash type analyzed	Crash count/frequency	Crash rate	Injury Severity	Casualty rate	Speed	Traffic volume	Vehicle distance traveled	Number of Trips - OD	Speed Limit	Curvature	Lane width	Lane number	Intersection nr./density	Roadway length	Population number/density	Road user/Population age	Modal distinction	Household/Personal Income	Employment percentage/density	Land use factor(s)	Regional level	Zonal level	Link/ segment/ intersection level	
Abdel-Aty et al.	2013	United States	TC	●		○				●	●	●				●	●	●	●		●		●		TAZ | CT | BG		Bayesian Multivariate Poisson Lognormal Regression
Abdel-Aty et al.	2011	United States	TC	●		○		○			●	●				●	●			○					TAZ		Negative Binomial Regression
Amoh-Gyimah et al.	2017	Australia	TC	●		○				●		●						●	●	●	●		●		SA1 | SA2 | TAZ | SED | ZIP		Random parameter negative binomial model | Semi-parametric Poisson GWR (also on custom grid cells)
Anderson	2007	United Kingdom	TC	●		○																			CT	Urban road segments	Kernel density estimation | Network analysis | Census Output Area estimation
Anderson	2009	United Kingdom	TC | P-V | B-V	●		○										○	●			●			●		Hotspot clusters		Kernel density estimation | K-means clustering
Bao et al.	2018	United States	TC	●		○				●	●	○				●	●	●	●		●	●			ZIP		Poisson GWR | Latent Dirichlet Allocation
Bao et al.	2017	United States	TC | V-V | P-V	●					●		●					●	●	●	●	○	●	●			TAZ		Geographically Weighted Regression (GWR)
Cai et al. (a)	2019	United States	TC	●						●					●	●	●	●	●		●		●		TAD		Bayesian Poisson Lognormal Regression: (1) at macro- level; (2) at micro- level; (3) integrated at macro- and micro- levels
Cai et al.	2018	United States	TC	●					●						●	●	●	●	●		●		○	County	TAD		Poisson-lognormal models: (1) Fixed param. univariate model; (2) Grouped random param. univ. spatial model; (3) Grouped random param. univ. spatial model with zonal factors; (4) Grouped random param. multiv. spatial model with zonal factors
Cai et al. (b)	2017	United States	TC | P-V | B-V	●						●						●	●	●	●	●	●	●			TAD		Bayesian Negative Binomial regression | Bayesian Logit regression model | Bayesian Joint model [of the two] | Elasticity analysis
Cai et al.	2016	United States	P-V | B-V	●						●	●					●	●	●		●		●	●		TAZ		Negative Binomial spatial and aspatial models (basic, zero-inflated & hurdle)
Cottrill & Thakuriah	2010	United States	P-V	●	●	○			●								●	●	●	○	●		●		EJ (CT)		Poisson Regression with heterogeneity | Poisson Regression with exogenous underreporting
Cui et al.	2015	Canada	TC (on boundary)		●											●	●							2 city areas	Neighborhoods		(1) Entropy-based histogram thresholding (2) Collision density probability distribution (3) Collision aggregation through density ratio
Delmelle & Thill	2008	United States	B-V	●												●	○	●	●	○	●		●		CT		OLS Regression | Kernel density

Study Characteristics				Dependent variables				Independent variables – parameters																Spatial aggregation approach			Analysis - Modelling approach
								Traffic				Road environment						Demographic			Socio-economic		Land Use				
Author(s)	Year	Country of study	Crash type analyzed	Crash count/frequency	Crash rate	Injury Severity	Casualty rate	Speed	Traffic volume	Vehicle distance traveled	Number of Trips - OD	Speed Limit	Curvature	Lane width	Lane number	Intersection nr./density	Roadway length	Population number/density	Road user/Population age	Modal distinction	Household/Personal Income	Employment percentage/density	Land use factor(s)	Regional level	Zonal level	Link/ segment/ intersection level	
Dong et al.	2016	United States	TC	●						●									●	●		●			TAZ		Bayesian Multivariate Poisson Lognormal Regression | Bayesian spatial-temporal interaction models
Dong et al.	2015	United States	TC	●						●	●	○				●	●				●		●		TAZ		v-Support Vector Machine with Correlation-based Feature Selector | Bayesian Multivariate Poisson Lognormal with CAR Prior
Dong et al.	2014	United States	TC	●						●	●	○				●	●	●			●		●		TAZ		Bayesian Multivariate Poisson Lognormal with CAR Prior Regression for boundary and non-boundary area models
Erdogan et al.	2008	Turkey	TC	●		●		○					○	○			●								Hotspot clusters		Poisson test | Chi^2 test | Kernel density analysis
Gomes et al.	2017	Brazil	TC	●		○										●	●		●		●		●		TAZ		Negative binomial regression | Poisson GWR | Negative Binomial GWR
Guo et al.	2017	Hong Kong	P-V	●		○		●	●		●					●	●		●	○			●		TAZ		Space Syntax | Poisson Lognormal Regression | Bayesian Poisson Lognormal with CAR Prior Regression with (1) contiguity (2) geometry-centroid distance and (3) road network connectivity
Hadayeghi et al.	2010	Canada	TC	●				●	●	●						●	●	●	●			●	●		TAZ		Poisson GWR | Negative Binomial Regression | Poisson regression
Hadayeghi et al.	2003	Canada	TC	●		○		●	●	●						●	●	●	●			●	●		TAZ		GWR | Negative Binomial Regression
Jiang et al.	2016	United States	TC | B-V | P-V	●		○				●						●	●	●	●	○	●		●		TAZ		Random Forest Models (CART trees) | Wiloxon Tests
Ladron de Guevara et al.	2004	United States	TC	●		○	○			●						●	●	●	●			●	●		TAZ		Negative Binomial Regression | Simultaneous equation estimation
LaScala et al.	2004	United States	P-V | B-V	●		○			●								●	●	●	○	●	●	●	Commu-nities	Geographic units		Linear regression models
LaScala et al.	2000	United States	P-V			●			●							○	●	●	●	○	●	●	●		CT		Spatial autocorrelation regression log-linear model
Lee & Abdel-Aty	2018	United States	B-V	●						●		●				●		●	●	●	●	●	●		ZIP		Bayesian Poisson lognormal CAR models
Lee et al. (b)	2018	United States	Crashes of 8 road user types	●	●					●						●	●	●	●	●	●	●	●		TAZ		Fractional Split Multinomial Model
Lee et al. (a)	2017	United States	TC | P-V | B-V	●		○			●							○		●	●	○	●		●	County | County Division	TAD | ZIP | TAZ | CT | BG | CB	Intersections	Mixed effects Negative Binomial models with: (1) micro-level variables, (2) micro- and macro-level variables and (3) micro- and macro-level variables with random-effects

Study Characteristics				Dependent variables				Independent variables – parameters																Spatial aggregation approach			Analysis - Modelling approach
								Traffic				Road environment						Demographic			Socio-economic		Land Use				
Author(s)	Year	Country of study	Crash type analyzed	Crash count/frequency	Crash rate	Injury Severity	Casualty rate	Speed	Traffic volume	Vehicle distance traveled	Number of Trips - OD	Speed Limit	Curvature	Lane width	Lane number	Intersection nr./density	Roadway length	Population number/density	Road user/Population age	Modal distinction	Household/Personal Income	Employment percentage/density	Land use factor(s)	Regional level	Zonal level	Link/ segment/ intersection level	
Lee et al. (a)	2015	United States	V/V-V | P-V | B-V	●						●		●						●	●	○		●	●		TAZ		Univariate and Multivariate Bayesian Poisson Lognormal with CAR Prior Regression
Lee et al. (b)	2015	United States	P-V	●						●		●				●	●	●	●	●	●	●	●		ZIP		Bayesian Poisson lognormal simultaneous equations spatial error model
Lee et al. (a)	2014	United States	V/V-V (at-fault)	●														●	●	●	●	●			ZIP		Bayesian Poisson-lognormal model
Lee et al. (b)	2014	United States	TC	●		○				●		○					●	●	●				●		TSAZ | TAZ		Brown-Forsythe test | Bayesian Multivariate Poisson Lognormal Regression
Levine et al.	1995	United States	TC	●		○										○	●	●				●	●		BG		Spatial lag regression model
Loukaitou-Sideris et al.	2007	United States	P-V	●		○			●							○	○	●	●	○	●	●	●		CT		OLS regression
Lovegrove & Sayed	2007	Canada	TC	●		○				●						●	●	●				●	●		Neighborhood - TAZ		Groups of Macrolevel Crash Prediction Models using GLMs
Lovegrove & Sayed	2006	Canada	TC	●		○		●		●	●					●	●	●			●	●	●		Neighborhood - TAZ		Groups of Macrolevel Crash Prediction Models using GLMs
Lovegrove et al.	2009	Canada	TC	●		○				●		○				●	●	●				●	●		TAZ		Groups of Collision Prediction GLMs | Modified T-tests
MacNab	2004	Canada	TC			●												●	●		●	●	●		Local health area		Bayesian spatial model with spatial autocorrelation
Naderan & Shahi	2010	Iran	TC	●		○					●							●							TAZ		Negative Binomial regression
Narayanamoorthy et al.	2013	United States	P-V | B-V	●		●											○	●	●	●	●		●		CT		Customized generalized ordered-response spatial multivariate count model
Nashad et al.	2016	United States	P-V | B-V	●						●						●	●	●		●		●	●		sTAZ		Negative binomial regression (copula-based)
Ng et al.	2002	China	TC | P-V	●		○												●		○			●		TAZ		Negative Binomial Regression with Empirical Bayes approach | Cluster Analysis
Noland & Quddus	2005	United Kingdom	TC | P-V	●		○										●	●	●		●	●	●	●		Enumeration District		Negative Binomial Regression | ANOVA
Noland & Quddus	2004	United Kingdom	TC	●		○			○							●	●	●	●		●	●	●		Ward		Negative Binomial Regression
Pirdavani et al. (a)	2014	Belgium	TC	●		○			●	●	●	●				●	●	●			●	●			TAZ		Geographically Weighted GLM | Negative Binomial Regression
Pirdavani et al. (b)	2014	Belgium	V-V P-V | B-V	●		○				●	●					●	○			●	●				TAZ		Geographically Weighted Regression (GWR)

Study Characteristics				Dependent variables				Independent variables – parameters																Spatial aggregation approach			Analysis - Modelling approach
								Traffic				Road environment						Demographic			Socio-economic		Land Use				
Author(s)	Year	Country of study	Crash type analyzed	Crash count/frequency	Crash rate	Injury Severity	Casualty rate	Speed	Traffic volume	Vehicle distance traveled	Number of Trips - OD	Speed Limit	Curvature	Lane width	Lane number	Intersection nr./density	Roadway length	Population number/density	Road user/Population age	Modal distinction	Household/Personal Income	Employment percentage/density	Land use factor(s)	Regional level	Zonal level	Link/ segment/ intersection level	
Pirdavani et al.	2013	Belgium	V-V P-V | B-V	●		○				●	●					●				○	●				TAZ		Negative Binomial regression Zonal Crash Prediction Models
Quddus	2008	United Kingdom	TC	●		○		●	●				●			●	●	●	●	○		●			Ward		Negative Binomial Regression | Spatial autoregressive model | Spatial error model | Bayesian hierarchical models for spatial units
Rhee et al.	2016	South Korea	TC	●		○				●	●	○			○	●	●	●	●		●	●	●		TAZ		OLS regression | Spatial lag regression | Spatial error regression | Poisson GWR
Siddiqui & Abdel-Aty	2012	United States	P-V (interior & boundary)	●								●				●	●	●		○		●	●		TAZ		Multivariate Negative Binomial regression | Multivariate Bayesian Negative Binomial regression for boundary and non-boundary area models
Siddiqui et al.	2012	United States	P-V | B-V	●				○				●				●	●	●		○	●	●	●		TAZ		Bayesian Multivariate Poisson Lognormal | Negative Binomial Regression
Soltani & Askari	2017	Iran	V-V	●		●												●		○			●		TAZ		Moran's I | Getis-Ord Gi* index
Tasic et al.	2017	United States	TC | V-V | P-V | B-V	●		○				●	●					●	●	●		●	●	●	●		CT		Generalized Additive Models
Ukkusuri et al.	2012	United States	P-V	●		○								●	●	●	●	●	●				●		CT | ZIP		Negative binomial regression | Negative binomial regression with heterogeneity in dispersion parameter | Zero-inflated negative binomial regression
Ukkusuri et al.	2011	United States	P-V	●											○	●	●	●	●	○			●		CT		Negative Binomial Regression with random parameters
Wang et al. (b)	2016	China	P-V	●		○										●	●	●		○			●		TAZ		Bayesian Conditional Autoregressive (CAR) models with seven different spatial weight features
Wang & Kockelman	2013	United States	P-V	●		○				●							●	●		○		●	●		CT		Multivariate Poisson Lognormal Regression with and without CAR Priors
Wei & Lovegrove	2013	Canada	B-V	●					●							●	●	●	●	●	●	●	●		TAZ		Negative Binomial Macrolevel Crash Prediction Models
Wier et al.	2009	United States	P-V	●		○			●							●	●	●	●	○		●	●		CT		Log-linear multivariate OLS regression model
Xu and Huang	2015	United States	TC	●		●		○		●		●				●	●	●			●				TAZ		Negative Binomial regression | Bayesian negative binomial model with CAR prior | Random parameter negative binomial model | Semi-parametric Poisson GWR
Xu et al. (a)	2017	United States	TC (interior & boundary)	●		○				●	●	●				●	●	●	●		●	●			TAZ		Bayesian spatially varying coefficients model

Study Characteristics				Dependent variables				Independent variables – parameters																Spatial aggregation approach			Analysis - Modelling approach
								Traffic				Road environment						Demographic			Socio-economic		Land Use				
Author(s)	Year	Country of study	Crash type analyzed	Crash count/frequency	Crash rate	Injury Severity	Casualty rate	Speed	Traffic volume	Vehicle distance traveled	Number of Trips - OD	Speed Limit	Curvature	Lane width	Lane number	Intersection nr./density	Roadway length	Population number/density	Road user/Population age	Modal distinction	Household/Personal Income	Employment percentage/density	Land use factor(s)	Regional level	Zonal level	Link/ segment/ intersection level	
Xu et al. (b)	2017	United States	TC	●					●		●					●	●	●	●	●	●	●	●		TAZ		Semi-parametric Poisson GWR | One-way ANOVA tests
Yasmin & Eluru	2016	Canada	B-V	●					●							●	●	●		●	●	●	●		TAZ		Poisson Regression | Negative Binomial regression (basic and Latent Segmentation)
Zhai et al. (a)	2019	United States	TC (interior & boundary)	●		●					●	●				●	●	●	●		●				BG | TAZ | CT | ZIP		Bayesian Poisson-lognormal models with Multivariate CAR priors
Zhai et al.	2018	United States	TC (interior & boundary)	●						●	●	●				●	●	●			●				TAZ		Bayesian Poisson-lognormal model with CAR prior

● Considered in the study design, ○ considered in the study process as filter/defining characteristic

TAZ approaches can conceptually include elements of segment approaches nested in them. An example is the study of Yasmin & Eluru (2016) that employed latent segmentation count models where TAZs are allocated probabilistically to different segments. This was in order to limit external factor impact and to classify segments within a TAZ to high- and low- risk based on empirical expected crash means. Studies have also developed models on several zonal systems for comparison purposes between them. Abdel-Aty et al. (2013) claimed that while TAZs and CBGs are equally desirable for spatial analysis, TAZs allow the examination of more transport-related factors, and thus are easier to integrate in transport contexts. Furthermore, the aggregation of TAZs into TSAZs with a rate of about 1:2 was found to be preferable for macroscopic safety modeling (Lee et al., 2014b). Cai et al. (2017a) conducted comparative Poisson lognormal models for three crash types with and without considering spatial autocorrelation effects, and recommended that CTs are better used for socio-demographic data collection, TAZs are used for transportation demand forecasting and TADs are used for transportation safety planning. Different zonal levels have also been used in conjunction for simultaneous aggregate and disaggregate modelling; it has been shown that aggregate models using ZIP codes were more volatile in parameter values and significance levels, while disaggregate CT models provided more consistent results (Ukkusuri et al., 2012). Lastly, it has been determined that separate considerations for crashes near TAZ boundaries revealed unique predictor variables (Siddiqui & Abdel-Aty, 2012), a finding worthy of examination in all spatial units.

2.1.2.3 Regional approaches

Regional areas (counties, cities, metropolitan areas, states) that are larger than the zonal ones examined above have also been implemented in the literature. Regional areas are administrative units, with often different governance laws and frameworks than their neighboring areas, as is often the case in US states. In the US, entire Metropolitan Statistical Areas (MSAs) have been used for the National Household Travel Survey, which has provided data for pedestrian trips (Lee et al., 2019a). The benefit of using regional units can lie in the interpretation of model results and possible evaluation of risk factors or road safety interventions, such as legislation changes. For instance, a study by Song et al. (2006) applied Bayesian multivariate spatial models in county-level data in Texas, and results indicated that eastern Texas counties had higher crash risks than western Texas counties, with less safe sites being near large city conglomerations. Studies have examined road safety indicators at the level of geographic units formed from communities (LaScala et al., 2001; 2004), at the city level (Moeinaddini et al., 2014), at the metropolitan area level (Bu et al., 2018), at the county level (Noland & Oh, 2004; Song et al., 2006; Erdogan, 2009; Huang et al., 2010; Li et al., 2013) or similarly at the state level (Atubi, 2012).

Regional-wide crash modification factors (CMFs) have also been developed for a single change affecting the traffic environment uniformly, e.g. for legal changes in some U.S. States or across the entire country (Lee et al., 2017b; 2018a), however this approach does not take spatial effects explicitly into account. As the area size increases, it is important to remember that unobserved heterogeneity is more difficult to capture, due to multiple unobserved parameters being introduced in the occurrence of events; as Wang et al. (2016b) state, it becomes more difficult to capture spatial trends and problems in a larger area. If differences in comparable units between remote areas such as different countries are taken into account, it is reasonable to assume that transferability of results for macroscopic spatial analysis is far from seamless. In a study seeking to examine transferability of results across regions of different countries (from US counties to Italian provincias) Lee et al. (2019b) employed negative binomial models using data from both countries and calculated the respective transferability indexes and calibration factors. Models for total crashes and bicycle crashes were transferable from Italy to the US; the opposite, however, was found to be untrue for most study areas. In addition, no model for pedestrian crashes was found to be transferrable between the two countries. It is important to note that this statistical disagreement emerged even while several significant variables were common across the two countries, and without accounting for spatial effects in the models of the study.

Reviewed studies that primarily focus on spatial analyses at the zonal level are shown on Table 2-3.

2.1.2.4 Conditional approaches

Apart from defined zones, conditional approaches have been adopted. As conditional is hereby defined any approach that does not utilize any of the previous segment, zonal or regional approaches but a more rigid ruleset set by researchers. An example is fix-distance grid structures, such as 0.1 square mile grids (Kim et al., 2006), 1 square mile grids (Ossenbruggen et al., 2009) and multiple grid sizes from 1 to 100 square miles (Cai et al., 2017a). While the impacts of grid-based characteristics on crash counts have been proven to be statistically significant, a grid of a particular size might be improper for certain areas, depending on spatial distributions of safety-related parameters (Kim et al., 2006).

An example of approaches that are conditional not by area, but by crash circumstance, are link-based approaches that utilize crash-mapping algorithms and assign crashes to each road segment, and assuming that the crashes happening on the same link have the same underlying conditions, which might not always be the case. Link-based approaches can be problematic in providing interpretable results, however. Conversely, crashes can also be grouped by pre-crash conditions, regardless of their actual location, for the purposes of spatial analyses. Pre-crash conditional approaches have appeared to be more transferable overall (Imprialou et al., 2016).

Reviewed studies that primarily focus on conditional spatial analyses are shown on Table 2-4.

2.1.2.5 Integration of different areal units

The aforementioned integration of characteristics of the corridor level to road segment or intersection level analysis by several studies (Zeng and Huang, 2014; Alarifi et al., 2017; 2018) is a considerable achievement in road safety. In these studies, the levels of analysis can be considered to be close in geographical characteristics (i.e. a segment is similar to a corridor). There have been other endeavors, however, to integrate factors from units of more different scales in spatial analyses, such as zonal-level characteristics to segment-level analysis.

Table 2-3: Studies with road safety spatial analyses primarily on the regional level

Study Characteristics				Dependent variables				Independent variables – parameters																	Spatial aggregation approach	Analysis - Modelling approach
								Traffic			Road environment							Demographic			Socio-economic			Land Use		
Author(s)	Year	Country of study	Crash type analyzed	Crash count/frequency	Crash rate	Injury Severity	Casualty rate	Speed	Traffic volume	Vehicle distance travelled	Number of Trips - OD	Speed Limit	Curvature	Gradient	Lane width	Lane number	Intersection nr./density	Roadway length	Population number/density	Road user/Population age	Modal distinction	Household/ Personal income	Employment	Land use factor(s)	Regional level	
Aguero-Valverde	2013	Costa Rica	TC	●		●				●								●	●	●		●			Canton	Full Bayes hierarchical approach Poisson multivariate CAR model for spatial random effects.
Aguero-Valverde & Jovanis	2006	United States	TC	●		○				●								●	●	●					County	Negative Binomial Regression | Full Bayesian hierarchical models
Atubi	2012	Nigeria	TC	●		○												●	●						State	Multivariate linear regression
Bu et al.	2018	United States	TC	●		●			●			●				●			●						Metropolitan areas	Simple Density distribution analysis
Erdogan	2009	Turkey	TC		●	●	●											●	●		●		●		County	Moran's I and Geary's c values, Z and G statistics
Flask & Schneider	2013	United States	MC	●		○							●	●		○		●	●	●		●			County | Township	Bayesian Negative Binomial Regression with mixed effects
Han et al.	2018	United States	TC	●					●							○	●	●							County (spec. road type)	Bayesian hierarchical random parameter model | Bayesian hierarchical random intercept model | Bayesian Poisson lognormal model
Huang et al.	2010	United States	TC	●		●			○	●							●	●	●	●	○	●	●	●	County	Bayesian Spatial CAR Priors regression
LaScala et al.	2001	United States	P-V			●	●		●								●	○	●	●	○	●	●	●	Communities	Spatial autocorrelation regression log-linear model
Lee et al. (a)	2019	United States	P-V		○	○	●				●								●	●	●	●		●	Metropolitan areas	Multiple linear regression model integrated in a Poisson Lognormal Model
Lee et al. (b)	2019	Italy, United States	TC | P-V | B-V	●															●	●	●	●			County | Provincia	Negative Binomial Regression | Calibration factors | Transferability Indexes
Lee et al. (a)	2018	United States	TC	●		○																			State	Crash Modification Factors
Lee et al. (c)	2018	United States	P-V | B-V	●		○					●								●	●	●	●			Metropolitan areas	Bayesian integrated and non-integrated Bivariate Models
Lee et al. (b)	2017	United States	MC	●		○													●			●	●	●	County | Parish	Before-and-After Study (1) with Comparison Group | (2) With Empirical Bayes | Safety Performance Functions | Crash Modification Factors
Li et al.	2019	United States	TC	●		○				●						●			●	○		●	●	●	County	Hierarchical Bayesian random parameters models (structured and unstructured spatio-temporal effects)
Li et al.	2013	United States	TC	●		○				●	●							●	●	●		●	●		County	Negative Binomial Regression | Poisson GWR
Liu and Sharma	2018	United States	TC	●		●				●												●	●	●	County	Hierarchical Bayesian random parameters models (struct/unstruct r.eff.)
Moeinaddini et al.	2014	20 Cities	TC	●		○											●	●							City	Gamma-distributed GLM
Noland & Oh	2004	United States	TC	●		○			●				●		●	●		●	●			●			County	Negative Binomial Panel Regression
Song et al.	2006	United States	TC	●		○				○			●				○								County	Bayesian Multivariate Poisson Lognormal Regression with and without CAR Prior
Zhai et al. (b)	2019	Hong Kong	P-V			●											●	●		●	●				City	Binary & Mixed logit models with and without variable interaction terms

● Considered in the study design, ○ considered in the study process as filter/defining characteristic

Table 2-4: Studies with road safety spatial analyses primarily by conditional approaches

Study Characteristics				Dependent variables				Independent variables – parameters																				Spatial aggregation approach			Analysis - Modelling approach
								Traffic				Road environment							Demographic		Socio-economic		Land Use				
Author(s)	Year	Country of study	Crash type analyzed	Crash count/frequency	Crash rate	Injury Severity	Speed	Traffic volume	Vehicle distance traveled	Number of Trips - OD	Speed Limit	Curvature	Gradient	Lane width	Lane number	Intersection nr./density	Roadway length	Population number/density	Road user/Population age	Modal distinction	Household/ Personal	Employment percentage/density	Land use factor(s)	Zonal level	Link/ segment/ intersection level	Condition-based level	
Bao et al.	2019	United States	TC	●	●	●			●							●	●	●		○			●			Multiple grids (approx. to ZIP areas)	Convolutional Neural Network augmented with a Long Short-term Memory Network
Bíl et al.	2013	Czech Republic	TC	●												○	●								Rural segments	Rural road network split into fundamental segments	Network Kernel Density Estimation with significance verification
Cai et al. (b)	2019	United States	TC	●					●		●				●	●	●			●			●			9-mi^2 grid structure divided to smaller cells	Convolutional Neural Networks (GLM and Artificial Neural Networks for benchmarking purposes)
Cai et al. (a)	2017	United States	TC | P-V | B-V	●		○			●		●					●	●		●	●				TAD | TAZ | CT		Multiple grids from 1 to 100 mi^2	Multivariate Poisson Lognormal Regression with and without spatial autocorrelation
Chung et al.	2018	United States	TC	●		○			●								○									Areas within 20 mi of 2271 weather stations	Categorical analysis (sensitivity, positive predictive value, Cohen's Kappa) | Negative Binomial Regression
Imprialou et al.	2016	United Kingdom	TC	●		○	●	●				●	●		●		●								Rural & Highway segments	Pre-crash conditions	Bayesian Multivariate Poisson Lognormal Regression
Kim et al.	2006	United States	TC | V-V | P-V | B-V	●														●		○		●	●			0.1-mi^2 grid structure	Negative Binomial Regression | OLS Regression
Loo et al.	2011	China	V-V | P-V	●										○			○			○					Urban & suburban segments	Urban and suburban network split into fundamental segments	Network Kernel Density Estimation
Mohaymany et al.	2013	Iran	TC	●										○			○	●							Rural segments	Rural road split into fundamental segments	Network Kernel Density Estimation
Ossenbruggen et al.	2010	United States	TC	●		○		●	●	●																1-mi^2 grid structure	Homogeneous Poisson process spatial testing
Xie et al.	2017	United States	P-V	○		●			●	●								●	●	●	●	●	●			300×300 feet^2 grid structure	Linear Regression Model | Tobit Model | Potential for Safety Improvement
Xie and Yan	2008	United States	TC	●										○			○									Urban network split into fundamental lixels	Network Kernel Density Estimation

● Considered in the study design, ○ considered in the study process as filter/defining characteristic

As stated before, the zonal level has become a promising medium during the more recent years for the exploration of new approaches of spatial analyses. Zonal factors, such as Vehicle Miles Traveled (VMT), are considered to be shared by segments of both segments and intersections of the same zone. It has been hypothesized that both observed and unobserved heterogeneity at the zonal level would influence crash frequency at both segments and intersections inside these zones. Cai et al. (2018) investigated crashes at the TAD level across three counties to determine the influence of any observed and unobserved zonal factors. Results indicate that including zonal factors improve model performance for both segment and intersection crash frequency prediction.

Another concept is incorporating macro-level variables into micro-level safety analysis. This has been attempted by Lee et al. (2017a) across seven areal units of varying sizes for intersection crashes. They determined that accounting for macro-level variables and introducing macro-level random-effects leads to models of better performance than the baseline, though performance varies when using data of different areal unit size. Additionally, there have been endeavors to link crash counts of micro- and macro-levels through their spatial interaction (Cai et al., 2019a). A spatial interaction matrix was created based on whether a road segment (micro-level) was inside a zone (macro level), and an adjustment factor was introduced to bridge the different estimates of expected crashes that would occur for the two levels. Once again, following an integrated approach increased model performance; moreover, the determination of both macro- and micro-level risk factors that influenced crashes were possible, as well as crash hotspots on both levels.

Conversely, road-level factors have been shown to influence safety by varying effects across regions, and can be considered to be correlated with unobserved heterogeneity, to an extent. To demonstrate this, a dedicated study examined specifically urban two-lane roadway segments in 34 counties in Florida, US. Regression coefficients of Poisson lognormal models and hierarchical models were found to fluctuate considerably for crash counts across the examined counties (Han et al., 2018). However, neither factors at the regional level nor spatial correlations at the microscopic level were taken into account in that particular study.

Huang et al. (2016) investigated a possible bridging of the macro- and micro-level approaches for an integrated crash prediction and hotspot identification approach. Crashes were analyzed both jointly at the micro-level (road segment/intersection level) and at the macro-level (TAZ level). The authors developed both a micro-level Bayesian spatial joint model and a macro-level Bayesian spatial model; as expected, the models included different statistically significant variables. Results reaffirmed the known model merits: micro-level modelling provided more informative and precise insights for directly improving road safety, while macro-level modelling allows for incorporating safety improvements in long term transportation planning. The authors acknowledge that TAZs may have unobserved scale and zonal effects and further, the boundary issue – explained in the following – needs to be accounted for.

2.1.2.6 Boundary problem and Modifiable areal unit problem

Apart from conducting studies across many different areal levels and bridging aspects and attributes of different spatial levels, researchers have also shown interest on how to define areas and areal units and how to treat events on their boundaries. The boundary problem, or boundary effect, refers to the manner in which crashes recorded on (or very close to) the borders of neighboring study areas are allocated and treated in statistical analyses. Fotheringham & Wegner (1999) claimed that neighboring zones influence crashes close to the borders of areal units. Since then, several studies have explored the problem, each proposing a solution. Delmelle and Thill (2008) mention simple solutions such as (1) assigning the locations as they were assigned by police records, (2) double-counting boundary crashes or (3) apportioning crashes, dividing the counts per neighboring zones.

Separate predictor sets have been prepared for boundary and interior pedestrian crashes per TAZ, introducing buffer zones around 2-D borders. This mutually exclusive separation and modelling within a hierarchical Bayesian framework has led to increased model fit. However, this approach was adopted due to the limited distance travelled by pedestrians, and accounting for additional road user types might differ due to higher amounts of areal units that are typically crossed (Siddiqui and Abdel-Aty, 2012). Instead of using a fixed buffer zone, Cui et al. (2015) introduced an entropy-based method applied on histogram thresholding, to obtain a variable buffer zone size. The crash density probability distribution was then calculated, and boundary crashes were aggregated into neighborhoods. The case study resulted in 6m and 9m buffer zones for central areas and south areas in Edmonton, Canada, respectively. The authors concluded that the entropy-based method was precise when compared to ground truth data, though more variables are required to verify this finding; especially traffic-related variables such as speed and traffic volume.

An alternative was proposed by Zhai et al. (2018), who adopted an iterative data aggregation approach to compensate for the boundary effect. The reasoning behind this method was the division of each zone into boundary and interior, the development of a crash prediction model for each zone based on interior crashes only, the aggregation of crashes based on crash model predictions, the assignment of boundary crashes to each zone based on the proportions of expected interior crashes, and, as a last step, re-run the prediction model until convergence. The crash assignment based using the CAR Poisson Lognormal Bayesian Spatial Model. It is notable that the impact of several independent variables were found to be influenced by the boundary effect in the case study in Florida, US. Both Cui et al. (2015) and Zhai et al. (2018) demonstrated that certain analytical approaches outperform conventional rules such as the various ratio methods that split boundary crashes based on numerical rules or exposure parameters). It is also worth noting that certain Bayesian statistical models can express the interaction of neighboring zones on crashes close to zone boundaries via the utilization of corresponding spatial weights (e.g. Wang et al., 2016b).

The modifiable areal unit problem (MAUP) occurs when boundaries are changed inside the study areas, causing possible influences on the statistical models and resulting inferences (Openshaw, 1984). The issue is particularly present in road safety when area boundaries are arbitrary or malleable, without any hard geographical borders, such as administrative areas or grids. Two studies did experiment with the discrepancies caused by MAUP on different aggregation levels (Ukkusuri et al., 2012; Abdel-Aty et al., 2013). While the areas which provided more accurate predictions were determined, no uniform solutions were proposed. When outlining MAUP, Xu et al. (2018) outlined four potential solutions. These were: (1) using disaggregate data as possible (2) capturing the spatial non-stationarity, which refers to capturing local space variation for each explanatory variable, (3) designing optimal zoning systems, an approach which presents its own limitations and (4) conduct sensitivity analysis for MAUP effects specifically.
A recent study has empirically highlighted the important effects of MAUP on four different zonal configurations using an identical dataset (Zhai et al., 2019a). It was determined that the impact of MAUP was significant on parameter estimates, model assessment and hotspot identification. Larger zones, such as CTs and ZIP codes led to models of higher predictive accuracy in that study. It has also been considered that the zonal systems may have inherent limitations by Lee et al. (2014b), who developed ten new zonal systems to tackle both the boundary and the MAUP problems. The Brown-Forsythe homogeneity of variance test was implemented to obtain the optimal zonal scale, which was found to be at the custom TSAZ level, as zones cannot be scaled up indefinitely to reduce boundary crash percentages. However,

the authors state that the boundary issue still needs to be accounted for in TSAZs, and that further research on additional crash types such as non-motorized (VRU) crashes is needed.

2.1.2.7 Examination of spatial proximity structures

A critical point that attracts researcher interest is the creation of different spatial proximity structures and the examination of the effects these structures have on model performance and fit. Various spatial proximity structures have been formulated both at the microscopic and macroscopic levels. Regarding the microscopic level, Aguero-Valverde & Jovanis (2010) concluded that by including route information in the neighboring structure, especially in a simple neighboring structure (direct adjacency), model performance is improved.

Regarding the macroscopic level, Dong et al. (2014) evaluated crash prediction models at the TAZ level using four different types of spatial proximity structures (0–1 first-order adjacency, common-boundary length, geometry-centroid distance, and crash-weighted centroid distance). The best model fit was provided when weighting the common-boundary length of neighboring TAZs, though cross-zonal spatial correlations was identified as present in crash occurrence for all four different configurations. The authors comment that the inclusion of all possible spatial correlations increases model complexity, thus resulting in decreased prediction performance.

Moreover, Alarifi et al. (2018) sought to investigate spatial weights configuration for a hierarchical spatial proximity structure, including intersection-, road segment- and corridor-level parameters. The authors examined four different types of conceptualization of spatial relationships and calibrated 13 Bayesian hierarchical Poisson-lognormal joint model with spatial effects. The adjacency-based first-order model (where directly adjacent road entities and feeding road entities are considered for each segment) was among the best performing models and once again significant variables were found in all configurations for all unit levels. The authors suggest that the sensitivity of AADT in the models is a matter for further investigation.

Another sophisticated approach was the utilization of the space syntax technique for modelling street patterns. Space syntax acknowledges the configuration of the urban grid itself is responsible for generation of movement patterns (Hillier et al., 1993), though its exact use for deriving certain route choices has been challenged in the past (Ratti, 2004). Guo et al. (2017) considered simple geographical proximity as inadequate to properly describe spatial relationships of crashes. Rather, they sought to integrate road network characteristics in a zonal level examination. They used space syntax to quantify road network structures in Hong Kong through three main parameters on the TAZ level: (1) connectivity, (2) local integration and (3) global integration. After calculating global integration for three road network patterns (grid, deformed grid and irregular), it was determined that global integration was positively related with increased pedestrian-vehicle crashes. Furthermore, the more structured patterns featured the highest global integration values, thus irregular patterns were found to be the safest, followed by deformed grids and lastly (regular) grids.

2.1.2.8 Further topics of areal unit analysis

In spatial analysis, study designs sometimes appear to be data-driven, conducted where there is availability of information instead of intuition or previous experience. Availability of data does not necessarily imply its fitness for use in studies. As an indication, weather data measured from stations may or may not describe the situation at crash sites accurately. A study was conducted to evaluate the effectiveness of coverage of weather stations for use in spatially analyzing traffic crashes (Chung et al., 2018). Hourly data which are observed from land-based stations was contrasted with data from fatal crash databases. Through categorical analysis, sensitivity, positive predictive value, and Cohen's Kappa were examined, and it was determined that there were agreements of data in rain and snow weather conditions but not in fog, which displayed a 91% rate of false alarm. The authors suggest that fog may present higher spatio-temporal sensitivity as a parameter. While the weather station data was found adequate overall for use in crash analyses, the finding regarding the fog parameter ought to make researchers carefully consider possible data sources for their studies.

Furthermore, instead of analyzing crashes collectively in each areal unit, or treating them as separate variables, different crash categories can be examined while taking their interactions into account. A study by Lee et al. (2018b) analyzed the proportions of crashes of each vehicle type at the TAZ level, using a fractional split multinomial model. The fractional approach ensures the summation of crash proportions of all categories to 100%, thus forcing interactions between each category. Findings showed considerable differences as to which variables were statistically significant for each vehicle type. Moreover, the spatial distribution of hot zones varied considerably per vehicle type considered. On that matter, hotspots have also been found to vary temporally. Soltani and Askari (2017) conducted a spatial autocorrelation analysis of crashes and hotspots at the TAZ-level in Iran. Moran’s I and Getis-Ord Gi* methods were used, and were found to provide significant clustering. The authors examined crashes based on location, time of day and injury severity, which is a very rare combination of parameters. This time, hotspots were found to vary considerably across the various times of day. Another important finding is that zones located at intersections connecting other zones were identified as clusters with high crash rates. Despite the hotspot identification, however, no other explanatory characteristics were introduced in the analysis. It appears thus reasonable to assume that the identified hotspots may vary considerably if certain elements are introduced to a study or omitted from it.

2.1.3 Modelling approaches

This section provides a brief overview of the various modelling approaches implemented so far in the literature of spatial analysis in road safety. A multitude of tools have been developed that endeavor to predict road safety indicators (Lord & Mannering, 2010; Mannering & Bhat, 2014) and explain spatial correlation and unobserved heterogeneity and to incorporate the effects of various spatial characteristics that are difficult to be represented individually. Several studies have been testing various advanced models against simpler ones for performance assessment (e.g. Miaou & Song, 2005; Chiou et al., 2014; Dong et al., 2016; Aguero-Valverde et al., 2016; Cai et al., 2019b).

Multivariate models are found to have better goodness-of-fit and precision due to correlation between dependent variables, such as crashes of different severity levels while accounting for spatial correlation (Barua et al., 2014) or simultaneous crash frequency and severity examination (Chiou et al., 2014). The benefits of multi-level data have been discussed in spatial analyses, for instance the multilevel structural hierarchy proposed by Huang & Abdel-Aty (2010) combining driver-level and site-level data with geographic region characteristics.

Spatial analyses often test for spatial autocorrelation or heterogeneity of events, and also consider size and structure for the various research areas and spatial units of analysis in the adopted approaches. For the precise examination of autocorrelation phenomena, various geo-spatial statistics have been adopted by scientists for decades, such as Moran's I, Local Moran's I, and Getis-Ord-Gi* statistics.

2.1.3.1 Generalized Linear Models

Generalized Linear Models (GLMs) have been used extensively in the road safety literature for decades, since they assume crashes are independent, random and sporadic countable events (Hauer et al., 1988; El-Basyouny & Sayed, 2009). Poisson and NB models they are the most common application forms of GLMs in spatial analyses in road safety. Poisson models are known to fail to calculate over-dispersion of data, which is often found in crash analyses, thus Negative Binomial (NB) models can be used to circumvent that limitation. There are well-known disadvantages when implementing GLMs to analyze road safety data.

There are assumptions that GLMs cannot capture underlying common unobserved effects, which can be remedied with random-effects models, as described in relevant literature (Lord & Mannering, 2010). When analyzing data over continuous time variables, a frequent problem is the rarity of crashes, with uneventful periods appearing as zero count. Thus Zero-inflated GLMs have also been applied to circumvent the problem, such as those utilized by Qin et al. (2005).

Both Poisson and NB models are known to ignore heterogenous impacts of variables, possibly skewing estimates. To circumvent this, researchers have been implementing clustering techniques (e.g. Karlaftis & Tarko, 1998) as well random parameter count models (e.g. Ukkusuri et al., 2011). Additional solutions can include calibration of latent-class models for crash counts (Yasmin & Eluru, 2016) and applications of different levels of classification via multivariate approaches (e.g. Wang et al., 2011) or hierarchical approaches (e.g. Yannis et al., 2007). These models can eventually become quite advanced, such as the EMGP model by Chiou and Fu (2013), further advanced by Chiou et al. (2014), which originated as an extension of the multinomial-Poisson regression model with added error components, to which spatial correlation effects were also added.

While GLMs in their basic form are aspatial, they can be extended to incorporate spatial effects in their structure, eventually becoming quite advanced. An example is the EMGP model by Chiou & Fu (2013), further advanced by Chiou et al. (2014), which originated as an extension of the multinomial-Poisson regression model with added error components, to which spatial correlation effects were also added. Better predictions have been obtained from GLMs including random effects rather from fixed effects, and from GLMs including zonal factors as opposed to those not including them (Cai et al., 2018).

2.1.3.2 Geographically Weighted Regression

A method that accounts for spatial variation is the simultaneous development of several localized models using Geographically Weighted Regression (GWR). First proposed by Fotheringham et al. (2002), these models extend the traditional regression framework to allow for a continuous surface of parameter values, with measurements at points that indicate the spatial variability of such a surface. A number of road safety GWR analyses have been published (Hadayeghi et al., 2003, 2010; Pirdavani et al., 2014a; 2014b; Rhee et al., 2016; Gomes et al., 2017; Liu et al., 2017). As Pirdavani et al. (2014b) note, GWR models offer explanatory and descriptive power and provide intuitive results that enable researchers and stakeholders to investigate varying effects of explanatory variables on crash occurrence throughout the study areas.

Gomes et al. (2017) compared the performance of GWR extended in a GLM context and highlight that Geographically Weighted Negative Binomial Regression (GWNBR) is appropriate for spatially analyzing crash data while accounting for their over-dispersion. Additionally, GWNBR models significantly reduced the spatial dependence of model residuals. GWNBR models were also utilized by

Liu et al. (2017) to produce localized models at the roadway segment level, without restrictions by jurisdiction boundaries. The variation of three calculated parameters (intercept, AADT and segment length) was found to be substantial in highway segments across Virginia, US, though the effects of several factors remain to be examined. Additionally, the introduced parameter of segment length is present in spatial structures, which might introduce bias to GWNBR estimations. The authors comment that GWNBR models are highly localized, thus the transferability of their predictions is limited and need to be reapplied to each area.

Xu & Huang (2015) extended GWR to semiparametric GWR (S-GWR), which combines geographically varying parameters with geographically constant parameters. Although their composite approach outperformed a random parameter negative binomial (RPNB) model, the authors claimed that S-GWR models are not transferable spatially, and that each region would need to develop separate S-GWR models (a common conclusion with the GWNBR method). S-GWR was compared again with RPNB by a study conducting crash analysis across six spatial units and three injury severity levels (Amoh-Gyimah et al., 2017). Again, results indicated that S-GWR performed better than the RPNB overall, based on mean absolute deviation (MAD) and Akaike Information Criterion (AIC) metrics, and had increased prediction accuracy. On the other hand, RPNB displayed increased sensitivity when examining the effect of variation of spatial units on unobserved heterogeneity compared to S-GWR. It should be noted that the latter study did not examine any geometrical characteristics such as segment length or intersection density.

S-GWR has also been employed to investigate possible correlations between jobs-housing balance and road safety, since disruptions in that balance have been found to lead to reduced road network efficiency (Xu et al., 2017b). The authors converted jobs-housing ratio to a categorical variable and then applied S-GWR models at the TAZ level. Considerable spatial variations were discovered for different jobs-housing ratio categories, through elasticity analysis of the model results for each jobs-housing ratio category. However, the study did not compare the S-GWR results with those of another baseline model.

2.1.3.3 Autoregressive prior models

A common problem in geographical studies with spatial dataset can be the selection of the appropriate size and scale units for analyses. This has a direct impact on results, as experience suggests that increasing granularity (i.e. spatial resolution) can weaken correlations between output areas and introduce spatial autocorrelation issues (Loo & Anderson, 2015). To counter this, studies have introduced spatial autocorrelation effects (e.g. Aguero-Valverde & Jovanis, 2006, 2008; Guo et al., 2010; Flask & Schneider, 2013; Chiou et al., 2014) or temporal autocorrelation effects in crash count models (e.g. Wang & Abdel-Aty, 2006). The respective models often use CAR or SAR with the former being more frequently implemented in road safety spatial analyses. A seminal study by Besag et al. (1991) presented a normal distribution for spatial autocorrelation effects using a CAR prior, which has been implemented in many studies since (e.g. Huang et al., 2016; Cai et al., 2018; Zhai et al., 2018; Wen et al., 2019).

CAR models have been found to perform better than Poisson models and Multiple Membership models (where higher level units are formed by each unit and its adjacent neighbors), by explaining a high degree of spatial heterogeneity and by being more lenient in spatial variable omission (El-Basyouny & Sayed, 2009). However, Yasmin & Eluru (2016) note that considering spatial autocorrelation effects and latent segmentation simultaneously can be analytically challenging. Autoregressive models can also be developed within a Bayesian Framework as shown in Aguero-Valverde et al. (2016); CAR models have been found to be convenient to compute while using a Gibbs sampler in the Bayesian inference (Huang et al., 2010). Bayesian CAR models have been shown as capable to function with a variety of customizable spatial weights (Aguero-Valverde & Jovanis, 2010; Alarifi et al., 2018). These weights can be calculated based on several different bases (e.g. by geometric distance of zone centroids or by land use type). Of these weight sets, it is natural that some will outperform others for a specific study configuration, though not always in the expected manner, as shown by Wang et al. (2016b), where a simple 0-1 configuration based on proximity outperformed land use type- and intensity-based weights for pedestrian crash prediction (population was used as exposure parameter for pedestrians only, without a corresponding parameter for vehicles).

2.1.3.4 Bayesian modelling

The process of Bayesian inference has led to the development of several interesting methodologies during more recent years. Bayesian hierarchical joint models have been developed in various complexities using regression and regression methods for parameter estimation, possibly with regression splines, as shown in an early Bayesian approach by MacNab (2004). Moreover, multivariate Bayesian models are capable of estimating excess crash frequencies at different severity levels in the same spatial analysis unit (Aguero-Valverde, 2013). Bayesian hierarchical joint models have been shown to highlight significant variables at both micro and macro levels while accounting for spatial correlations between entities (e.g. in Cai et al., 2019a). Such an application by Wang & Huang (2016) determined higher AADT, more lanes and accesses for segments on the micro level, signal control, more intersection legs, and higher speed limit for segments for intersections on the micro level and higher road network and trip generation densities as significant risk factors, among others.

As studies often report, models with Bayesian approaches have been found to perform consistently better than their non-Bayesian counterparts (e.g. Miaou & Song, 2005; Siddiqui et al., 2012; Wang & Huang, 2016). Bayesian models with CAR effects have been shown to simultaneously account the spatial correlation and uncorrelated heterogeneity present in aggregated crash count data, and to reveal more significant variables with the same signs as frequentist modelling (Quddus, 2008). However, Bayesian models are not without drawbacks, as a main strength of their applications is reduced in cases without any solid basis of prior knowledge (uninformed priors). Furthermore, they require a considerable amount of calibration cases (sometimes mentioned as burn-outs) which leads to some loss of information and might require considerable computational time and power to obtain.

A noteworthy development is the recent investigation of spatiotemporal heterogeneity using multivariate hierarchical Bayesian models across injury severity categories. Relevant studies have endeavored to capture data heterogeneity with spatial and temporal effects, with the hierarchical framework serving to predict crash counts of different severities simultaneously. Spatial and temporal components are specified with several structured and unstructured components, and random effects can be inserted in the models to address the underlying data structure. Specifically, Ma et al. (2017) aggregated crash counts from 100 homogenous US highway segments into injury/no injury crash categories using high temporal resolution (daily intervals). They identified vehicle-distance travelled and some geometric characteristics as significant crash predictors, as well as variables that are more sensitive temporally, such as wet pavement and average speed.

In a recent study by Liu and Sharma (2018) examining injury crashes, both spatial and temporal effects were bound to be important in approximately the same magnitude across spatial, temporal and spatio-temporal structures. Crash frequencies showed significant spatial, but not temporal, autocorrelations. Similarly, Li et al. (2019) mentioned the issues of spatio-temporal instability in crash data, apart from the typical unobserved heterogeneity that is inherent to data collection. They calibrated Bayesian random parameters models (with both structured and unstructured spatio-temporal effects) which show that daily VMT, proportion of males, unemployment rate and education are found to positively increase crash frequency and are normally distributed across crash severities for crashes related to substance consumption. There have been studies where the application of spatially structured and unstructured effects were separated, such as in a series of spatial analyses conducted for road crashes and fatalities in Greece which also took into account maritime connections on a county-level (Papadimitriou et al., 2013).

2.1.3.5 Empirical Bayes and Full Bayes methods

Since several decades, Empirical Bayes (EB) methods have been implemented in road safety by contrasting crash counts of a road segment with sites with comparable true crash risk, which are the reference population. EB estimations have displayed better predicting capabilities and eliminate regression to the mean issues than Naive before-after comparisons (Hauer, 1997; Geurts, & Wets, 2003). EB methods have been also used in a before-after study in complementarity to a before-after study with a comparison group in order to obtain more reliable CMFs (Lee et al., 2017b).

Further to that direction, Full Bayes (FB) extended models can be used to account for heterogeneity due to unobserved road geometric characteristics, traffic characteristics, environmental factors and driver behavior (El-Basyouny & Sayed, 2011; Ma et al., 2017). The FB approach has also been shown to be more reliable empirically in hotspot identification compared to EB (Huang, 2009). The advantage of FB over EB is that it takes into account that model parameter estimates include an amount of uncertainty and can provide a quantitative measure of said uncertainty (Miaou & Lord, 2003). The FB approach is the basis of several recent developments discussed in the following.

2.1.3.6 Alternative Prior Distributions

Apart from the widely used CAR model, other approaches can be implemented to account for spatial effects in models through different prior distributions. Mitra (2009) adopted a hierarchical Full Bayes spatial model to investigate the presence of possible influences of spatially structured factors on injury crashes at intersections. The reasoning behind such an approach is an attempt to capture both heterogeneity from spatial effects (implying a common global structure) and excess heterogeneity (originating from spatially unstructured effects). The first level of the hierarchy is a Poisson-lognormal specification. The Poisson rate then included the typical intercept and covariates, and also two separate effect terms, spatially structured and unstructured, to capture spatial and excess heterogeneity respectively. The spatially structured effects used a multivariate normal joint prior. Results indicated considerable spatial autocorrelation effects at the intersection level, while a comparison with aspatial Negative Binomial regression revealed similar coefficient estimates but increased model precision.

A similar jointly-specified approach was adopted by Aguero-Valverde (2014), to determine the effective range after which no lingering correlation is found at the road segment level. The Poisson rate function featured one parameter for heterogeneity among segments, using a normal distribution, and one for spatially correlated random effects per segment, using a jointly specified prior. Additionally, a temporal indicator for the evolution of crashes in years in covariate values and predicted crash counts was included. Ultimately, the joint prior model outperformed a random-effects model and a CAR prior model and the effective range was determined (at about 168m). The author states that the manner in which distance is measured (e.g. Euclidean distance, ground route distance or any other way) also has an impact on model predictions.

A different form is the Full Bayes Multiple Membership (MM) spatial model proposed by El-Basyouny & Sayed (2009). The approach includes similar spatially structured and unstructured effects as the previous studies. In addition, MM models consider each site as a member of a higher-level unit that contains its nearest neighbors. They also include a parameter measuring the strength of association between structured and unstructured spatial effects. The authors further extended MM models by adding an additional component to allow for variance in the values of crash risks and characteristics between mutually exclusive corridors. When tested, the extended MM model slightly outperformed a CAR model, which in turn outperformed a basic MM model, though the overall DIC metrics showed quite close values.

Xu et al. (2017a) introduced another methodological alternative in the form of a very detailed Bayesian spatially varying coefficients approach, based on the hierarchy proposed by Huang and Abdel-Aty (2010). The process again started with a Poisson function in a Full Bayesian framework, and the parameters were modelled using a CAR prior. The innovation of the study lied in the utilization of a single set of random effects ranging from purely unstructured to purely spatially structured effects; this simultaneous process is considered superior by the authors, however it features a mathematical structure that is quite complicated.

2.1.3.7 Spatial spillover effects

An emerging aspect of spatial analyses is the examination of spatial spillover effects. Spatial spillover effects are the effects that exogenous observed variables have on the dependent variable at both the target and the neighboring locations. Spatial spillover effects differ from spatial autocorrelation (or error correlation) effects, which entail unobserved exogenous variables at one location affecting dependent variables at the targeted and neighboring locations (Narayanamoorthy et al., 2013; Cai et al., 2016; Lee et al., 2018b).

Past studies have utilized spatial lag regression models in an effort to capture spillover effects. LaScala et al. (2000) and Quddus (2008) converted count variables into continuous approximations for their analyses. They then used an explanatory variable in the expression of a spatially lagged dependent variable to form a spatial autoregressive (SAR or spatial lag) model.

Cai et al. (2016) included spatial spillover effects in the examination of pedestrian and bicyclist crashes. Via the application of dual-state GLMs, it was determined that taking observed spatial spillover effects into consideration results to models with better performance consistently. The zero-inflated negative binomial models were found to have the best fit for pedestrian and bicycle crashes, though unobserved spatial autocorrelation effects were not simultaneously examined in the study. To evaluate the impacts of significant factors, marginal effects were calculated as well.

In addition, Wen et al. (2019) aimed to capture both spatial autocorrelation and spillover effects using a hybrid model. The hybrid model featured the traditional Poisson-lognormal basis. The authors expressed spatial autocorrelation effects as the CAR prior and spillover effects as exogenous variables of neighboring road segments. Homogeneous highway segments were used for the analysis. Both of spatial autocorrelation and spatial spillover effects were found to be significantly correlated with the respective crash data. This hybrid approach yielded better estimates than both of its individual components, with coefficients that showed lower standard deviations. The authors suggest that accounting for spatial heterogeneity may further refine the model, but a much more complex structure would be required.

2.1.3.8 Machine learning & Deep learning approaches

Given their popularity as a powerful, data-driven family of prediction tools, machine learning (ML) methods have been implemented for spatial and spatio-temporal road safety analyses. Indicative methods used in road safety spatial analyses are outlined below. ML methods can operate with increased degrees of freedom without requiring traditional assumptions as regression models do, and are more resilient to data outliers. They are methods typically used in conjunction with big data in transport and road safety.

Random forest (RF) models are collections of numerous superimposed decision trees that emerge from a selection and validation process, as described in Chang and Wang (2006). RF models have been used in road safety studies by researchers. For instance in Jiang et al. (2016) the feasibility of RF models for ranking hot-zones on a TAZ level and identifying critical parameters for crash occurrence when utilizing big data was investigated. Road network distribution (density) and socio-economic features such as school enrollment and car ownership percentages were found as the most statistically significant variables for crash occurrence. The study concludes that RF models provide classification with about 80% accuracy in hotspot identification.

Support Vector Algorithms (SVM) have been successfully implemented as alternatives to traditional statistical-regression modelling. In a relevant study, SVMs were employed together with a coactive neuro-fuzzy inference system (CANFIS) algorithm (Effati et al., 2015). SVMs were found to be considerably better performing when examining crash injury severity, especially when utilizing a radial basis kernel function (RBF). The researchers propose the enhancement of spatial analyses with machine learning algorithms as the key to unveiling significant factors affecting crash injury severity while accounting for spatial correlation and heterogeneity effects. The study of Dong et al. (2015) implemented SVMs as a tool for handling big and complex data structures. They examined zone-level crash prediction while taking spatial autocorrelation into account, and SVMs were found to perform better when including a spatial weight feature with an RBF kernel as opposed to SVM models. SVMs have been also used in conjunction with Bayesian methods, though, to the author's knowledge, not yet in a spatial analysis framework; for instance, Wang et al. (2019) used Bayesian logistic regression to detect factors contributing to highway ramp crashes.

Latest technological progressions make neural network implementation much more feasible than past years. Bao et al. (2019) utilized a deep learning approach for short-term crash risk prediction for crash risk on an urban level. They augmented a convolutional neural network (CNNs) with a long short-term memory network in order to examine variables that varied spatially, temporally or spatio-temporally, proposed by earlier research for traffic speed and congestion prediction (Ma et al. 2015a; b). Weekly, daily and hourly prediction models with varying spatial grids were produced as a result. The authors mention that prediction performance of the proposed model decreases as the spatiotemporal prediction outcome resolution increases towards the hourly level. It is noteworthy that machine learning models exhibited better performance on the daily level, while benchmark econometric models generally performed better on the weekly level, suggesting that neither approach is clearly superior. Another interesting application is described in Zhu et al. (2018); the CNNs developed in the study take into account spatio-temporal network and traffic structure. However, they are used for traffic incident detection/identification, and not road safety prediction or causation analysis.

Cai et al. (2019b) explored that research direction by applying CNNs for road safety prediction by collecting and utilizing high-resolution data: 3mile x 3mile grids with crash counts and data, each grid containing 100×100 cells with width and height of 158.4 feet, examined in 17 layers of data matrices. By feeding data of a higher resolution into a CNN, the authors allowed variables to fluctuate across locations more freely, thus increasing the model accuracy. It was stated that the hierarchical structure enables better understanding of the circumstances of crash occurrence. While the authors demonstrated a viable approach for crash prediction, it is obvious that extra effort is required for the creation of this high-resolution grid and the complementing database. Some variables might be readily available for calculation in high-resolution or inferred via the existing road geometry (such as segment lengths), while others may be harder to obtain in case of missing data (such as land uses). Approaches such as CNNs might require custom, tailor-made data collection frameworks in order to provide their full potential, as the authors suggest. Furthermore, no specific framework is established for assigning the values of required hyperparameters during the CNN training phase.

2.1.3.9 Kernel Density Estimation

Another crash and hotspot analysis method is kernel density estimation (KDE), which allows the generalization of incident locations to an entire area. It should be noted that this is not a direct analytical method, but rather an interpolation technique (Anderson, 2007) mainly used for the identification of clustering patterns of traffic collisions. KDE can be advantageous in predicting the spread of crash risks, though the kernel radius has been a matter of debate in several scientific fields (e.g. Raykar & Duraiswami, 2006; Hart & Zandbergen, 2014). It appears that bandwidth determination influences the outcome of the hotspots (Fotheringham et al., 2000; Anderson, 2009; Loo & Anderson, 2015). Furthermore, the fact that KDE treats discrete events as a continuous area effect can be presented as a limitation (Anderson, 2009). Erdogan et al. (2008) conducted an analysis of hotspot clusters in a province of Turkey and utilized KDE together with a repeatability analysis of hotspot crashes for a decade. The authors reported considerable overlap of the outcomes, though KDE determined less hotspot locations overall. An interesting approach by Mountrakis & Gunson (2009) investigated the development of KDE spatially (determining varying density peaks among roads) and temporally (determining an exponentially increasing trend with annual periodicity and a seasonal cyclic component) for animal-related crash hotspots in Vermont, US.

Kernels are projected over 2-D spaces, while road crashes usually occur in a 1-D linear area, which most road environments approach, as Xie and Yan (2008) note. In order to overcome this discrepancy, KDE has been expanded to network KDE approaches, in which the network is represented as fundamental units of equal network length (termed lixels). Xie and Yan (2008) investigated this method and how fundamental lengths and regular kernel bandwidth affect its performance for road crash prediction. They conclude that network KDE describes crash densities and network borders more precisely than regular KDE, and that lixel length appears more important than Kernel function selection. However, Loo et al. (2011) implemented network KDE in areas of varying land use and found that kernel bandwidth critically affects the spatial distribution of resulting density estimates. Furthermore, wider bandwidths appeared to be more appropriate for non-urban areas where crash density is lower.

Similarly, Mohaymany et al. (2013) applied network KDE to a rural road in order to determine hazardous segments; apart from static spatial autocorrelation of crashes they also investigated its temporal evolution through a three-year period. Bíl et al. (2013) also used KDE in a 1-D area by separating the network into sections. They explored an alternative venue for better refining KDE results by providing a method to test their statistical significance. The proposed method utilized relative spatial positions of crashes and roadway length to calculate kernel strength, which allows detection and prioritization of the most hazardous locations, which included classifying clusters with values above the 95th percentile of the kernel density function as hazardous.

2.1.4 Vulnerable Road Users

In road safety, vulnerable road users (VRUs) include pedestrians, bicyclists and other road users who are often children, elderly, people with impairments and disabilities. Due to their vulnerability to injuries or fatalities compared to vehicle users, VRUs have increased safety needs. The use of spatial analyses, or approaches in a spatial context, to examine aspects of road safety concerning VRUs warrants specific examination. A notable example is the study of Tasic et al. (2017) which investigated crashes involving vehicles and VRUs by using models that accounted for spatial correlation effects. Data was aggregated on a CT level for a large array of about a hundred variables for vehicle-only, pedestrian and bicycle crashes. The data were analyzed using an extension of GLMs, Generalized Additive Models (GAMs), which included a two-dimensional smooth function to account for spatial correlation. A remarkable finding was that the expected pedestrian or bicyclist crashes increased less than proportionally with the exposure variables of vehicle, pedestrian or bicyclist trips, confirming the safety-in-numbers effect on a macroscopic level while accounting for spatial correlation effects.

Analyzing pedestrians' walking exposure and crashes in an integrated manner was proposed in a dedicated study on the MSA level (Lee et al., 2019a). For estimating exposure, multiple linear regression models were calibrated, followed by a Poisson-lognormal regression model for fatality estimation using the estimated exposure as input. Walking hours was determined as the best performing exposure variable. The proposed integrated model outperforming the non-integrated ones. Spatial correlation of trips was not investigated in the study, however, and pedestrian safety features were not examined either. VRU exposure, in the form of trips, has also been estimated at a macroscopic level in an integrated manner. These trip numbers were used to calibrate VRU crash prediction models in a study across 23 Metropolitan areas, and it was found that estimated exposure (VRU trips) led to models with calibrated performance compared to observed exposure for both pedestrians and cyclists (Lee et al., 2018c).

Pedestrian crash hotspots have been examined through spatial processing of their respective costs using big data from multiple sources such as taxi trips and social media (Xie et al., 2017) by employing a grid structure divided in higher resolution cells, similar to Cai et al. (2019b). Crash costs were assigned to cells using a kernel density estimation function, and sites were identified using tobit models with potential safety improvements (PSIs) and ranked as potential hotspots based on the potential of pedestrian crash cost reduction. The authors claim that their method can be transferred to less populated regions by adjusting kernel bandwidths.

Pedestrian crashes do not necessarily occur in the zone of residence of the pedestrians involved; Lee et al. (2015b) sought to identify zones where pedestrian crashes occur, and zones where pedestrian crashes originated from. Using different exposure variables, a variation of a Bayesian lognormal model with Poisson structure was applied. The occurrence of crashes with pedestrian involvement was revealed to be significantly affected by more location-related factors, while pedestrian origin was revealed to be significantly affected by more demographic-related factors. A similar concept of investigating both ZIP codes of crash locations for bicyclists and the number of crash-involved bicyclists in their ZIP of residence was explored in a study by Lee & Abdel-Aty (2018). Bayesian Poisson lognormal CAR models were used to examine bicycle crashes, and the contributing factors were not identical in each case. For instance, increases in the number of schools per mi2 were only found to lead to increases in bicycle crashes in the crash location ZIP. Conversely, lower income areas were found to be a contributing factor overall through the significance of many related variables. Again, PSI was used to identify VRU crash hotspots in both studies.

A noteworthy finding is that of Siddiqui et al. (2012), who produced Bayesian models for pedestrian and bicyclist crashes at the TAZ level, noting the necessity of accounting for spatial correlation while examining VRU crashes at the macroscopic level, which is also corroborated by Guo et al. (2017). In addition, spatial spillover effects have also been examined in a VRU context, as mentioned before (Cai et al., 2016).

Apart from methodological and modelling approaches, the influence of parameters for pedestrian crashes have also been examined in high resolution. Specifically, the effects of weather conditions have been investigated using GIS within a spatial context (Zhai et al., 2019b). Binary and mixed logit models were used in the study, in a basic form and in a more advanced form including terms of interaction between weather conditions and risk factor variables. Both high temperatures and precipitation were found to be associated with pedestrian crashes of increased severity. Hotter weather and the presence of rain were also found to exacerbate the effect of risk factors, such as jaywalking or unsafe driver behavior.

2.1.5 Discussion on spatial approaches

2.1.5.1 Summary of findings

The examination of the studies that was carried out in this research has led to some noteworthy conclusions for spatial analyses in road safety. It appears that a multitude of different approaches and modelling methodologies has been adopted in the literature, with a trend towards advanced Bayesian models and methods in the past decade. This has led to the development of powerful tools that provide accurate predictions for crash counts per area with increasingly complex model configurations. However these approaches also lead to a lack of a common established methodology or framework to compare results of spatial analyses. Additionally, this finding does not imply that more traditional functional/econometrics methods, such as GLM models or GWR are not found useful still, at least for benchmarking purposes. Functional models appear to be more straightforward in their interpretation and assessment of results. In both cases, results of spatial studies have also been reported to have limited transferability as well.

Recently, machine learning approaches have come to challenge the dominance of Bayesian models by being implemented alongside or instead of them. It should be noted that these are mostly data-driven approaches, which have also been reported as containing inherently biased samples, especially when examining big data (e.g. Bao et al., 2017; 2019). While the aforementioned transferability issues are mostly solved with machine learning methods, there are often difficulties in the interpretation of results: A commonly cited example is the hidden layers of neural networks and the meaning of each contributing factor. Approaches such as SVM are subpar in determining the significance of revealed patterns in the data they examine or the utility each variable offers in prediction tasks.

Further on the results of spatial studies, another important finding is the revelation of sensitivity of hotspot locations. Researchers have shown that hotspots are radically different across users of different vehicles and ages, and that hotspots display significant variation throughout the time of day. It can be reasonably surmised that many elements that are introduced to an analysis radically change the hotspot map. Naturally, the employed methodologies also affect the final outcome of spatial studies. Researchers should be vigilant and try to convert unobserved factors into observed ones, in order to receive more substantial and precise hotspot maps.

Though studies have been published internationally, spatial analyses have been more common in more modernized and developed countries (especially USA), while developing countries are considerably less represented. The use of different sizes of spatial units as basis for spatial analyses has been examined extensively, and it appears that apart from information and data availability, spatial areas of each size have different advantages and disadvantages. Several studies include exposure parameters in order to establish a common baseline for crash risk comparisons between models (Imprialou et al., 2016). When exposure parameters such as road length, AADT and vehicle distance travelled are examined, they are found to increase crash risk overall, as expected, however there are particular cases where these results might not apply or even be reversed (e.g. Dong et al., 2014).

It has been demonstrated that the parametrization of the spatial correlation term, namely, its inclusion as a variable in models, can aid in situations where data are scarce or difficult to obtain. Its use can be further expanded, however, as a complementary feature to even variable-rich models, in order to explain parts of variation in the data.

That being said, data availability remains a critical issue, and lack of consistent data across a respectable duration of time can be a critical obstacle in conducting spatial and spatio-temporal analysis. Spatial analyses in road safety appear data-driven most of the time, stemming from the drive of researchers to prove or test a concept. There are variables that have not been extensively tested due to lack of data, for instance pavement condition. Similarly, there are study areas that merit more attention, such as extensive urban network environments formed by roads of lower categories.

Traffic speed does not appear to be as frequently used as in past decades, though speed limits are taken into account as network characteristics, rather than traffic characteristics. Moreover, it can be observed that certain geometrical features seem to be used less frequently, such as road gradient, curvature and lane width. As an indication, the 'gradient' column on **Table 2-2** was blank at the end of the reviewing process and was thus removed. This decline in use can be attributed to missing data for many study areas, or to difficulty in data acquisition. Another reason may be the lower prioritization of geometrical features from researchers: studies often seek to include crash data, traffic data, socio-economic data, demographic data and land-use data. Therefore traditional road geometry data examination is receiving less attention in comparison to past decades.

2.1.5.2 Future research directions and challenges

This section outlines research directions that do not appear to be adequately investigated from the present literature of road safety spatial analyses and can constitute meaningful future research endeavors. An important aspect that was does not appear to be adequately investigated is that of micro-level road safety and event analysis with spatial modelling considerations. A small number of studies has been found to explore concepts such as automated conflict extraction via trajectory analyses using automated data (Saunier and Sayed, 2007; St-Aubin et al., 2015). The inclusion of spatial effects in such design concepts would be very interesting for the determination of the influence of spatial effects at a small-unit level.

While crash counts have been examined extensively, their distributions over several categories have received less focus within a spatial context. The recent fractional approach by Lee et al. (2018b) that examines crash distribution across vehicle types is an example towards that direction, as is the examination per crash type proposed by Aguero-Valverde et al. (2016). Nonetheless, more research is needed on the manner in which various categories of crashes occur across study areas. The distribution of exact crash proportions and the factors that affect them needs to be researched within a spatial context.

For instance, injury severity distributions have not been investigated as frequently as crash counts; rather, they have mostly been used as a categorization mechanism. By jointly examining crash severities and occurrence while taking spatial effects into account, more informative results can be reached for practitioners. Similar potential exists for studies aiming to examine casualty rates. In addition to the previous, it would be interesting to spatially analyze other road safety indicators, such as those related to driver behavior: conflicts, near-misses, harsh events and traffic law violations. These can aid in determining high crash concentrations and locations of poor road safety performance (hotspots).

Hotspot detection, or problematic region identification in greater scales, is a crucial advantage typically provided by spatial analyses for locating problems. Therefore, the determination of the spatial impacts of implemented road safety measures would also be very beneficial. Before-after studies within a spatial context (or even a spatiotemporal context, if a dedicated data collection scheme can be set) would allow observation of crash reductions due to targeted observations from the initial analyses. Such study designs would also allow the examination of the variation of spatial autocorrelation of events (and whether any exists) before and after interventions, and would offer interesting insights in any possible crash mitigation phenomena. Another promising research direction is the transfer and application of more focused spatial analysis methods for the examination of segments of a contiguous road network, similar to network KDE approaches, so that segments are assessed instead of areal units, but in the form of an extended and complex road network, as an expansion of the segment analysis approaches that were previously mentioned.

Some spatial issues, while proven to exist, need to be further analyzed to increase comprehensiveness. The specific effective range of spatial correlation among analysis units, as studied by Aguero-Valverde (2014) and Wang et al. (2016b) needs to be expanded upon. Again, there is a need for results for different road environments, road users, crash types and injury severities in order to obtain measures of the extent that spatial dependency needs to be accounted for. In addition, different countries are expected to produce varying results, possibly due to differences in driving culture or other unobserved factors.

Another direction that would increase the low transferability of results of spatial analysis is the creation of common frameworks for the two famous problems (boundary and MAUP), preferably on the international scale. The establishment of an acceptable boundary value in order to address boundary issues under different conditions, as suggested by Zhai et al. (2018b), is such an example. More effort is needed to be devoted to understanding the impacts of both the boundary issue and MAUP across areal unit sizes as well, especially if different contributor variables are found in boundaries. Similarly, methods to obtain more homogeneous road segments or areal units need to be developed, in an effort to reduce heterogeneity. They would have to be comprehensible and straightforward in order to be more widely accepted and applied by practitioners worldwide.

Yet another finding from the reviewed studies is that built environment is not very strictly defined in the sense that every study selects some of its characteristics to examine. In a dedicated study, Ukkusuri et al. (2012) include in the term built environment factors such as land use patterns, population characteristics such as age profiles and professional driver percentages, road infrastructure and transit characteristics. This review section has not exhausted all built environment parameters, and the investigation of more specific variables such as the presence of refuge islands or crosswalks or proximity to health or education buildings merit additional investigation, and can be a future direction of targeted road safety spatial analyses.

These endeavors can all be further augmented by new technological developments, such as transport applications of big data, cloud computing and connected & autonomous vehicle technologies that can be used to provide a more connected spatial environment (e.g. as in Bao et al., 2018). For instance, it has been found that smartphone technology sampling can provide a vast amount of driving data in real conditions, including risk factors such as distraction and speeding (Papadimitriou et al., 2018), while achieving a seamless transition from data collection to data analysis (Yannis et al., 2017). This framework could enable not only a collection of a wealth of real-time information across several spatial unit levels, but also allow for easier calibration of spatial models without the doubt of transferability that is often present in spatial analyses.

2.2 Meta-regressions of exposure parameters used in spatial analyses in road safety

2.2.1 Introduction

In order to further suppress crash occurrence, it is critical to examine road crashes while taking into account as much information as possible, and proximity and relative position within the road layout, namely the dimension of space, are an important aspect of that information.

As past literature has indicated, in order to establish a common baseline for crash risk comparisons between models, it is informative to include at least one exposure parameter (Imprialou et al., 2016). The exposure variable can be either introduced to the statistical model as another independent variable or it can be already included in the data (for instance when analyzing crash rates normalized by vehicle-distance travelled instead of crash counts). When analyzing road crash frequencies, three of the most prevalent exposure parameters are traffic volume, vehicle-distance travelled and roadway length, though other variations have been also utilized, such as road network density (also per road type or speed limit) or trip generation.

The examination of studies that use exposure parameters as independent variables can offer interesting insights in road safety spatial analyses, as influences on exposure variables can heavily skew study results. By taking influencing parameters into consideration at the process of study design and establishing a common framework, result transferability can be improved.

Apart from the classic modelling approach of independent/dependent variables, rate-based models have been also developed in the literature – (e.g. Cottrill & Thakuriah, 2010). This approach incorporates exposure as an independent variable with its parameter estimate constrained to one. This is also achieved in count data models via the inclusion of an offset term, which has a parameter estimate constrained to one. The use of offset terms have been debated by researchers, with some support for the constraint of offset coefficients to one and others adopting a more unconstrained approach, as effects may be inelastic in some circumstances (for instance, congested conditions where crashes increase at decreasing rates with respect to traffic volume). As such, the current research focuses on classic modelling approaches.

This section includes sections providing the methodological outline of meta-regression, and afterwards three sections are provided, one for each exposure parameter. Therein, a brief overview of the literature results for each exposure parameter is provided, to allow for a brief introduction for each parameter. Afterwards, the respective meta-regression results are shown. These results provide insights on which study characteristics influence the coefficient values of the exposure variables which in turn predict crash outcomes.

The present section can be considered as a quantitative continuation of the previous rigorous and extensive literature review of 132 spatial road safety studies conducted in Section 2.1, focusing on exposure parameters.

2.2.2 Meta-regression methodology

The aforementioned common framework can be established by the application of meta-analytic and meta-regression techniques to road safety studies with spatial analyses. The methodology of meta-analysis can be used to qualitatively combine the results of a number of input studies using the inverse-variance technique. A rich theoretical background for meta-analyses and applications in transport studies is available in the literature (Elvik, 2001; 2005; Van Houwelingen et al., 2002; Viechtbauer, 2010; Caird et al., 2014a; 2014b; Elvik & Bjørnskau, 2017; Theofilatos et al., 2017; 2018a; Ziakopoulos et al., 2019).

In the present research, a meta-analysis application was explored but ultimately was not found to be possible. This is due to the fact that a meta-analysis requires similar sampling frames, comparable methodologies (developed models etc.) and dependent variables. The examined studies display large dissimilarities in sampling frames, as they investigate different regions, and their dependent variables comprise a multitude of crash variations on several crash severities, as listed by Abdel-Aty et al. (2013). The most inhibiting factor, however, is considered to be the different methodologies/models that have been proposed in the literature, as spatial analyses is a fertile field for advanced statistical methods to be used, which unfortunately limits any aggregation attempts in a meta-analysis. Methodological differences can constitute reason for study exclusion as shown by Roshandel et al. (2015).

These limitations can be circumvented in the case of meta-regression, which offers further explanation of the heterogeneity in the existing effects reported in the literature. A recent study by Elvik & Goel (2019) underlines that meta-regression can be used to identify sources of differences in coefficient estimates of studies. The authors employed this method to identify factors that explain the large heterogeneity of estimates. They determined stronger safety-in-numbers effects for pedestrians than for motor vehicles and cyclists, and stronger safety-in-numbers effects at the macro level (e.g. a city) than at the micro level (e.g. in junctions).

In a meta-regression, effects such as study characteristics are assessed for their influence on coefficient estimates, as aggregated information can describe the differences between studies (Van Houwelingen et al., 2002). However, this method has the drawback of not providing direct model estimators, but rather outline the effects that influence existing the estimates of existing models. The inverse variance technique is utilized herein, which considers an overall estimate (or summary mean) of effects based on n input estimates as proposed by Elvik (2001):

$$\bar{Y} = \frac{\sum_{i=1}^{n} Y_i * W_i}{\sum_{i=1}^{n} W_i} \qquad \text{Eq. (1)}$$

Where:

- i is the number of input studies ($i = 1,2, \dots, n$)
- $\bar{Y}$ is the overall estimate (or summary mean) of effects
- W_i are the statistical weights such that:

$$W = \frac{1}{SE_i^2} \qquad \text{Eq. (2)}$$

Where:

- SE_i^2 is the standard error for coefficient i

In the present framework, Y, which is the dependent variable, expresses the overall estimate of the coefficient of each exposure parameter. If the Y_i are considered as the observed effects in the i-th study, θ_i as the corresponding true effects and ε_i as the corresponding sampling error following a normal distribution, then:

$$Y_i = \theta_i + \varepsilon_i \qquad \text{Eq. (3)}$$

The inverse variance technique allows two model specifications: (i) the fixed effects model and (ii) the random effects model. Fixed effects models provide results as an overall estimate of the included study sample, while random/mixed effects models assume the used sample of studies are a random part of a greater group of effects. In other words, the main target of fixed effects analysis is to provide a conditional estimate exclusively from those studies provided in the meta-analysis. On the contrary, mixed effects are considered as random samples of a greater set, therefore inferences made from them are unconditional (Viechtbauer, 2010).

For meta-regression, fixed effects models use the following structure:

$$\theta_i = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_k x_{i,k} \qquad \text{Eq. (4)}$$

Where:

- θ_i are the true effects of model coefficients
- β_i are potential influencers of the true effect of model coefficients
- $x_{i,1}$ is the value of the independent variable $j, (j = 1, 2, \ldots, k)$ in study i

In this case, the independent variables also known as moderator variables, are the different study characteristics of each study such as the areal unit of analysis.
Mixed effects models can better account for potential heterogeneity between studies, using the previous structure while adding a representative random effects term u_i.

$$\theta_i = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_k x_{i,k} + u_i \qquad \text{Eq. (5)}$$

Where:

- u_i is the random effect term following a normal distribution $(u_i \sim N(\mu, \tau^2))$.

As Theofilatos et al. (2017) state, τ^2 is the amount of residual heterogeneity (the variability among true effects that cannot be explained by the moderators entered in the meta-regression model). Obviously, if $\tau^2 = 0$, then the θ_i are homogenous and there is a reversion to a fixed effects model. In order to determine the proper model specification, the Q-test is used to verify the presence of systematic variation between results provided by studies.

The Q-test for meta-analysis/regression is a form of Cochran's Q-test. The Q-test is a non-parametric statistical test to verify whether a number of factors have identical effects. In other words, the null hypothesis is that there is no systematic variance in the selected group of studies, and fixed-effects models can be used. As Elvik (2011) states, the Q-test follows a chi-squared distribution with $g - 1$ degrees of freedom, where g is the combined number of estimates, which are used to determine its significance.

Furthermore, Viechtbauer (2010) notes that the Q-test usually keeps better control of the Type I error rate and therefore should be preferred for hypothesis testing over likelihood ratio tests.

If the Q-test is significant, the variance between studies is larger than would be expected on the basis of the within-study variation, and the use of mixed-effects models over fixed-effects models is warranted. The utility of the Q-test extends to within-study heterogeneity; namely the possibility that several of the effects reported in the same study are strongly heterogeneous with each other. In that case, random effects are included in the equations to allow for meta-regression.

$$Q = \sum_{i=1}^{n} \left[W_i * Y_i^2\right] - \frac{\left(\sum_{i=1}^{n} W_i * Y_i\right)^2}{\sum_{i=1}^{n} W_i} \qquad \text{Eq. (6)}$$

Lastly, a funnel plot can be used to visualize results of meta-regressions by showing the symmetry of the estimate value on the horizontal axis vs. the reported standard errors on the vertical axis, and can aid in detecting possible publication bias (Elvik & Bjørnskau, 2017). The term publication bias refers to the exclusion of relevant studies from meta-analyses, which reduces their robustness. These studies might have not been published or have counterintuitive effects (Høye & Elvik, 2010).

On the processing part, meta-regressions in this doctoral dissertation were conducted in R-studio using the metafor package and following Viechtbauer (2010). From the value of the t-statistic standard error values could be obtained, provided that the beta-coefficients are known, using the common conversion for regression testing:

$$t = \frac{\hat{\beta}_i - \beta_{i,0}}{se(\hat{\beta}_i)} = \frac{\hat{\beta}_i}{se(\hat{\beta}_i)} \qquad \text{Eq. (7)}$$

A common reference framework was also established, transforming for common units of roadway length (miles), logarithmic and non-logarithmic estimates were transformed on the same scale and similar adjustments were made (e.g. AADT was examined per lane). If there were more than one suitable models reported from each study, only the one with the best reported fit was included (by assessment of AIC, AICc or similar indicators). This is because a particular participant should only contribute data once when calculating the observed outcomes (Viechtbauer, 2010). Thus the funnel plots display the adjusted coefficient estimates plotted by the respective adjusted standard errors.

An equally important decision was the exclusion of results of studies conducting Bayesian modelling, as they utilize fundamentally non-frequentist approaches (posterior distributions, rather than parameter estimates, Bayesian credibility intervals rather than frequentist confidence intervals and so on). The meta-analysis and meta-regression methods considered do not currently offer robust ways of integrating Bayesian and non-Bayesian study results. This decision was supported by the process reported in the study of Roshandel et al. (2015) as well.

Similarly to study assessment, meta-regression models with the lowest AICc values are considered to accrue minimum information losses and they are the ones selected. Meta-regression attempts were made on all detected studies for combinations of the moderator variables (study characteristics) that were reported in each case.

Therefore the established criteria for inclusion of a study in the meta-regression are:

1. The study is published in a scientific reference source (journal or conference, in English)
2. The examination of the considered parameter by the study with functional-econometric statistical models in a logarithmic format.
3. Correlation of the parameter with road crashes by the study (as opposed to injury severity). This is completed by the reporting of beta coefficient.
4. Reporting of the respective standard error in order to acquire the corresponding sampling variances – essential as per Viechtbauer (2010).

2.2.3 Meta-regression on traffic volume/AADT estimators

In studies conducting spatial analyses, traffic volume (often used as AADT) has been found to be positively correlated with higher overall crash risk (Wang & Huang, 2016) and with higher non-local driver crash risk. Higher traffic volume has been found to be positively correlated with both severe and property damage only collisions (Barua et al., 2014). Interestingly, AADT was found to be negatively correlated with local driver crash risk (Wang et al., 2016), with the authors of the study suggesting that local drivers cope better with higher driver conditions compared to foreign ones. In intersections, Huang et al. (2017) found that major road AADT positively contributes to crash occurrence at a significant level for motor vehicle, bicycle and pedestrian crashes (minor road AADT was found positive as well, but not statistically significant).

After examining the literature with the established 4 criteria, 4 spatial analysis studies contributed to the meta-regression for AADT with a total of 8 effects (Abdel-Aty & Wang, 2006; Wang & Abdel-Aty, 2006; Wier et al., 2009; Lee et al., 2017a). The transformed coefficient values used as input for the meta-regression are provided in a forest plot format on Figure 2-1.

The Q-test for Residual Heterogeneity was not found to be statistically significant ($Q_{[df = 6]}$ = 4.1614; p-value = 0.6548), suggesting no considerable heterogeneity among the true effects. Therefore, there is justification for using the fixed-effects meta-regression model. The outputs of the fixed-effects meta-regression appear on Table 2-5. Both study characteristics are treated as binary variables, based on whether they were considered in an input study or not, the latter being reference categories. For clarification, it is again noted that the estimates provided here are moderator variable (independent variable) impacts on regression coefficients and not the direct effect of AADT/traffic volume on crash occurrence.

Table 2-5: Parameter estimates of meta-regressions coefficients for the effect of AADT estimators on crash occurrence

Moderator Variable	Estimate	Standard Error	p-value
Speed Limit [ref. cat.: 0]	1.6479	0.6297	0.0089
Age [ref. cat.: 0]	1.8031	0.6031	0.0028

In practice, these coefficients denote that in studies which take the speed limit into account, the beta coefficient for AADT is increased by 1.65 on average. Similarly, in studies which take road user age into account, the beta coefficient for AADT is increased by 1.80 on average.

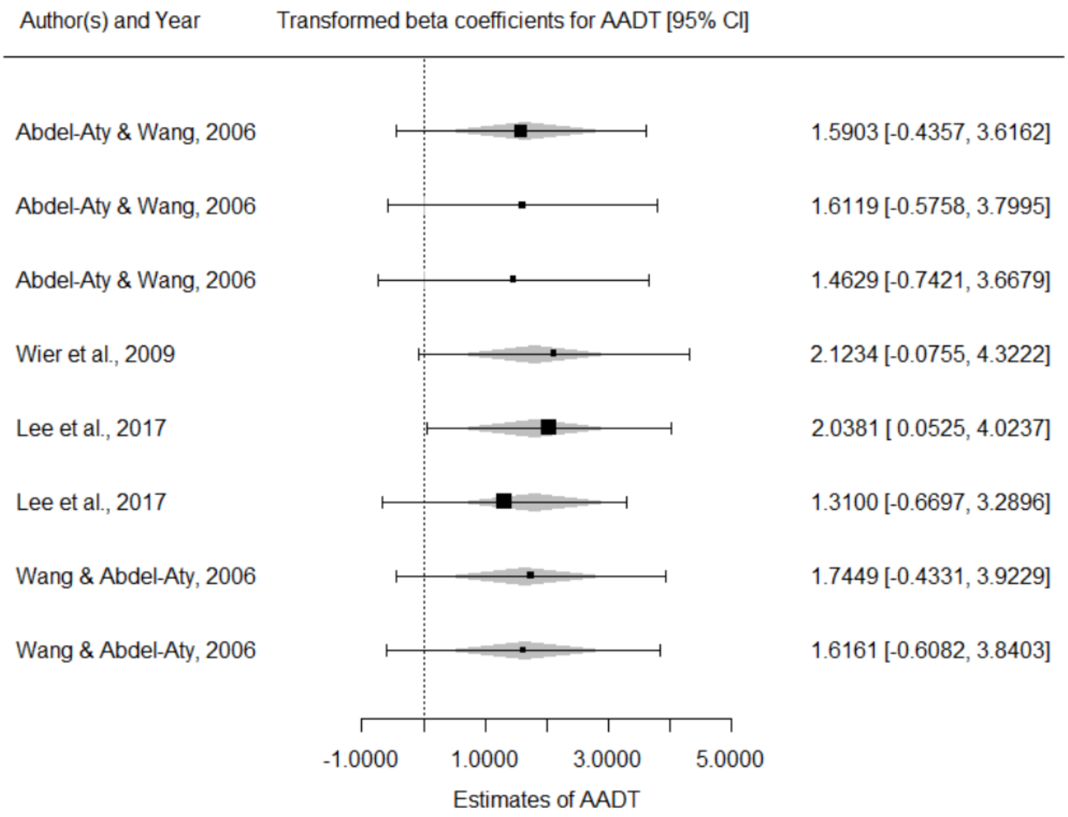


Figure 2-1: Forest plot of AADT beta coefficients on crash occurrence

The respective funnel plot is shown on Figure 2-2. The test for funnel plot asymmetry was not statistically significant (z = 1.9580; p-value = 0.0502), which suggests no indication of publication bias amongst the studies, though with a small statistical margin. Results indicate that from all considered study characteristics, the main moderator variables (study characteristics) affecting the overall estimate of traffic volume on crash occurrence are the examination of the presence of a speed limit and road user age. More specifically, the impact that AADT has on crash occurrence is increased if researchers consider the speed limits present in the study areas (as opposed to not considering them). A slightly higher impact of AADT on crash occurrence is found if researchers consider the age categories of road users present in the study areas (as opposed to not considering them).

Additional study characteristics that were considered but where not found to be statistically significant for AADT were the types of dependent variable (with categories: total crashes, pedestrian crashes and rear-end crashes), modal distinction (with categories: total crashes, motorized vehicle crashes, vulnerable road user-vehicle crashes and pedestrian-vehicle only crashes), regional approach (with categories: intersections and census tracts) and the examination of the number of lanes, all as defined in the respective studies.

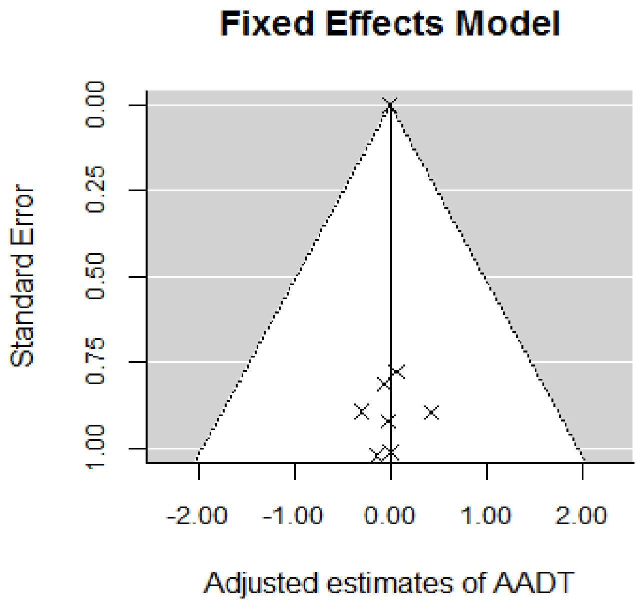


Figure 2-2: Funnel plot of AADT beta coefficients by standard errors

2.2.4 Meta-regression on roadway length estimators

In studies conducting spatial analyses, roadway length is one of the most traditional exposure variables. Increased roadway length has been found to significantly and positively contribute to slight, serious and fatal crashes (consistently with different models) in segments of an English motorway (Wang et al., 2009). Noland & Quddus (2005) reported that minor road length did not have an effect on serious injuries (and even decreased slight injuries), while it increased serious injury occurrence in roads of a higher category ('B' roads). Abdel-Aty et al. (2011) developed spatial models for 1349 Traffic Analysis Zones (TAZs) in Florida and determined that roadway lengths with higher speed limits (e.g. 45 & 65 mph) were positively correlated with increased crash frequency and severity in general, while lower speed limits (e.g. 25 mph) were negatively associated with crash frequency during peak hours.

Considering vulnerable road users (VRUs), Nashad et al. (2016) note an increase to crash likelihood involving VRUs if sidewalk lengths are increased in a zone, indicating a transfer of effect across transport modes. A similar result is reported in Wang & Kockelman (2013), albeit via a highly non-linear, two-stage relationship. Lastly, a study in Canada developed advanced urban models that revealed that bicycle-car collisions are directly associated with total lane and bicycle lane kilometers (Wei & Lovegrove, 2013).

After examining the literature with the established 4 criteria, 7 spatial analysis studies contributed to the meta-regression for road length with a total of 29 effects (Hadayeghi et al., 2003; Noland & Oh, 2004; Quddus, 2008; Cottrill & Thakuriah, 2010; Atubi, 2012; Yasmin & Eluru, 2016; Gomes et al., 2017). Transformed coefficient values used as input for the meta-regression are provided in a forest plot format on Figure 2-3.

The Q-test for Residual Heterogeneity was statistically significant ($Q_{[df = 20]}$ = 39.2066; p-value = 0.0063), suggesting considerable heterogeneity among the true effects. Therefore, there is justification for using the mixed-effects meta-regression model. The outputs of the mixed-effects meta-regression appear on Table 2-6. It is noted that the estimates provided here are moderator impacts on regression coefficients and not the direct effect of roadway length on crash occurrence. (KSI crashes: Killed and Serious injury crashes).

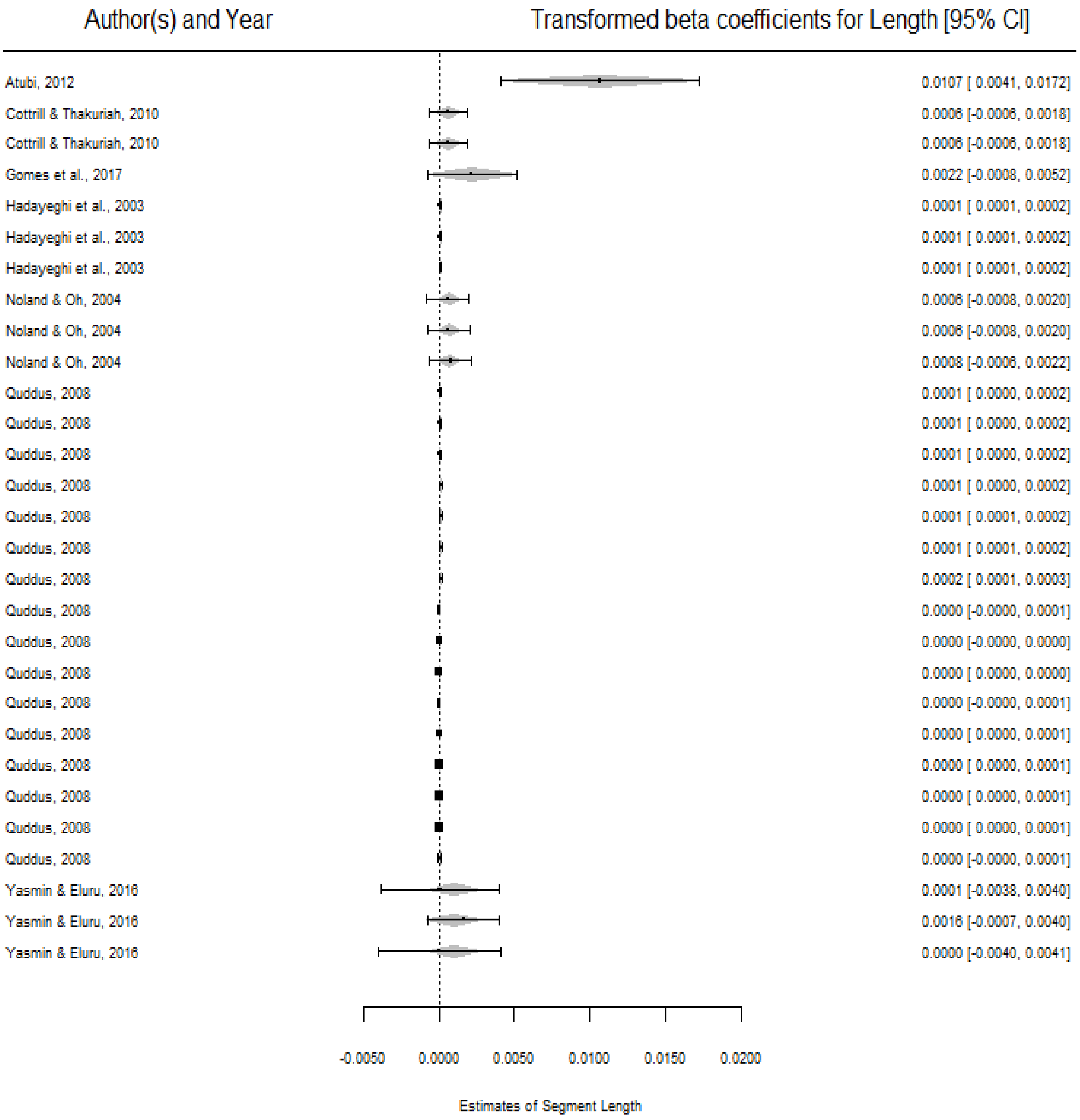


Figure 2-3: Forest plot of segment length beta coefficients on crash occurrence

The respective funnel plot is shown on Figure 2-4. The test for funnel plot asymmetry was not statistically significant (z = 3.9332; p-value <0.0001), which suggests no indication of publication bias amongst the studies.

Table 2-6: Parameter estimates of meta-regressions coefficients for the effect of segment length estimators on crash occurrence

Moderator Variable	Estimate	Standard Error	p-value
Constant term	0.0024	0.0015	0.1204
Unit of analysis: KSI crashes [ref. cat.: fatal crashes]	-0.0001	0.0000	0.0022
Unit of analysis: Serious injury crashes [ref. cat.: fatal crashes]	-0.0001	0.0000	0.0067
Unit of analysis: Slight injury crashes [ref. cat.: fatal crashes]	-0.0001	0.0000	0.0052
Unit of analysis: Total crashes [ref. cat.: fatal crashes]	-0.0001	0.0001	0.0127

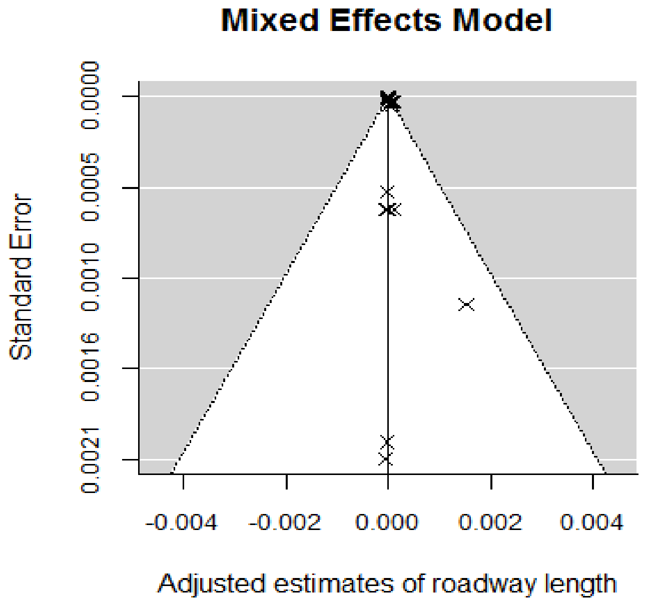


Figure 2-4: Funnel plot of segment length beta coefficients by standard errors

Results indicate that from all considered study characteristics, the main moderator variable (study characteristic) affecting the overall estimate of roadway length on crash occurrence is the severity of considered crashes (with fatal crashes as reference category). More specifically, analyzing crashes in any other severity level than fatal crashes decreases the impact of roadway length on crash occurrence in comparable levels across severities.

In practice, these coefficients denote that in studies which do not examine fatal crashes only, the beta coefficient for roadway length decreases by -0.0001 on average. The impact is non-negligible considering the larger scales of roadway length, by which this coefficient is multiplied.

Additional study characteristics that were considered but where not found to be statistically significant for roadway length were traffic speed, road user age, modal distinction (with categories: total crashes, motorized vehicle crashes, VRU-vehicle crashes and pedestrian-vehicle only crashes), the presence of intersections, regional approach (with categories: County/State, TAZ/Census Tract (CT)/Census Ward) and road type (with categories: Aggregate, A-road, B-road, Motorways, Minor-road, Low-risk road), all as defined in the respective studies.

2.2.5 Meta-regression on vehicle distance travelled estimators

Vehicle distance travelled, usually expressed in miles (VMT) or kilometers (VKT), is another classic road safety exposure indicator. In studies conducting spatial analyses, some controversial results have been found for vehicle distance travelled. Lee et al. (2017a) conducted univariate and multivariate CAR analyses and found a positive correlation of VMT at a significant level for the occurrence of motor vehicle, bicycle and pedestrian crashes. An interesting result is reported by Cai et al. (2017a), who presented several crash models for at CT, TAZ and Traffic Analysis District (TAD) levels for Florida, USA, for total, severe and non-motorized crashes. VMT, when significant, was found to be positively correlated with crash occurrence. However, heavy vehicle mileage in VMT was found to reduce crash occurrence across all severity levels.

On the other hand, results of a spatial analysis in Florida, USA showed that the crash rate decreased with daily VMT increases, which the authors attribute to higher levels of traffic density and lower travel speed with higher daily VMT, or better maintained and safer road environments (Dong et al., 2014); a similar result was reported by another study as well (Aguero-Valverde & Jovanis, 2006). A study in Belgium showed that the crash occurrence contributions of non-motorway VKTs were more than twice of motorway VKTs for crashes between cars amongst all severity levels. However, the sign of motorway VKTs was reversed when examining crashes between cars and VRUs, which can be explained from lack of intermodal interaction (Pirdavani et al., 2013).

After examining the literature with the established 4 criteria, 7 spatial analysis studies contributed to the meta-regression for road length with a total of 8 effects (Hadayeghi et al., 2003; Lee et al., 2013; Xu & Huang, 2015; Rhee et al., 2016). It was decided to perform a meta-regression on a VMT level. Transformed coefficient values used as input for the meta-regression are provided in a forest plot format on Figure 2-5.

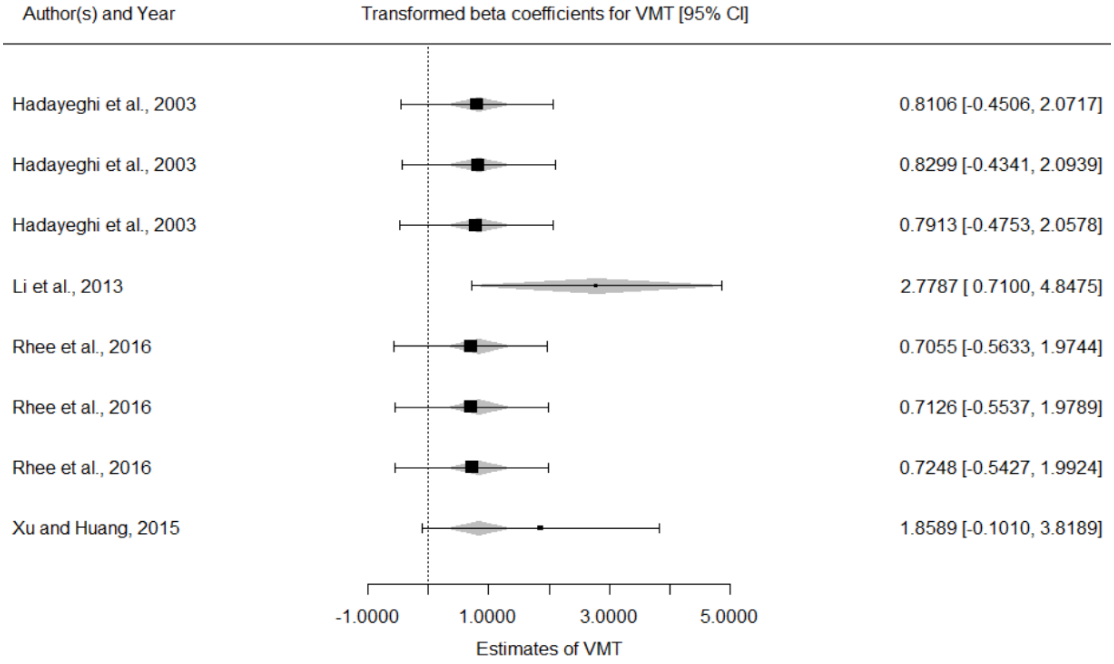


Figure 2-5: Forest plot of VMT beta coefficients on crash occurrence

Table 2-7: Parameter estimates of meta-regressions coefficients for the effect of VMT estimators on crash occurrence

Moderator Variable	Estimate	Standard Error	p-value
Constant term	2.7787	1.0555	0.0085
Regional approach: TAZ [ref. cat.: County]	-1.9449	1.0858	0.0733

The respective funnel plot is shown on Figure 2-6. The test for funnel plot asymmetry was statistically significant (z = 1.0579; p-value = 0.2901), which suggests a degree of publication bias amongst the studies. Results that are published cause the plot to appear asymmetrical; consequently there are results from studies that are unpublished, missing or have counterintuitive effects (Høye & Elvik, 2010). The trim-and-fill method cannot be applied to improve meta-regression models with moderators (Viechtbauer, 2010; Theofilatos et al., 2017).

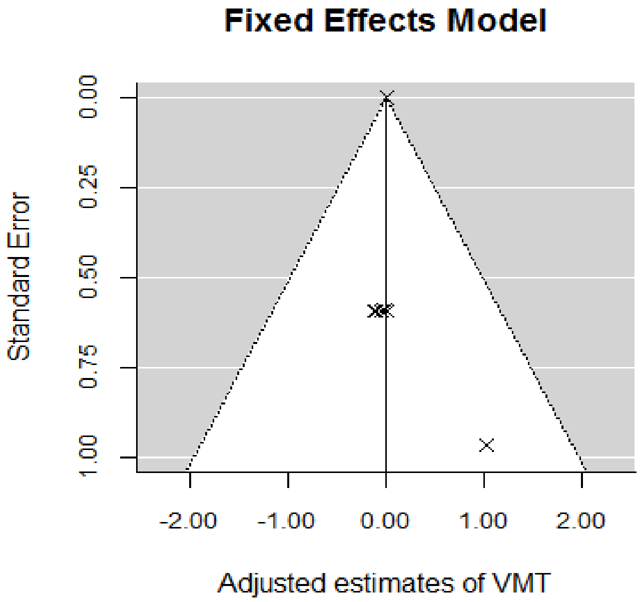


Figure 2-6: Funnel plot of VMT beta coefficients by standard errors

Results indicate that from all considered study characteristics, the main moderator variable (study characteristic) affecting the overall estimate of the effect of VMT estimators on crash occurrence is the level of regional approach. More specifically, the impact that VMT has on crash occurrence decreases in studies considering a TAZ-level approach as contrasted to a county-level approach. This finding is an indicator of how the levels of units when conducting spatial analysis might influence the final outcomes, though more studies are needed to verify it. Its p-value lies between the thresholds of 0.05 and 0.10, and the correlation is also largely described by the constant term.

In practice, this coefficient denotes that in studies which examine crashes on a TAZ level, the beta coefficient for VMT decreases by -1.94 on average, compared to studies which examine crashes on a county level.

This is a particularly interesting result, because it hints at the effect of the modifiable areal unit problem (MAUP). As previously stated, MAUP occurs when boundaries are changed inside the study areas, affecting the coefficients of statistical models, a problem that has been manifesting and studied in road safety as well (Ukkusuri et al., 2011; Abdel-Aty et al., 2013; Xu et al., 2018). The manifestation of MAUP

in VMT analysis indicates a particular sensitivity of the results obtained for this parameter based on the size of the boundary area for each study. In areas with high VMT influences on crashes, such as urban road networks, these effects may be exacerbated. Therefore researchers and authorities need to establish common zonal frameworks before comparing results of spatial analysis from different areas.
Additional study characteristics that were considered but where not found to be statistically significant for VMT were the presence of intersections, regional approach (with categories: County, TAZ), unit of analysis (with categories: Total, Fatal or Serious crashes) and road user age, all as defined in the respective studies.

2.2.6 Discussion on meta-regression findings

The meta-regression approach followed in this section provided evidence of how the various study characteristics can affect the calculated coefficient values, especially for exposure variables. The impact of traffic volume on crash counts was found to be positively correlated with taking speed limit and road user age into consideration in spatial analyses, while the impact of road length on crash counts in spatial analyses was found to be higher in studies considering only fatal crashes. Similarly, the impact of vehicle distance travelled on crash counts was found to be more important in county-level approaches as opposed to TAZ-level approaches, indicating vehicle distance travelled as more prone to statistical bias from fluctuations in boundary definition.

The findings presented in the present section are meant as an indicator of how study design affects outcomes of spatial analyses. It is not suggested that researchers who do not take into account, speed limits, for example, in spatial road safety analyses produce biased results. Rather, the implication of these findings is the quantification of those discrepancies between the studies that do include speed limits and the studies that do not, as well as an identification of the most significant factors that cause the discrepancy. The current results might serve to bridge differences between outcomes of studies of different designs in the future.

Apart from the academic exercise, this research has value for practitioners and authorities as well. Road management authorities commission several road safety studies in order to detect problematic spots. Road safety measures have different costs and effectiveness (Daniels et al., 2019), thus the allocation of limited administrative funds requires scientifically informed decisions to achieve the maximum possible benefits. This research highlights the circumstances under which estimates change, and therefore aid in prioritizing road safety measures and interventions. As an example, results imply that traffic management measures aimed to reduce AADT will have increased impacts in reducing crash frequency an area determined as hotspot without accounting for speed limits, all other parameters held constant.

Naturally, the followed approach has some limitations. Literature findings indicate that Bayesian models frequently appear to offer more precise predictions and to have better performance overall. As discussed previously, the meta-analysis and meta-regression methods considered do not currently offer proper ways of including Bayesian study results. However, since the meta-regression results provided herein are indicative of the effects of study design and environment in spatial exposure parameters, they can provide insights to future studies, even in Bayesian frameworks, for instance by influencing priors.

In addition, models may consider exposure parameters on crash frequencies but not the influence of all other variables, which then become unobserved factors. Elvik (2011) mentions that if studies do not account for all contributing factors, estimates of risk likely reflected not just the examined risk factor but the confounding ones as well. Such differences, along with differences in sampling frame, were an added reason why no meta-analysis was possible in this research.

However, in spatial analyses, the majority of unobserved heterogeneity can be accounted for: this is achieved by introducing spatially structured and unstructured random effects in the spatial models. Furthermore, by conducting the Q-test, bias related to unpublished/counterintuitive results is detected and reported (which was the case for vehicle distance travelled). Therefore, the author is confident that while imperfect, the results of these paper are fruitful, especially regarding their qualitative aspect: there is clear indication, for instance, that accounting for speed limit alters predictions for the influence of AADT on crash occurrence. Realistically, no study will ever flawlessly account for all confounding factors, at least in the foreseeable future, let alone a sufficient number of similar enough studies to conduct a meta-analysis.

The number of studies included in the meta-regressions is indeed small. This is expected as specialization of each study increases, and researchers pursue innovative designs and results. In the science of epidemiology, Terrin et al. (2005) mention that more than 50% of meta-analyses include 10 or fewer studies. Elvik (2011) who has addressed the issue distinguishes two groups: precise (or reliable) studies and imprecise (or unreliable studies). Based on the criteria set in that paper, the studies that are included here are precise/reliable. The funnel plots and tests for funnel plot asymmetry were conducted to test for publication bias; no indication of publication bias was found amongst the studies. The funnel plot symmetry and the overall convergence of more effects towards the central axis further support the reliability of the studies (Terrin et al., 2005) despite their small numbers. Therefore the author retains the produced results as informative.

Certain research directions can be derived from the meta-regression analyses provided in the present research. The meta-regression results (and the fact that not enough studies were found for a meta-analysis) is an indicator of the strong diversity present in published studies. More studies can contribute to the current knowledge and state-of-the art and further hone the results presented here (especially by consistently reporting standard errors). Dedicated studies that utilize case-control or cross-sectional designs assessing road safety interventions or other measures that extend to the parameters can be used to clarify why the factors found significant in this research influence the exposure parameter coefficients. Additionally, more research is needed to produce studies that do not examine crash frequency, but crash injury severity, and to determine if the influence caused by the same parameters persists for that dependent variable as well.

As previously stated, by establishing a common framework, result transferability can be improved. An important expectation is the introduction of an assessment method that would be akin to meta-regression, for Bayesian studies, which, to the author's knowledge, is yet to be presented. Furthermore, it is worth noting that as the world enters the era of big data, the large information available to transport and road safety spatial analyses is expected to lead to the convergence of Bayesian and frequentist parameters.

2.3 Overview of driver recording tools

2.3.1 Introduction

As driver distraction and behavior analyses became more researched topics in road safety and transport overall, a wide array of tools and methods has been developed in order to record driving behavior and measure various aspects of driving performance. Technological advancements in data recording systems, software and programs, cloud computing services and increase in pure processing power have allowed progress in several fronts. These fronts range from the more traditional methods of interviewing, surveying or simulators to car instrumentation for naturalistic driving, data collection from on-board diagnostics (OBD) ports, in-vehicle data recorders (IVDRs) and smartphone applications.

The aim of the present section is to present and comparatively assess the various driver recording tools that researchers have at their disposal, and furthermore present future challenges for their applications and overall integrations with the driving task. However, the list of driver recording tools presented here is not exhaustive. A wide array of tools was selected in an effort to provide an overview of many popular driver recording methodologies. While additional tools have been developed, such as controlled on-road studies or field-operational tests as presented by Carsten et al. (2013), the scope of the present research is not the consideration of every single one. Furthermore, an attempt was made to include mostly recent studies, namely those published after 2000. Relevant studies showcasing various driver recording tools were located using scientific databases and repositories (Science Direct/Scopus, Google Scholar, and TRID).

Firstly an overview of traditional survey methods is provided for completeness. Subsequently, the advantages and disadvantages of driving simulators are examined, while the following section investigates the various naturalistic driving data tools, including the utilization of on-board diagnostics, in-depth incident analyses, and the exploitation of smartphone data. The study is concluded with a critical discussion of the various characteristics of the tools that were examined.

2.3.2 Traditional survey methods

The first and more traditional survey methods to be implemented in the fields of transport and road safety are: (i) questionnaires, (ii) police reports and (iii) direct observer methods; an outline for each of them is provided in this section.

2.3.2.1 Questionnaire studies

Traditionally, studies recruit participants that respond to prepared questionnaires in order to gauge their experience or outlooks on aspects of road safety. The design and purpose of questionnaires often involves exploring self-reported experiences from people involved in crashes (Backer-Grøndahl & Sagberg, 2011), self-reported (also known as stated) driver or road user behavior (Şimşekoğlu & Lajunen, 2009; Rowe et al., 2015) or general road safety perceptions or opinions of participants (Vardaki & Karlaftis, 2011).

A very well-known tool is the Driver Behavior Questionnaire (DBQ), which has been introduced in a seminal study by Reason et al. (1990) and used widely in several variations ever since. The DBQ distinguishes between violations, dangerous errors and harmless lapses. In a meta-analysis that quantitatively summarized the findings of 174 studies using the DBQ, it was determined that the DBQ can predict self-reported crashes, both proactively and retroactively in a comparable range of significance (De Winter & Dodou, 2010).

The DBQ has been used to compare populations of specific driver subgroups (e.g. young drivers, professional drivers or populations from different countries). An adequate goodness-of-fit has been determined for DBQ application across sub-groups of different characteristics: a better fit has been discovered for elderly drivers compared to young drivers, while fits for male and female drivers were found to be more or less similar. The authors of that research state the importance of conducting exploratory analyses first in driver populations (Martinussen et al., 2013). When examining three neighbouring countries (Romania, Bulgaria and Serbia), it was found that the DBQ was a reliable tool to measure driving behavior and that a common underlying two-factor structure of driving errors and violations was shared by all three country samples (Stanojević et al., 2018). It is worth noting that the DBQ has also been validated using simulator data and pseudo-naturalistic driving data (obtained from an instrumented vehicle) successfully. It was found that the DBQ can be reasonably reliable when gauging driver speed choice, as its sub-scale is overall correlated with objectively-measured speed choice both in simulated and in pseudo-naturalistic driving (Helman & Reed, 2015).

However, the DBQ was not found to be a tool to produce strong predictions of recorded crashes. Furthermore, a number of study biases (publication bias, consistency motif bias, common scale anchor bias) have been identified as well; accounting for these biases may help increase the robustness of DBQ predictions (De Winter & Dodou, 2010).

The main disadvantage of questionnaires/surveys of stated behavior and opinion is that questions are often hypothetical and the actual behavior cannot be observed, while data produced are likely to lack from important details or depth on the topic that is investigated (Kelley et al., 2003). Moreover, driver lapses, errors and violations (classified as aberrant behaviors in the literature, e.g. in Reason et al., 1990) have been shown as reported less frequently in public than in private settings (Lajunen & Summala, 2003).

Questionnaires and similar forms of participant interviews do endure in research and are often used in conjunction with more sophisticated methods such as simulators and naturalistic driving experiments (Toledo et al., 2008; Birell et al., 2014) that are discussed in the following sections, or have even been tested by more sophisticated approaches, as previously mentioned (Helman & Reed, 2015).

Questionnaires can also be used for surveying/polling in order to obtain public opinion on a more general matter, such as automated vehicles (Kyriakidis et al., 2015). In cases where participation criteria are not numerous or not very strict, internet-based surveys are helpful tools to increase sample sizes in small amounts of time or reaching participants remotely.

2.3.2.2 Police report studies

An alternative to questionnaires is examining reports from police, health services or similar authorities in order to acquire snapshots of driving behavior at the time of crash or shortly preceding it (Yannis et al., 2013; Yannis et al., 2017b). Police data are typically collected by more or less specialized personnel attending the scene of a road crash or constitute records of traffic law enforcement (such as tickets or violation records) that can be detailed or aggregated. Police data are the primary sources of countries and organizations for road safety statistics and research and their use is well documented in the literature. The fact that they are issued and used by authorities often grants an increased sense of credibility and official status to the use of police data. A recommended approach is that police data are complemented with hospital data to acquire a more informed overview of the actual numbers (IRTAD, 2011).

However, if a country does not possess the required data analysis infrastructure (such as a formal linkage system of police with hospital data) or if researchers wish to combine said data sources, discrepancies may occur in the absolute number of crashes or similar events as well as underreporting (Petridou et al., 2009). A relevant study in Australia contrasted police data with multiple hospital databases. Results indicated considerable under-reporting with about two-thirds of road crash injuries not appearing in police data. The authors cite low injury severity and misclassification as primary reasons for police underreporting (Watson et al., 2015). It is reasonable to assume that such extended underreporting also affects driving behavior records, whether they result to crashes or not.

While country-level reports include crash contributing factors, they usually provide data in case of crashes and not for driving behavior overall. Police-reported data have also been reported to suffer from discrepancies of the accounts of the police officers that attended each crash scene. A study in the UK investigating police data quality concluded that officers underreport driver distraction activities, especially mobile phone use (Regev et al., 2017).

Moreover, police officers themselves have been reporting difficulties in determining human factors affecting drivers as crash contributing factors. Proxy definitions are a debatable solution to that issue; proxy definitions entail the classification of a human factor as a crash contributing factor based on related attributes of the crash (e.g. single-vehicle crashes in highways at night are classified as sleep-related). Proxy definitions of human factors are used by road safety authorities to augment police reporting, as they offer a practical guideline, and this is obviously reflected in the produced crash statistics. However, the use of proxy definitions has raised some debate because of their restrictions and lack of matching of definitions with the experience of drivers (Armstrong et al., 2013). The research of Filtness et al. (2017) is indicative of inaccuracies in police data when examining the recording of human/driver factors in crashes. This research investigated different approaches of reporting sleep-related driving which included the use (or not) of a proxy definition. Results showed that when not using a proxy definition, males were more likely to be involved in sleep related crashes with high severity. Conversely, when the proxy definition was employed, these results were not found. These results challenge the accuracy of police data used in road safety research and practice, and highlight the need for standardization and increased training of personnel recording and handling road crash data.

2.3.2.3 Direct observer method

One more traditional survey approach that eliminates third-party intermediaries and the respective uncertainties of self-reporting by road users is surveying by direct observer method. This approach involves direct observation on the roadway by researchers or other data collectors and the subsequent recording of the examined driver behavior and relevant parameters. The observers stand at the roadside, at intersections or segments and record driver characteristics as vehicles pass, such as seatbelt or mobile phone use (Yannis et al., 2011; Yannis et al., 2015). This reveals driver behavior in real circumstances, as driving occurs. Furthermore, trained observers can very easily become specialized in a very specific task as instructed (e.g. recording instances of partially wrong helmet use). There also lies a demand in observer person-hours, especially if observers are few and larger sample sizes are required. Indicatively, two independent observers in Spain worked for 63 daylight hours of observation in order to record a sample of 6578 drivers (Prat et al., 2014). In some cases, the obscurity of the observers has to be ensured, to avoid alerting drivers that they are being recorded and thus skewing relevant measurements (Manan & Varhelyi, 2015).

However, due to the lack of randomization in the selection of sites and in the observation schedule, results are not easily generalized or transferrable (Eby & Vivoda, 2003). When measuring more continuous variables, observers need to use their judgement which is inevitably imprecise. Observer attributes, such as age, have been shown to affect their estimates in a systematic manner. For instance, when investigating the observer judgment for time-to-arrival of bicycles to predefined obstacles, older observers provided systematically lower estimates (Schleinitz et al., 2016). Finally, observers can only use finite neural resources for mentally noting observations; increased visual complexity of a scene has been shown to lead to high detection failure rates for vulnerable road users, even though their presence was expected at the scene (Sanocki et al., 2015).

As a side note, caution is required to avoid confusing this method with observational research methods in general, which include any empiric investigation of effects caused by interventions when random experiments are not ethical or feasible (Rosenbaum, 2010); the latter category of studies can utilize several different arrays of tools.

2.3.3 Driving simulators

Another methodological setup apart from interviewing and using questionnaires of individuals involved in crashes is the employment of driving simulators. Driving simulators are devices designed to emulate the activity of driving, either fully or partially, and they provide a safe, virtual environment for testing driver behavior characteristics. They have been used for at least two decades in studies (Lenné et al., 1997; Desmond et al., 1998).

Driver simulators can be used to measure a large number of driving parameters, such as speed (speeding, speed variance), lane position (lane keeping or departure events), response time to events, time to complete tasks, headway distances, instrumentation use (e.g. brake pedal press counts, signaling) and others. Eye-tracking (ocular movements and fixations) can also be measured within a simulator setup through the utilization of additional devices, such as eye-tracking devices (Palinko et al., 2010; Benedetto et al., 2011; Nabatilan et al., 2012). More rarely, simulators can include integrated purpose-built systems for eye-tracking that are non-invasive, such as the one developed by Balk et al. (2006), which requires specialized knowledge and effort. In other studies, a camera-based eye-tracking system that records the eye movements of the driver is installed inside simulators (Victor et al., 2005).

A common purpose of utilizing a driver simulator is the investigation of the effects of a risk factor compared to baseline (normal) driving conditions. For instance, Yannis et al. (2016) explored the impact of texting on young drivers' behavior and safety on motorways using a driver simulator in different driving scenarios (moderate/high traffic, good/rainy weather). More uncommon risk factors can be analyzed as well: Hughes et al. (2013) used a simulator to investigate how singing while driving affects driver performance. Another use for simulators is performance assessment of drivers (e.g. Rosenbloom & Eldror, 2014) or a combination of common and more uncommon risk factors (e.g. Beratis et al., 2017). Furthermore, the effectiveness of road safety measures can also be tested, such as in a study by Dumitru et al. (2018), who investigated the influence of a smartphone warning advanced driver assistance system (ADAS) application on driver behavior.

Simulators can also be used to examine a particular aspect of the driving process. An example of isolated examination of a particular activity in a simulator is the setup utilized by Consiglio et al. (2003). The goal of the study was to examine braking behavior of the participants, who drove in a simulated environment, using a setup including a red brake lamp, an accelerator pedal and a brake pedal. This apparatus allowed measurements of mean reaction time under various conditions of distraction (baseline, radio listening, conversation with passengers and handheld and hands-free mobile phone uses). The authors refer to identified interference, namely the act of determining the implications for real world driving to be problematic. A similar setup was used by Bellinger et al. (2009) for their study as well.

From the two previous studies, and others as well, it can be determined that another useful feature of driving simulators is that they also allow the examination of the reaction time of drivers with a very high precision, which can be difficult to obtain otherwise. Reaction time is an established metric of the driving performance level which provides insights on driving in real conditions (Jackson et al., 2011). For instance, reaction time measurements during incursion effects on a driving simulator have been shown to predict safety errors associated with turns in the on-road drive, albeit with moderate robustness (Aksan et al., 2009). However, different methodological approaches on measuring reaction times have been established, and measurements are influenced by factors related to the vehicle, the driver and the environment (Jurecki et al., 2014). Therefore simulator use does not constitute a guarantee that the aforementioned precision yields values which are close to the true values of reaction times.

Another scientific area that has been recently attracting the attention of researchers is the use of driving simulators for the investigation of the effects of resuming vehicle control from Autonomous Vehicles (AVs). A transition phase where the driverless AV will yield control to the driver is expected to entail some risk, which is currently investigated using driving simulators (e.g. Merat et al., 2014; Zeeb et al., 2015). Simulators have been proven critical in allowing hypothetical transition situations to be tested although AVs are not widely circulated yet, even combined with environmental effects as faded lane markings or gusts of wind on the roadway (Zeeb et al., 2016).

Driving simulators have several advantages and disadvantages as reported in the international literature. The most dominant advantage is the safe, isolated driving environment provided by simulators. This ensures safety for all participants at all times, allowing individuals that would be unfit for driving under normal circumstances to be evaluated, such as drivers with Mild Cognitive Impairment (Beratis et al., 2017). A controlled environment also allows the exploration of parameters that would be difficult or unethical (due to possible increases in crash occurrence probability) to investigate otherwise, such as the impacts of mobile phone use or listening to music (for instance in Consiglio et al., 2003; Bellinger et al., 2009) or driver workload and eye blink duration (Benedetto et al., 2011).

A similar advantage is the standardization that driving simulators can provide. The type and difficulty of driving tasks can be precisely specified, and any potentially confounding variables, such as weather and road layout, can be eliminated or controlled. Establishing a common driving environment, even in different physical locations, increases the precision of any driving assessment results (De Winter et al., 2012).

Apart from driver recording, driving simulators can also be used to assess, train and provide feedback to drivers, thus allowing more constructive evaluation procedures. Simulators can be used for guided error training that improves driving behavior (Ivancic & Hesketh, 2000) or more specific tasks, such as lane-changing (Petzoldt et al., 2011). They can also be used for the elimination of driving impairment from cell phone use in a group of experienced drivers (Shinar et al., 2005). Driver education workshops with simulators have been conducted for newly licenced drivers. Statistical analyses, however, showed that drivers that underwent simulator training had more relaxed approaches to safe driving (less safe driving intentions), possibly due to increased driving skill confidence (Rosenbloom & Eldror, 2014).

On the other hand, driving simulators do have potential disadvantages such as the following: While simulators do provide a safe driving environment for experiment purposes, this might cause participants to become more aggressive and risk-taking when they realize any potential consequence is eliminated, skewing any measurements. Blana & Golias (2002) investigated differences in lateral displacement when driving on curved and straight road sections in real and simulated road conditions. Their results showed that the mean vehicle lateral displacement is in higher overall on the real road than in the simulator. However, these differences decreased for higher speeds at curved sections and for lower speeds at straight sections.

Furthermore, acquiring, calibrating, operating and maintaining a high-resolution driving simulator can be a very costly process and at times time-consuming. Smaller-scale on-road experiments could be less costly than simulator operation, which decreases simulator competitiveness. Moreover, in large-scale experiments, data collection is easier and more voluminous from on-road driving than simulator operation. Simulator sickness is another problem encountered when conducting simulator experiments and appears particularly when older drivers participate and should be systematically reported (Papantoniou et al., 2015).

When conducting studies with driving simulators, the issue of learning effects must be considered as well. The term learning effects refers to the process of participants becoming accustomed with aspects of the simulated experiment and adjusting their performance, even subconsciously. While familiarizing with aspects of driving on simulators is a positive use (as discussed before), the negative aspect of learning effects can skew results if unaccounted for (in as many as 30% of the studies as Papantoniou et al. (2015) mention). Weiler et al. (2000) state that learning or habituation effects are periodical and as such unrelated to previous treatments. Therefore, researchers have to be mindful of learning effects when conducting driver simulator experiments and adjust their study designs accordingly to avoid them as much as possible.

2.3.4 Naturalistic driving

2.3.4.1 Vehicle instrumentation

More recently, naturalistic driving experiments began to emerge as an option for driver behavior recording. Naturalistic driving experiments involve instrumentation and other relevant equipment installed in the vehicles of the participants that provides capabilities of recording the vehicle (maneuvers, trajectories etc.), the driver (hand motions, eye movements, distraction), external driving conditions (weather, obstacle) or any combination of the previous (Yang & Morton, 2012). Participant drivers then drive their vehicles as normal for a sufficiently long time period in a real-circumstance environment.

Certain naturalistic driving experiments have been seminal in transport and road safety research and their results redefined the way driver distraction and inattention parameters are viewed. Some of the most famous studies are the 100-car naturalistic study (Neale et al., 2005; Dingus et al., 2006) and the SHRP2 experiment (Dingus et al., 2015; Victor et al., 2015; Dingus et al., 2016), both conducted in the USA. For the 100-car study, sensors included accelerometer information, headway and sideway obstacle detection systems, incident flagging as well as five camera videos monitoring surrounding views as well as the driver's face (Dingus et al., 2006) were used. The SHRP2 database comprised multiple video images, machine vision-based applications, accelerometers and rate sensors in three dimensions, GPS, forward radar, illuminance and passive cabin alcohol presence sensors, turn signal state, vehicle network data (as available), and an incident push button (Dingus et al., 2016). It is important to note that the nature of data collection for the 100-car study (90-second segments of crash and near-crash events and 6-second baseline epochs) prevented obtaining information relevant to the duration of several types of distraction (interaction with passengers, eating or drinking etc.), that were not an issue due to continuous video recording.

Naturalistic driving experiments can provide advantages over several driver research aspects, including detailed exposure data that enable crash risk odds ratio calculations, capabilities of examination of real crashes and near-crashes, better understanding of traffic violations and evaluation of the impacts of road safety measures (Regan et al., 2012).

Overall, on-road driving at first and naturalistic evaluations later have been considered as the optimal method for assessing fitness to drive (Di Stefano & Macdonald, 2003) and possibly research of driver behavior (Backer-Grøndahl et al., 2009), due to their flexibility and control of several variables affecting driver behavior. Moreover, this type of study provides the opportunity to examine driver fitness as it involves actual driving activities, and includes aspects of driving (including physiological stimulation, traffic interaction, and tactical planning) that may not be easily replicable by other testing means (Reimer et al., 2006) and can be implemented in driver training and environmentally friendly driving training (Sagberg & Backer-Grøndahl, 2010). Naturalistic driving research attracts interdisciplinary teams examining wide topics and research questions (Wadley et al., 2009; Bowers et al., 2013), for instance driver psychology, vehicle mechanical conditions, novel mathematic and algorithmic approaches etc.

As expected, naturalistic driving studies can present some drawbacks as well. There are significant costs relevant to the equipment of on-road driving studies (Ball and Ackerman, 2011). Overall, costs are expected to raise as the duration of the experiment increases, and naturalistic studies can last from 6 to 12 months or more (Regan et al., 2012). In smaller-scale experiments of on road driving (which are types of experiments related to naturalistic driving), usually the presence of researchers is required in the vicinity for navigation instructions and for the possible recording of additional behavioral parameters. Naturalistic methods are resource demanding in terms of sample recruitment, data gathering, data storage and data analysis (Backer-Grøndahl et al., 2009). There is also the uncertainty of driving bias, namely drivers adjusting their behavior to be more careful since they know they are recorded, but this is assumed to be largely reduced because drivers forget they are being monitored after a while (Tselentis, 2018c).

Apart from traditional vehicle instrumentation, additional methodologies and sophisticated devices have been developed in order to obtain data from naturalistic driving, such as the exploitation of OBD and IVDRs. The more sophisticated methods typically include remote data processing, storage and analysis. Data obtained from GPS, OBD devices and smartphone use detection circuit are usually acquired and processed from an in-vehicle device. Data are transmitted via a mobile telephone connection to a control center (CC), where individual crash risk for each vehicle is estimated. Mobile telephone connection is used for data transmission between the on-board system (OS) and the CC. CCs can also be equipped with other databases that contain additional parameters, such as road environment data (Boquete et al., 2010), that is not mandatory for inclusion in basic smartphone applications.

2.3.4.2 On-Board Diagnostics and In-Vehicle Data Recorders

On-Board Diagnostics (OBDs) refer to an array of systems developed to provide vehicles with self-diagnostic and reporting capabilities, originating in the USA (they were standard issue in California since 1991).

The first OBD standard, known as OBD-I, defined only a few parameters to monitor, and thus, failures resulted in just a visual warning to the driver and the storage of the error. The second generation of OBD, known as OBD-II, standardizes different elements such as the connector used for diagnostic, the electrical signaling protocols, and the message format. Several operating modes are defined by the OBD-II standard to allow for an easier interaction with the system, and defining the desired functionality (Zaldivar et al., 2011).

The European version of the OBD-II standard, known as EOBD, is mandatory for all gasoline and diesel vehicles since 2001 and 2003, respectively, and closely resembles OBD-II. Several other versions have been developed globally, such as JOBD in Japan or ADR 79/01 & 79/02 in Australia. HDOBD (heavy duty OBD) specification has been made mandatory for selected commercial (non-passenger car) engines sold in the United States in 2010.

In order to obtain results for driving behavior, recorders that are connected to the car engine have been examined by past research. Zaldivar et al. (2011) proposed an Android-based application that monitors the vehicle through an OBD interface. The application was reported as being able to detect crashes via the estimation of G-forces that passengers would experience in the event of a crash, in a time margin less than 3 seconds. The authors acknowledge the low bandwidth of Controller Area Network as a limitation to the amount of sensors that can be simultaneously monitored, however they support the retrieval of only critical data from a low number of sensors to detect a crash.

A broader example of instrumentation used in naturalistic studies to classify events is In-Vehicle Data Recorders (IVDRs). IVDRs are devices installed in vehicles in order to record crash-related parameters in such an event. Examples include vehicle speed, engine speed (rotations per minute – RPM), throttle use, brake use, airbag sensors etc. (Chidester et al., 2001; Correia et al., 2001). IVDRs originated in the aviation industry in the 1950s ('black boxes') and have spread to all modes of transport, providing data for road safety studies, among other uses.

In their study, Taubman-Ben-Ari et al. (2016) used IVDRs assessing G-force based events to classify 20 types of events and break them down into five categories (i.e., braking, accelerating, handling turns, changing lanes, speeding) and three levels of risk (i.e., low, medium, high). Furthermore, events classified as risky by IVDRs have been shown in the past as an indicator of crash involvement probability (Prato et al., 2010; Toledo et al., 2008).

Similarly, Jensen et al. (2011) used OBD to provide longitudinal, lateral and vehicle power data. OBD-II/IVDR recorders gathered the vehicle speed, engine speed (RPM), mass air flow rate, coolant temperature and throttle use percentage. Handheld recorders and a Controlled Area Network (CAN) data recorder were used in conjunction, and the authors conclude that IVDR applications offers valuable information on driver behavior through the examination of vehicle operation data. Determining driving behavior (accelerations, decelerations, cruising) from observing a technical parameter (engine RPM fluctuations) is an example of innovative application of IVDRs.

The required amount of sampling is unclear for IVDRs, with limited research published on the matter. Shichrur et al. (2014) had IVDRs installed in participant vehicles that recorded detailed information about undesirable events during trips, including vehicle position, speed, vertical and horizontal acceleration, and maneuvers. The authors concluded that collecting a sample of about 300 h per driver should result in a relatively stable and reliable measure for assessing the driver's average event rate.

2.3.5 Non-intrusive driver recording

The methods described in this section entail the recording of driver behavior under real-world traffic circumstances but do not involve considerable instrumentation and are non-intrusive.

2.3.5.1 In-depth incident analysis

In-depth incident analysis involves the microscopic examination of records of cases by trained experts from several disciplines. The teams investigate either on-site or records of traffic incidents, which are typically crashes or collisions (Ziakopoulos et al., 2018) but can also be other parameters, such as overtaking maneuvers (Barmpounakis et al., 2016a). The investigations are usually conducted on a small-case, non-massive basis. This allows increased detail in determining injury mechanisms and how the interaction between different vehicle types affects injury outcome.

As expected, in-depth analyses are time consuming and involve increased expert workload and often there are missing evidence for a complete reconstruction of some cases (Hill et al., 2012). In-depth incident analyses can be augmented by algorithmic analyses of video recordings for trajectory extraction and clustering (Nikias et al., 2012; Orfanou et al., 2012; Barmpounakis et al., 2016a).

Complementary or even as an alternative to in-vehicle recording, external venues offered by new technological advancements such as Unmanned Aerial Vehicles (UAVs, also known as drones) have been also investigated for real time traffic monitoring. Relevant research has concluded that the possibilities are enticing, with low cost cameras having been used to successfully extract kinematic characteristics. However, certain limitations have to be bypassed first, such as low battery duration and susceptibility to weather events (Barmpounakis et al., 2016b).

2.3.5.2 Smartphone data

On another note, smartphone data are also utilized in studies published more recently. Utilizing smartphone data allows for massive data collection via sensors embedded in mobile phones, which makes this method continuous, inexpensive and rapid. Moreover, smartphones can be programmed and their sensors have expanded to a wide array, many of which can be exploited for transport and road safety research, such as accelerometer, digital compass, gyroscope, GPS, microphone, and camera, which enable sensing applications, even without user engagement (Mantouka et al., 2018). The most usual approach is the establishment of a central database for data cleaning, processing and further analysis (Iqbal & Lim, 2006), which does not only provide a common reference framework but also allows for the development of specialized indicators and for increased processing power to databases that quickly become big-data problems.

Vlahogianni & Barmpounakis (2017) examined the use of smartphones as an alternative for driving behavior analysis. Their research was based on data collection from a smartphone application and the respective platform by OSeven Telematics. After implementing re-orientation algorithms to raw smartphone data, they detected critical patterns based on a rough set theory framework. The authors concluded that smartphones can accurately detect harsh longitudinal and lateral driving patterns as accurately and reliably as OBD-II devices. However, they do cite differences between sensor technologies due to diverse brands and devices as a possible issue of the approach.

Similar data obtained from smartphone sensors have also been used to create driving profiles describing degrees of environmental driving from naturalistic driving trips according to the respective range of accelerations (Adamidis et al., 2020). Through simulation, smartphone data were shown to provide a solid basis for achieving reductions in vehicle emissions by controlling the range of acceleration and braking characteristics. Data from smartphone sensors have also been used for trip profiling through clustering methods based on road safety criteria (Mantouka et al., 2019). Relevant findings indicate that drivers behave differently during every trip, and are all prone to risky driving, albeit in different time percentages – as opposed to an inflexible separation of drivers to safe and unsafe. Furthermore, data from smartphone sensors have been successfully used to identify transport mode via advanced machine learning techniques such as Gradient Boosting and Random Forests, with the latter exhibiting a better performance (Efthymiou et al., 2019).

Despite rapid improvement, smartphone data still exhibit quality issues. When examining the implications of smartphone-based insurance telematics, reliability of smartphone measurement data was found as a major hindrance to the development of smartphone usage-based motor insurance (UBI). The discrepancy in quality is such that state-of-the-art algorithms implemented in UBI tailored hardware measurement probes initially could not be directly transferred to a smartphone application (Handel et al., 2014). Ever since, there have been constant quality improvements such as those mentioned in a relevant study allowing the collection of GPS trajectory and speed data per second, along time-varying traffic and roadway dynamics (Ma et al., 2018).

Furthermore, Lee (2014a) describes a process to store and analyze real-time collision data in a distributed processing framework. The proposed work was presented to analyze 'near-real time big data', including collision data and road traffic data from a section of 400 km in South Korea. The framework included a traffic event cloud, traffic big data storage and processing parts, among others. Both studies mention the high cost of real-time driving data recording systems, data programs, cloud computing services, the inability to accumulate and exploit massive data bases (big data) for transport and traffic management purposes, as well as the low penetration rate of smartphones as barriers to the collection and management of real-time data. However, smartphone penetration rate keeps increasing with time (by 2020 approximately 70% of the world’s population will be using smartphones as mentioned by Kanarachos et al., 2018). Research has indicated that the other barriers can be overcome when consumers are given an incentive such as a monetary rewards (Reese & Pash, 2009).

Tselentis et al. (2017) have conducted a review on studies concerning motor insurance schemes, exploring concepts that would lead drivers to pay based on their travel and driving behavior (UBI schemes). In terms of the indicators mostly used in today’s UBI models, mileage, speeding, road network type and risky & rush-hour driving predominate among them. The authors mention that UBI is expected to improve traffic safety as the impact of behavioral indicators is contrasted for various road safety parameters (such as crash risk).

It has been proven that the required amount of sampling when using smartphone devices for driving behavior assessment varies for each road type, driving characteristic and driving aggressiveness. Overall, the values range at less than ten times lower than those for IVDRs, from 16.3 h to 23.0 h per driver (Tselentis et al., 2018b), which showcases a clear advantage of utilizing smartphones.

From a road safety perspective, one of the biggest advantages of the smartphones is the measurement of driver distraction due to mobile phone use, which is a critical road safety risk factor when used handheld, hands-free or for message texting (Horrey & Wickens, 2006; Caird et al., 2008, 2014; Elvik, 2011; Simmons et al., 2016). Overall, relevant research indicates that naturalistic driving experiments, especially those conducted with smartphone data, are appropriate for the assessment of driving behavior, providing a wealth of real-life data on driving behavior and related risks such as distraction and speeding (Papadimitriou et al., 2018), enabling a smooth transition from the data collection to the data analysis procedure (Yannis et al. 2017a) and exploiting a variety of smartphone Application Programming Interfaces (APIs) to read and transmit sensor data, and most importantly, the capability to provide feedback to drivers (Tselentis et al., 2018a). It has been proven that informing drivers through personalized feedback about their speeding is also effective at encouraging drivers to improve their driving behavior, mainly on the aspect of speeding (Ellison et al., 2015). Further emphasis was placed on the personalization of feedback by Vlachogiannis et al. (2020), who claim that even drivers of an initial driving state (i.e. road safety behavioral level) were found to require different types of policies for their successful transition to a safer state. Some were required to confine a single driving behavior, whereas others were required to achieve overall improvements.

A similar alternative to smartphone data collection is the use of data collected by smart passenger cards for public transport. Smart cards are plastic contactless cards with similar functionalities to credit cards and can be used in public transport in lieu of traditional paper tickets to enable faster and easier transitions. Smart cards could be clustered using temporal activities and researchers can partition passengers into smaller sets of clusters based on their usage habits of the transportation network (Medina, 2018). Though they are not directly driver-related, frameworks developed from smart-card analysis can be used to analyze smartphone data, thus being relevant for driver analysis as well. De Romph (2013) mentions the introduction of public transport smart cards as a venue for the creation of large databases with public transport movements, possibly in order to create a source of useful matrix building data.

2.3.5.3 Discussion on driver recording tools

When compiling the information gathered from studies as discussed in the previous sections, a series of helpful insights can be achieved. It is evident that technological advancements constantly allow for the development of more sophisticated tools that in turn provide more rich and rapid data acquisition. An overview of the main advantages and disadvantages of each driving assessment method and recording tool was obtained from the international published literature and can be presented on Table 2-8.

Overall, questionnaire surveys (Vardaki & Karlaftis, 2011) and driving simulators (Kaber et al., 2012) can aid in the evaluation of the impact of various human factors or distraction in driver behavior, yet they suffer from the known limitations of self-reported information. On the contrary, naturalistic driving experiments are considered to be more appropriate for the assessment of driving behavior (Regan et al., 2012; Tselentis et al., 2017; Yannis et al., 2017a; Tselentis et al., 2018a; Papadimitriou et al., 2018). This is because behavior is recorded under normal driving conditions and without any influence from external parameters such as the presence of an experimenter, prior knowledge or possibility for participants to observe or predict conflicts, near crashes or even actual crashes in real time without potential biases on the recording. Furthermore, if drivers are monitored for an appropriate amount of time, driving under normal conditions will be recorded and no bias will appear because of the fact that drivers are aware that they are being recorded.

Researchers have to solve what evolves into a multi-parametric problem regarding the selection of driver recording method, experimentation and instrumentation. One of the most obvious issues is, as expected, the monetary cost of each application. Available (or required) capital can be considered as proportional to data quality and result in robustness and transferability when examining driver behavior, but the choice must be made in conjunction with the research questions or task at hand. For instance, it can be surmised that naturalistic driving is overall more accurate than questionnaire surveys. Despite that, expending a quarter of the budget of a naturalistic driving study to conduct a high-quality questionnaire survey might prove to be more cost-effective for a specific behavioral problem (e.g. alcohol consumption or lane-changing behavior). Thus it is inaccurate to assume that older, traditional research methods will disappear from practice, at least in the immediate future.

In several transport studies, timing is also of the essence. Temporal factors might increase research budget to allow for more rapid data collection in case of urgent demand, for instance when wanting to capture specific behavioral impacts such as harsh braking behavior after infrastructure interventions have been implemented (and might pose possible road hazards).

Table 2-8: Comparative overview of driver recording tools, experiments and methods

Experiment Type	**Method – Driver Recording Tools**	**Advantages**	**Disadvantages**
Surveys on opinion and stated behavior	Interviews and/or Questionnaires	• Investigation of new situations • Can finish in a short time • Low cost • Some forms [e.g. DBQ] are well established and validated	• Hypothetical questions • Data lack details • Self-reported data • Low response rates • Numerous bias sources
Past police or hospital record investigation	Existing Database investigation	• Relatively easy to obtain • Low cost • Official data used/issued by authorities and organizations	• Required databases must be functional and maintained • Missing data for specific times / regions • Underreporting issues • Difficult to acquire driver behavior variables • Results may appear considerable later than event records
Direct observer method	Roadside observations by researchers/data collectors	• Recording of real traffic and behavior as it occurs • Trained observers can be purpose-specialized • Third party elimination	• High person-hours may be needed for larger samples • Lack of randomization leads to transferability limitations • Observers may fail to detect events/parameters • Observer obscurity may need to be ensured • Observer attributes can skew data records

Experiment Type	Method – Driver Recording Tools	Advantages	Disadvantages
Driving simulator	Driving simulator	• Safe environment • Greater experimental control • Large range of test conditions • Measurement of the reaction time	• Learning effects • Simulator sickness • Very high cost • Recalibration needs
Naturalistic driving - Vehicle instrumentation (& On-road driving)	Instruments installed in participant vehicles who drive normally (On-road driving experiments are shorter and researchers are present for the duration)	• Understanding real traffic conditions • Conflict observations • Excellent for assessing driving fitness • Capabilities of use for driver training • Interdisciplinary extensions	• Traffic incidents can be rare • Long experiment time period • High cost • Demanding in recruitment, data gathering, data storage and data analysis
OBD/IVDRs	Specific diagnostic subsystems of vehicles	• Can indicate crash involvement probability accurately • Can be exploited for real time traffic monitoring	• Unclear sampling frame
In-depth incident investigation	Trained experts investigate records & causes of past crashes or other incidents	• Identification and reconstruction of crash factors • Allows research into injury prevention	• Insufficient reconstruction evidence • Long analysis time period • Demanding in data analysis
Smartphone data exploitation	Smartphone applications	• Easy to recruit drivers • Continuous and rapid data collection • Wide application capabilities • Upfront costs during development, low cost and ease of use in data collection • Seamless course from data collection to data storage and analysis • Lower data sampling hours per driver required • Measurements of mobile phone distraction	• Demanding in data storage • Demanding in data analysis (big data) • Quality and reliability issues compared to OBD need to be accommodated with sophisticated data filtering and cleaning methods

From a managerial standpoint, there is also the consideration of human resources: the more state-of-the-art methods require increasingly specialized personnel (data scientists and engineers, programmers, front & back-end developers etc.) that are not necessarily readily available in a transport research facility, and their possible employment costs and delays have to be similarly accounted for.

Newer technologies (such as OBD uses) are always dependent on their market penetration. This affects not only the feasibility of the study at a fundamental level, but it may also skew results if it is not accounted for. For instance, collecting OBD or smartphone data inevitably causes results to higher participation of young users and new cars in the sample. Therefore, increased study periods or areas might be needed to meet the necessary statistical sample quota, as well as statistical adjustments in the results.

Driver recording tools are expected to receive increased development with the advent of automated vehicles (AVs). In order to achieve complete driverless automation, the transition to a moderate level is required first; in this level drivers will be required to take over control in situations where the AV cannot navigate and react safely. This will require the development of several connected vehicle technologies which will allow better augmentation and monitoring of driving activities to ensure driver readiness to resume control (Zeeb et al., 2015).

However, drawbacks might even arise from developing too good driver recording tools. Data protection and cybersecurity issues might arise when monitoring drivers in great detail. It has been shown that drivers can be successfully identified with 100% accuracy when analyzing trained sensor data against test data (Enev et al., 2016) and, in theory, this caveat could be turned into law enforcement measures. Greater threats of user privacy can also arise in the implementation of systems that actively seek more input from their environment (Acharya, 2014).

From a traffic management and road safety perspective, the overall future trend appears to be more interconnected devices and real-time recording with richer data that will allow smoother and seamless integration both for traffic flow and management and road safety purposes.

2.4 A note on the conceptual merit of analysing harsh driving events

Since road crashes or road crash casualties are a traditional focus in the science of road safety, one might argue: What is the value of analyzing harsh events?

Harsh events have been adopted as a parameter for measurement of road safety in the past, as they are strongly correlated with reduced spatial and temporal headways (unsafe distance) from neighboring vehicles, near misses with road users or stationary objects, and also include additional behavioral parameters such as lack of concentration or experience. Harsh events have been determined as inherently linked with driving risk (Tselentis et al., 2017), while research has also documented harsh driving behavior as critical for driving risk assessment (Bonsall et al., 2005; Gündüz et al., 2018). Harsh accelerations and decelerations, and their correlations with crash risk, have been investigated by the insurance industry as well (Paefgen et al., 2012; 2014).

While harsh events are ultimately driver behavior metrics, they have the potential to be analyzed as point-data (locations), much like road crashes. The examination of patterns in the distribution of harsh event points does have the potential to reveal interesting underlying mathematical and spatial relationships that show dependencies with the same parameters that lead to crashes and casualties, with similar causality. An aggressive driver will have elevated harsh events not only in a particular trip, but in all trips made across the map. Thus a large enough driver sample, leading to a sizeable trip sample, can be reasonably expected to convey useful information about problematic road segments with high road safety risk (hotspots). Moreover, harsh events constitute pro-active road safety parameters and can thus disclose these hotspots preemptively, before crashes occur and their respective consequences manifest. The aforementioned potential increases in light of recent research results which indicate that harsh braking incidents are influenced by traffic and geometric variables in a similar way to total and truck-related crashes (Kamla et al., 2019).

Furthermore, harsh events are expected to be increasingly employed as an important driver classification metric in usage-based motor insurance (UBI), as they appear to be more representative of crash occurrence probability (Tselentis et al., 2017). However, to the experience of the author, studies focusing on factors influencing harsh event occurrence and similar characteristics are significantly outnumbered by studies analyzing crashes, indicating significant research gaps in this field.

At this point it should be underlined that harsh accelerations and harsh brakings are two different phenomena occurring during different situations. As such, it is recommended that they are not analyzed collectively in principle. Indicatively, drivers with higher anger, frustration and anxiety levels display higher acceleration values and apply increased physical pressure on the accelerator pedal (Stephens & Groeger, 2009). Harsh braking events are thought to indicate reaction in anticipation of a safety-critical event (e.g. near-miss) or crash, and are used as indicators for that purpose in naturalistic driving data (e.g., Hanowski et al., 2005; Olson et al., 2009; Zohar et al., 2014; Jansen & Wesseling, 2018). Harsh acceleration events have been shown to be influenced by similar but not identical variables compared to harsh braking events (Ziakopoulos et al., 2020).

Harsh events have been found to have environmental and energy efficiency impacts as well. Aggressive driving has been found to be more than 40% more costly in terms of fuel consumption and gas emissions compared to calm driving (Alessandrini et al., 2012). Transport interventions have become more multifaceted as time passes, and the intent of reducing environmental footprints is now integrated alongside road safety measures. To that end, calmer, safer and more environmentally friendly driving is strongly promoted during the past years. Targeted reductions of harsh events, which will be achieved by preceding analyses, are a promising venue to achieve this (Yamakado et al., 2009).

In conclusion, harsh events are promising, understudied phenomena that are expected to aid in proactive road safety improvements. As such, they merit further investigation with application of techniques that have shown noteworthy and informative results in crash analyses, such as hotspot detection and infrastructure assessment by prediction of crash frequencies. When viewed in the context of spatial analyses, harsh events are expected to be able to adequately substitute crashes, while providing more voluminous, and more accurate information that is available for the majority of segments in an urban network area. This will provide the basis for a high-resolution, data-rich framework that can support road segment assessment and hotspot detection adequately.

As a final note on this issue, harsh braking events can be thought to denote a potential crash avoidance maneuver more directly than harsh accelerations, which can happen for a variety of reasons, such as haste or aggressiveness on behalf of the driver. Henceforth in the present doctoral dissertation, harsh brakings are reported first, followed by harsh accelerations, in descriptive statistics and statistical model results.

2.5 Critical Synthesis

From the exploration and subsequent critical evaluation of the literature, a number of key points can be highlighted. Spatial analyses have been utilized in the literature for crash examination on a basis of zonal, regional and conditional spatial units. The most popular type of examination appears to be frequentist crash analysis, while injury severity examination is considerably less popular. Areal units are considered significantly more frequently than segments for analysis, and when segments are considered, they are usually isolated environments such as rural roads.

Conversely, urban network analyses are far scarcer due to lack of proper data collection schemes and increased structure complexity. There are underexplored research directions, such as spatial analyses of other road safety aspects such as harsh events instead of crashes. Similarly, the spatial correlation of crashes with the presence of specific road safety measures that can be expressed in a similar point-data format (for instance locations of red-light cameras or specific road markings) has not been yet attempted to the author's knowledge.

Spatial approaches have shown more and less apparent comparative advantages over non-spatial approaches. Amongst the more apparent comparative advantages, spatial approaches offer the capability of intuitive presentation of their results across the examined areal units. This allows complex mathematical structures and dependencies to yield high-quality results that are easily communicated to individuals without the respective backgrounds, similar to weather or election maps. In addition, results are more precise across space, due to the fact that the dimensions and span of the study area are not arbitrarily assigned; rather, they are integrated variables and/or parameters of the investigation at hand. These advantages can lead to more precise awareness campaigns for the public and to more informed decision-making processes from stakeholders and road management authorities.

A critical comparative advantage of spatial analyses is the inclusion of location-specific effects, known as spatial effects, which allow model predictions to vary locally. Spatial effects are not just random numbers; their fluctuations may reflect unmeasured variables which, if omitted, may increase uncertainty in the models. This is particularly useful for fields such as road safety, which study road crashes and similar phenomena that are hard to observe and that contain many possible contributing factors in complex environments. In a study area such as a city, spatial effects may reflect changes in weather, population, parking regimes, economic status, education etc. (Aguero-Valverde & Jovanis, 2006). MacNab (2004) also mentions that spatial analysis offer a way of spatial smoothing and data pooling in areas of small 'at risk' population.

As shown by the meta-regressions, several factors can influence the results of spatial analyses, such as the inclusion or exclusion of specific variables, the unit of analysis and its categories and the areal unit of examination. The complexity of spatial analysis introduces additional problems such as the border problem and the modifiable areal unit problem (MAUP). These problems can be circumvented in segment-based spatial approaches. Road segments have greatly reduced and more clearly defined borders, thus simple rules can address the border problem. Respectively, examining road segment centroids and altering the road network extent to shape study areas is expected to address the MAUP if performed consistently.

From all the previous, it is concluded that spatial analyses of harsh events on urban networks is a novel unexplored, and presumably informative research direction. Smartphone sensors can provide the core trip data reliably and consistently, while offering additional information such as mobile use and speeding parameters. Such an approach would be 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.

2.6 Research Questions

Based on the multifaceted literature review that was conducted in the previous sections, the following research questions are 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 a statistical impact on harsh event frequencies?

The following sections of the present doctoral dissertation endeavor to meaningfully answer these research questions with substantial results and findings.

3 Methodological Approach

3.1 Overall framework

The current section aims to outline the methodology that was followed for the segment-based spatial analyses of harsh events in the present doctoral dissertation. An overall summary of the framework of the dissertation is provided in the present section. Following this outline, the theoretical background and explanatory statistical framework of the underlying spatial theory and the utilized models is provided. Afterwards, the methodological steps that were applied for data preparation and execution of the statistical spatial analyses are elaborated upon.

The core goal of the present doctoral dissertation is the spatial analysis of harsh event frequencies across road segments using high resolution naturalistic driving data from smartphones. In order to achieve that goal, several variables and parameters originating from different sources are combined.

Initially, urban network study areas were selected as the more complex and less studied road environments. They constituted the main training and testing areas for the produced spatial models. After the area selection, primary geometric and infrastructure characteristics were extracted for each road segment. From these primary characteristics, secondary ones were calculated as well, to formulate the segment dataset that serves as basis for the spatial analysis. This dataset is enhanced by high-resolution naturalistic driving trip data collected via smartphones. The trajectories of these trips are analyzed and assigned to each segment: both normal driving and harsh events are recorded in order to include exposure and normalize harsh event frequencies in the models.

At this point, the data preparation phase was completed for urban networks. A training area was initially used to calibrate (train) the models which disclosed underlying spatial relationships of harsh event frequencies with the independent variables. Subsequently, the predictions of these models were tested in a test area with known event frequencies, and the models were assessed on their performance, thus evaluating the overall transferability.

It is widely accepted that the main known pillars in road safety are (i) road user, (ii) vehicle and (iii) infrastructure (Papadimitriou et al., 2019a; Martensen et al., 2019). In an effort to include a wider range of these parameters, apart from solely geometrical/infrastructure characteristics, an additional part of this dissertation concerns the introduction of traffic and road user behavioral parameters to the models of spatial analysis of harsh events. The overall philosophy and setup, i.e. the segment-based spatial analysis of harsh event frequencies produced from smartphone recorded driver trips, remains largely similar to the spatial analysis conducted on the urban networks as outlined previously. In other words, additional analyses were conducted to determine possible statistical relationships of these traffic and road user parameters with harsh event frequencies in road segments and their magnitude.

However, a conundrum arises 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 (a process also known as imputation) because they are snapshots of a particular instant. This essentially means that while road user behavior and traffic variables, such as mobile use and speeding percentages, can be utilized in the particular test and training areas, any produced models would be ill-founded and unreliable when moving to areas with no data. This limitation does not arise with geometric/infrastructure data which are fixed and not temporally varying attributes.

There were two possible solutions to this conundrum: either to exclude these parameters from spatial analysis completely, or to create statistical models that would include them without intending to transfer them. Ultimately, the second solution was selected: to create additional models which use road user behavior and traffic variables. This approach was further enhanced by analysis of similar data in earlier non-spatial research, which highlighted the importance of traffic parameters for harsh event frequencies (Petraki et al., 2020).

Therefore, causal models including road user behavior and traffic input data were created to investigate additional underlying correlations in an effort to further understand the phenomena of harsh braking and harsh accelerations, and to explore whether there are noteworthy spatial correlations between segments regarding these phenomena. It is accepted that traffic data are more clearly defined in linear environments such as urban arterials. Moreover, higher-resolution traffic data was available in urban arterials. An urban arterial was thus selected as an additional study area, and the process was repeated with the inclusion of road user behavior and traffic input data and the exclusion of predictions.

Several mathematical tools and machine learning algorithms were examined in order to approach the problem. Since the variables of interest are count data (frequencies of harsh events), all models were developed within a Poisson framework. It was decided that since there were no similar past studies approaching such issues, to the author’s knowledge, the application of a range of tools was more appropriate. These are:

1. Geographically Weighted Generalised Linear Regression (GWPR models), a frequentist regression method
2. Spatial Generalised Linear Mixed regression with conditional autoregressive priors (CAR models), a Bayesian regression method
3. Extreme Gradient Boosting (XGBoost), a potent machine learning algorithm which was implemented with (i) random and (ii) spatial cross-validation

As demonstrated in Section 2.4, previous research has shown that harsh brakings and harsh accelerations are two different phenomena, therefore separate models were developed for each category.

The first two regression methods integrate spatial data in their functional forms by exploiting existing prior knowledge of the distribution of the dependent variables in the study area. This approach allows for explicit examination of the impact of spatial effects and improves result interpretability across an area. However, it does not allow for unbiased transferability of results in other areas. In other words, to gain knowledge of the spatial effects in a new area, the dependent variable is required to be known there as well, and the models ought to be recalibrated – therefore events would not be predicted. Thus prediction can only be conducted with the respective aspatial versions of the models, which are Poisson models in a frequentist or Bayesian GLM framework.

On the other hand, XGBoost, which is an advanced machine learning algorithm, specializes in creating rules from data and conducting predictions in a 'black-box' manner. This algorithm can integrate spatial effects directly from the training data for more accurate predictions, at the cost of low interpretability of the output rules.

The predictions of all three methods were then averaged for road segments, and the best-functioning combination was sought out. This allowed the mitigation of the different inherent errors of the methods and will yield final, optimal prediction harsh event frequencies for each road segment.

The entire process, from geometric and naturalistic driving data gathering and merging to spatial modelling was then repeated for urban arterials with the inclusion of road user behavior and traffic input data and the exclusion of the predictive processes.

The framework of the present doctoral dissertation is depicted in Figure 3-1.

The remainder of Section 3 is structured as follows: Section 3.2 provides a thorough explanation of the theoretical background of spatial analysis. The underlying theory and exploratory tools of spatial analysis are discussed. Afterwards, the mathematical structure of the three statistical models implemented in this dissertation is provided (GWPR, CAR and XGBoost). In Section 3.3, the data sources and the multi-parametric data (geometric, road characteristic, naturalistic driving behavior and traffic data) that were utilized are described in detail. Section 3.4 showcases the exact methodological steps that were undertaken to obtain and extract the data on a primary step, and to merge and combine it in a compatible manner so that spatial analyses are conducted on informative and meaningful datasets – a process of critical importance for this dissertation. Harsh braking and harsh acceleration analyses produced numerous informative results in both mathematical model and map formats that are presented in Sections 4 to 7 which follow subsequently.

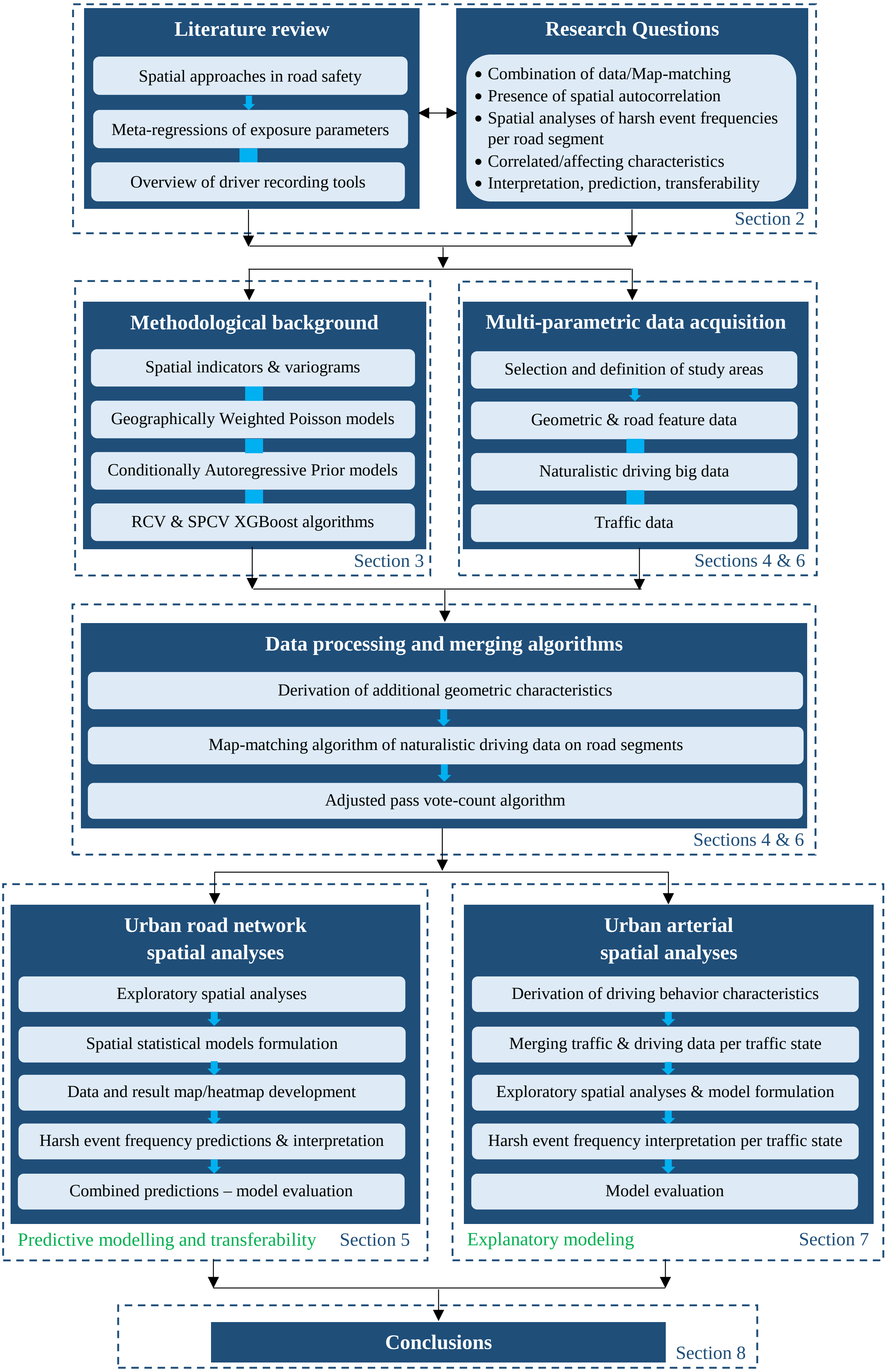

Literature review
Spatial approaches in road safety
Meta-regressions of exposure parameters
Overview of driver recording tools
Research Questions
• Combination of data/Map-matching
• Presence of spatial autocorrelation
• Spatial analyses of harsh event frequencies per road segment
• Correlated/affecting characteristics
• Interpretation, prediction, transferability
Section 2
Methodological background
Spatial indicators & variograms
Geographically Weighted Poisson models
Conditionally Autoregressive Prior models
RCV & SPCV XGBoost algorithms
Section 3
Multi-parametric data acquisition
Selection and definition of study areas
Geometric & road feature data
Naturalistic driving big data
Traffic data
Sections 4 & 6
Data processing and merging algorithms
Derivation of additional geometric characteristics
Map-matching algorithm of naturalistic driving data on road segments
Adjusted pass vote-count algorithm
Sections 4 & 6
Urban road network spatial analyses
Exploratory spatial analyses
Spatial statistical models formulation
Data and result map/heatmap development
Harsh event frequency predictions & interpretation
Combined predictions – model evaluation
Predictive modelling and transferability
Section 5
Urban arterial spatial analyses
Derivation of driving behavior characteristics
Merging traffic & driving data per traffic state
Exploratory spatial analyses & model formulation
Harsh event frequency interpretation per traffic state
Model evaluation
Explanatory modeling
Section 7
Conclusions
Section 8

Figure 3-1: Overall methodological framework of the doctoral dissertation

3.2 Theoretical background

3.2.1 Introductory concepts

The nature of spatial analyses is succinctly captured in the First Law of Geography presented by Tobler (1970):

"Everything is related to everything else, but near things are more related than distant things."

This simple sentence provides the intent between including the dimensions of space in the analysis and subtly hints at two fundamental geographical issues: (i) spatial dependence /autocorrelation and (ii) spatial heterogeneity.

Spatial dependence essentially refers to events at a location being highly influenced by events at neighboring locations. Spatial dependence is measured via spatial autocorrelation metrics. In turn, spatial autocorrelation refers to the influence of variable values of given points on variable values of adjacent points. While several parallels have been drawn between temporal autocorrelation and spatial autocorrelation historically, spatial autocorrelation rises to be much more complex. Not only do the correlations occur simultaneously, they occur in more directions (2 or 3) and in each direction, they occur in a bidirectional manner (like a two-way street, to put it in a transport context).

Spatial heterogeneity is a spatial fraction of unobserved heterogeneity which refers to the non-stationarity of model parameters. In other words, spatial heterogeneity occurs in the modelled relationships as the coefficients between random parameters and observed events are not fixed spatially. The reasons for this variation are not directly known, nor are they described by the available data – thus they are unobserved. Spatial heterogeneity can vary from non-existent, in cases where the relationship between dependent and independent variable is explained entirely by a global model, to extreme, in cases where the relationship between dependent and independent variable varies widely at a local level, and there is a different parameter for each data observation.

A related problem, known as the inverse problem, refers to the difficulty of distinguishing spatial dependence from spatial heterogeneity in practice. While cross-sectional data do allow the identification of clusters and patterns, typically they do not provide information that suffices to pinpoint the generation process of these clusters and patterns. Therefore, spatial analyses attempt to reduce some of the uncertainty by detecting spatial dependence and spatial heterogeneity and allowing for its inclusion in the calibration of statistical models.

3.2.2 Detection of spatial dependence

The first step in addressing spatial dependence is to detect the degree in which it exists in a particular phenomenon, through its manifestation in a particular dataset. The most widely used measure of spatial autocorrelation is Moran's I coefficient, introduced by Moran (1950), though additional ones have been developed, such as Geary's C (Geary, 1954) and Getis-Ord ${G_i}^*$ tests (Ord & Getis, 1995) or more recently the approximate profile-likelihood estimator (APLE) (Li et al., 2007). Variograms are another widespread tool used for that purpose.

3.2.2.1 Global Moran's I

If a population with a particular characteristic is considered, Moran's I is given for that characteristic as per Moran (1950)

$$I = \frac{n}{W} * \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} (x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \qquad \text{Eq. (8)}$$

Where:

- n is the population
- x is the characteristic of interest of the population
- $\bar{x}$ is the average value of the characteristic of interest of the population
- i and j are location indices
- w_{ij} is a matrix of spatial weights given by a selected geographical criterion with diagonal elements equal to zero ($w_{ii} = 0$)
- W is the sum of all w_{ij} across the study area so that:

$$W = \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} \qquad \text{Eq. (9)}$$

There have been several proposed geographical criteria for the specification of w_{ij}. Some alternatives are, indicatively:

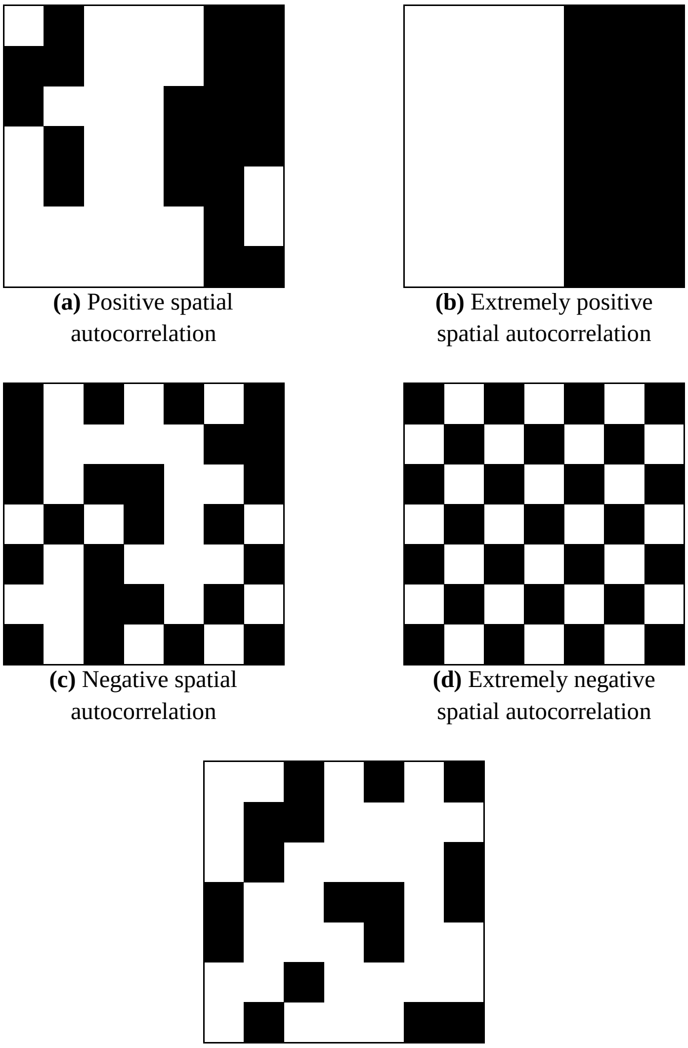
- Common border criterion (weights of neighboring locations with common borders are 1, the rest are 0) – a variation also exists in the form of k-nearest neighbor criterion.
- Critical distance criterion (weights of neighboring locations within a certain distance are 1, the rest are 0).
- Shared edge distance criterion (weights of neighboring locations are assigned based on the length of a shared edge among locations).
- Distance decay function criterion (weights of neighboring locations are assigned based on the distance of centroids from the location of interest, with more distant locations being reduced by an inverse power function).

Several other weighting functions exist in the literature (e.g. Bertazzon & Elikan, 2009). However, it is understood that the construction of the weighting matrix with a particular criterion, or, in other terms, by the application of a particular weighting function, should be based on the underlying theory of the phenomenon at hand. One can imagine the critical distance criterion as reasonably possible for analyzing house prices and crime rates, for instance. Misspecification of the weighting function may also introduce estimation bias (Smith, 2009).

The expected value of Moran's I, $E(I)$, is given by Equation (10); as sample sizes increase, dispersion is expected and $E(I)$ tends 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.

$$E(I) = \frac{-1}{n-1} \qquad \text{Eq. (10)}$$

Values of I considerably above $E(I)$ indicate positive spatial autocorrelation, and values considerably below $E(I)$ indicate negative spatial autocorrelation. As an intuitive rule, one can consider that positive autocorrelation implies clustering and negative autocorrelation implies dispersion. Simple patterns are often used to visualize typical Moran's $\boldsymbol{I}$ values, such as the ones shown in Figure 3-2 (a, b, c, d, e):



(a) Positive spatial autocorrelation

(b) Extremely positive spatial autocorrelation

(c) Negative spatial autocorrelation

(d) Extremely negative spatial autocorrelation

(e) Zero spatial autocorrelation

Figure 3-2: Spatial autocorrelation examples

In road safety, Moran's I has been quite widespread in research because its distributional characteristics are more desirable and it displays greater general stability and flexibility (Mitra, 2009).

3.2.2.2 Local Moran's I

While global autocorrelation is examined across the study area as a whole, focusing on local areas is also meaningful. Spatial autocorrelation can manifest in a specific area in the dataset, which is a form of local autocorrelation. Furthermore, in areas with significant global autocorrelation, local autocorrelation measures can provide a means to assess the contribution of smaller parts of the area to global autocorrelation (Loo & Anderson, 2015).

Local Moran's I is such an indicator, developed by Anselin (1995) in a seminal research paper concerning Local Indicators of Spatial Association (LISA). Local Moran's I, noted as I_i, is derived from the global Moran's I coefficient for each observation i as per Equation (11):

$$I_i = \frac{(x_i - \bar{x}) * \sum_{j=1}^{n} w_{ij}(x_j - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad \text{Eq. (11)}$$

The interpretation remains similar to the global Moran's I coefficient as well. However, Local Moran's I is not as standardized as its global counterpart, and can assume values significantly outside the -1 to 1 range Anselin (1995). Researchers may interpret it by examining the values of its quartiles (e.g. Waller and Gotway, 2004; Bivand et al., 2008).

3.2.2.3 Geary's C

Similarly to Moran's I, Geary's C coefficient is used to detect correlation in neighboring characteristics in the study area. Geary's C is inversely related to Moran's I, but the two are not identical. Geary's C is defined as per Equation (12), using the previous notation (Geary, 1954):

$$C = \frac{n-1}{2W} * \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\,(x_i - x_j)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad \text{Eq. (12)}$$

Geary's C assumes positive values ranging from 0 to over 1. Thus values lower than 1 indicate increasing positive spatial autocorrelation, and values higher than 1 indicate increasing negative spatial autocorrelation. Geary's C is considered as more sensitive to local rather than global spatial autocorrelation.

3.2.2.4 Variograms

Another way to determine the degree of spatial dependence of a variable in spatial analysis is the creation of a variogram. In essence, semi-variance is a measure of the spatial dependence between two observations as a function of the distance between them. Variograms are plots of the semivariance of field values of a variable as a function of distance; in essence, they are the spatial equivalent of a correlogram or covariogram. In other words, variograms are tools used quantify spatial autocorrelation, in 2D or 3D space.

Variograms are components of a greater process of analyzing stochastic (non-deterministic) phenomena spatially, known as kriging. Kriging has relatively complex underlying mathematical theories that are not the scope of this section; the reader can refer to Matheron (1963) who first explored the concept, as well as Cressie (1993). Here an outline of the fundamentals of variograms is presented.

Following Bivand et al. (2009), a variable x which is observed in space at different locations p is assumed to be generated from a random function which includes a mean and a residual:

$$X(p) = \bar{x} + e(p) \qquad \text{Eq. (13)}$$

Where:

- x is the random variable
- $\bar{x}$ is the mean of the random variable
- p is the location of observation
- $e(p)$ is the residual
- $X(p)$ is the random function generating the random variable

The underlying assumption in this case is that the mean is constant spatially:

$$E(X(p)) = \bar{x} \qquad \text{Eq. (14)}$$

The variogram can be then defined as:

$$2\gamma(h) = E(X(p) - X(p+h))^2 \qquad \text{Eq. (15)}$$

Or, the semi-variogram can be defined as:

$$\gamma(h) = \frac{1}{2}E(X(p) - X(p+h))^2 \qquad \text{Eq. (16)}$$

Where, additionally:

- h is the separation distance between locations

Semivariograms are obtained by dividing variogram functions by a factor of 2, to isolate the function, which is the parameter of interest. Thus the terms have come to be frequently used interchangeably (here only the term 'variogram' is used).

The above form implies that the spatial correlation of $X(p)$ is dependent only on the separation distance and not on location. In order to introduce the influence of location, multiple location pairs $\{X(p_i), X(p_j)\}$ are considered. The variogram takes the following form, known as the sample or experimental variogram:

$$\hat{\gamma}(h_j) = \frac{1}{2n_h}\sum_{i=1}^{n_h}(X(p_i) - X(p_i + h_j))^2 \qquad \text{Eq. (17)}$$

Where, additionally:

- n_h is the number of total data pairs
- h_j is now considered as the separation distance, or lag vector, between each pair of locations $\{p_i, p_j\}$ so that:

$$h_j = p_i - p_j \qquad \text{Eq. (18)}$$

When investigating a spatial variable, the usual case is that only certain observations are available, and not its entire distribution across space. It is usual practice to obtain empirical variograms drawn from observation point pairs, as described above. Afterwards, a theoretical variogram, $\gamma(h_j)$, can be fitted based on the trend provided by the empirical data. This technique is called variogram modelling and it can be achieved using a weighted least squares method (Cressie, 1985) or a restricted maximum likelihood method (Bivand et al., 2009).

Variogram plots are described by three parameters, shown graphically in Figure 3-3:

- The nugget n, describing the initial noise or random deviations in measurements; it is the height of the jump of the semivariance from the horizontal axis to the start of the trendline.
- The (partial) sill s, which is maximum semivariance, asymptotically limiting the variogram to non-infinity values at the largest distances
- The range r, describing the distance at which the semivariance is about ~95% of the sill variance. It is a cutoff after which any difference from the sill is negligible.

The full sill can be also obtained by adding the (partial) sill s and the nugget n. Moreover, nugget inclusion is optional in the theoretical model.

Several theoretical model forms are available, though Bivand et al. (2009) mention that some models, such as the exponential, spherical, Gaussian, or power models are more widely used in practice. Mathematical forms of these configurations are shown in Equations 19 to 23, found in various literature sources (e.g. Kitanidis, 1997):

Non-stationary models:

- Linear model

$$\gamma_L(h_j) = c * h_j + n \qquad \text{Eq. (19)}$$

- Power model

$$\gamma_P(h_j) = c * h_j^k + n \qquad \text{Eq. (20)}$$

Stationary models:

- Exponential model

$$\gamma_E(h_j) = (s-n) * \left(1 - \exp\left(\frac{-h_j^2}{\frac{r}{3}}\right)\right) + n \qquad \text{Eq. (21)}$$

- Gaussian model, defined for max distance at which autocorrelation is 0.05.

$$\gamma_G(h_j) = (s-n) * \left(1 - \exp\left(\frac{-h_j^2}{\left(\frac{4}{7}r\right)^2}\right)\right) + n \qquad \text{Eq. (22)}$$

- Spherical model

$$\gamma_S(h_j) = \begin{cases} (s-n) * \left(\frac{3h_j}{2r} - \frac{h_j^3}{2r^3}\right) + n, & for\ d \le r \\ s, & for\ d > r \end{cases} \qquad \text{Eq. (23)}$$

Where, additionally:

- c is a scaling factor or slope, $c > 0$
- k is a power exponent, $0 < k < 2$

The thin blue line in Figure 3-3 represents an indicative form of a spherical model, fitted in a synthetic dataset for a random variable. Interestingly, the use of variograms and the term 'nugget' originate from the mining industry, as these techniques were used to determine spatial distributions of gold and other mineral concentrations. In practice, empirical variograms are created using classes created from the dataset values, therefore even very large datasets will only display relatively few points.

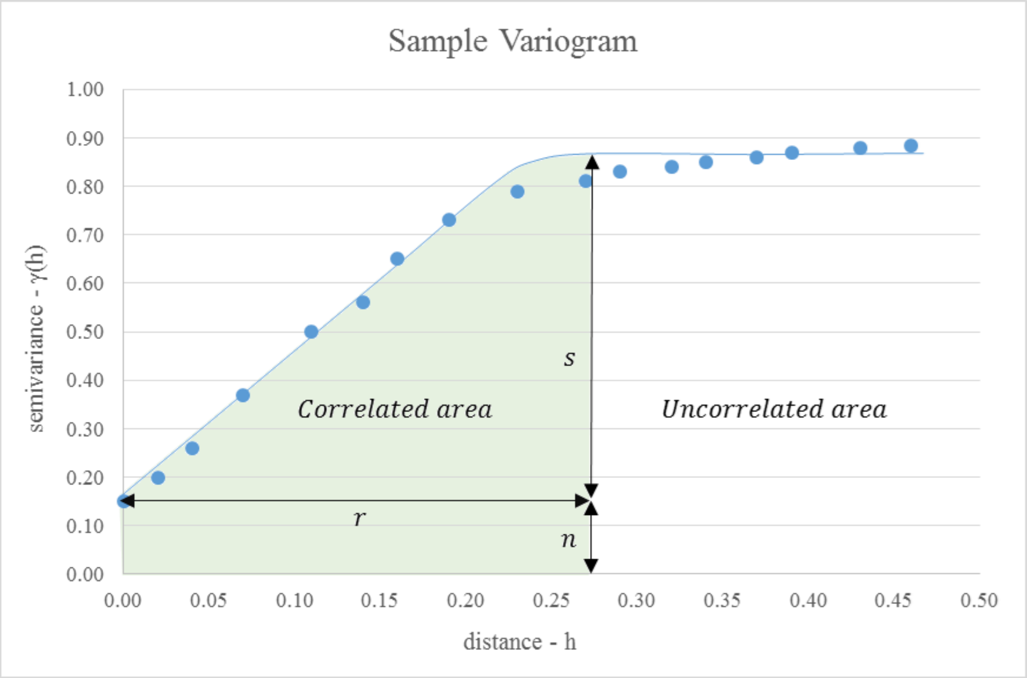


Figure 3-3: Indicative sample variogram with a spherical theoretical function

Lastly, some of the properties of variogram functions are the following:

- Non-negativity, as variograms are expectations of squares: $\gamma(p_i, p_j) \geq 0$
- Variograms at identical locations (distance=0) are always 0: $\gamma(p_i, p_i) = 0$
- Symmetricity: $\gamma(p_i, p_j) = \gamma(p_j, p_i)$
- Parity: Variograms are even functions in all directions: $\gamma(p_i) = \gamma(-p_i)$

3.2.3 Detection of spatial heterogeneity

Traditionally, spatial heterogeneity has been detected by the use of the aforementioned indicators (global and local Moran's I). Additional techniques have been used in other sciences as well. These can include additional indicators, such as Oden's I_{pop} (Oden, 1995) and Tango's maximized excess events test (MEET) (Tango, 1995) in epidemiology (e.g. Laohasiriwong et al., 2017). Moreover, in geography, the examination of static Very High Resolution (VHR) images for the exploitation of texture indices and anisotropy of urban patterns has been implemented as well (e.g. Wang et al., 2019).

However, certain barriers arise when trying to transfer these additional techniques to road safety science. The nature of crashes, which are random, point-type, non-static phenomena makes it hard to depict them in zones and examine them as a characteristic of the landscape, transferring the methods used in geography. Furthermore, certain zonal characteristics, such as population, while needed in Oden's and

Tango's indicators, are not available or possibly not as meaningful when examining a road safety aspect which refers to individuals, such as crash involvement or driving behavior metrics.

Road safety is a predominantly empirical science, relying on past records for the study of crash parameters. Therefore, it becomes more meaningful to acknowledge that spatial heterogeneity, similar to spatial dependence, exists in a varying degree in any spatial dataset, and then take it into account by implementing the proper statistical modelling techniques, as explained in the following sections. This is the main manner with which road safety researchers have been endeavoring to tackle the issue of spatial heterogeneity, and multiple techniques have emerged to that end (Ziakopoulos & Yannis, 2019; 2020).

3.2.4 Accounting for spatial dependence and spatial heterogeneity

Moving a step further from diagnostic techniques, various forms of spatial models have been developed in order to account for spatial dependence and spatial heterogeneity; a brief outline is provided herein. For further details on these topics, the reader is referred to scientific books for spatial analysis in general (e.g. Bivand et al., 2009; Brunsdon & Comber, 2015) and for its applications in road safety in particular (e.g. Loo & Anderson, 2015).

Spatial dependence is captured in spatial analysis as a form of interaction between a specific location and its neighbors. For instance, this effect can be mathematically inserted in the models using spatial weights, and captured in spatially lagged variables. Models including spatially lagged variables are known as spatial autoregressive (or spatial lag) models, which were some of the first spatial statistical models implemented. For a given location i, the weights are given with a similar reasoning as in Equation (9), and similar weighting techniques can be used as in Moran's I calculations (Anselin et al., 2014):

$$W_{y_i} = \sum_{j=1}^{n} w_{ij} y_i \qquad \text{Eq. (24)}$$

Where:

- W_{y_i} is the spatially lagged variable
- w_{ij} are the spatial weights, usually row-standardized so that:

$$\sum_{j=1}^{n} w_{ij} = 1 \qquad \text{Eq. (25)}$$

Spatial models aim to capture a degree of spatial heterogeneity by imposing a form of structure. This can be achieved either by including fixed effects in the regression process (discrete spatial heterogeneity) or by varying model coefficients (continuous spatial heterogeneity). The frequentist or Bayesian statistical models that have been widely used in the field consider forms of continuous spatial heterogeneity. Spatial heterogeneity does not require a separate set of methods, as spatial dependence does (Anselin et al., 2014). The following sections describe the theoretical background of the models utilized in this dissertation.

3.2.5 Geographically Weighted Regression

3.2.5.1 Geographically Weighted Regression overview

The Geographically Weighted Regression (GWR) family of models constitutes an extension of traditional linear regression models across a study area. In particular, GWR is a technique mainly used to indicate points or areas on a map where local regression coefficients vary from their global average. Thus, GWR allows for a continuous surface for variable coefficient values which take specific values diverging depending on the local conditions. In other words, GWR is used for the exploration of non-stationarity in the examined parameters. GWR has been developed and extensively documented by Fotheringham et al. (2002).

A global model (i.e. applying to the entire area) is initially established as a starting point. If one considers the traditional multivariate linear regression framework, a linear predictor is provided for the dependent (or response) variable, which is correlated with several independent (or explanatory variables):

$$y_i = b_0 + \sum_{k=1}^{n} b_k * x_{ik} + \varepsilon_i \qquad \text{Eq. (26)}$$

Where:

- y_i is the dependent (or response) variable at a given point i
- x_{ik} are the independent (or explanatory) n variables at a given point i
- b_k is the globally stable coefficient of a particular x_k
- b_0 is the constant term
- ε_i is the error term of the model at a given point i

Following Fotheringham et al. (2002), this basic linear framework is extended so that local parameters are calibrated instead of their global values. Equation (26) is therefore re-written as:

$$y_i = b_0(u_i, v_i) + \sum_{k=1}^{n} [b_k(u_i, v_i) * x_{ik}] + \varepsilon_i \qquad \text{Eq. (27)}$$

Where, additionally:

- $u_{i,}$ v_i are the coordinates of a given point i in space

With this modification, the above equation allows the constant term and the beta coefficients to vary across the surface of the study area in which GWR is applied. In essence, b_0 and each of the b_k are now continuous functions which manifest specific values at each point i. They are in turn estimated via weighting-based matrices similar to the ones used for Moran's I calculations; matrices are square with $n\ x\ n$ dimensions, namely equal to the total data-points at hand.

3.2.5.2 Cross-validation: Bandwidth selection

To proper formulate a statistical model, it is good practice to conduct cross validation (CV). Cross-validation is a resampling method (James et al. 2013), which involves calibrating a model in test datasets and evaluating its performance by using test datasets, with known values of the dependent variable. Cross-validation enables the assessment of the generalization capabilities of the created model and its transferability to new data. There are several types of cross-validation, the description of which is beyond the scope of this section. Indicatively, three of the most predominant types are:

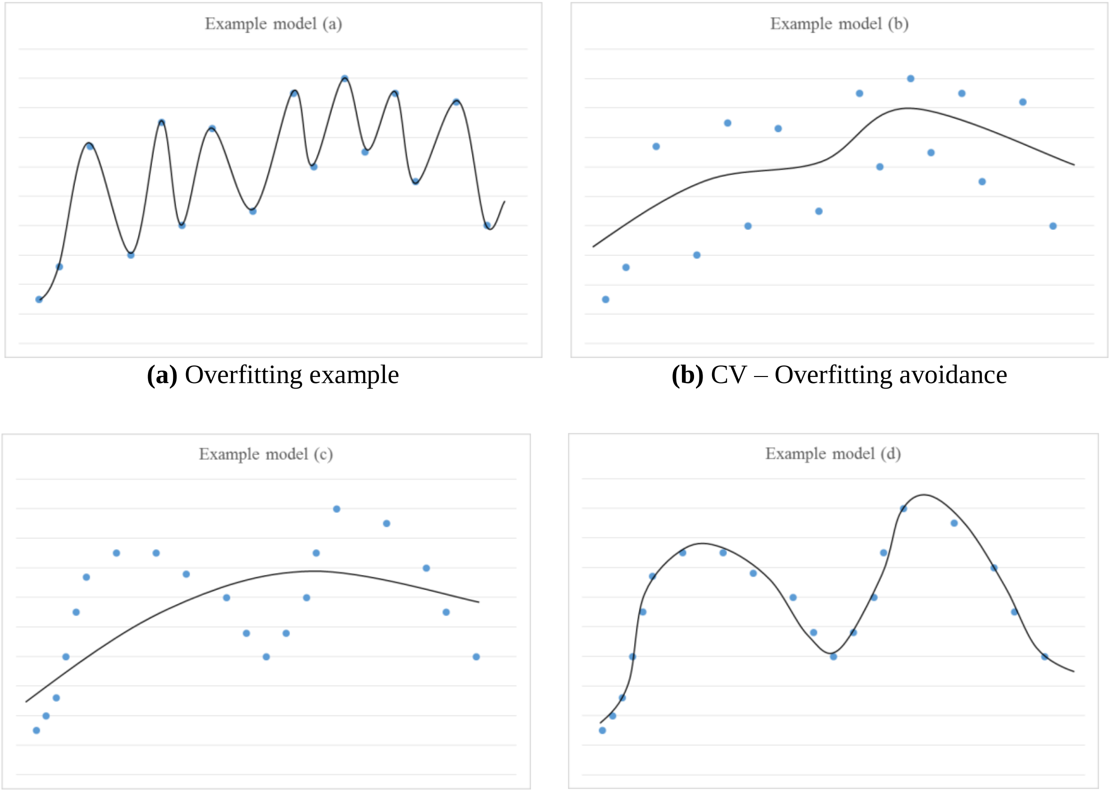
- Test/train cross validation, which involves splitting the dataset randomly into two subsets: (i) the training set, used to calibrate a model and (ii) the test set, used to evaluating model performance.
- k-fold cross validation, which involves splitting the dataset randomly into k folds (usually 5 or 10), calibrating the model using $k-1$ folds and testing it using the last remaining fold, and repeating this process until every fold has been used both for calibration and testing.
- Leave-one-out cross validation is an extreme case of k-fold cross validation, where each fold consists of a single data point $(k = n)$. In other words, the model is calibrated for every record in the data set except one, which serves as the test point.

Cross-validation aims to counter two usual errors in statistical model calibration: overfitting and underfitting. Overfitting refers to models being too closely fit to a particular dataset, following each point. While this might mean an accurate depiction of a relationship with the particular dataset, the predictive/transferability capabilities of an overfitted model are poor. Conversely, underfitting to models too loosely fit to a dataset, which might lead them to ignore trends in a relationship, again diminishing the predictive/transferability capabilities of the model. These errors can be countered with cross-validation techniques, such as the aforementioned three. Examples of overfitting and underfitting errors and their remedies are visually presented in Figure 3-4 (a, b, c, d).

Normally, in aspatial datasets, test/train divisions or k-fold cross validation is used based on the nature of the data as described by Bengio & Grandvalet, (2004), for instance. However, spatial data are not typically compatible with separations of datasets in randomly split fractions; random separation would lead to unrealistic and misleading spatial configurations, and provide false relations of neighbours that are not in reality such. This would lead to misspecification of spatial models in turn.

Based on Section 3.2.5.1, when conducting GWR, data values that are proximal to the regression point are taken into account, weighted by their relative distance from that regression point. The weighting function takes into account data points from up to a certain kernel bandwidth, which is typically a Gaussian kernel. In other words, data points in distances greater than the bandwidth have a weight of zero.

The first step in applying GWR is the selection of an appropriate kernel bandwidth. This is a crucial step in the GWR process which serves both for model calibration and for cross-validation. Fixed width kernels are normally used, and the optimal values are obtained by leave-one-out cross validation (Bivand, 2017) across all of the data points n.



(a) Overfitting example

(b) CV – Overfitting avoidance

(c) Underfitting example

(d) CV – Underfitting avoidance

Figure 3-4: Overfitting and underfitting examples and avoidances by CV

Computationally, the optimal bandwidth for GWR is the one which minimizes the following function:

$$CV = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \qquad \text{Eq. (28)}$$

Where:

- n is the sum of the data points
- y_i is the true value of the dependent variable
- $\hat{y}_i$ is the predicted value of the dependent variable

Bandwidth selection is a form of leave-one-out cross validation. This means that n models have to be fitted, which can be a computationally demanding process, as Bivand (2017) mentions.

Another noteworthy point that Fotheringham et al. (2002) raise is that GWR results are insensitive to the selection of a specific weighting function, while being more sensitive to the selection of a specific bandwidth value. Furthermore, they state that adaptive spatial kernels might be employed instead of fixed kernels in areas of sparse data, which provide similar results but result to somewhat smoother maps where the dependent variables are continuous over a surface.

3.2.5.3 Geographically Weighted Regression in a Generalized Linear Model context

After the initial form was established, the concept of GWR was extended to include regression techniques in a Generalized Linear Model context (GWGLM), again described by Fotheringham et al. (2002). This was an essential step which provided an econometrics method for spatially modelling count models. Linear models tend to fit count data inadequately, and often project negative predictions. While the second issue can be circumvented by modelling the logarithm of the dependent variable, log-linear models cannot handle zeros in the dependent variable. Therefore the need for spatial count modelling is evident.

The primary GWGLM distribution for crash analysis is the Poisson distribution, which has been employed in a number of studies as described in Section 2. Theoretical GWGLM negative binomial model concepts have been developed as well. However, comparatively, Geographically Weighted Poisson Regression (GWPR) is a far more widely applied method in the literature, and thus more robustly scrutinized. In addition, it has better algorithmic refinement options and support available, and it is also more comparable with Bayesian Poisson-lognormal models. As such, it was consider the more promising method to be applied for the segment-based investigation of harsh events in this doctoral dissertation.

The general form of a GLM models the log odds of the frequency (count) of the y_i via a linear predictor. Following McCulloch (2003), if λ_i are the expected frequencies of the y_i, a count variable, then an aspatial Poisson GLM is specified as:

$$y_i \sim Poisson(\lambda_i) \qquad \text{Eq. (29)}$$

Which translates to the probability of location i having y_i events, $P(y_i)$:

$$P(y_i) = \frac{{\lambda_i}^{y_i} * e^{-\lambda_i}}{y_i!} \qquad \text{Eq. (30)}$$

The linear predictor is formulated with the same notation as Equation (26):

$$ln(\lambda_i) = b_0 + \sum_{k=1}^{n} [b_k * x_{ik}] \qquad \text{Eq. (31)}$$

Following Fotheringham et al. (2002), by allowing spatial variation in the coefficients, the GWPR model is obtained, given by Equation (32):

$$ln(\lambda_i) = b_0(u_i, v_i) + \sum_{k=1}^{n} [b_k(u_i, v_i) * x_{ik}] \qquad \text{Eq. (32)}$$

In transport and road safety research, a good practice involves including relevant exposure parameters in the models, in order to establish a common baseline for parameter comparison (such as crash risk or event risk) between different observations or different model specifications (e.g. Imprialou et al., 2016). In a Poisson framework, exposure parameters or the constant term are often presented in a logarithmic form, similar to the response variable, shown in Equation (33):

$$ln(\lambda_i) = ln(b_0(u_i, v_i)) + b_e(u_i, v_i) * ln(x_{ie}) + \sum_{k=1}^{n}[b_k(u_i, v_i) * x_{ik}] \qquad \text{Eq. (33)}$$

Where, additionally:

- x_{ie} are the independent exposure variables e at a given point i

An example of the application of this form of GWPR in road safety can be found in Hadayeghi et al. (2010). As in all Poisson-based models, marginal effects can be used to examine the effect of a single-unit change in the independent variables x_{ik} on the dependent variable y_i. Following Washington et al. (2010), marginal effects – ME – are computed as:

$$ME_{x_i}^{y_i} = \frac{\partial y_i}{\partial x_{ik}} = b_k EXP(b_i x_i) \qquad \text{Eq. (34)}$$

Marginal effects can be more comprehensive when dealing with integer variables or when dealing with binary "flag" categorical variables. Since the derivative is still a function of the independent variables x_{ik}, an input value is required. A commonly used value is the mean, yielding Marginal Effects at the Means (MEM).

3.2.5.4 Semi-parametric Geographically Weighted Regression

Complementary to the previous, semi-parametric variations of GWR and GWPR models (termed SGWR and SGWPR respectively) have been developed as well, developed firstly by Nakaya et al. (2009). To obtain an SGWR model, the GWR model is further extended by allowing some variable coefficients to vary locally (group l) while others retain their global regression averages (group g). Combining Equations (27) and (32), Equation (35) describes a SGWR model:

$$y_i = \ b_0 + \sum_{k=1}^{l}[b_k * x_{ik}] + \sum_{k=1}^{g}[b_k(u_i, v_i) * x_{ik}] + \varepsilon_i \qquad \text{Eq. (35)}$$

Similarly, Equation (36) describes a SGWPR model:

$$ln(\lambda_i) = b_0 + \sum_{k=1}^{l}[b_k * x_{ik}] + \sum_{k=1}^{g}[b_k(u_i, v_i) * x_{ik}] \qquad \text{Eq. (36)}$$

Exposure parameters can also be integrated in a logarithmic form as in the baseline GWPR. There is a number of pseudo-R^2 metrics available for GWPR, including Cox & Snell, Nagelkerke and McFadden pseudo-R^2. Ultimately, researchers agree that there is no single metric that is absolutely better overall. The usual course of action is to select one and acquire an indication of goodness-of-fit, but equally importantly to compare its values across competing models for the same data, similar to AIC.

An important point is that, despite some contentions, GWR/GWPR has been proven robust against multicollinearity issues from a correlated mix of independent variables in relevant general-topic literature (Fotheringham & Oshan, 2016) and on specialized road safety research (Gomes et al., 2017).

As a final remark on the model family of geographically weighted regression, it should be noted that GWR/GWPR is predominantly used as an exploratory technique, and its use for prediction is somewhat contested. In that capacity, it makes sense to introduce additional models in the assessment and prediction processes of harsh events across urban network segments. SGWPR particularly has been reported to suffer in terms of result transferability in a road safety context (Xu and Huang, 2015). Moreover, Lu et al. (2014) state that it is important to conduct the respective baseline global regression before the respective GWR/GWPR application, so that any acquired local value is compared to its global benchmark counterpart.

3.2.6 Conditional Autoregressive Prior models

3.2.6.1 Conditional Autoregressive Prior model overview

Residual spatial autocorrelation in a regression model can violate the assumption of independence, which is a common prerequisite for many regression analysis. This can be caused if critical spatially autocorrelated covariates are omitted from the model (Lee, 2013). Autoregressive models overcome this obstacle by enhancing the linear predictor with a set of spatially autocorrelated random effects in a hierarchical Bayesian framework. These random effects can be represented by conditional autoregressive Bayesian priors (CAR priors) which are integrated in baseline models, such as Bayesian GLM models, resulting in Bayesian CAR GLM models.

As evident from Section 2.1.3, Conditional Autoregressive Prior Models (also known as Conditional Autoregressive Models or CAR Models) have become a very popular tool of spatial analysis in road safety. Alongside simultaneous autoregressive (SAR) models, CAR models have been used for: (1) model selection, (2) spatial regression, (3) estimation of spatial autocorrelation, (4) investigation of connectivity interactions, (5) spatial prediction, and (6) spatial smoothing (Ver Hoef et al., 2018). CAR models can handle both latent and observed variables, and statistical inference can be conducted via either: (1) maximum likelihood estimation (MLE) or (2) a Bayesian approach framework (de Oliveira, 2010).

In essence, CAR priors act as proxies that allow for substitution of unmeasured or unobserved risk factors which can vary spatially. CAR priors, which constitute the aforementioned spatially autocorrelated random effects, can be formulated from several different configurations. Seminal configurations include those proposed by Besag et al. (1991), Stern and Cressie (1999) and Leroux et al. (2000). In this dissertation, Bayesian CAR models with Poisson-based distributions are implemented, and thus they are the sole focus of this section, after a brief introduction to Bayesian analysis.

3.2.6.2 Bayes' Theorem and Bayesian Inference

Bayesian analysis and modelling has been an extensive part of statistics, gaining popularity during the second half of the past century. The cornerstone of Bayesian modelling is Bayes' theorem, expressed by Reverend Thomas Bayes in 1763, briefly outlined in the following. Mathematical proofs and additional material can be found in many relevant sources (e.g. Bolstad, 2007).

In its simplest form, for discrete events A and B, provided that the probability of B, $P(B) \neq 0$, Bayes' theorem states that:

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)} \qquad \text{Eq. (37)}$$

Where:

- $P(A|B)$ is the conditional probability, the likelihood of event A occurring given that event B is true.

In Bayesian analysis, every component of the theorem has a specific term. Based on the notation of Equation (37):

- $P(A|B)$ is the posterior probability
- $P(B|A)$ is the likelihood
- $P(A)$ is the prior probability
- $P(B)$ is the marginal likelihood

The process of Bayesian inference refers to constant updates of a prior hypothesis in light of more information for the examined events. When conducting Bayesian inference, the probability $P(B)$ is considered a fixed quantity. This means that the shape of the probability distribution is given by the numerator of Equation (37), while the denominator is fixed in the problem under consideration. In other words, various outcomes in multiple possible instances of A and $P(A)$ are of interest. Therefore one can use proportionality instead of equality:

$$P(A|B) \propto P(B|A) * P(A) \qquad \text{Eq. (38)}$$

Which effectively means that the posterior probability is proportional to the prior probability times the likelihood. The above framework can be extended from events to random variables. If it is considered that A and B now refer to the event that random variables X and Y assume specific values x and y so that:

$$A = \{X = x\} \qquad \text{Eq. (39)}$$

And

$$B = \{Y = y\} \qquad \text{Eq. (40)}$$

Then Equation (37) can be re-written as:

$$P(X = x|Y = y) = \frac{P(Y = y|X = x) * P(X = x)}{P(Y = y)} \qquad \text{Eq. (41)}$$

Equation (41) is also known as Bayes' rule. Additional variations can be derived in case of discrete or continuous random variables, as well as joint variables. To link to a functional form, the utilized function is linked to the conditional probability for particular values, using notation relating closer to the following sections:

$$f(y_i|x_i) = P(Y = y_i|X = x_i) \qquad \text{Eq. (42)}$$

Therefore, if X and Y are two continuous variables, with probability density functions $f(X|Y)$ and $f(Y)$, respectively, then:

$$f(Y|X) = \frac{f(X|Y) * f(Y)}{f(X)} = \frac{f(X|Y) * f(Y)}{\int_{-\infty}^{+\infty} f(X|Y) * f(Y)dY} \qquad \text{Eq. (43)}$$

Bayes' rule is used for calibrating Bayesian models via Bayesian inference, a process also known as Bayesian regression. The marginal likelihood $f(X)$ can be considered as a non-zero constant, c, indicating the initial form of X, therefore not important for detecting what influences the values for Y. It can be thus omitted if equation is substituted for proportionality:

$$f(Y|X) = \frac{f(X|Y) * f(Y)}{c} \propto f(X|Y) * f(Y) \qquad \text{Eq. (44)}$$

In a Bayesian setting, regressing a count function involves integrating over the posterior distribution posterior distribution of the coefficient of interest. The exact form varies depending on the utilized distribution. As stated previously, Poisson-based models are examined in this dissertation. Mathematically, predicting the posterior distribution for any coefficient β is given by Equation (45) for Poisson-based regression, following Chan and Vasconcelos (2009):

$$P(\beta|X,Y) = \frac{P(Y|X,\beta) * P(\beta)}{\int P(Y|X,\beta) * P(\beta) d\beta} \qquad \text{Eq. (45)}$$

Where:

- $\{X, Y\}$ is the training dataset (independent and dependent variables, respectively)
- β is the coefficient, the distribution of which is being calculated. Usually the coefficient prior distribution is a Normal distribution (Gaussian prior), so that:

$$\beta \sim N(0, \Sigma_p) \qquad \text{Eq. (46)}$$

Where:

- Σ_p is the covariance matrix of the weight prior.

Therefore by obtaining the coefficient distributions β, along with any other specified effects, the Bayesian model is obtained.

3.2.6.3 Conditional Autoregressive Prior model formulation

Following several sources (Bivand et al., 2009; Lee, 2013; Lee, 2014b; Ver Hoef et al., 2018), the formulation of a CAR model is given for a count model in the following. In road safety, CAR models have been used for modelling event frequencies (typically crash counts), therefore the response variable is specified as following the Poisson distribution, similar to Equation (30). Baseline Poisson models are known to handle over-dispersion poorly, and thus a preferable alternative is Negative Binomial-based models (Lord & Mannering, 2010). However, when conducting spatial analysis, the introduction of spatially structured effects cause additional demands in the models. Cai et al. (2018) have determined that Bayesian Poisson-lognormal models are best suited to handle both over-dispersion and spatially structured effects simultaneously. Thus, this dissertation utilized Bayesian Poisson-lognormal models with CAR priors, which are described below.

The Hierarchical Bayesian setting is:

$$y_i \sim Poisson(\lambda_i) \qquad \text{Eq. (47)}$$

Where, again, λ_i are the expected frequencies of the dependent variable y_i in a location i. Thus the likelihood function of Equation (37) can be written as:

$$f(y_i|\lambda_i) = P(y_i|\lambda_i) = \frac{\lambda_i^{y_i} * e^{-\lambda_i}}{y_i!} \qquad \text{Eq. (48)}$$

The relative risk is:

$$ln(\lambda_i) = b_0 + \sum_{k=1}^{n}[b_k * x_{ik}] + \varphi_i + \theta_\iota \qquad \text{Eq. (49)}$$

Where:

- y_i is the dependent (or response) variable at a location point i
- x_{ik} are the independent (or explanatory) n variables at a given location i
- b_k is the coefficient distribution of a particular x_k
- b_0 is the coefficient distribution of the constant term
- θ_i is the spatially unstructured error term at a given point i
- φ_i are spatially autocorrelated (structured) random effects

As evident from Equation (49), in Bayesian statistical models, beta coefficients are not fixed effects (single optimum values), but are rather sampled from distributions which are normal distributions in this case. A popular quote by Weiss states that *"If you're not using a proper, informative prior, you're leaving money on the table."* However, the nature of road safety events urge researchers to avoid making prior assumptions about the extent of the effect of each independent variable. Therefore, coefficients are given non-informative prior values to eliminate any bias during the calibration phase (e.g. Mitra 2009; Lee et al. 2015a; Alarifi et al. 2017). A sample non-informative prior distribution has a mean of zero and a very large variance, such as:

$$b_k \sim N(0, 10^6) \qquad \text{Eq. (50)}$$

Known offsets, namely parameters that are known to influence the response variable with a direct relation may also be specified and integrated in the model without a coefficient. In that case, $b_{offs} = 1$ (fixed effect).

The spatially unstructured error random effects, θ_i, are set to follow a normal distribution:

$$\theta_i \sim N\left(0, \frac{1}{{\tau_{\theta i}}^2}\right) \equiv N(0, {\sigma_{\theta i}}^2) \qquad \text{Eq. (51)}$$

Where:

- ${\tau_{\theta i}}^2$ is the precision parameter that is the inverse of the distribution variance of the spatially unstructured effects

The precision parameter ${\tau_{\theta i}}^2$ is assigned a Gamma (Γ) distribution prior; indicatively, Wakefield et al. (2000) suggest values of ${\tau_{\theta i}}^2 \sim Gamma(0.5, 0.0005)$, while Cai et al. (2018) use values of ${\tau_{\theta i}}^2 \sim Gamma(0.001, 0.001)$. Abdel-Aty et al. (2013) explain that this variance, $1/{\tau_{\theta i}}^2$, provides the amount of variation not explained by the Poisson assumption (Lawson et al., 2003), which states that events are independent, homogeneous and occur during a fixed time period.

Regarding spatial random effects, there are several available configurations. The intrinsic form of CAR models is a widely adopted one, and is defined as follows. The spatially autocorrelated (structured) random effects, φ_i, are set to follow a normal distribution as proposed by Besag (1974; 1991) so that:

$$\varphi_i \sim N\left(\overline{\varphi_\iota}, \frac{1}{{\tau_{\varphi i}}^2}\right) \equiv N\left(\overline{\varphi_\iota}, {\sigma_{\varphi i}}^2\right) \qquad \text{Eq. (52)}$$

The mean $\overline{\varphi_\iota}$ is defined as:

$$\overline{\varphi_\iota} = \frac{\sum_{i \neq j}^{n} \varphi_i * w_{ij}}{\sum_{i \neq j}^{n} w_{ij}} \qquad \text{Eq. (53)}$$

Where:

- w_{ij} is a matrix of spatial weights given by a selected geographical criterion with diagonal elements equal to zero ($w_{ii} = 0$)
- ${\tau_{\varphi i}}^2$ is the precision parameter that is the inverse of the distribution variance of the spatially structured effects

Once again, for a given location i, spatial weights are assigned with a similar reasoning as Equation (9), and similar weighting techniques can be used as in Moran's I calculations. It can be easily gleaned that the weighting function and the resulting values have a direct impact on the spatially autocorrelated effects. In the literature, ${\tau_{\varphi i}}^2$ is assigned Gamma (Γ) distribution priors similar to the aforementioned parameter ${\tau_{\theta i}}^2$.

The structure for the spatially correlated term described in Equations (52) & (53) is also known as the Besag-York-Mollie CAR (or BYM CAR) model form and has been implemented in many road safety studies since its inception (e.g. Huang et al., 2016; Cai et al., 2018; Zhai et al., 2018; Wen et al., 2019). The equivalent values of ${\sigma_{\theta i}}^2$ and ${\sigma_{\varphi i}}^2$ are also provided for spatially unstructured and structured effects respectively, because they are often reported in this form due to computational reasons instead of the inverse effects.

Regarding model calibration, Lee (2014b) mentions that θ_i values are contained in the posterior probability $P(\theta|y)$, thus Equation (45) is rewritten as:

$$P(\theta|X,Y) = \frac{P(X,Y|\theta) * P(\theta)}{\int P(Y|X,\theta) * P(\theta) d\theta} \qquad \text{Eq. (54)}$$

3.2.6.4 Conditional Autoregressive Prior model evaluation

Computationally, Bayesian inference is conducted using Markov chain Monte Carlo (MCMC) simulation. Instead of cross-validation, MCMC processes utilize the "burn-in" practice. This practice involves the creation of a "burn-in" period at the start of the simulation. The initial iterations that are performed during the "burn-in" period are discarded because the simulation output is still exploring the convergence path, and as such has not started to converge yet. In practice, the number of iterations on every model varies in order to fulfill a requirement that the MCMC error is less than 5% of the standard deviation of the parameter being estimated, and can reach typical values of 50,000 to 100,000 iterations or higher (Aguero-Valverde, 2014; Guadamuz-Flores & Aguero-Valverde, 2017).

In order to evaluate model performance, there are several Bayesian measures of model complexity and fit. For b_0 and b_k, the significance of the estimated coefficient distributions is determined based on Bayesian Credible Intervals (BCI), which reflect of the value of the true parameters of the distributions.

For example, a BCI of 95% will contain the true parameters 95% of the time. The selection between candidate variable sets is calculated using the Deviance Information Criterion (DIC), which is a hierarchical modeling generalization of the Akaike Information Criterion (AIC) provided by Spiegelhalter et al. (2002):

$$DIC = 2 * \overline{D} - \widehat{D} \quad \text{Eq. (55)}$$

With the deviance D being defined as:

$$D = -2\log(P(y|\theta)) \quad \text{Eq. (56)}$$

Where:

- $\overline{D}$ is the posterior mean of D
- $\widehat{D} = 2 * P(y\middle|\bar{\theta})$
- $\bar{\theta}$ is the posterior mean of θ

For a given dataset, models which minimize DIC are preferred to the alternatives, but only when referring to identical areas (Abdel-Aty et al., 2013). Additionally, Lee (2013) mentions that Watanabe's modified AIC, noted as WAIC, and the Log Marginal Predictive Likelihood (LMPL) can both be used in conjunction with DIC. Specifically, the model with overall best fit is one that minimizes DIC and WAIC but maximizes LMPL. As stated in Section 3.1, similar to GWPR, spatial effects obtained from CAR priors in one area are not typically transferrable to other areas. Therefore Bayesian predictions in other areas have to be conducted with baseline Bayesian Poisson models without spatial effects.

3.2.7 Extreme Gradient Boosting – XGBoost

3.2.7.1 XGBoost overview

Extreme Gradient Boosting, henceforth referred to as XGBoost, is a machine learning (ML) technique, encompassing multiple Classification And Regression Trees (CART). Additionally, XGBoost belongs to the family of supervised ML techniques, meaning that it utilizes labeled training data, the structure of which is defined by the researcher. In practice, this means that the independent/dependent variable division is known and present in the examined variables, and the outcome is a mapping function to the effect of $y = f(x)$.

Being a ML technique, it is inherently data-driven, and thus free of any prior assumptions concerning underlying relationships in the data. Any obtained relationships are products of what lies strictly within the provided dataset. XGBoost originated from the seminal work of Chen & Guestrin (2016) – also presented in Chen et al. (2015) – who expanded known tree-boosting techniques to handle sparse data, approximate problems for better memory handling and ultimately be more scalable. Furthermore, in one of the very few, as of the time of writing, published applications of XGBoost in road safety, the algorithm was shown to outclass other ML techniques regarding the accuracy of injury severity classification tasks, especially with an increased number of features (Ting et al., 2020).

The aforementioned properties of XGBoost urged the selection of this technique despite it not being a technique used in spatial analysis or, to the extent of the author's knowledge, currently featuring any way to explicitly integrate spatial effects in model building. This essentially means that while calibration and prediction can be conducted for all road segments, the influence and characteristics of neighboring segments are not a separate factor, but rather integrated in the data.

3.2.7.2 XGBoost algorithm

XGBoost applies the gradient boosting decision tree algorithm, also known as multiple additive regression trees, stochastic gradient boosting or gradient boosting machines. The learning process of the algorithm is iterative and includes correction of previous errors in future iterations of the algorithm. XGBoost can be applied in both classification and regression problems. An overview of the algorithm is described in this section based on Chen & Guestrin (2016), more detailed explanations can be found in that study.

As previously mentioned, if a mapping function is considered between variables:

$$\hat{y} = f(x_i) \qquad \text{Eq. (57)}$$

Where:

- y is the dependent (or response) variable
- $\hat{y}$ is the predicted value of the dependent (or response) variable
- x_i are the independent (or explanatory) n variables across I observations

Then a regression tree ensemble model uses a number of functions K additively to predict y, so that:

$$\hat{y} = \varphi(x_i) = \sum_{k=1}^{K} f(x_i) \qquad \text{Eq. (58)}$$

Then the difference between a prediction $\hat{y}_i$ and a true value y_i can be measured via a loss function which is differentiable and convex, defined as $l(\hat{y}_i, y_i)$. In other words, the loss function expresses the distance between predictive values and the training data. A common choice of l is the mean squared error for a set of parameters φ_i (XGBoost developer team, 2019):

$$l(\varphi_i) = \sum_{\iota=1}^{I} (\hat{y}_i - y_i)^2 \qquad \text{Eq. (59)}$$

Furthermore, a penalizing term, $\Omega(f)$, is introduced for model complexity control such that:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \|c\|^2 \qquad \text{Eq. (60)}$$

Where:

- γ, λ are penalizing coefficients
- T is the number of leaves in the regression tree. Each leaf represents a value of the target variable given the values of the input variables represented by the path from the root to the leaf, creating a flowchart, as shown in Figure 3-5.
- c is the weight assigned to each leaf

Figure 3-5 depicts an example of a single decision tree from the ensemble utilized in XGBoost, and the respective components are labeled. The input is a single typical road safety variable, average traffic speed, $\overline{v_{tr}}$, and the output is the observed crash numbers at a single segment over the course of one month, n_{cr}. All numbers are hypothetical. In this example, the defining value of average speed is 50 km/h, and the different results are represented in the tree leaves.

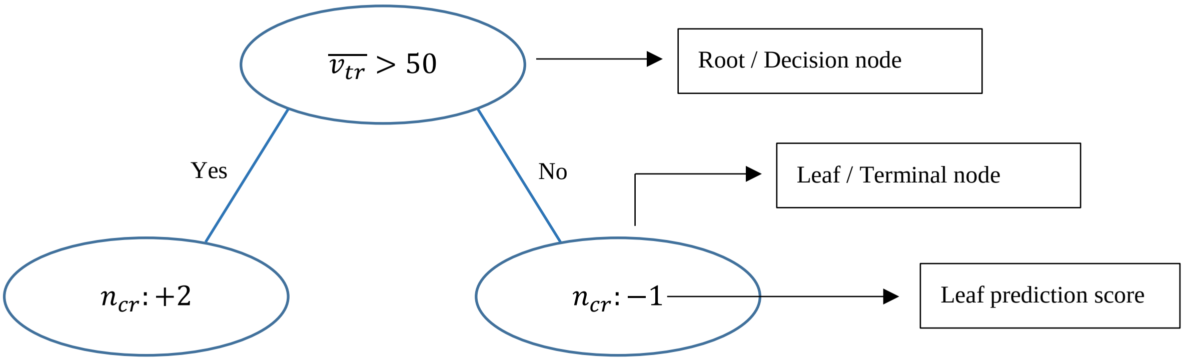


Figure 3-5: Example of a single decision tree from the XGBoost ensemble

It is evident that if $\overline{v_{tr}}$ exceeds the threshold of 50 km/h, the tree increases the prediction of crashes, and reduces it in the opposite case.

Continuing towards the algorithm formulation, having obtained the loss function, $l(\hat{y}_i, y_i)$, and the penalizing term, $\Omega(f)$, the objective function can be formulated as:

$$L(\varphi_i) = \sum_{i=1}^{I} l(\hat{y}_i, y_i) + \sum_{k=1}^{K} \Omega(f) \qquad \text{Eq. (61)}$$

At its core, the XGBoost algorithm relies heavily on the minimization of an objective function, which is in essence a function of functions. If this process is considered iteratively, then for the t-th iteration the additive function f_t is added to minimize the objective function formulated until that point $t-1$:

$$L_t(\varphi_i) = \sum_{i=1}^{I} l(y_i, \hat{y}_i{}^{(t-1)} + f_t(x_i)) + \Omega(f_t) \qquad \text{Eq. (62)}$$

Chen & Guestrin (2016) note that since nested functions exist inside the objective function, traditional optimization methods cannot be used. They then circumvented this obstacle by using second-order approximation from Taylor's Theorem to transform L to a function in the Euclidean domain, so that optimization methods can be now used. The function ensemble then reverts to simple quadratic functions that can be minimized normally. The second-order Taylor approximation is:

$$f(x) \approx f(m) + f'(m)(x-m) + \frac{1}{2}f''(m)(x-m)^2 \qquad \text{Eq. (63)}$$

Applied to Equation (62):

$$L_t(\varphi_i) \approx \sum_{i=1}^{I} \left[l\left(y_i, \hat{y}_i{}^{(t-1)}\right) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) \qquad \text{Eq. (64)}$$

Where:

$$g_i = \partial_{\hat{y}_i(t-1)} l\left(y_i, \hat{y}_i{}^{(t-1)}\right) \qquad \text{Eq. (65)}$$

And:

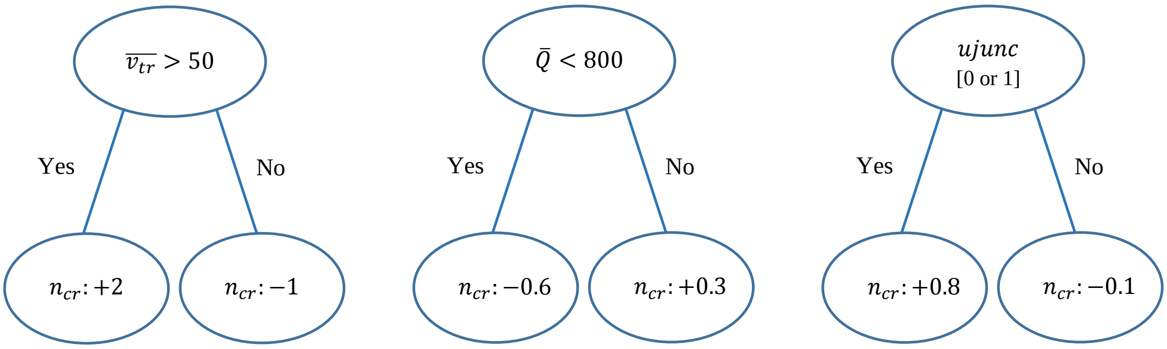
$$h_i = \partial^2{}_{\hat{y}_i(t-1)} l\left(y_i, \hat{y}_i{}^{(t-1)}\right) \qquad \text{Eq. (66)}$$

By removing constant terms, Chen & Guestrin (2016) reach the simplified objective function at iteration t:

$$\widetilde{L_t}(\varphi_i) = \sum_{i=1}^{I} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) \qquad \text{Eq. (67)}$$

The exact mathematical formulae from this point onward depend on the structure of the loss function. XGBoost supports count regression modelling with the Poisson cost function, as described in Equation (30), and that specific form is the ML model ultimately implemented in this dissertation. To visualize XGBoost, an ensemble of three decision trees is provided in Figure 3-6. The first is the previous example tree of Figure 3-5. The second and third trees refer to examining the expected crashes, n_{cr}, in the same hypothetical segment with two other variables: average traffic volume of the segment, $\bar{Q}$, which is continuous like speed, and the presence of uncontrolled junctions in the segment, $ujun$c, which is binary (yes/no variable):

Therefore, in a segment m that the aforementioned variables had values of $\overline{v_{tr}}_m = 67$ km/h, $\bar{Q}_m = 407$ veh/h and some uncontrolled junctions so that $ujunc$: yes, the respective ensemble would predict a crash number of $+2 - 0.6 + 0.8 = 2.2$ crashes in the segment overall.



(a) Decision tree for average traffic speed

(b) Decision tree for average traffic volume

(c) Decision tree for presence of uncontrolled junctions

Figure 3-6: Example of three decision trees forming an XGBoost ensemble

Moreover, in ML processes like XGBoost, it is recommended to engage in hyperparameter tuning (or optimization) before executing the final algorithm. This involves the selection of the optimal algorithmic hyperparameters for the problem at hand, and in a sense can be considered the equivalent of conducting exploratory trials in statistical models. To conduct hyperparameter tuning, a range is given to the ML hyperparameters, which are parameters that govern the structure of the algorithm, rather than describing the data. The process depends heavily on the dataset under consideration, and although it can be manual, typically it is automated. Various combinations of said parameters are selected from the respective available ranges and used to create algorithms which are tested on their performance on a specific metric, such as root mean square error, which is explained in Section 3.2.8.2.

For XGBoost, some of the typical hyperparameters that can be tuned are:

- Learning rate (also known as ETA), governing the magnitude of iterations for minimizing the cost function
- Gamma, governing the minimum loss reduction that can justify making a partition on a tree
- Maximum tree depth, greatly governing ensemble complexity and overfitting
- Evaluation metric, which is a target for minimization such as RMSE, RMSLE, MAE and others
- Number of k-folds for each cross-validation task
- Number of rounds that are tested before convergence of the cost function is finalized

Therefore, following good ML practices, the hyperparameters of XGBoost algorithms were initially tuned before their final executions, and their predictions were subsequently evaluated.

When dealing with large numbers of predictor variables, XGBoost algorithms have functions that can calculate the importance of each predictor variable. This is known as Gini feature importance, or, equivalently, Mean Decrease in Impurity (MDI), and was proposed in a seminal study by Breiman (2001). One definition for Gini Importance for tree-based algorithms is the following: Gini Importance is the value obtained as the sum over the number of splits that include the feature across all trees, optionally divided by the number of samples it splits. This allows for powerful and accurate models to be created by utilizing only the most important predictor variables from a given dataset. Several feature importance calculation methods exist, as outlined in Hastie et al. (2009).

In XGBoost, three particular variable importance metrics are observed (XGBoost developer team, 2019):

1. Gain, describing the improvement in accuracy added by a feature to the branches it is on.
2. Cover, describing the relative quantity of observations (or number of samples) concerned by a feature.
3. Frequency, describing the number of times a feature is used in all generated trees.

These variable importance metrics used by the XGBoost algorithms were calculated in the analysis and examined to reveal which variables are informative to describe harsh event frequencies.

3.2.7.3 Spatial Cross-Validation in machine learning models

Regarding XGBoost algorithmic performance, Shi et al. (2019) mention that cross-validation is recommended for XGBoost, and that various metrics are available; the three cross-validation methods mentioned in Section 3.2.5.2 can be applied to XGBoost predictors as in any functional predictor.

Another way to circumvent the aforementioned restrictions in cross-validation of spatial data is the separation of the spatial dataset in location clusters, a concept known as spatial cross-validation (SPCV). SPCV has emerged more recently and has been primarily applied to augment machine learning algorithms (Schratz et al. 2018; Lovelace et al., 2019), though there are no considerable barriers in transferring the method to other statistical methods.

As previously mentioned, spatial/geographic data require cautious treatment, in the sense that spatial analyses take the attributes of both the target location and its neighborhood into account. Therefore, it is much more meaningful to split the training dataset into k spatial folds (k neighborhoods) instead of k random folds, and then train the model in each as per normal k-fold cross-validation. The concept can be intuitively visualized, as shown on Figure 3-7.

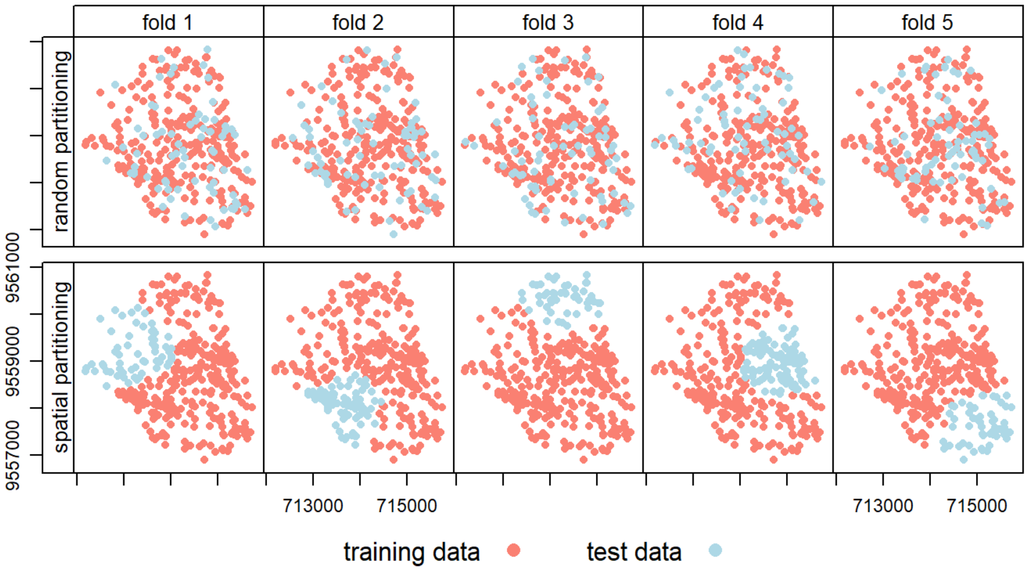


Figure 3-7: Visualization of random and spatial k-fold cross-validation
[Source: Lovelace et al., 2019]

By using SPCV, the model is trained while retaining local characteristics that are integrated in independent variable values in each of the spatial folds. This has the potential to lead to more accurate predictions, and respects the spatial structure of the data. In the present doctoral dissertation, SPCV was implemented to augment the XGBoost algorithm to better preserve spatial relationships that might be underlying in the road segment data. Performance comparisons between SPCV XGBoost, random cross-validation (RCV) XGBoost and functional methods were made as well.

3.2.8 Assessment and integration of model predictions

3.2.8.1 Theoretical assessment of utilized statistical models

Based on the existing literature on the various statistical methods that are examined, a list of their respective main advantages and disadvantages is provided on **Table 3-1**. It should be noted that this assessment is within the context of the scope of the current doctoral dissertation and the contents of the table refer to the research questions at hand:

Table 3-1: Main advantages and disadvantages of the statistical methods of the dissertation

Method	Method type	Main advantages	Main disadvantages
GWR	Frequentist spatial analysis	• Easy interpretability • Intuitive assessment of spatial heterogeneity in estimated relationships	• Weak result transferability & generalization • Linear form does not handle counts or rates well overall, Poisson form has issues with local clusters of zeroes
CAR	Bayesian spatial analysis	• Coefficient distributions offer increased flexibility compared to fixed optimal values • Estimates the probability that a hypothesis is true given the data (and not the opposite) • No bias from reduced sample size	• Absence of informative priors in road safety • Conditional on observed data • Can be computationally demanding depending on requested simulations
XGBoost – random CV	Machine learning	• Data-driven approach • High execution speed • One of the most potent known ML algorithms • Lower data-point requirements than other ML algorithms • Integrated count modelling with Poisson functional features	• "Black box"; no clear interpretability of independent variable influences • No way at present to integrate spatial effects separately • Random CV may distort relationships spatially
XGBoost – spatial CV		• Preservation of spatial relationships in data with SPCV	• Adequate observations required in each spatial fold, increasing data demand

3.2.8.2 Evaluation of model predictions - Performance metrics

After calibrating a model on a test dataset, good practice, and, frequently, research demands dictate that predictions are made by applying the model on a training dataset. Several metrics can be then calculated to determine the performance of the model in the prediction task, which typically measure the difference of the model predictions from the true values reported on the test dataset. Road safety studies have utilized several forms of metrics in the past without considering one as optimal over the others, but rather some as more appropriate to specific data over others (e.g. Dong et al., 2015).

At this point, it is necessary to note that prediction based on calibration of a spatial model in a training area and its application on a testing area is not widely used with spatial methods; this mostly stems from the limited transferability of spatial effect terms without known counts /concentrations of the dependent variable. Therefore predictions on the test area had to be made from the GLM (Poisson) component for the two methods. This was another major motivator for the inclusion of XGBoost algorithms in the analysis, since a main strength of machine learning methods lies in their predictive power, though this often incurs a trade-off in interpretability.

The problem tackled in this dissertation is a regression problem with a frequency (count) dependent variable. Amongst the several applicable metrics, three have been determined as appropriate, intuitive and informative for the developed spatial models. These are: (a) (Root) Mean Squared Error (RMSE/MSE), (b) Mean Absolute Error (or Deviation) (MAE/MAD) and (c) (Root) Mean Squared Log Error (RMSLE/MSLE). Since they all represent errors, the smaller their values are, the better the predictive power of the model. The three chosen metrics are not exhaustive; several similar metrics can be devised and monitored, based on researcher preferences and the specifics of the datasets at hand.

If the notation of Equation (26) is followed so that:

- n is the sum of the data points
- y_i is the true value of the dependent variable
- $\widehat{y_i}$ is the predicted value of the dependent variable

In performance metrics, the term 'error' usually denotes a measure of the difference between true and predicted values of the dependent variable.

The Mean Squared Error is defined as:

$$MSE = \frac{1}{N}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2 \qquad \text{Eq. (68)}$$

Respectively, the square root of MSE is the RMSE:

$$RMSE = \sqrt[2]{\frac{1}{N}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2} \qquad \text{Eq. (69)}$$

The reason for selecting RMSE over MSE is that RMSE is a metric on a same scale as the dependent variable, and not squared. This can be viewed as similar to preferring standard deviation over the variance of a variable.

Then the Mean Absolute Error, also known as Mean Absolute Deviation (MAD), is defined as:

$$MAE = MAD = \frac{1}{N}\sum_{i=1}^{n}|y_i - \widehat{y_i}| \qquad \text{Eq. (70)}$$

MAE is the simplest metric which expresses the average between the true and predicted values. An advantage of MAE, especially over (R)MSE, is that MAE is more robust against outlier values. This is due to the fact that outlier errors are not squared in MAE, and thus their contributions are not distorted. The squared error structure of RMSE grants it an interesting property: large errors have individually more pronounced effects on the metric compared to an equal amount of error spread across more observations. This grants RMSE the capability to indicate isolated larger errors, but it can also cause discrepancies between RMSE and MAE.

Therefore, a third metric is also introduced to aid in model optimization, the Mean Squared Log Error (MSLE), which is defined as:

$$MSLE = \frac{1}{N}\sum_{i=1}^{n}(log(\widehat{y_i}+1)-log(y_i+1))^2 \qquad \text{Eq. (71)}$$

Respectively, the square root of MSLE is the RMSLE:

$$RMSLE = \sqrt[2]{\frac{1}{N}\sum_{i=1}^{n}(log(\widehat{y_i}+1)-log(y_i+1))^2} \qquad \text{Eq. (72)}$$

In both cases, the predicted and actual values are increased by one to avoid errors due to zeros in the dataset. If the predicted and true values are small, (R)MSE and (R)MSLE converge. RMLSE is robust to outliers as well, due to its square-root of logarithmic difference structure that can absorb high values. The main argument of this metric is the logarithmic difference, which can be equally transformed to its fractional form, which in turn is fundamentally a relative calculation error (Inagaki et al., 2019):

$$log(\widehat{y_i}+1)-log(y_i+1) = \frac{log(\widehat{y_i}+1)}{log(y_i+1)} \qquad \text{Eq. (73)}$$

Since the present problem is a count problem, model predictions can be safely rounded as natural numbers (positive integers) and zeros. It should be mentioned that additional metrics popular in the field of Machine Learning, such as the Mean Absolute Percent Error (MAPE), cannot be used if zeros are present in the dataset, because they migrate to the fraction denominator. However, in lieu of these unusable metrics, another metric is hereby devised which bridges the gap between regression and classification metrics: custom accuracy (CA).

The reasoning behind custom accuracy is to denote as acceptable each prediction that is up to one count removed from the truth. In the context of the present dissertation, if one considers a road segment with 6 harsh brakings, predictions of 5, 6 and 7 counts would all be characterized as 'accurate'. As the term implies, custom accuracy grants an intuitive percentage of correct count predictions. In mathematical terms, CA is expressed as follows.

First the total number of predictions is classified into accurate and not accurate:

$$N_{\widehat{y_i}} = \begin{cases} accurate, & if\ |y_i-\widehat{y_i}|\ \leq 1 \\ not\ accurate, & otherwise \end{cases} \qquad \text{Eq. (74)}$$

And then CA is obtained as the percentage of accurate predictions from the total:

$$CA = \frac{N_{\widehat{y_i}}\ [accurate]}{N_{total}} \qquad \text{Eq. (75)}$$

As CA is a percentage of accuracy, higher values indicate models with better predictive ability.

Finally, when assessing model predictions it is recommended to examine the plot of predicted values against the true values of the predicted variable. In certain algorithms that feature rule-based calibration, such as XGBoost, it is important to ensure that no rules that limit model predictions horizontally are enforced, such as stunting model predictions under a certain value in favor for 'blind' gains in RMSE.

3.2.8.3 Combining model predictions

Statistical models are abstract descriptions of the underlying phenomena of reality; as such, they will always inherently contain flaws, biases and errors from the truth. Models from different methodologies can only be tested in terms of predictive performance against specific datasets, and are not directly comparable with metrics such as AIC. The two alternatives then is to adopt one model that appears to be better in the circumstances of its testing, or to combine model predictions.

Combined model predictions have been used in several cases of the literature for averaging of predictions in the field of transportation engineering (e.g. Ma et al., 2020) and other sciences, such as economics (e.g. Dash & Cooper, 2004; Berge, 2015). Specifically in economics, model combination was found to alleviate model misspecification and even surpass individual models (Hendry and Clements, 2004; Timmermann, 2006).

There have also been more sophisticated approaches that involve various models supplying different components that lead to a combined prediction instead of an average. For instance Li et al. (2020) combined a long short-term memory neural network with a convolutional neural network for real-time crash risk predictions.

Some of the studies that involve model combination and are arguably most relevant to the topic of the present doctoral dissertation are the ones by Lovegrove and Sayed (2006; 2007) and concern the development of micro- and macro-level crash prediction models (CPM) and subsequent combination of their predictions.

Following the investigation of the average prediction approach in the literature, the author adopts the approach suggesting that model discrepancies can be mitigated by averaging the predictions of different methodologies in order to minimize the errors of each methodology.

The outputs of the models used in this doctoral dissertation are all harsh event frequencies, in other words, count data. Therefore, final predictions will be conducted as the equal-weight average of the predictions of (i) frequentist GLMs, (ii) Bayesian GLMs, (iii) RCV XGBoost and (iv) SPCV XGBoost models.

3.3 Data sources and tools

3.3.1 Digital map data

In its core, the present doctoral research constitutes infrastructure-based analysis and assessment. This process is conducted with road geometry and infrastructure data that is either directly extracted from digital online maps or inferred from data extracted from these maps, as explained in this section. After the analysis, the depiction of several model results, as well as statistical predictions, were also displayed on maps. Therefore a map system that is (i) flexible, (ii) reliable and (iii) accessible is required for the purposes of the present doctoral dissertation. With the previous criteria in mind, and following consideration of several alternatives, a decision was made to use maps from the open-source platform OpenStreetMap.

3.3.1.1 OpenStreetMap background

The OpenStreetMap (OSM) project is a knowledge collective that provides user-generated street maps. In other words, OSM is a project that exploits Volunteered Geographical Information (VGI) (Goodchild, 2008). The OSM project originated from University College London in 2004, and exponential crowdsourcing contributions with constant, iterative additions and corrections, have created a reliable Open Geodata repository suitable for high quality research needs (Haklay & Weber, 2008). Several corporations, projects and medium or smaller businesses regularly used OSM, granting further credence to the OSM project (OSM, 2019). OSM coverage started from England and has progressed from capturing 29% of England in 2009 (Haklay, 2010) to worldwide coverage (Zhang & Malczewski, 2019).

As noted on the official OSM website, OSM is a free, editable map of the whole world that is being built by volunteers largely from zero basis, and released with an open-content license. The OpenStreetMap License allows free access to the map images and all of the underlying map data, and one of the core aims of the project is to actively promote new and interesting uses of map data (OSM, 2019). The open-source nature of the data and the corresponding freedom to utilize them without charges, copyright concerns or limitations from locked interfaces (such as application program interfaces – APIs). OSM data had an accuracy of 80% to 90% in segment length and a measurement error of about ±6 m a decade ago, and has constantly been improving since (Haklay, 2010; Zhang & Malczewski, 2019). OSM uses the WGS84 coordinate system, common to most GPS units and services.

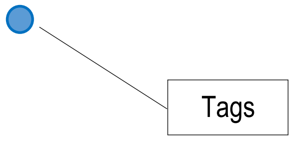
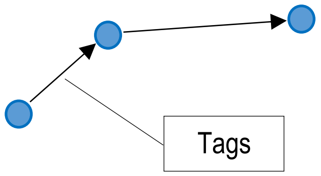
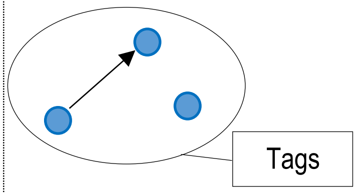
3.3.1.2 OpenStreetMap core elements

The OSM project consists of three core, fundamental elements which comprise the conceptual data model of the physical world (OSM, 2019):

1. Nodes, which are used to define points in space
2. Ways, which are used to define linear features and area boundaries
3. Relations, which are used to note interactions and relationships between elements

These core elements can have any number of corresponding associated tags. Example types of relations are 'bus route' and 'fire hydrant', denoting the locations of all bus stops of a single route or all fire hydrants in the examined area, respectively. Graphical examples of these fundamental elements are depicted in **Table 3-2**.

Table 3-2: Graphical representation of OSM core elements

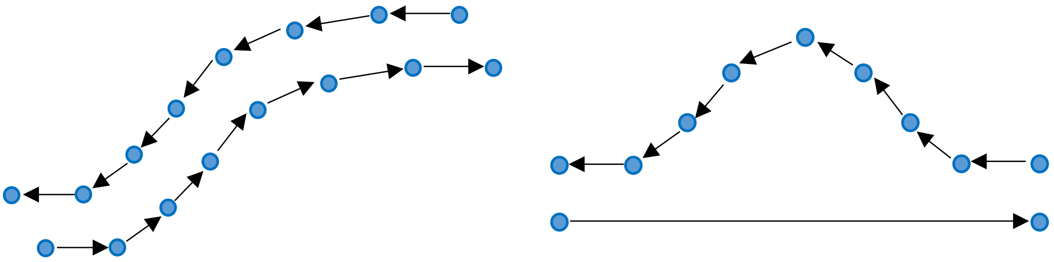
Hierarchy level	Core Element	Schematic representation	OSM symbol	
1	Node		Node	
			Tag	k=v
2	Way		Open polyline	
			Closed polyline	
			Area	
3	Relation		Relation	

Nodes are determined, at the minimum, by a pair of latitude and longitude coordinates and an identification (ID) number. While they can be used to determine point features, such as a well, the primary use of nodes is the definition of ways. Ways are an ordered list of two or more nodes. When ordered openly and linearly, ways are used to represent linear features, which are mainly roads, though they are used for rivers or borderlines as well. When ordered non-openly, they are used to determine area boundaries, such as suburbs, buildings or forests, which are also known in OSM as closed ways. The spatial analyses that were conducted in the following sections mainly concern, and draw information from, the first two core elements, and not relations.

OSM features an inherently hierarchical data structure. The hierarchy levels are important when handling OSM data in a programming environment. If, for instance, a researcher knows the node ID of a road, and wishes to find the node IDs of a neighboring road, then they would (i) shift up a level to find the way ID from the related node ID, (ii) find the adjacent ways (iii) determine the neighboring way of interest among them and (iv) shift down a level from the neighboring way to find the node IDs of interest.

Road segments are ways tagged with a $'highway =*'$ tag. The particularities of the road classification system of every country are also taken into account by the contributors. As ways, road segments are formed from at least two nodes. Based on OSM standing practice, road segmentation is performed with homogeneity in mind. In other words, road segments are split when there is a reason to, such as a change of signage or lanes. The guidelines mention that if a road segment is completely straight, with no adjoining ways, then it can be described with just two nodes no matter its length.

Good segmentation practices allow for the avoidance of redundant nodes. For instance, in dual carriageways, segments are parallel and aligned when both directions are in reality so (Figure 3-8a). If one direction does not curve then additional nodes are not assigned on it like they have to be in the curved one (Figure 3-8b) (OSM, 2019). The essence of OSM guiding rules is that ways are split when data belong to separate groups or follow different structures. For instance, a justified split might be a speed limit change or the introduction of a median barrier. These practices reduce node numbers, thus neighborhood complexity estimations, explained in the following, were closer to the true environment.



(a) Parallel dual carriageway segments **(b)** Node economy in non-parallel segments

Figure 3-8: Examples of good practices in OSM dual carriageway segment creation

In addition to the previous, OSM data include a sequence of nodes based on traffic direction. Therefore nodes are noted in the way the segment is travelled upon. This enables easier calculations of gradient and curvature values as explained in following sections, as well as unambiguous determination of road bearing (direction). Examples of the OSM environment can be seen on the extracted map segments of the selected study areas in Sections 3.4.1.1 and 3.4.1.2.

3.3.2 Elevation data acquisition

Despite its numerous benefits, OSM does not generally include elevation data, namely altitudes of all nodes; when it does, values are recorded as rounded to the nearest meter (i.e. no decimal values). However, for the present doctoral dissertation, elevation data with some accuracy are critical for realistic calculation of gradient values for each road segment. The solution and fulfillment of this requirement is provided by the Shuttle Radar Topography Mission of NASA (SRTM). Therefore, SRTM is presented in this section because it offers a source of primary data that is complementary to OSM node data. SRTM data are used in tandem with OSM data in order to produce two secondary geometrical characteristics, namely gradient and neighborhood complexity, as explained in the following sections.

The SRTM project constituted an effort to obtain digital elevation models on a near-global scale from 56°S to 60°N. Data were collected by a purpose-modified radar system aboard the Space Shuttle Endeavour during an 11-day mission in February 2000. The United States government released the highest resolution results to the public domain during 2014 and 2015. SRTM coverage encompasses most of the developed areas of the world, as seen in Figure 3-9.

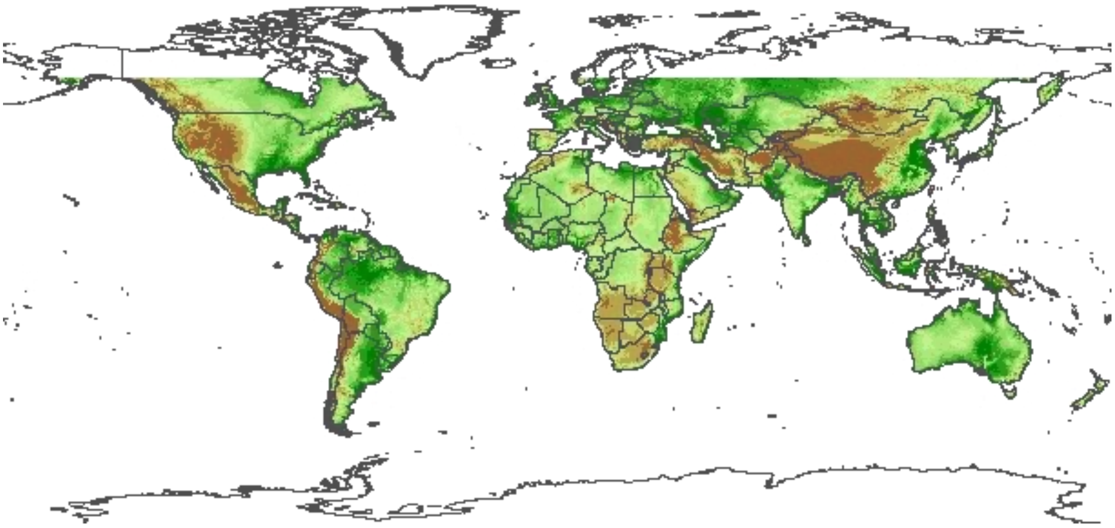


Figure 3-9: SRTM worldwide coverage
[Source: SRTM official website. Retrieved in November 2019]

The latest version available, SRTM Version 4, features a resolution of 3 arcseconds, corresponding to 90m x 90m at the Equator (0°N). Specific countries, like Australia, enjoy even higher resolutions of 1 arcsecond (30m x 30m at the Equator) but these data are unprocessed and generally not available. Altitude data are provided in mosaiced 5 deg. x 5 deg. tiles for easy download and use. SRTM Version 4 includes new interpolation algorithms that are used to fill voids in the original SRTM data and auxiliary digital elevation models (DEMs) (Reuter et al., 2007; Jarvis et al., 2008). Therefore, by extracting the required number of tiles for a study area, altitudes can be obtained for any point in that area by supplying latitude and longitude coordinates. SRTM altitudes have a precision of up to 10 cm (a single decimal point). These values were deemed satisfactory for the purposes of the dissertation.

The original SRTM project and its subsequent enhanced versions resulted in a very high quality dataset, lauded as the best open-access DEM, which has also been verified in Greece (Nikolakopoulos et al., 2006). This praise does not imply that SRTM is free of measurement errors, as geographical processes always are (e.g. Patel et al., 2016). Nonetheless, due to its free access, similar to OSM, SRTM has thousands of active users exploiting the data for research and educational purposes (SRTM, 2019).

3.3.3 Naturalistic driving data from smartphone sensors

Since the core of the present dissertation is infrastructure assessment, then the naturalistic driving trip data obtained from smartphones can be considered to constitute the fuel of the spatial analyses. This section provides an overview of the OSeven application that was used to provide the naturalistic trip data in this dissertation, and its respective digital infrastructure.

3.3.3.1 OSeven driving application

OSeven Telematics is a high-tech startup company, active in the field of Driving Behaviour Analysis, Telematics, Road Safety and Usage Based Insurance. Since 2015, OSeven has been actively and continuously developing and supporting the OSeven application, which can be installed in driver smartphones and seamlessly and non-intrusively record driving trips when users drive their vehicles normally without any user involvement. The application enables the recording of driving related data through the use of smartphone sensors, the calculation of several driving behaviour / road safety / eco metrics and scores and it also includes several coaching, gamification and rewarding features to provide feedback and motivation to the users to improve their driving behaviour. It is important to note that no other instrumentation on driver vehicles is required (e.g. OBDs)

When a smartphone with the application is in a vehicle that starts driving, data recording is initiated automatically, requiring no user involvement – the same applies to the end of a trip. A trip is defined as the time period from the beginning of driving until a stop of driving of at least five minutes according to the OSeven algorithms. Data recording is conducted to a minimum of 1 Hz frequency. Data are stored locally in the device, until it is wirelessly transmitted to the OSeven backend infrastructure through WiFi or mobile network data (3G/4G), based on user choice.

Recorded data are provided by the smartphone sensors and the Operating System of the smartphone devices (Android or iOS). Indicatively the recorded data are GPS values (indicatively longitude, latitude, speed, heading), Accelerometer^{xyz} values, Gyroscope^{xyz} values and device orientation data (i.e. yaw, pitch, roll). The notation ^{xyz} refers to the collection of each parameter in x, y, z axes of the smartphone device. For average drivers, the total transmitted volume of data is estimated at about 50 MB/month (Papadimitriou et al., 2019b).

3.3.3.2 OSeven trip data processing

Once trip raw data have been transmitted by the app to the OSeven backend cloud infrastructure, it undergoes significant cleaning and processing. This includes a toolkit of filtering, signal processing, Machine Learning (ML) and scoring algorithms that: detect harsh driving events (such as harsh braking and acceleration events), mobile phone use, determine exceedance of speed limits, identify the transportation mode (car, motorcycle and mass transit), recognize if the user is the driver or a passenger, calculate driving behaviour and eco scores and display driving data spatiotemporally to help the users identify risk related behaviours (OSeven, 2019). The OSeven data recording and collection scheme is visualized in Figure 3-10.

The produced trip data feature a very high spatial and temporal resolution. As a result, a combination of ML algorithms featuring a range of methods for filtering, clustering and classification, including Big Data approaches and Data Fusion integration is required.

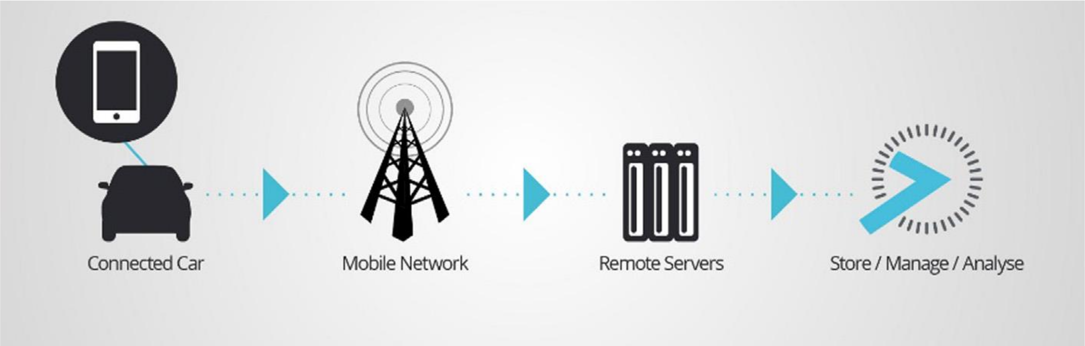


Figure 3-10: OSeven data recording and collection scheme
[Source: OSeven, 2019]

The algorithms used for event detection are agnostic; this means they can analyze data from several devices such as OBDs, smartphones and connected vehicle sensors (4G/5G) and produce equivalent results. Driving pattern recognition from backend processing is depicted in Figure 3-11.

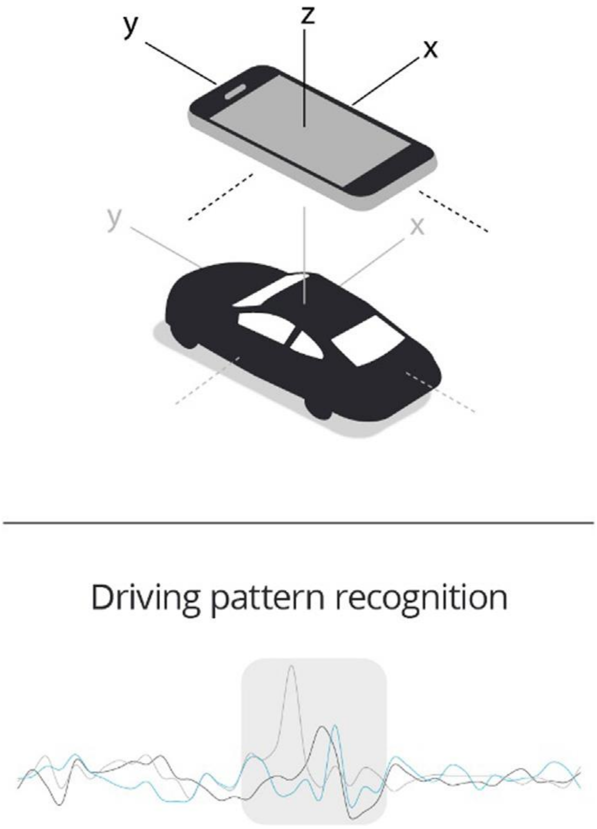


Figure 3-11: OSeven naturalistic driving pattern recognition
[Source: OSeven, 2019]

The steps of data processing and calculation have been outlined in Papadimitriou et al. (2019b):

1. Data outlier detection and removal
2. Data smoothening when required
3. Identification of speeding duration/regions based on speed limit data from map providers
4. Identification of harsh events (acceleration/braking/cornering)
5. Identification of mobile phone use duration
6. Identification of driving distance during 'risky hours' (indicatively from 00:00 to 05:00)
7. Identification of transportation mode (e.g. car, powered-two wheeler, public transport)
8. Determination of driver or passenger status of the user
9. Calculation of driving behaviour and eco scores

A number of road safety-related trip and user metrics are calculated from the processing of trip data. These indicatively include:

1. Harsh braking events (longitudinal deceleration)
2. Harsh acceleration events (longitudinal acceleration)
3. Severity of harsh events (categorical scale in the form of 1: low, 2: medium, 3: high)
4. Speeding (duration of speeding, speed limit exceedance etc.)
5. Harsh cornering (angular speed, lateral acceleration)
6. Driving aggressiveness (e.g. braking, acceleration)
7. Duration of mobile phone use (any type of mobile phone activation by the driver e.g. talking, texting, gaming etc.)

An important pillar of the OSeven application is the provision of feedback and incentives to drivers, so that they can perceive their weaknesses and subsequently improve their driving behavior. This includes the calculation and display of an aggregate driver score and separate scores for the driver behavior indicators based on instances of harsh braking, harsh acceleration, speeding and mobile phone use. Trips are also plotted on maps across networks (highway, rural, urban) so that the drivers have a chance to reflect on events that occurred during each trip. As a side note, OSeven uses both OpenStreetMap and Google Maps for all the relevant map related information (snap to roads, speed limits, geographical information etc.) and all OSeven data is in the WGS84 ellipsoid coordinate reference system.

As of the start of 2020, the databases of OSeven included millions of trips from > 50 countries. It is imperative to stress that all of the OSeven fully complies with the requirement of the General Data Protection Regulation (GDPR) and it also applies state-of-the-art Information Security procedures. Therefore, all data for the needs of this research has been provided in a completely anonymized format. As such, demographic characteristics about the drivers’ sample (i.e. age, gender, driving experience) cannot be obtained, and personalized driving models cannot be created from independent researchers. Nonetheless, the data can be exploited for research and development purposes under this anonymized format.

Regarding road safety research, the OSeven framework results in rich, high-quality datasets that are either event-based (i.e. one row represents a harsh event recorded during a driver trip) or trip-based (i.e. one row represents a second of a driver trip, including normal driving conditions). Such naturalistic driving datasets have been used in a number of diploma theses in Road Safety at the Department of Transportation Planning and Engineering in the National Technical University of Athens (21 diploma theses as of June 2020). Furthermore, a number of scientific studies with papers published in journals and conferences have exploited OSeven data (indicatively Mantouka et al., 2019; Papadimitriou et al., 2019b; Stavrakaki et al., 2019; Tselentis et al., 2019; Ziakopoulos et al., 2020; Petraki et al., 2020).

3.3.3.3 Harsh event determination

Since the cornerstone of this doctoral dissertation is the spatial analysis of harsh driving event frequencies, in the naturalistic trip data provided by OSeven, the determination process of whether a driving maneuver constitutes a harsh event merits some discussion. Harsh events are determined by the OSeven algorithms where the author does not have access due to intellectual property (IP) protection of these algorithms, therefore the exact detection mechanism is not disclosed. However, after additional feedback was requested from OSeven, the following information was made available.

The harsh events are calculated via data fusion and machine learning algorithms and not a rule based approach using as input accelerometer values, gyroscope values, orientations values and GPS related values. Therefore, there is not a specific threshold of the acceleration value for the determination of the harsh events. The reliability of the OSeven algorithms has been evaluated against literature data, OBD data, on-road experiments on the assessment of driving behaviour, and experiments on driving simulators. It is noted that the OSeven product has already been adopted and used by major insurance companies in several countries (Greece, Cyprus, Kingdom of Saudi Arabia, Qatar, Oman, Kuwait, United Arab Emirates, Brazil, USA, Switzerland, Russia, Egypt, Jordan, Thailand, Malaysia, Indonesia, Australia, Singapore, Philippines); this serves as evidence regarding the acceptance of the OSeven algorithms.

3.3.4 Traffic data

3.3.4.1 Traffic Management Centre description

As explained in Section 3.1, traffic data were integrated in the spatial analyses of harsh events in urban arterials. Traffic data were collected from the Traffic Management Centre of Athens and the Region of Attica, which is described in this section. The Traffic Management Centre (TMC) launched in July 2004 and its operation is continuous ever since (24h per day, 365 days per year). The new technical headquarters of the TMC are located in Amerikis Square in Athens. The interior of the main control room of the TMC is shown in Figure 3-12.

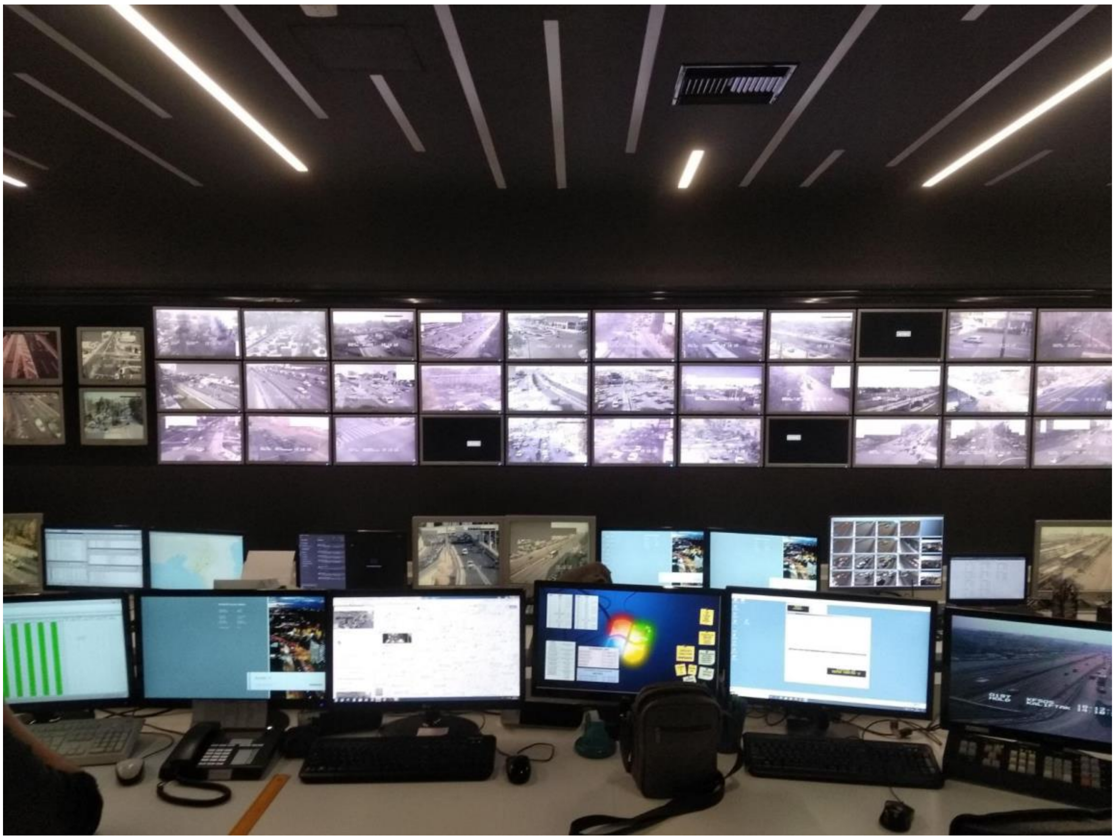


Figure 3-12: Traffic Management Centre of Athens

According to the Region of Attica, which is the supervisory authority, the main aims of the TMC operation are as the following (Region of Attica, 2012):

- The optimization of traffic conditions and road safety across road networks through quick response to emergency events, informing drivers of major traffic conditions and traffic signaling interventions
- The recording, processing and analysis of traffic data obtained along the main road network, as well as collaborating with academic institutions to conduct relative studies
- The provision of real-time traffic data to third parties to support telematics applications
- The cooperation with other traffic control centers (Traffic Police, Attica Tollway Traffic Control Center, Fire Brigade, National Emergency Center, Tram Center etc.).

3.3.4.2 Traffic Management Centre equipment

The primary equipment of the TMC consists of 550 inductive loop detectors, 217 traffic cameras, 24 variable message signs (VMS) and 75 specialised Autoscope systems for vehicle detection. Siemens SI-Traffic Concert is used as central software of the traffic management system. This software features decision-making algorithms that can support decision making in traffic and can act automatically via the VMS. Supported by the above instrumentation, the TMC regulates about 1500 traffic signals in 850 intersections in Athens and in the greater Region of Attica (Theofilatos, 2015; Road Traffic Technology, 2019). The TMC and its equipment famously handled the daunting traffic demands in Athens successfully during the Olympics of 2004. An inductive loop detector situated in Kifisias Avenue is shown in Figure 3-13 as an example of TMC field equipment. This particular loop is situated in a southbound segment; the directional separation is also visible.

Similarly with OSeven data, only anonymized vehicle speed and traffic flow data are recorded, and camera image and video data are not stored. Data protection is ensured by partial blocking of images (e.g. when cameras would peer into buildings), while the TMC does not report any traffic violations to the authorities. As per TMC aims, several scientific studies using TMC data have been conducted and published in journals and conferences (Minis & Tsamboulas, 2008; Yannis et al., 2014; Theofilatos, 2015; Theofilatos et al., 2017b; Petraki et al., 2020).

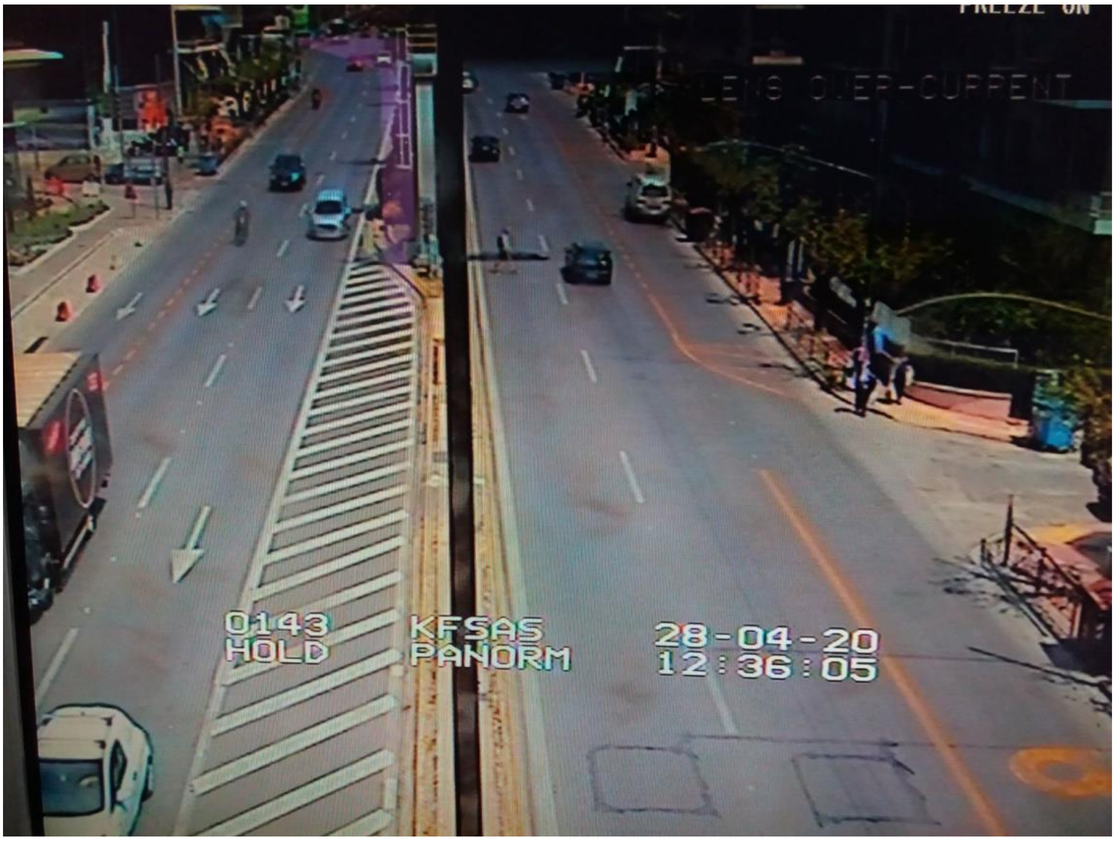


Figure 3-13: Inductive loop detectors in Kifisias Avenue

TMC collects traffic occupancy as a primary quantity, measured as the percentage of time during which vehicles occupy measurement positions [%]. Moreover, vehicle numbers are measured to obtain traffic flow counts. Traffic flow is measured every 90 s, and aggregate datasets of 5 m or 1 h temporal intervals can be obtained from the TMC as well. Regardless of interval, traffic flow data are transformed to vehicles/h in the provided datasets. Measurement Quality is recorded, and is classified categorically as

High or Low. Traffic speed, averaged over time (i.e. mean-time speed), is also obtained either by image processing by the TMC software or by indirectly inferring it by accounting for the time required for vehicles to cross the loop, as per Equation (76):

$$\overline{v_{tr}} = \frac{m+l}{1000} * \frac{3600 * q}{t * k} \quad \text{Eq. (76)}$$

Where:

- $\overline{v_{tr}}$ is the traffic speed averaged over time [km/h]
- m is the length of the inductive loop [m]
- l is the average vehicle length [m]
- q is the current traffic flow [veh/h]
- k is the current traffic occupancy [%]
- t is the time interval [s]

In the present doctoral research, the other data sources (naturalistic driver trips and map data) feature high spatial and temporal disaggregation and resolution. Therefore, the highest available resolution was pursued for traffic data as well, namely 90 s. For purposes of consistency and mitigation of uncertainty, spatial analysis will be conducted only in road segments that have TMC measurement locations. For Kifisias Avenue, a span of 7.90 km from the starting point features TMC measurement locations, and this is the road length that is considered for the analyses of Sections 6 and 7.

3.4 Methodological steps

The main methodological steps that are followed in this dissertation are presented in the present section. Emphasis is placed on critical parts which include calculation processes and algorithms used to merge data from the various sources.

All analyses in this doctoral dissertation have been conducted in R-studio (R Core Team, 2019) using a multitude of add-on function libraries, known in the R community as packages. In total, more than 45 packages were utilized. The maps shown in this doctoral dissertation were created with the OSM/R-studio interface package and JavaScript library 'leaflet' (Cheng et al., 2019).

3.4.1 Selection of study areas

Since the spatial analyses that were conducted for this dissertation utilize high-resolution, large-scale trip data as input, areas where these data would be richer in events were investigated. Harsh braking data have been found to be significantly higher in urban roads than rural roads and highways in recent research (Jansen & Wesseling, 2018). In addition, spatial effects, namely unobserved parameters in the network and the influence of neighboring road segments in a given road segment, were considered to be more pronounced in an urban network environment, and as such worthy of investigation. Additionally, as outlined in the critical synthesis of the literature and the respective research questions, in Sections 2.5 and 2.6, urban network analyses are far scarcer than those in larger road classes due to lack of proper data and increased structure complexity. If spatial analysis of harsh events can be successfully conducted in an urban network, then transferability of the methods used here to simpler network structures, such as rural road networks, is reasonably possible.

Urban arterials are relatively more isolated driving environments than unban networks, with entrances and exits predominantly on the longitudinal dimension (1-D degrees of access) in the form of ramps rather than both longitudinal and lateral dimensions (2-D degrees of access). Additionally, there exist several principal urban arterials in Athens (such as Kifisias Avenue, Mesogeion Avenue, Siggrou Avenue) which feature median barriers throughout their length and essentially are dual carriageways, where directions are separated, as shown in Figure 3-14. In such environments it is easier to obtain traffic volume and speed measurements from traffic management centers (such as the Athens TMC) that are do not vary greatly from the true value as distance increases from the measurement point. In other words, there is less loss of information, and thus less introduced uncertainty, from the measurement location to the incident location.

This configuration enables transportation researchers to acquire a faithful data description of naturalistic conditions regarding traffic volume, occupancy and speed at a specific time (at the time of a crash or harsh behavior event, for instance) and to conduct various types of analyses while taking these traffic parameters into consideration (as in, for instance, Yannis et al., 2014; Petraki et al., 2020).

Urban networks areas with high road density incorporating different road types were sought after; a distribution of roads throughout the rectangular area was also desirable. Naturally, in both urban road network and urban arterial area selection, data availability for the selected areas played an important part in any decision regarding area selection as well.

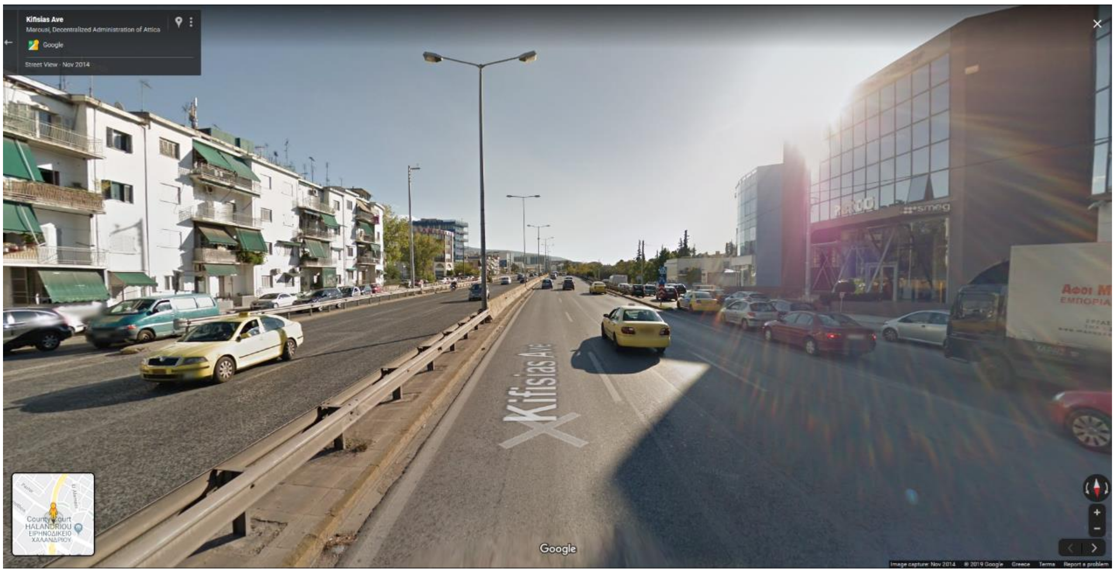

Kifisias Ave
Marousi, Decentralized Administration of Attica
Google
Street View · Nov 2014
Kifisias Ave
Country Court HALANDRIOU
ΕΙΡΗΝΟΔΙΚΕΙΟ ΧΑΛΑΝΔΡΙΟΥ
Google
Image capture: Nov 2014 © 2019 Google Greece Terms Report a problem

Figure 3-14: Directional separation and ramp exit on the left side in Kifisias Avenue
[Retrieved in November 2019 from Google Maps, © Google]

With the research questions of this doctoral dissertation in mind, and following discourse with OSeven Telematics and the Traffic Management Centre of Athens, the following study areas were consolidated. It should be noted that any segments representing walkways are subsequently removed from the data, in order to only consider roads with vehicle traffic.

3.4.1.1 Urban network study areas

Two urban network study areas were selected for this dissertation. The first urban network is the training area, which were used for training (calibration) and cross-validation of the statistical models used in the study. The training area is an urban road network situated in a section of Chalandri, a northern suburb of Athens, Greece. The training area ranges between latitudes of 38.0135 (south) to 38.0307 (north) and longitudes of 23.7835 (east) to 23.8148 (west), corresponding to a 2.743 km by 1.913 km rectangle with total area of 5.247 km^2. The training area comprises 48.56% of the municipality of Chalandri, which has a total area of 10.805 km^2, and contains one-way and two-way road segments. The surrounding areas have primarily commercial and residential land uses.

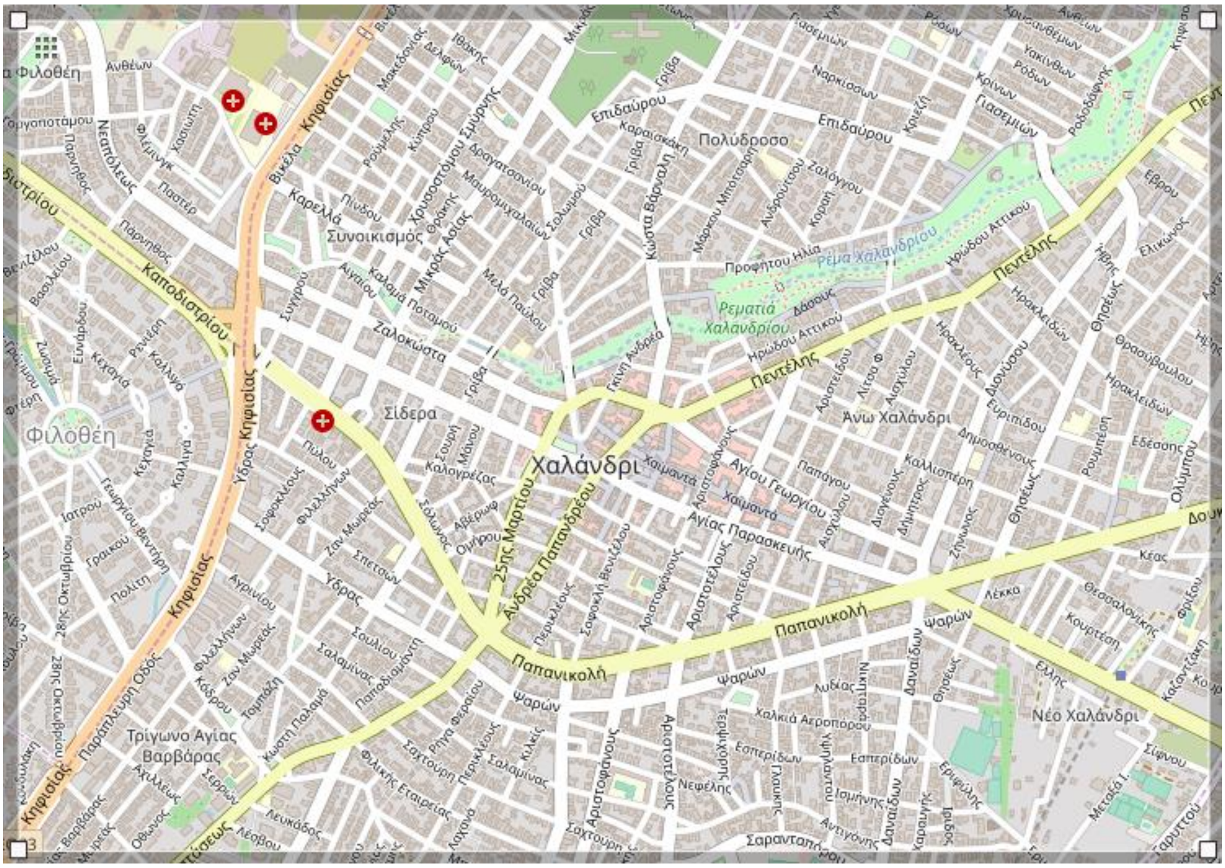


Figure 3-15: Chalandri urban network training area
[Retrieved in February 2020 from OpenStreetMap]

The second urban network is the test area, which was used to check the transferability of the models created in the training area. The test area is situated in a highly commercial section of the city center of Athens, containing the central Omonoia Square and its adjacent area to the North; it too contains one-way and two-way road segments. The surrounding areas have primarily commercial and residential land uses, with some educational and touristic uses from some university and museum buildings. The test area includes Attiki, Victoria, Metaxourgeio and Panepistimio, which are important central locations, and constitutes one of the oldest parts of the city of Athens that have been functioning in such a form without significant disruptions. For brevity, in this dissertation it is collectively referred to as 'Omonoia area' henceforth. The test area ranges between latitudes of 37.9783 (south) to 38.0020 (north) and longitudes of 23.7148 (east) to 23.7397 (west), corresponding to a 2.636 km by 2.183 km rectangle with total area of 5.754 km^2. The streets appear with a higher density in this central area, which comprises 14.77% of the larger Athens municipality that has a total surface area of 38.96 km^2.

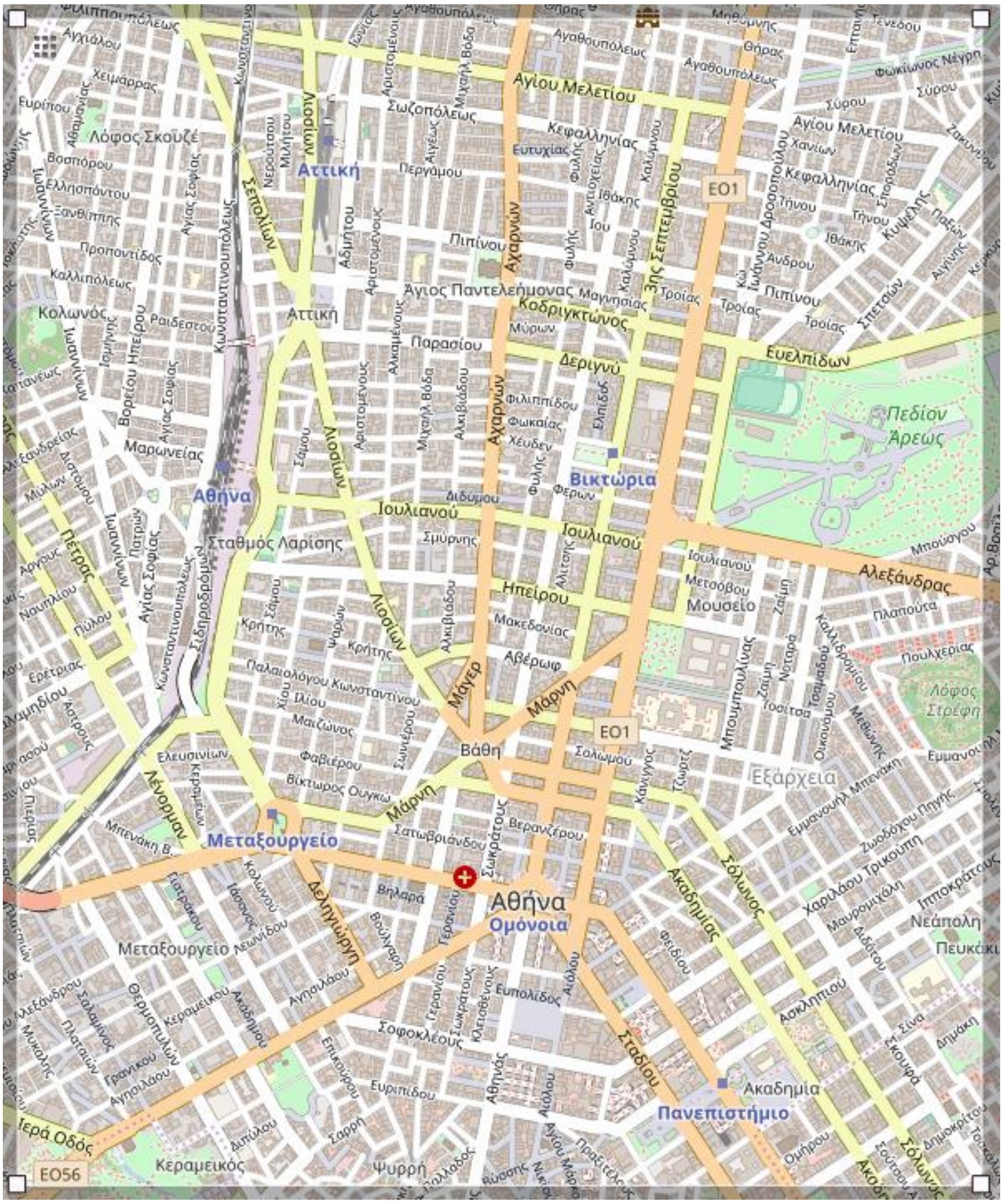


Figure 3-16: Omonoia urban network test area
[Retrieved in February 2020 from OpenStreetMap]

3.4.1.2 Urban arterial study area

Apart from spatial analyses conducted on the networks described in the previous section, additional spatial analyses were performed on an urban arterial area. This constitutes an effort to integrate traffic parameters and road user behavior parameters in the spatial analyses and identify additional underlying trends, as outlined in Section 3.1, with the drawback of limiting transferability.

Ultimately, Kifisias Avenue (also known as Leoforos Kifisias) was selected as the training area for the analysis. Kifisias Avenue is an urban arterial featuring a median barrier, thus classifying as a dual carriageway. In its entirety, Kifisias Avenue spans 19.34 km and connects directly to El. Venizelou Street to the north and to Vasilissis Sofias Avenue to the south. At the majority of this length, it features three lanes per direction, though lanes can drop to two locally. This number includes a bus lane for a significant fraction of its length, which taxis are legally allowed to use as well. Kifisias Avenue is connected to a series of entry and exit access ramps across its length as well. Kifisias Avenue is one of the major arterials of Athens, with high-value third sector service, office and commercial land uses, as well as several embassy complexes along its length. For the analyses, a span of 7.90 km from the starting point which is the intersection with Alexandras Avenue to the south (latitude 37.9865, longitude 23.7614) up to the intersection with Agiou Konstantinou Street to the north (latitude 38.0464, longitude 23.8074) was chosen. This selection was made due to the presence of conductive loop detectors and the respective traffic data availability there (as explained in Section 3.3.4.2). This section constitutes 40.85% of the total length of the Avenue and features a simple roadway axial design overall. This study area crosses or borders a number of municipalities, from South to North: Athens, Chalandri, Filothei & Psychiko and Maroussi. The examined length is shown on Figure 3-17 (municipality boundaries are shown in blue).

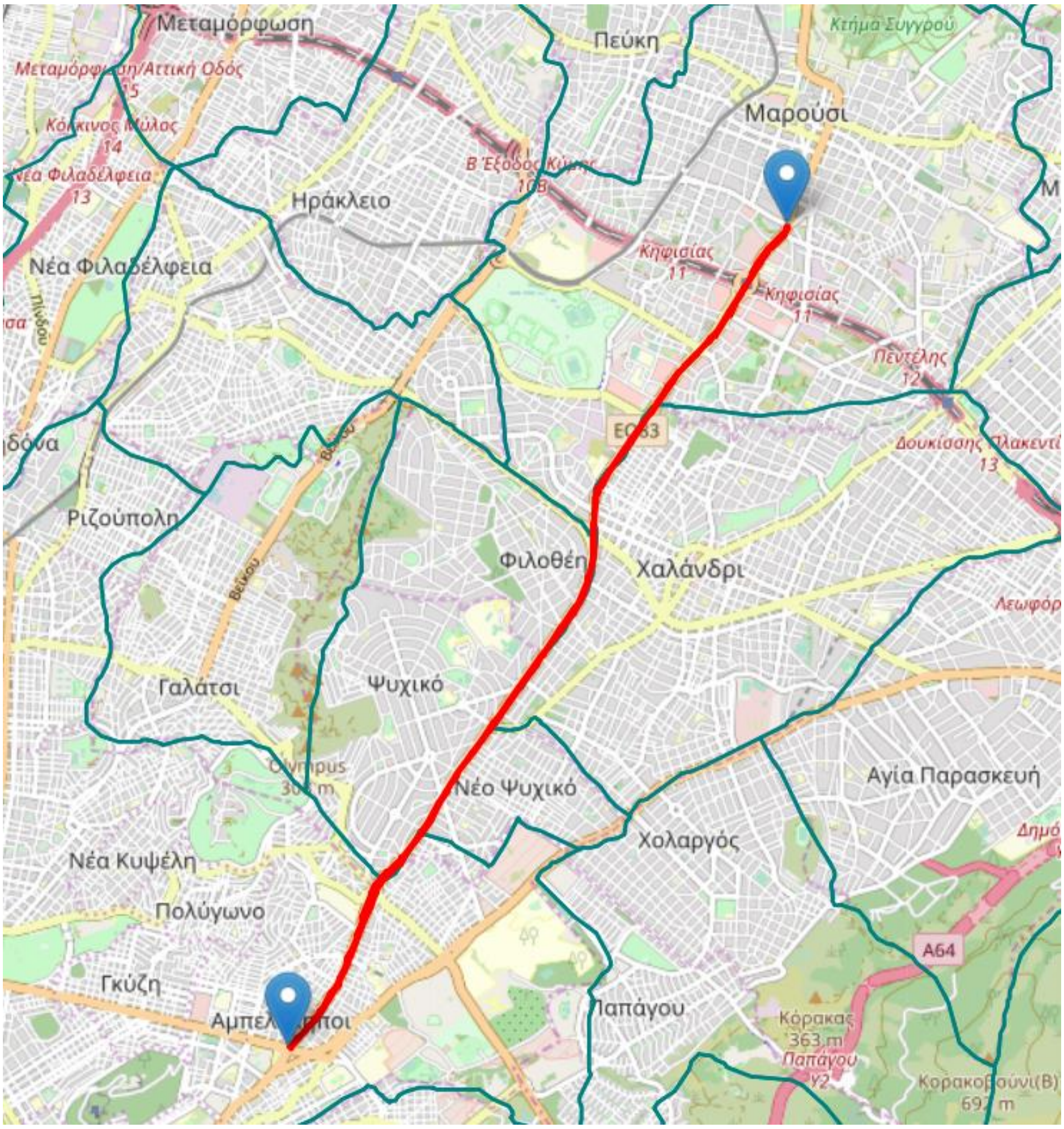


Figure 3-17: Kifisias Avenue urban arterial study area
[Retrieved in February 2020 from OpenStreetMap, processed with R-leaflet]

The relative positions of all three study areas within the capital of Athens can be seen in Figure 3-18:

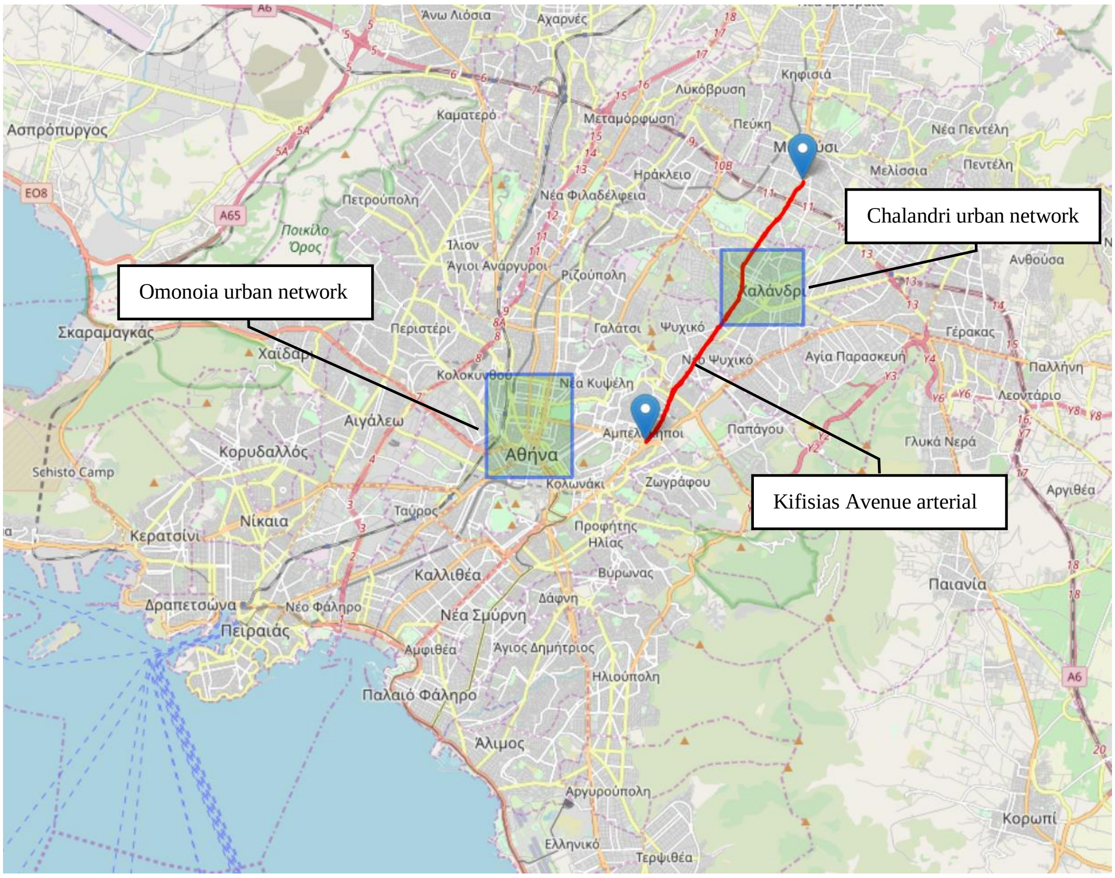


Figure 3-18: Relative position of study areas within the city of Athens.
[Retrieved in February 2020 from OpenStreetMap, processed with R-leaflet]

3.4.2 Derivation of geometric characteristics

After the study areas are defined, the respective data are obtained from OSM. Therefore, node, way and relation data become available for processing and preparatory analysis. These data enable the calculation of several geometric characteristics for each road segment. It is important to note that these features all refer to a particular road segment and are considered to be common across the entirety of its span for analytical purposes. In practice, it was determined that calculations relevant to these features should preferably precede data merging with any other source for reasons of computational speed and simplicity. That is also the reason that walkways are initially removed from segment datasets, so that they do not burden subsequent calculations.

3.4.2.1 Road segment length

The first and most intuitive quantity to calculate for each road segment is its length. Segment length is calculated from node latitude and longitude coordinates. In Geography, the shortest path between two points across the Earth's surface is called a geodesic. Several methods have been developed over the years for geodesic calculation on the surface of Earth, indicatively:

- Great Circle or Haversine distance, which simplifies Earth to a sphere
- Vincenty distance, based on Vincenty's more accurate formulae that consider an oblate spheroid shape for Earth. Vincenty's formulae were made to be iterative and thus programmable (Vincenty, 1975).
- More sophisticated and accurate geodesic calculations with increased precision, for instance the one proposed by Karney (2013).

It should be noted that, for transport engineering and road safety assessment purposes on the road segment scale, Great Circle/Haversine distance produces results that are very accurate and more than adequate. Accordingly, in this dissertation, most road segments have length of a magnitude in the tens of meters, and few have a length of roughly 500m.

Haversine distance calculation is provided below; the formula demands latitudes and longitudes to be in radians.

$$d_H = 2R * arc\sin\left(\sqrt{\sin^2\left(\frac{\varphi_2 - \varphi_1}{2}\right) + \cos(\varphi_1) * \cos(\varphi_2) * \sin^2\left(\frac{\lambda_2 - \lambda_1}{2}\right)}\right) \quad \text{Eq. (77)}$$

Where:

- d_H is the calculated Haversine distance
- R is the Earth's radius ($R \approx 6{,}378\ m$)
- φ_1, φ_2 are the latitudes of points 1 and 2 [rad]
- λ_1, λ_2 are the longitudes of points 1 and 2 [rad]

However, modern programming environments allow for rapid calculations while enjoying the precision granted from modern ellipsoid-derived algorithms, such as the work of Hijmans et al. (2017), who utilized Karney's algorithms. Furthermore, ellipsoid approaches grant higher transferability to results, without fear of additional accuracy loss when examining very long segments. This more sophisticated approach was therefore preferred. The exact equations are iterative and too lengthy to be replicated here; interested readers can refer to Karney (2013).

Following Hijmans et al. (2017), the shortest distance between two points on an ellipsoid is provided. The default ellipsoid for both map and smartphone data is WGS84, so that is used for distance calculations. Road segment lengths were thus obtained. As expected, in road segments with more than two nodes, the respective fundamental two-node distances are calculated for each consecutive pair and then summed. This can apply to the significantly more complex geometrical design curves for rural road segments as well.

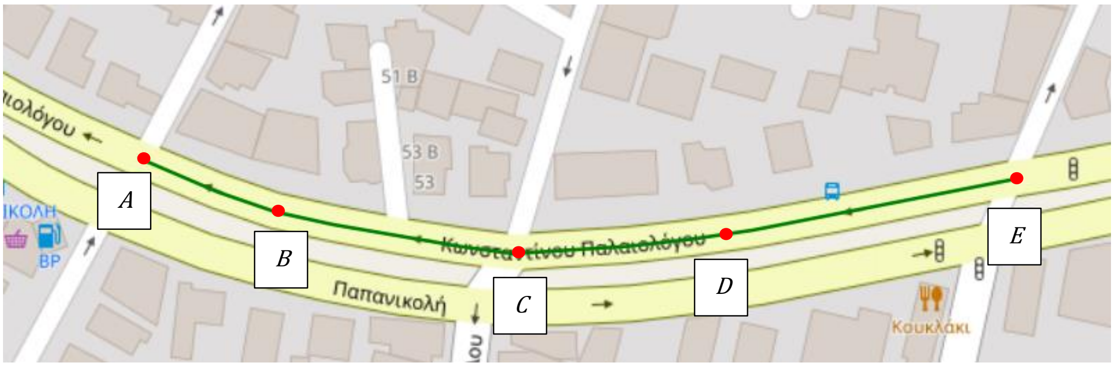


Figure 3-19: Indicative length calculation from fundamental distances

For example, in the segment of Figure 3-19, the total segment length is obtained as $d_s = AB + BC + CD + DE$.

3.4.2.2 Road segment centroids

Having calculated segment lengths, the next segment geometric characteristic to be calculated is road segment centroids. Centroids are a dimensionless, point-type quantity; they represent a core location of each road segment that can be used to identify proximity and relative position of different segments. Another way to regard centroids is as the 'label' of each road segment, with centroid coordinates being nominal coordinates for the entire segment. In simple two-node segments the centroid falls on the way axis, while in more complex segments typically it falls outside of the axis due to shape irregularity.

While the process of centroid identification does not directly offer a variable to be used for modelling, it is helpful to treat segments as points computationally – for instance when wishing to assign spatial weights in neighboring segments.

Segment centroid coordinates are calculated as the mean of all node coordinates for each segment so that:

$$\varphi_c = \frac{1}{n} * \sum_{i=1}^{n} \varphi_i \qquad \text{Eq. (78)}$$

And:

$$\lambda_c = \frac{1}{n} * \sum_{i=1}^{n} \lambda_i \qquad \text{Eq. (79)}$$

Where:

- n is the number of total nodes in the segment
- φ_i are the latitudes of each node
- φ_c is the centroid latitude
- λ_i are the longitudes of each node
- λ_c is the centroid longitude

An alternative to segment centroids would be the utilization of segment midpoints. The term midpoint defines the middle point inside a polyline that is equidistant to its borders, also known as endpoints. However, the process of midpoint definition is iterative and computationally demanding. Furthermore, due to the relatively simple structure of OSM segments, there would be little actual difference between midpoints and centroids leading to uncertain gain. More importantly, the use of segment centroids, has been proposed in the literature before by Aguero-Valverde (2014). Following the definitions of that study, this dissertation utilizes the 'aerial' distance of segment centroid, namely the direct line between midpoints as a bird flies.

3.4.2.3 Road segment gradient

Subsequently, the calculations of gradient values are conducted for each road segment using SRTM altitudes and ellipsoid geodetics. The main reasoning is similar to the length calculation: the gradient of each fundamental two-node segment is determined, and then the mean gradient is obtained for the segment by averaging the fundamental ones. Fundamental gradients are defined as:

$$s_{f,i} = \frac{h_i - h_{i+1}}{d_{i\,,i+1}} \qquad \text{Eq. (80)}$$

Where:

- $s_{f,i}$ is the fundamental gradient (calculable for a total of $n - 1$ nodes)
- h_i, h_{i+1} are the altitudes of two subsequent nodes
- $d_{i\,,i+1}$ is the geodesic of two subsequent nodes

Referring to the segment of Figure 3-19, the overall gradient would be the average of gradients weighted by the distance of the fundamental segments:

$$s_{total} = \frac{\sum_{i=1}^{n-1} s_{f,i} * d_{i\,,i+1}}{\sum_{i=1}^{n-1} d_{i\,,i+1}} \qquad \text{Eq. (81)}$$

As mentioned in Section 3.3.1.2, OSM segmentation follows the direction of traffic. Therefore the values of gradients follow traditional highway engineering convention: positive gradients s are used to denote uphill slopes and negative gradients $-s$ are used to denote downhill slopes.

3.4.2.4 Road segment curvature

In geometry, curvature is a measure of the instantaneous rate of change of direction. Since there are no official geometric guidelines for urban road networks, a more mathematical approach is adopted in this dissertation: road curvature is assumed equal to the Menger curvature. The Menger curvature of a set of three of points is defined as the reciprocal of the radius of the circle (O, R_f) that passes through all three points, as shown in Figure 3-20 for the node set $\{A, B, C\}$.

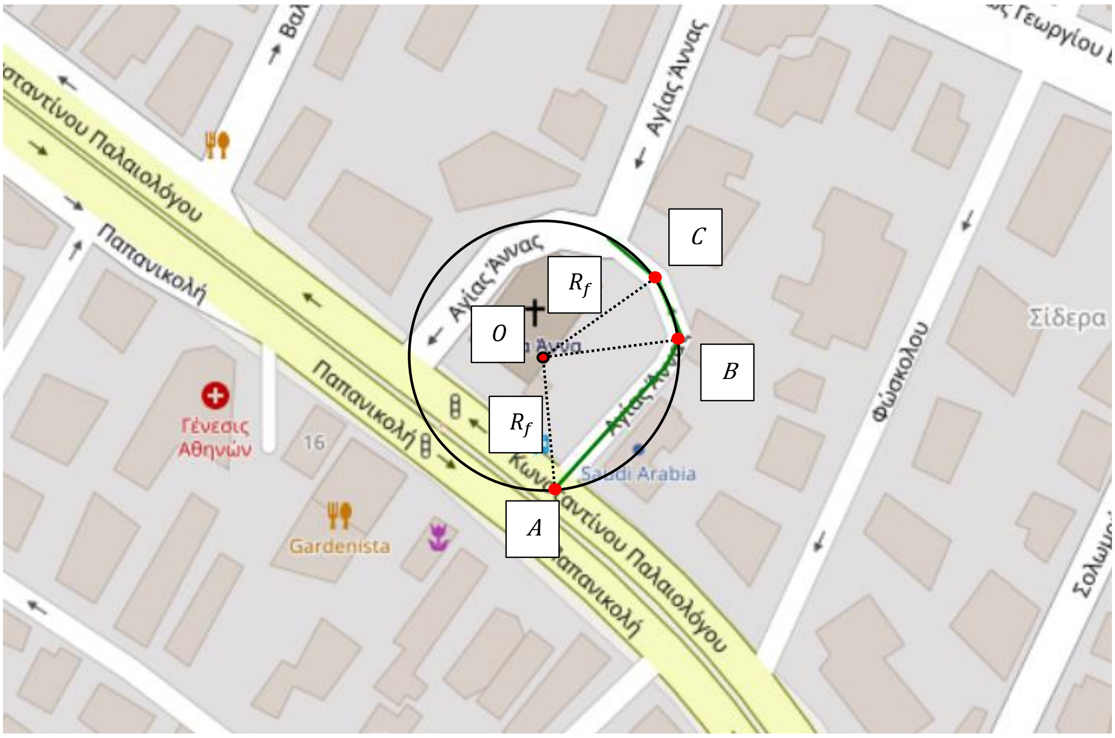


Figure 3-20: Indicative curvature calculation from fundamental distances

$$c_f = \frac{1}{R_f} \qquad \text{Eq. (82)}$$

If the three points happen to coincide across a straight line, $R_f \to \infty$; therefore $c_f \to 0$. When using point coordinates, Menger's curvature transforms for a fundamental set of three points to:

$$c_{f,i} = \frac{1}{R_f} = \frac{4 * A_{\widehat{ABC}}}{AB * BC * CA} \qquad \text{Eq. (83)}$$

Across a segment, each fundamental curvature is calculable for $n-2$ nodes. The area $A_{\widehat{ABC}}$ can be calculated from Heron's formula, eschewing further distance calculations:

$$A_{\widehat{ABC}} = \sqrt{\tau(\tau - AB)(\tau - BC)(\tau - CA)} \qquad \text{Eq. (84)}$$

Where:

- τ is the semi-perimeter of the $\widehat{ABC}$ triangle, equal to:

$$\tau = \frac{AB + BC + CA}{2} \qquad \text{Eq. (85)}$$

Regarding implementation is this dissertation, curvature calculation follows a similar logic with gradient. The main difference is that curvature is not always present: simple two-node segments are assigned a curvature of zero by default. In more complex segments, the algorithmic process is conducted by accounting for consecutive triangles which are created by consecutive sets of three nodes, and then averaging the curvature values across these sets of nodes, weighted by their individual distances:

$$c_{total} = \frac{\sum_{i=1}^{n-2} c_{f,i} * (d_{i\,,i+1} + d_{i+1\,,i+2})}{\sum_{i=1}^{n-1} d_{i\,,i+1}} \qquad \text{Eq. (86)}$$

Theoretically, more complex calculations are possible for curvature, but it was decided that they would fall outside of the scope of the present research, especially due to OSM segmentation which breaks too complex curves into smaller segments. Regarding algorithmic implementation, the subroutine concerning gradient and curvature calculation is shown in Figure 3-21. It is important to note that these two characteristics are embedded in a common loop to mitigate computational demands by reducing the loading and re-loading instances of different node groups.

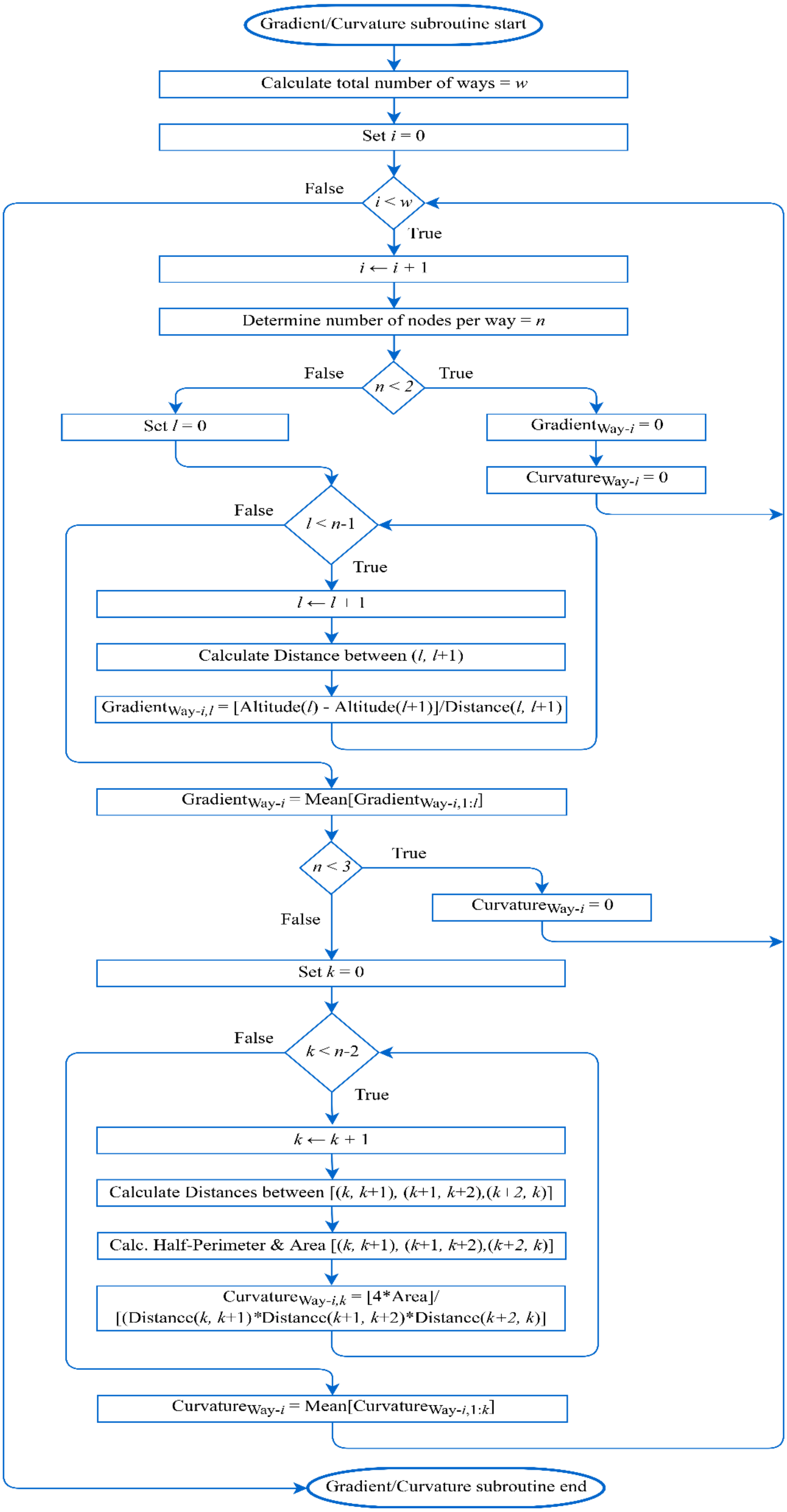

Gradient/Curvature subroutine start
Calculate total number of ways = w
Set $i = 0$
$i < w$
False
True
$i \leftarrow i + 1$
Determine number of nodes per way = n
$n < 2$
False
True
Set $l = 0$
$\text{Gradient}_{\text{Way-}i} = 0$
$\text{Curvature}_{\text{Way-}i} = 0$
$l < n\text{-}1$
False
True
$l \leftarrow l + 1$
Calculate Distance between $(l, l{+}1)$
$\text{Gradient}_{\text{Way-}i,l} = [\text{Altitude}(l) - \text{Altitude}(l{+}1)]/\text{Distance}(l, l{+}1)$
$\text{Gradient}_{\text{Way-}i} = \text{Mean}[\text{Gradient}_{\text{Way-}i,1:l}]$
$n < 3$
True
False
$\text{Curvature}_{\text{Way-}i} = 0$
Set $k = 0$
$k < n\text{-}2$
False
True
$k \leftarrow k + 1$
Calculate Distances between $[(k, k{+}1), (k{+}1, k{+}2), (k{+}2, k)]$
Calc. Half-Perimeter & Area $[(k, k{+}1), (k{+}1, k{+}2), (k{+}2, k)]$
$\text{Curvature}_{\text{Way-}i,k} = [4{*}\text{Area}]/$
$[(\text{Distance}(k, k{+}1){*}\text{Distance}(k{+}1, k{+}2){*}\text{Distance}(k{+}2, k))]$
$\text{Curvature}_{\text{Way-}i} = \text{Mean}[\text{Curvature}_{\text{Way-}i,1:k}]$
Gradient/Curvature subroutine end

Figure 3-21: Gradient & Curvature subroutine flowchart

3.4.2.5 Neighborhood complexity

Another interesting quantity is neighborhood complexity. This attribute takes advantage of the OSM segmentation process, and refers to the general density of the road network surrounding the area of the event. The main reasoning here is that more complex roads around the location of the event, with more twists and turns, lead to a more unpredictable and complicated road environment, with increased concealment of vehicles, more traffic rules to observe, limited reaction margins and more distractions. Therefore the metric of neighborhood complexity is constructed as per the following.

For each road segment centroid, all nodes inside a specific moving window with dimensions $\{W_x, W_y\}$ are examined. The window is an area that moves per selected centroid. There is no comparable metric in the literature, to the extent of the author's knowledge, therefore window dimensions are empirically assigned due to lack of precedent. For the present dissertation, the moving window was assigned values of 400*650 pixels on the OSM grid, corresponding to about 470m*470m on the ground. All nodes inside that area are recorded and listed. After that, nodes that have an altitude difference of more than 3.5 m from the centroid are removed from the list, to exclude completely unrelated features such as overpasses, tunnels or similar locations that are vertically remote from the segment and thus do not affect its neighborhood complexity. Lastly, the neighborhood complexity is defined as the natural logarithm of the number of remaining nodes in the list, which are proximal to the centroid.

$$ncom_i = \ln(n_{prox}) \qquad \text{Eq. (87)}$$

Where:

- $ncom_i$ is the neighborhood complexity metric of road segment i
- n_{prox} is the number of nodes inside the moving window with an altitude difference that is less than 3.5 m from the centroid i

3.4.2.6 Additional road features

Additional network features are drawn from OSM and are added as variables that can augment the analysis. The first set of features are used to describe the classification of a road: (i) number of lanes ranging from 1 to 4, (ii) road traffic direction (one/two way roads) and (iii) road type (primary, secondary, tertiary, residential roads and footways). The first two features are self-explanatory, but the third merits some elaboration. Definitions for all categories are provided on **Table 3-3**.

Additional rare types or categories are available in OSM, though typically not used for roads in the study areas. For the purposes of this doctoral dissertation, road categories were merged when a road was assigned to a rare category, as shown on **Table 3-3**. As per the OSM Wiki, slip roads/ramps which are usually considered to belong to the through highway they exit and enter, for instance Kifisias Avenue ramps belong to Kifisias Avenue. This is usually the higher classification of the intersecting highways because on and off ramps almost always have the same kind of restrictions as the main road.

Table 3-3: Initial and merged OSM road categories

Initial road types (OSM tags)	OSM definition	Merged road types
Primary	A major highway linking large towns, in developed countries normally with 2 lanes.	Primary
Secondary	A highway which is not part of a major route, but nevertheless forming a link in the national route network.	Secondary
Secondary link	Used to identify slip roads/ramps and channelised (physically separated by an obstruction or painted island) at-grade turning lanes connecting the through carriageways /through lanes of a secondary class highway to other minor roadways.	
Tertiary	Used for roads connecting smaller settlements, and within large settlements for roads connecting local centers. In terms of the transportation network, tertiary roads commonly also connect minor streets to more major roads.	Tertiary
Tertiary link	The link roads (sliproads/ramps) leading to/from a tertiary road from/to a tertiary road or lower class highway.	
Residential	Used for roads accessing or around residential areas but which are not normally used as through routes.	Residential
Service	Generally for access to a building, service station, beach, campsite, industrial estate, business park, etc. This is also commonly used for access to parking, driveways, and alleys.	
Living Street	These type of roads have lower speed limits, and special traffic and parking rules compared to residential streets. Legislation either grants pedestrians the right of way over or at equal rights to other road users	
Footway	Used for mapping minor pathways which are used mainly or exclusively by pedestrians.	Footways (discarded from analysis)
Pedestrian	A road or an area mainly or exclusively for pedestrians in which some vehicle traffic may be authorized (e.g. emergency, taxi, delivery). Typically found in shopping areas, town centers, places with tourism attractions and recreation/civic areas, where wide expanses of hard surface are provided for pedestrians to walk.	
Track	This tag represents roads for mostly agricultural use, forest tracks etc.; often unpaved (unsealed) but may apply to paved tracks as well, that are suitable for two-track vehicles, such as tractors or jeeps.	

Footways were excluded by removing their respective ids from the road segment id vector. It should be mentioned that the assigned categories are not absolute, as OSM users without official training utilize them to characterize roads accordingly. Thus there may be some ambiguity for specific roads in the smaller categories – a user may be unsure whether to characterize a minor road as tertiary or residential. However, the variable is informative for the general size and traffic intensity of each road, and as such is retained in the analysis. A manual verification and quality control via online satellite and street-view images from Google Maps was conducted. This ensured that the correct footways were removed and that no categories were determined as out of place in the examined road segments.

The second set of additional features concerns traffic management features. Specifically, the presence of traffic lights and pedestrian crossings can be thought to influence harsh event frequency, as they interrupt the free flow of traffic in road segments. The respective information is drawn from OSM data, and the presence of these network features in denoted in a binary fashion for each segment {0,1}.

A location in Omonoia area featuring traffic lights and a pedestrian crossing is shown on Figure 3-22.

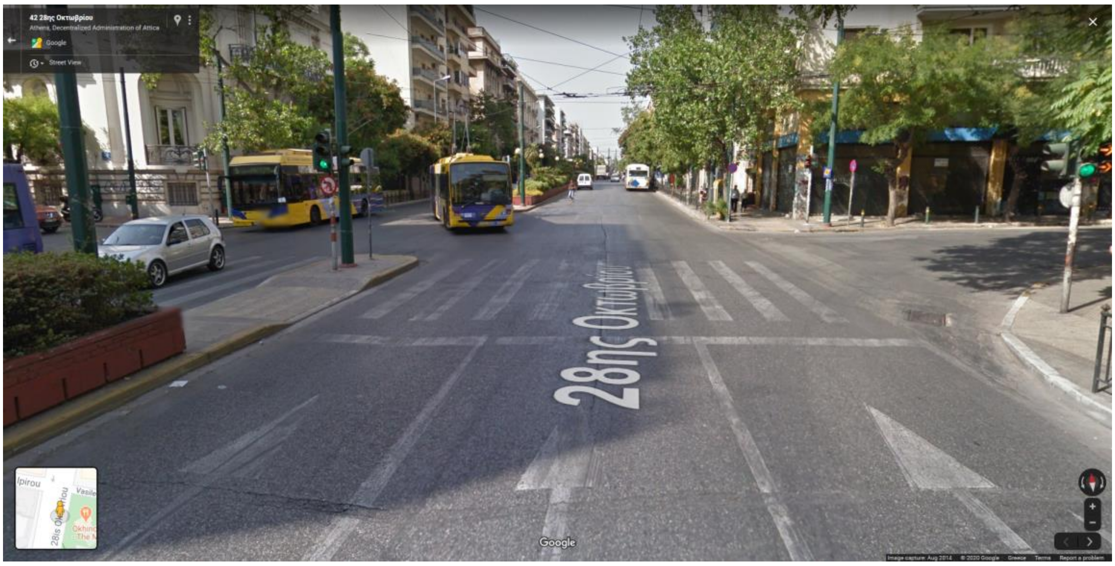


Figure 3-22: Road segment with presence of traffic lights and pedestrian crossing in Omonoia area [Retrieved in March 2020 from Google Maps, © Google]

The respective road segment (with OSM id number: 168725546) is shown on Figure 3-23.

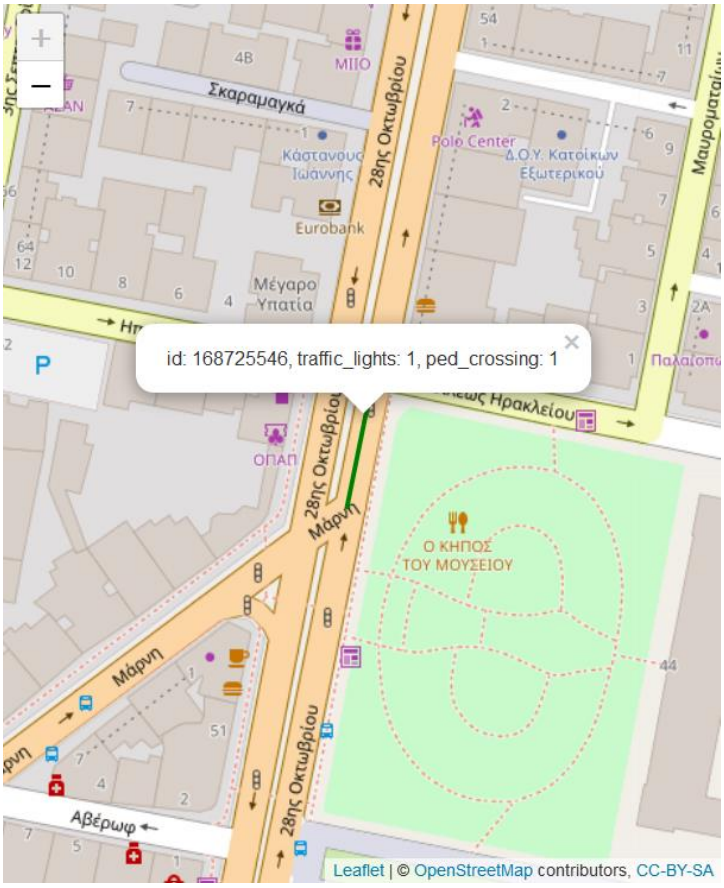

id: 168725546, traffic_lights: 1, ped_crossing: 1

Figure 3-23: Accounting for presence of traffic lights and pedestrian crossings

3.4.3 Integration of smartphone and map data – Map-matching

3.4.3.1 Concept presentation

A critical part of the methodology of the present doctoral dissertation is the integration of naturalistic driving data that is provided from smartphones, and their projection and matching with map data, including the geometrical characteristics previously calculated. This process is conducted by analyzing each second of a trip, noting baseline driving as well as event locations as they are brought up in the dataset. Through this process, the location of harsh events and relevant metrics are matched on map coordinates and assigned to the appropriate segment centroids.

3.4.3.2 Accounting for risk exposure

Apart from the aforementioned problem of converting driver-behavior metrics to segment-based metrics, a simultaneous challenge to be tackled is taking exposure into account. The smartphone data obtained can be used to measure exposure to risk. Exposure to risk measures the likelihood of being involved in a dangerous or hazardous situation and is a critical factor in estimating crash risk (Pei et al., 2012).

In simple terms, one should consider a busy urban arterial that has 10 harsh brakings per day, and a small residential road that has 7 harsh brakings per day. Claiming that the urban arterial is more dangerous than the residential road would appear intuitively wrong; the missing link is exposure! In practice, exposure is a form of accounting for the amount of travel involved in transport processes or phenomena, such as road crashes (Hakkert et al., 2002). Spatial analysis studies in road safety usually include traffic volume, roadway length and vehicle distance traveled as exposure parameters, as shown extensively in the literature review conducted in Section 2.2.

A synthesis but also a departure from this precedent is attempted for the segment-based analyses in the present dissertation. Two exposure parameters were included for a given road segment: (i) the length of the segment and (ii) the amount of driver trips on the road segment; i.e. the number of times drivers passed from the segment, also called pass count. These parameters are augmented by traffic volume exposure variables in urban arterial analysis. This selection was due to the fact that each exposure parameter offers different information regarding the variance of frequencies of harsh events: segment length measures geographical (spatial) exposure, pass count measures naturalistic driving exposure and traffic volume measures traffic exposure.

3.4.3.3 Description of the map-matching process

The trips are recorded from the OSeven applications and provided in datasets in a format such that every row represents a second of a single trip, including latitude and longitude. Therefore the position of the drivers can be inferred as the trip progresses. It is highlighted again that all data have been provided in a completely anonymized format and therefore they cannot be linked with any particular natural person. For the analysis, this changing position is considered momentarily static, as a ping indicator on a radar, and its per-second attributes are analyzed. In an algorithmic environment, this is realized via a for-loop implementation for each of the rows (seconds) of the trip dataset.

The initial target now becomes the matching of driver position coordinates with their nearest road segment node. This was achieved by finding the distance to a small number of nearby node, subsequently finding the minimum distance and lastly finding the way (road segment) that the node with the minimum distance belongs to. It should be highlighted that while the OSeven application derives road network types from GPS position, for the purposes of this dissertation, vehicle location was only defined from latitude and longitude coordinates instead of other location information.

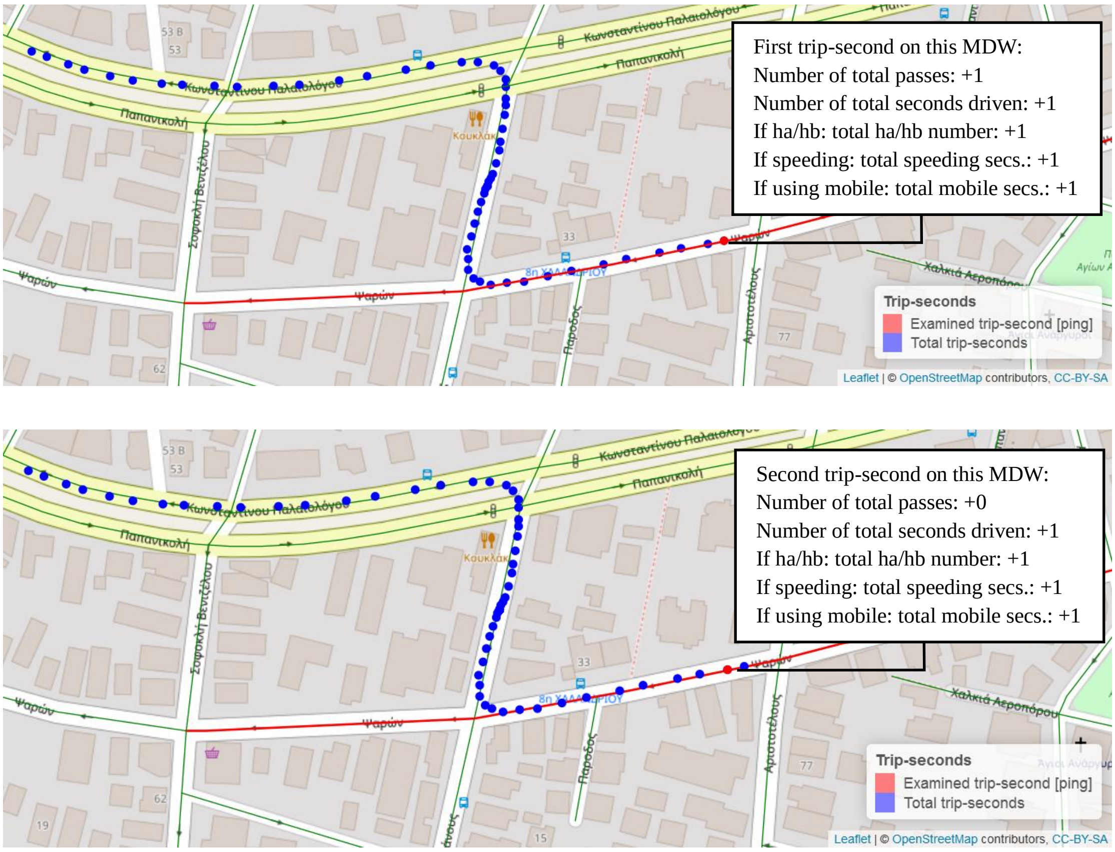
A small moving polygon is drawn around the coordinates of driver position, similar to neighborhood complexity, but without the altitude difference restriction. In practice, dimensions of 350*350 on the OSM grid were used, corresponding to about 280m*380m. The moving window this time saves considerable computation time, as only the most proximal nodes/ways are checked instead of the whole map. The minimum distance is calculated much faster as well, due to limited number of ways. As every row represents a trip-second, the importance of avoiding redundant calculations becomes quickly apparent. The grid could not be shrunk further as large surfaces without roads caused errant behavior from the algorithm. In fact, for urban arterial segments the window had to be raised to dimensions of 500*500 on the OSM grid, corresponding to about 400m*540m. This demand was due to parks and other large areas without roads that prevented map-matching with any road segment at all.

It should be mentioned here that finding the node with the minimum distance from the position is not enough; nodes that are on segment endpoints belong to multiple (≥2) starting and ending segments. To avoid misclassification, all ways that are relevant to the node are listed and then a second distance minimization is conducted. This time, point-to-polyline distances from the position to each of the ways are calculated. Finally, the way that the driver is on is selected as the way with the minimum distance (or min distance way – MDW) from the driver position. The reason that minimum ways are not determined straight from the start is, once again, computational. Comparing distances of a point with the ways of the entire dataset is a computationally demanding process that was mitigated by searching for proximal nodes first.

Having determined the way on which the driver has moved, several metrics that have been required from the previous can now be obtained, namely:

1. Number of seconds driven on the road segment
2. Total number of passes on the road segment
3. Total number of events on the road segment
4. Number of seconds that drivers were speeding that were driven on the road segment
5. Number of seconds that drivers were using mobile phones that were driven on the road segment

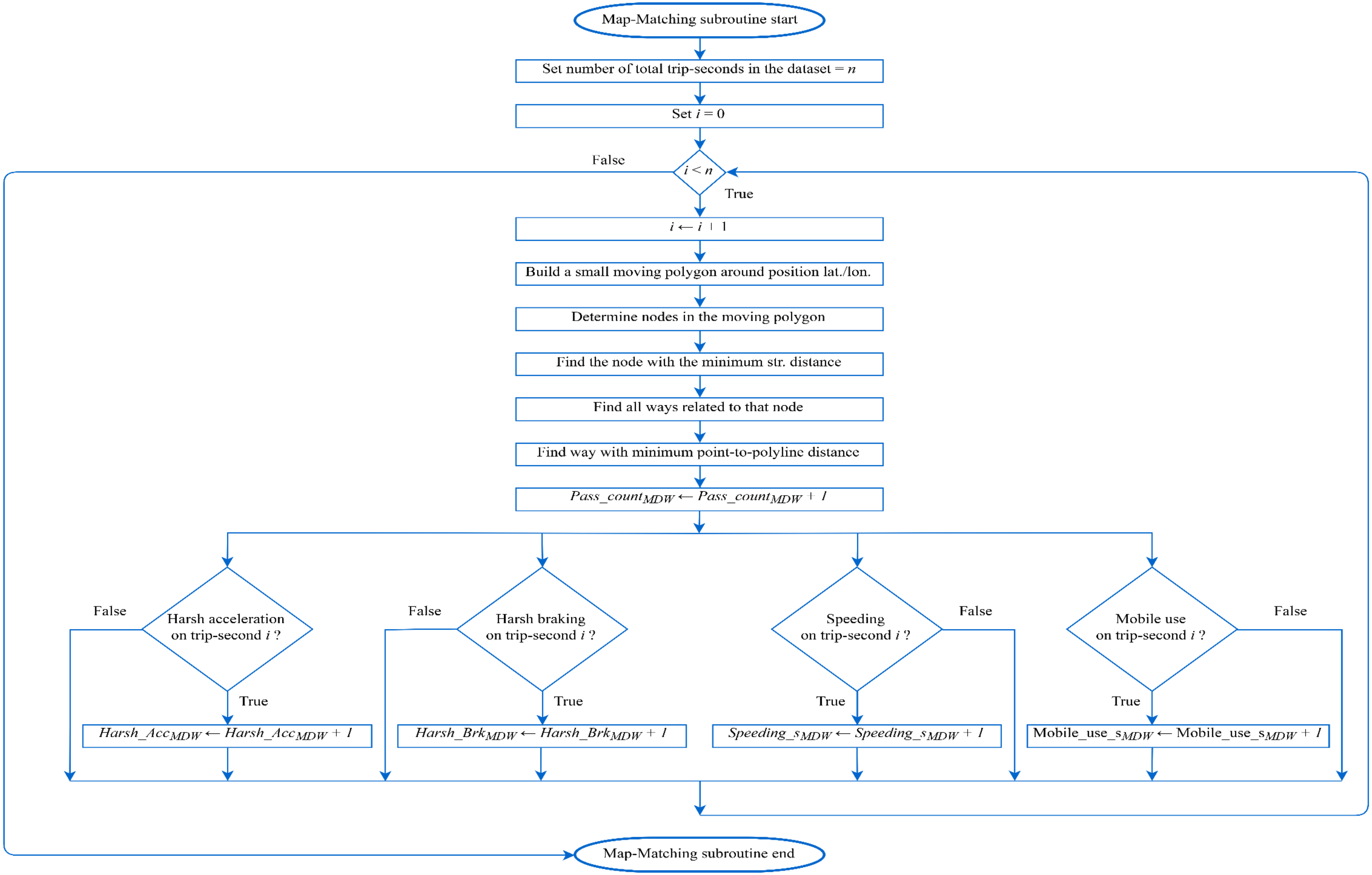
The map-matching process is shown graphically in Figure 3-24 for a sample MDW – drawn in red. The reasoning of the subroutine that is used for the map-matching process is also depicted on a flowchart on Figure 3-25.



First trip-second on this MDW:
Number of total passes: +1
Number of total seconds driven: +1
If ha/hb: total ha/hb number: +1
If speeding: total speeding secs.: +1
If using mobile: total mobile secs.: +1

Second trip-second on this MDW:
Number of total passes: +0
Number of total seconds driven: +1
If ha/hb: total ha/hb number: +1
If speeding: total speeding secs.: +1
If using mobile: total mobile secs.: +1

Figure 3-24: Parameter assignment on the MDW as a trip progresses



Map-Matching subroutine start
Set number of total trip-seconds in the dataset = n
Set $i = 0$
False
$i < n$
True
$i \leftarrow i + 1$
Build a small moving polygon around position lat./lon.
Determine nodes in the moving polygon
Find the node with the minimum str. distance
Find all ways related to that node
Find way with minimum point-to-polyline distance
$Pass\_count_{MDW} \leftarrow Pass\_count_{MDW} + 1$
False
Harsh acceleration on trip-second i ?
True
$Harsh\_Acc_{MDW} \leftarrow Harsh\_Acc_{MDW} + 1$
False
Harsh braking on trip-second i ?
True
$Harsh\_Brk_{MDW} \leftarrow Harsh\_Brk_{MDW} + 1$
Speeding on trip-second i ?
False
True
$Speeding\_s_{MDW} \leftarrow Speeding\_s_{MDW} + 1$
Mobile use on trip-second i ?
False
True
$Mobile\_use\_s_{MDW} \leftarrow Mobile\_use\_s_{MDW} + 1$
Map-Matching subroutine end

Figure 3-25: Map-Matching subroutine flowchart

3.4.3.4 Pass count adjustment

Initial exploratory analyses and map-matching of data revealed an additional issue in the map-matching process: initial results showed considerable double-counting for the number of passes in segments where several road axes travel in parallel. This was particularly pronounced in some segments of Kifisias Avenue, which were allocated up to 5068 pass counts. This number is more than the total trips in the train area dataset, which amount to 3294. Clearly, the issue merited additional attention.

Beginning with a form of diagnosis, this was possibly due to two reasons: Firstly, because some drivers revisited some segments during their trip, therefore contributing more than one pass counts within the same trip. Repeated routes seemed unlikely to happen to a large extent in a confined area, however. The other reason is that, unfortunately, smartphone GPS sensors have an accuracy of ± 5 m at best conditions. This could lead to a 'jump' of the GPS trajectory from one road to another, especially in the aforementioned segments where several road axes travel in parallel. This led to considerable double-counting in trip sections, inflating the calculated pass count unrealistically.

As a result, a vote-counting system was introduced in the algorithm for those contesting segments after visual identification of problematic pairs. For each trip, the instances of assignment to each of the contested segments would be counted as votes. The pass count is then assigned exclusively to the segment winning the majority of the votes, thus limiting double-counting.

The concept is presented visually below. From Figure 3-26, it is evident that Kifisias Avenue and its parallel auxiliary road are both candidate segments for the assignment of the considered trip, based on the wavering of the GPS locations. As shown in Figure 3-27, the specific location has pavement separation, therefore this is not the result of lane-changing. The algorithm recognizes more points in the auxiliary road for that trip, therefore that trip is assigned to the auxiliary road by vote-count. As a note, GPS accuracy, or limitation thereof, led to the exclusion of too microscopic variables from the analysis, such as lateral position, distance from curb/shoulder or lane of preference, which could not be reliably obtained.

Therefore, the pass counts are adjusted for each road segment by the vote-counting system. Henceforth, all mentions of 'pass counts' refer to the pass counts after adjustment.

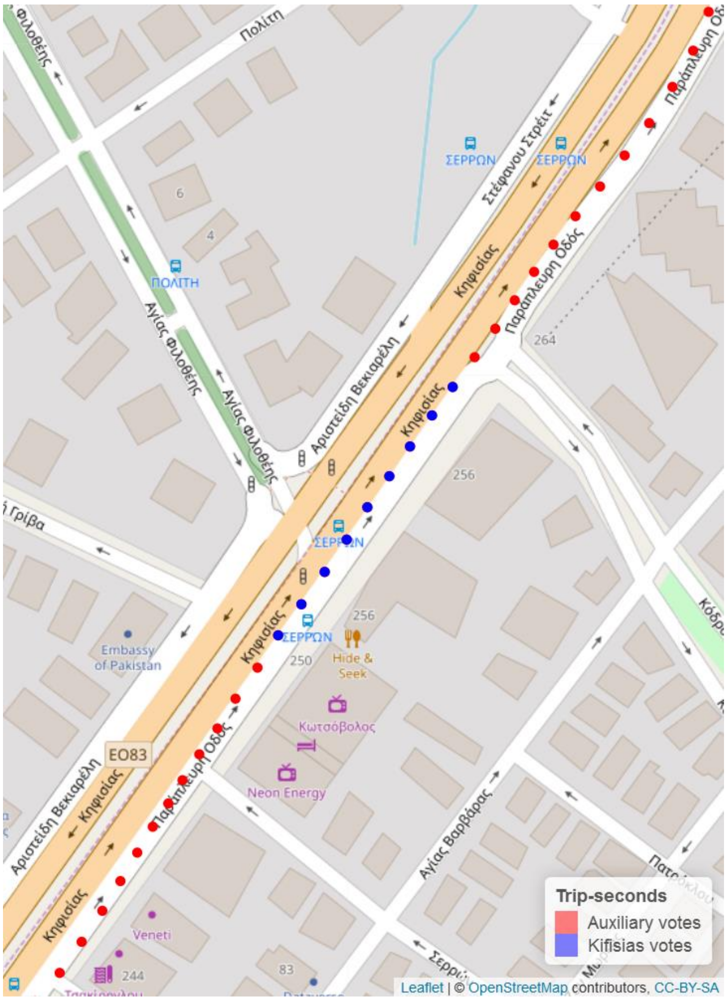


Figure 3-26: Visualization of the vote counting algorithm for pass count adjustment

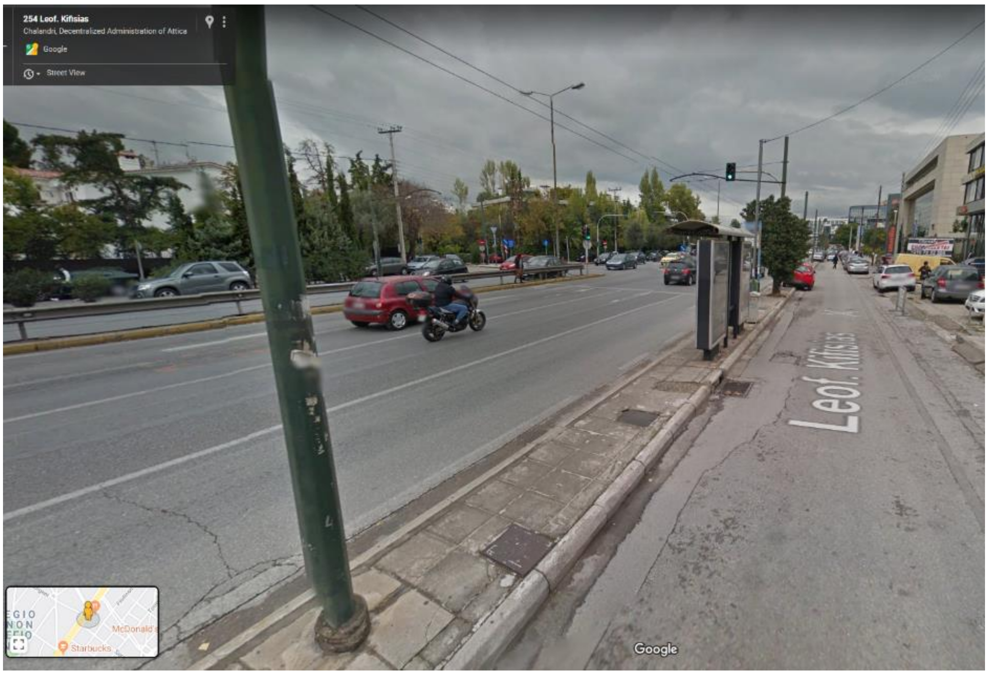


Figure 3-27: Presence of pavement preventing lane-changing in a specific location

3.4.3.5 Harsh event rates

Finally, from the adjusted pass counts, harsh event rates can be also calculated for each road segment. The harsh braking rate for a segment is defined as the number of rates per segment length per trips on the segment. The equivalent mathematical form for a specific road segment w is:

$$hb\_rate_w = \frac{hb_w}{Adj\_pass\_count_w * d_w} \qquad \text{Eq. (88)}$$

And respectively, the harsh acceleration rate for a segment is defined as:

$$ha\_rate_w = \frac{ha_w}{Adj\_pass\_count_w * d_w} \qquad \text{Eq. (89)}$$

Where:

- hb_rate_w and ha_rate_w are the calculated rates of harsh brakings and harsh accelerations in the specific road segment w respectively
- hb_w and ha_w are the count numbers of harsh brakings and harsh accelerations in the specific road segment w respectively
- $Adj\_pass\_count_w$ is the number of passes from the segment adjusted after the application of the vote-count algorithm
- d_w is the segment length in meters

While event rates are typically not compatible with the count models which are widely used in spatial analysis, they can be analyzed in certain similar model forms. For instance, calibrating GWPR is not possible with event rates, but calibrating standard GWR is. Furthermore, harsh event rates can be used to normalize harsh events for descriptive statistics, as they express occurrence of events in number of events per meter per trip for each road segment. Within the framework of this presentation, some modelling attempts were attempted with rates, instead of counts. These attempts ultimately proved unsuccessful, as described in Sections 5.2.4 and 5.4.1.

3.4.4 Derivation of additional behavioral characteristics

The process described so far provides adequate data for purposes of training and testing analyses in urban networks. The previously calculated variables are fixed for each road segment and initially appear to be enough to endeavor transferable analyses that can be also used for prediction. As stated in Section 3.1, however, the occurrence of harsh events can be explored further. These additional analyses were conducted in urban arterials and included additional traffic and road user parameters. The integration of these parameters, which are supplementary to the previous geometric characteristics, is described in this section. These parameters are collected by the algorithm for all study areas, but they are used only in causation models of urban arterials.

3.4.4.1 Speeding percentage

The smartphone application utilized for this dissertation includes information noting whether a driver was speeding at a given second in a binary form $\{0,1\}$. In an attempt to determine road segments which offer ripe ground for speeding, two temporal metrics are considered: (i) the total seconds that a driver spends on the segment, in other words, pass count seconds and (ii) the seconds that a driver spends on the segment while speeding. It is then possible to add speeding seconds passed on the road for all drivers, and total seconds passed on the road for all drivers. By dividing the former by the latter, a speeding percentage per segment can be obtained, which is a dimensionless quantity for a specific road segment w, as described in Equation (90):

$$Speeding_w\ [\%] = \frac{Speeding_w\ [s]}{Pass\ count\ seconds_w\ [s]} \qquad \text{Eq. (90)}$$

In theory, a similar result could be calculated by using total and speeding length units (meters), instead of seconds, yielding another dimensionless quantity. However, temporal units were more clearly defined – as different rows of the dataset – and would also allow for more rapid calculations. Therefore, speeding percentage is determined using trip seconds instead of meters.

3.4.4.2 Mobile use percentage

Similar to speeding, the smartphone application also includes information noting whether a driver was engaged with their mobile phones at a given second in a binary form $\{0,1\}$. Mobile use percentage is calculated with a reasoning mirroring that of speeding percentage, in an attempt to highlight road segments in which drivers prefer to engage their phones. Thus mobile use percentage is determined as the seconds that drivers spend on the segment while using their phones divided by the total seconds that drivers spend on the segment, for a specific road segment w, as per Equation (91):

$$Mobile\ use_w\ [\%] = \frac{Mobile\ use_w\ [s]}{Pass\ count\ seconds_w\ [s]} \qquad \text{Eq. (91)}$$

3.4.5 Integration of traffic data

3.4.5.1 Theoretical background of traffic states

The integration of parameters describing the state of traffic in the examined road segments for the duration of the naturalistic driving was a particularly challenging part of the doctoral dissertation. The intricacies of the problem lied in the fact that traffic data, which describe an instantaneous state of the road, needed to be combined with and augment fixed data, such as geometric road data.

On a higher level of analysis, traffic can be categorized into three states, also known as regimes: (i) free flow, (ii) synchronized flow and (iii) congested flow. This separation is also known as three-phase traffic theory (Kerner, 2012). In brief, in free flow vehicles are free to move with the desired speed and there is a positive correlation between traffic flow and traffic density. There is then a transition to synchronized flow by either slow variations in volume or abrupt increases in occupancy (Vlahogianni et al., 2008). There are reductions of speed, and different vehicles synchronize their speeds, but without vehicle stoppage. If the traffic flow further increases, there is a speed breakdown and a transition to congested flow. In congested flow, vehicle speed is lower than the lowest vehicle speed found in free flow.

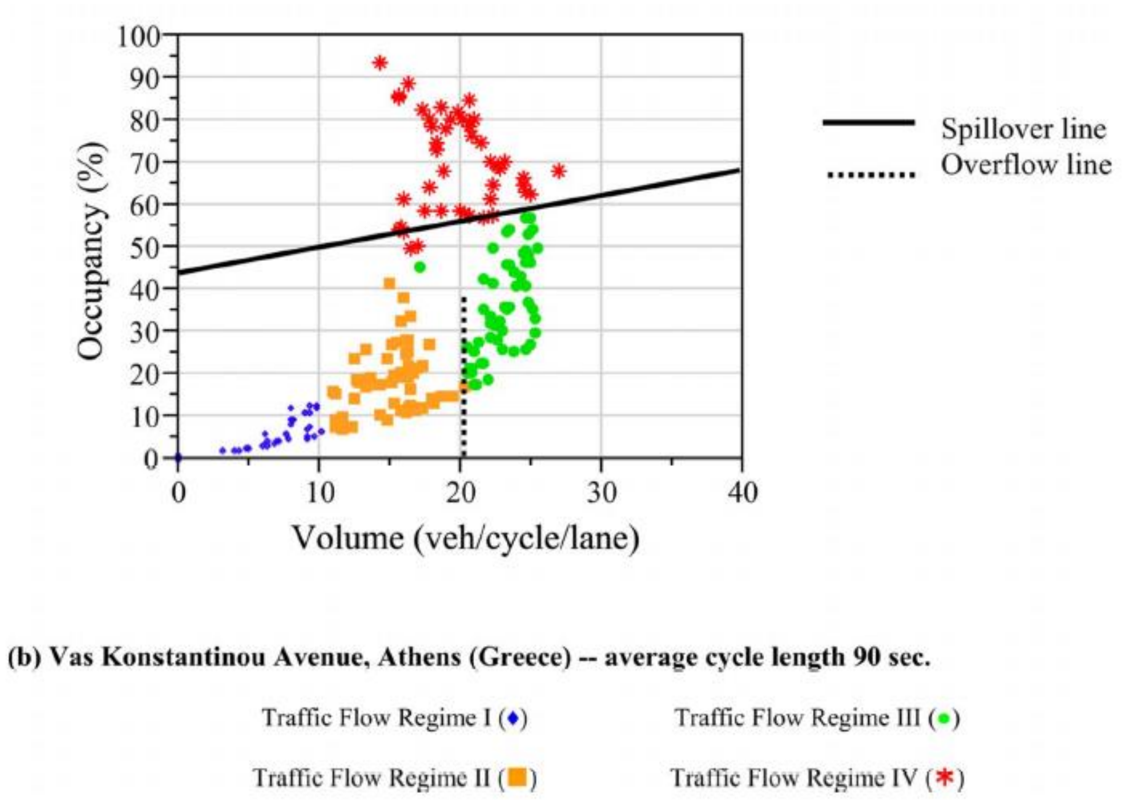
Several seminal papers have examined the particularities of traffic flow mathematics and transitions between states (e.g. Lighthill & Whitham, 1955; Kerner, 2012). It has been widely established in traffic engineering and traffic flow theory that traffic behaves in a highly non-linear manner, and that traffic flow can display volatile behavior in microscopic scales (Kamarianakis et al. 2005; Vlahogianni et al. 2006). Many complex mathematical approaches and machine learning algorithms have been developed in traffic theory studies for short-term traffic forecasting (e.g. Hu et al., 2016; Xia et al., 2016; Polson & Sokolov, 2017). On the other hand, many road safety-oriented studies – including studies conducting spatial analyses – eschew these intricacies and treat traffic parameters as instant measurements (e.g. Wang & Abdel-Aty, 2006; Lee et al., 2017a).

An over-simplistic approach would be inappropriate and undermine the effort made for the collection and combination of high-resolution data. On the other hand, delving into the explicit mathematic intricacies for short-term traffic flow are outside the scope of the present dissertation. Therefore, in an effort to bridge the two extremes, an approach involving the determination and separate modelling of each traffic state was chosen for urban arterial analysis.

Specifically, the existing traffic parameters were obtained for every trip-second. Based on the values of these parameters, the state of traffic flow was determined. Subsequently, the naturalistic driving dataset was split into three smaller subsets, corresponding to each traffic flow regime. Finally, the selected models were calibrated for each state separately.

The limit values for traffic flow categorization can be quite sensitive and dependent to the particular location. Fortunately, earlier work by Vlahogianni et al. (2008) had determined limit values for traffic states in Vasileos Konstantinou Avenue, which is the extension of Kifisas Avenue and Vasilissis Sofias Avenue to the south. The geographical location is not only very proximal but an actual extension of the arterial, the directional separation is retained, the number of lanes and overall road design is similar and the traffic light cycle calibration and management is the same – conducted by the TMC. For all these reasons, it was decided to adopt the limits determined by that study. Therefore, in Kifisias Avenue traffic states can be determined based on traffic occupancy and traffic volume per cycle per lane as shown in

Figure 3-28. Traffic flow regime (i) refers to free flow, traffic flow regimes (ii) & (iii) refer to synchronized flow and traffic flow regime (iv) refers to congested flow.



(b) Vas Konstantinou Avenue, Athens (Greece) -- average cycle length 90 sec.

Figure 3-28: Traffic flow regimes for Vas. Konstantinou Ave.
[Source: Vlahogianni et al., 2008]

It is noteworthy that, in the same study, Vlahogianni et al. (2008) calculate traffic occupancy and traffic volume limits for Lincoln Boulevard in Los Angeles as well. These values are significantly different, therefore the limits between traffic states are not widely transferrable to all areas indiscriminately. The stability of the selected approach is enhanced by the fact that only a single urban arterial is considered, specifically Kifisias Avenue.

3.4.5.2 Merging of traffic data with naturalistic driving data

As previously stated, traffic data was acquired for the enhancement of analyses in urban arterial segments from the Athens TMC. The traffic dataset had to be then matched to the naturalistic driving dataset. Another map-matching algorithm was required, similar to the one used for the integration of smartphone data.

The objective was slightly different this time: each row representing a trip-second was examined and the coordinates representing the vehicle position were extracted. Based on these coordinates, the most proximal measurement location was obtained for each trip-second. This practice has been followed in previous road safety research utilizing TMC data in the past (e.g. Yannis, et al., 2014; Theofilatos et al., 2018b). Naturally, this approach required the determination of road segment bearing, namely a separation based on whether segments are northbound or southbound for Kifisias Avenue. This was conducted in advance to avoid the sampling of locations that are from the opposite direction of traffic. TMC measurement locations are shown on Figure 3-29.

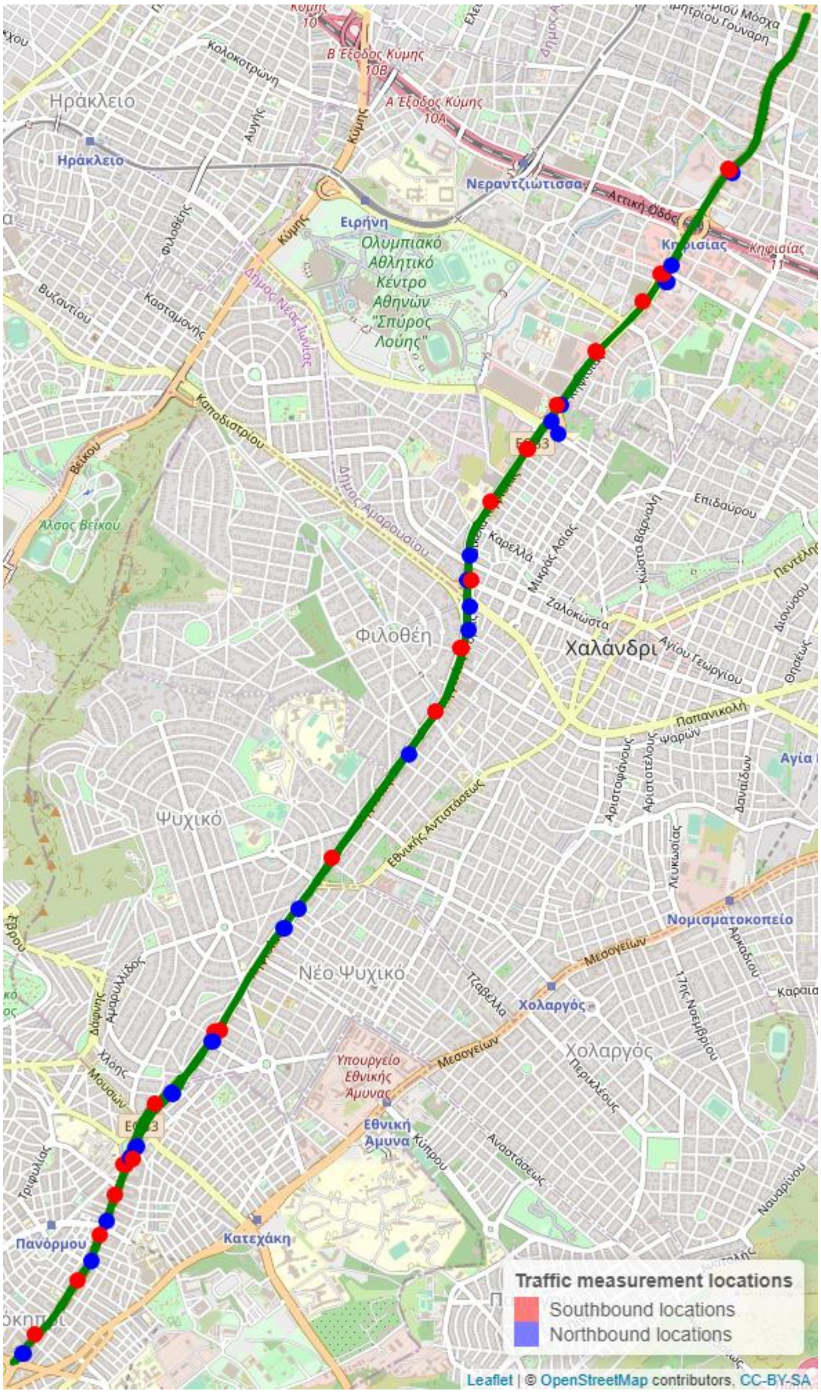


Figure 3-29: Traffic measurement locations in Kifisias Avenue

The bearing and type (main road or non-main road) of measurement locations are shown on **Table 3-4**.

Table 3-4: Measurement locations per type and bearing

Road segment bearing	**Type of measured segment**		
	Main road	Non-main road	**Total**
Northbound	13	14	27
Southbound	14	13	27
Total	27	27	**54**

On average, there is one measurement location every 293 m in Kifisias Avenue for both northbound and southbound segments if measurements locations on ramps and bus lanes are included. This density is reduced to one measurement point per 607 m in northbound segments and one measurement point per 564 m in northbound segments if only measurements on the main road are considered. Overall, this is considered a very satisfactory measurement density that can meaningfully augment high-resolution big data from naturalistic driving.

However, TMC detection systems are subject to constant physical strain and can fail temporarily or provide erroneous recordings. This can lead to the recording of unrealistic values, such as occupancy exceeding 100 %, speed exceeding 200 km/h or speed equal to 0 along with traffic volume equal to 0, which are discarded from the traffic dataset.

When that occurs, the measurements from the second nearest measurement point are sought out, or the third nearest after that and so on. A spatial-based reduction from the entire measurement location dataset was implemented, only traffic measurements within 3 kilometers of the naturalistic data coordinates were considered. A temporal-based reduction was also implemented; only traffic measurements within 15 minutes of the naturalistic data timestamp were considered. These reductions were applied for two reasons: (i) to ensure that a realistic representation of traffic conditions was acquired and (ii) to further reduce computational times for the algorithms. The term timestamp refers to a character string variable containing the information of the precise date and time of the measurements.

In practice, the matching of traffic and naturalistic dataset process was based on the following steps:

1. Examination of each trip-second and acquisition of its coordinates and timestamp
2. Determination of the nearest measurement location from these coordinates
3. Examination of a list of candidate measurements within 15 min
4. Acquisition of measurements from the timestamp that is closest to the trip-second temporally
5. If the list of candidate measurements of step 3 was empty or erroneous, the next-closest measurement location was sought, and so on, within a distance of 3 km.

After the merging process, the following traffic characteristics are extracted from the TMC measurements and matched to each trip-second:

- $\overline{v_{tr}}$, current proximal traffic speed averaged over time [km/h]
- q, traffic flow projected to traffic flow per hour by the TMC [veh/h]
- k, current traffic occupancy [%]

Following the described process, a naturalistic driving dataset enhanced with traffic parameters was obtained. In this dataset, the previous characteristics allowed for the calculation of three more meaningful parameters, namely:

- $\hat{q}_{lane}$, current traffic flow of the 90s cycle and measurement interval standardized to traffic flow per lane [veh/lane/cycle].
- Tr_State, the current traffic state calculated from the limits of Vlahogianni et al. (2008) based on k and $\hat{q}_{lane}$
- $\Delta Speed$, the mathematical difference of the average traffic speed and the naturalistic driving speed for each driver collected from smartphone data:

$$\Delta Speed = \overline{v_{tr}} - \overline{v_{nd}} \qquad \text{Eq. (92)}$$

As a note, the speed difference $\Delta Speed$ was selected as a particular parameter of interest. Larger speed differences have long been determined as related to a higher rate of crashes from relevant literature (Aarts & Van Schagen, 2006), therefore the examination of its effect on harsh event frequency was considered a fruitful pursuit.

Traffic state was then used as a filter label variable to obtain enhanced naturalistic driving subsets for free, synchronized and congested flow conditions, as described in the following. For each subset, the map-matching process that was described in Section 3.4.3 was followed again. The averages (arithmetic mean) of traffic parameters were then obtained for each road segment per traffic state. In other words, 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. This was an essential information compression since the final analyses of harsh events are conducted on a road-segment basis.

The spatial datasets were ready for the calibration of models. A series of trials followed with all geometric, traffic and road user parameters to determine the most informative combination in the respective statistical analysis for urban arterials described on Section 7.

4 Urban road network data collection and processing

This section provides technical information on the process of data collection, descriptive statistics, exploratory parameters and various additional information for the urban network data describing both the training area (Chalandri) and the test area (Omonoia).

4.1 Training area – Chalandri

4.1.1 Initial study area examination

The relevant exports were conducted via a purpose-made Application Programming Interface (API) which receives a user-selected area as input and provides raw OSM data for that area (https://overpass-turbo.eu/). An initial visual exploratory check was conducted to determine any discrepancies between the map image and the raw OSM data import; no discrepancies were detected. The map with the axes of the imported segments (in green) is shown on Figure 4-1 for the selected Chalandri area. Walkways and similar footpaths have been removed from the segments, hence they are not appearing on the processed maps apart from the baseline.

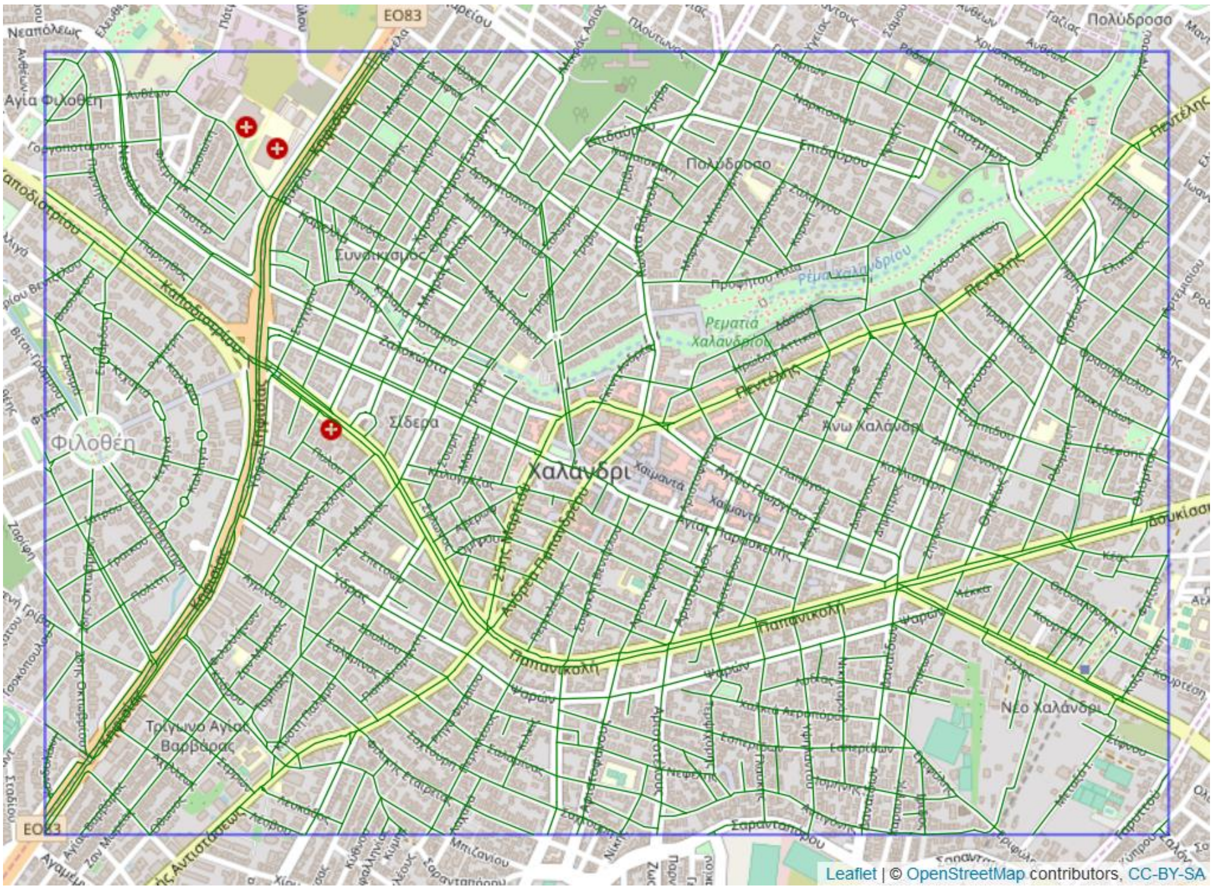


Figure 4-1: Chalandri road segments following import from OSM and removal of footways

Having exported the raw OSM data, and after enhancing them with SRTM data for the training area, the processing phase was ready to begin. The stages outlined in Sections 3.4.2 – 3.4.3 were followed consecutively.

4.1.2 Road geometry characteristics

The initial raw data for Chalandri show that in the examined area 883 segments (ways in OSM terms) are initially included which are consisted of 4293 nodes; relation data were not exported since they are not needed in the analyses in order to reduce processing times. After the exclusion of 14 footway segments, 869 segments with vehicle traffic remained. There were no other exclusions necessary, for instance due to missing data. All remaining road segments are utilized. The various road type frequencies of the segments per road direction and lane number appear on **Table 4-1**:

Table 4-1: Road type frequencies in Chalandri area

Road type	**Road direction**									
	One-way segments				Two-way segments				**Total**	**Total [%]**
	Lanes: 1	Lanes: 2	Lanes: 3	Lanes: 4	Lanes: 1	Lanes: 2	Lanes: 3	Lanes: 4		
Primary	0	0	29	1	0	0	0	0	30	3.45 %
Secondary	6	77	24	2	1	11	2	0	123	14.15 %
Tertiary	103	39	0	0	4	13	0	0	159	18.30 %
Residential	521	5	0	0	31	0	0	0	557	64.10 %
Total	630	121	53	3	36	24	2	0	**869**	**100.00 %**

By examining the table, it is evident that residential one-way, one-lane roads comprise the majority of the segment sample. That being said, there is considerable representation of other segment values, which is expected from a dense urban area. There is a relatively good balance of primary, secondary, tertiary and residential road segments in the area, and only primary or secondary rows have more than two lanes, while residential roads feature strictly one lane.

Descriptive values for the obtained geometric and road network characteristics appear on **Table 4-2**. As a reminder, gradient and neighborhood complexity are dimensionless quantities, and negative gradient values refer to downhill slopes.

Table 4-2: Descriptive statistics for the obtained geometric characteristics for road segments in Chalandri area

Geometric characteristics	**Descriptive statistics**						
	Average	Min	Median	Max	St. Dev.	Skewness	Kurtosis
Segment Length [m]	144.2795	0.7913	96.8131	963.8541	147.8674	2.0124	4.7595
Curvature [m^{-1}]	0.0053	0.0000	0.0002	0.1295	0.0144	4.5182	23.8712
Gradient [–]	0.0007	-0.1779	0.0010	0.1641	0.0458	-0.0001	1.8157
Neighborhood Complexity [–]	4.8311	2.0792	4.8903	5.5410	0.4016	-1.4386	4.6303

It is again evident that all geometric characteristics have significant fluctuations and dispersion of values. This is a desirable image overall; a diverse urban network serving as training area ought to lead to more transferrable results. Most of the metrics (with the exception of gradient) assume positive values, and can vary greatly between road categories, explaining the larger values of standard deviation.

It should be noted that there was a minor number of segments with very small lengths; this was due to cropping of the training area as shown in Section 3.4.1.1. Trip data are cropped in a similar – rectangular – manner. Therefore discarding these segments would lead to erroneous assignment of trip-seconds to neighboring roads by the map-matching algorithm, thus introducing bias.

The kurtosis of most geometric characteristics is comparable to the value of 3 featured by univariate normal distributions in magnitude; the kurtosis of curvature is larger, suggesting more frequent outliers with higher value in that variable. Additionally, since neighborhood complexity is a logarithm of the number of proximal nodes of each segment, negative skewness is expected.

To conclude with the fixed network characteristics of the training area, the presence of traffic lights and pedestrian crossings detected from tags in the OSM data is provided on **Table 4-3**. In total, 49 segments featuring traffic lights and 80 segments featuring pedestrian crossings were detected. It is possible that segments feature both network characteristics, as evident from the example of Figure 3-22.

Table 4-3: Fixed network characteristics in Chalandri area

Road type	**Network characteristics**							
	Presence of traffic lights				**Presence of pedestrian crossing**			
	No		Yes		No		Yes	
Primary	26	2.99 %	4	0.46 %	28	3.22 %	2	0.23 %
Secondary	115	13.23 %	8	0.92 %	108	12.43 %	15	1.73 %
Tertiary	151	17.38 %	8	0.92 %	144	16.57 %	15	1.73 %
Residential	528	60.76 %	29	3.34 %	509	58.57 %	48	5.52 %
Total	820	94.36 %	49	5.64 %	789	90.79 %	80	9.21 %
Grand Total			**869**	**100.00 %**			**869**	**100.00 %**

Several maps and heatmaps can be produced from the above characteristics. Figure 4-2 to 4-6 provide an intuitive presentation of existing network features, such as road type, and derived road features, such as gradient.

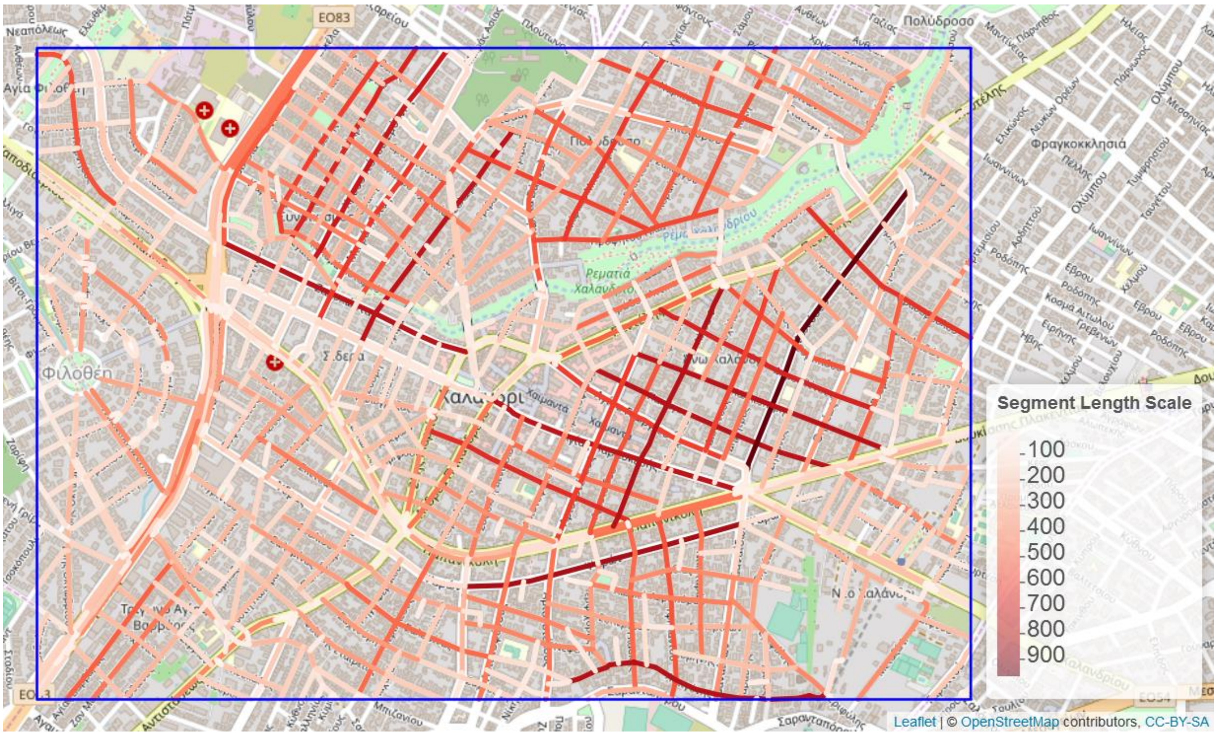


Figure 4-2: Heatmap of road segment lengths in Chalandri area

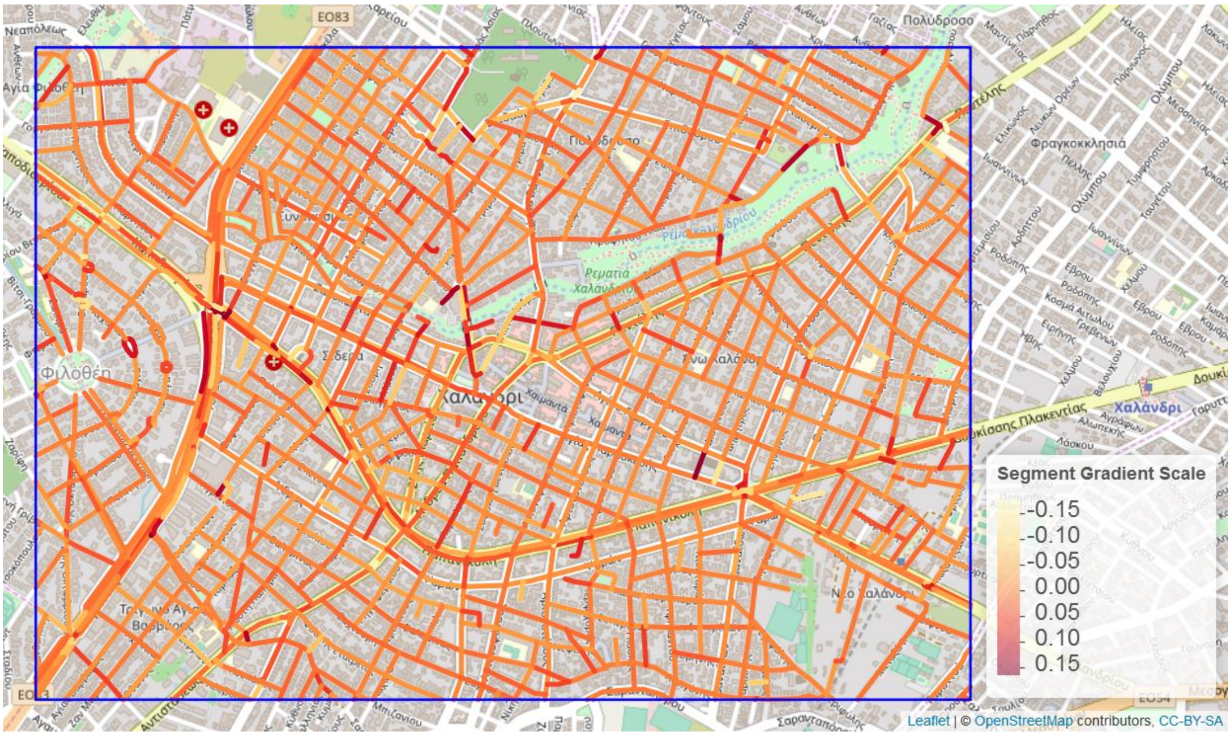


Figure 4-3: Heatmap of road segment gradients in Chalandri area

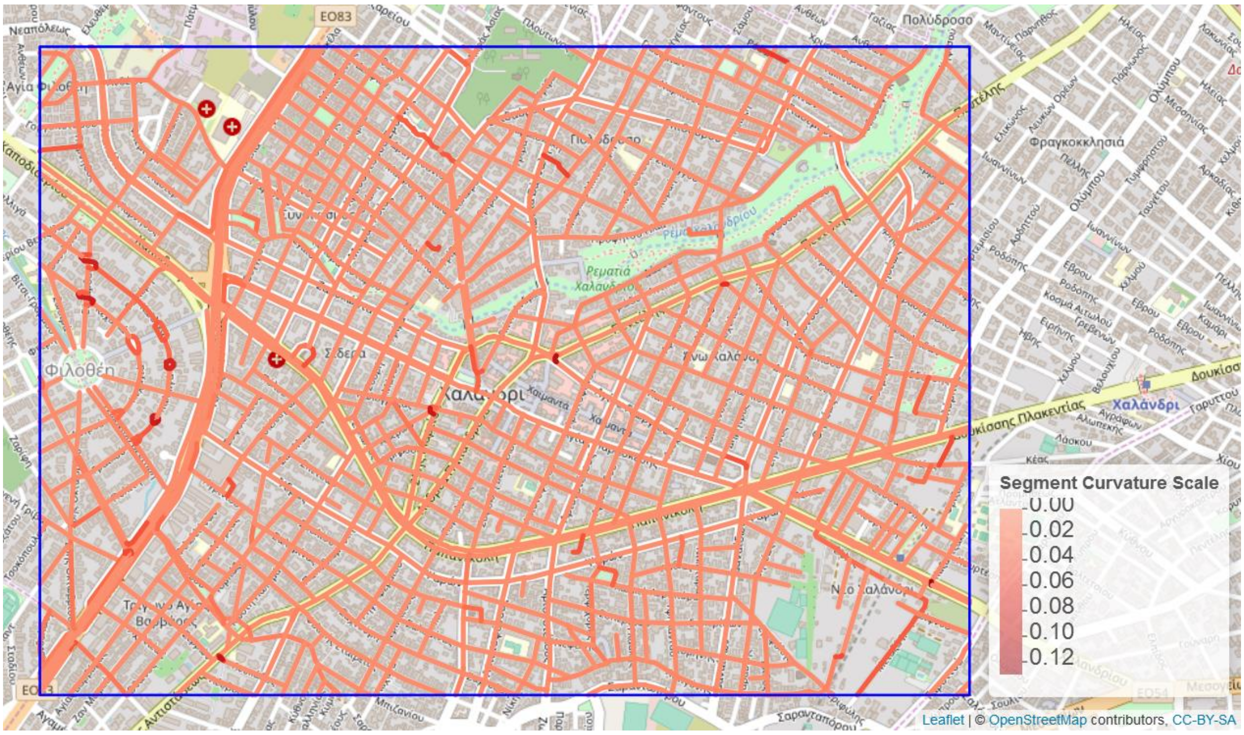


Figure 4-4: Heatmap of road segment curvatures in Chalandri area

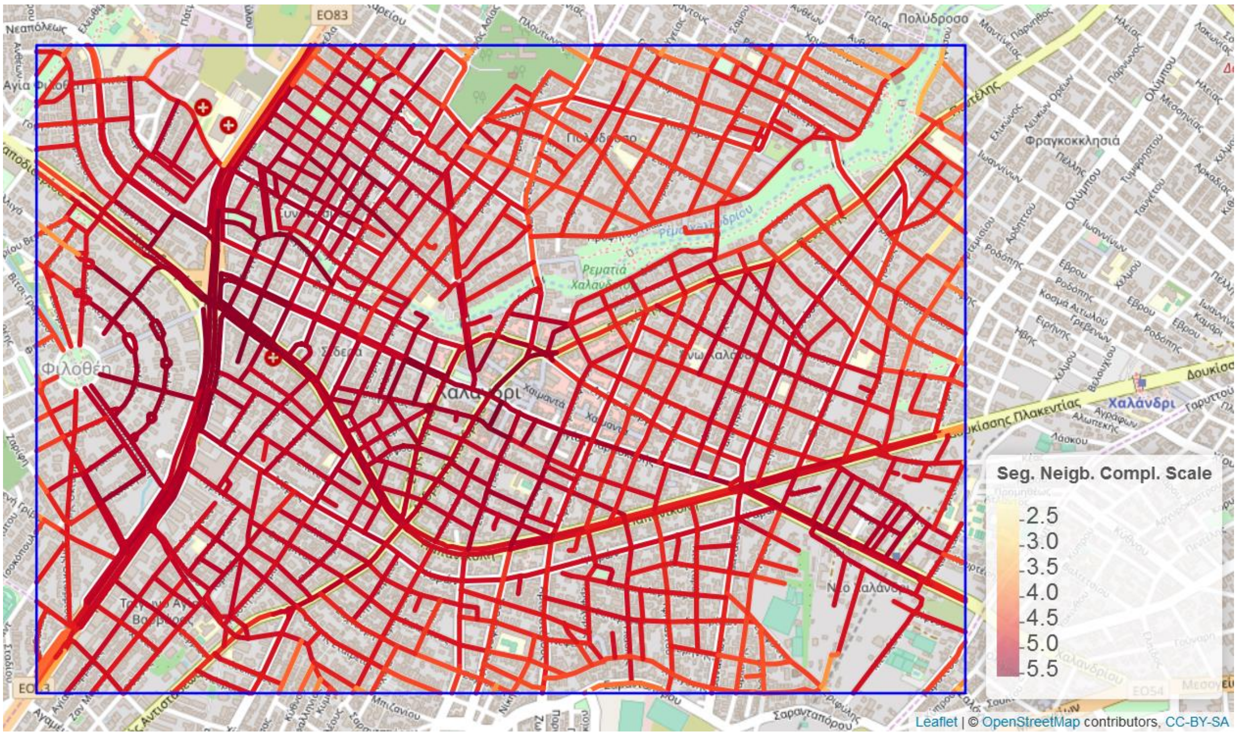


Figure 4-5: Heatmap of neighborhood complexity in Chalandri area

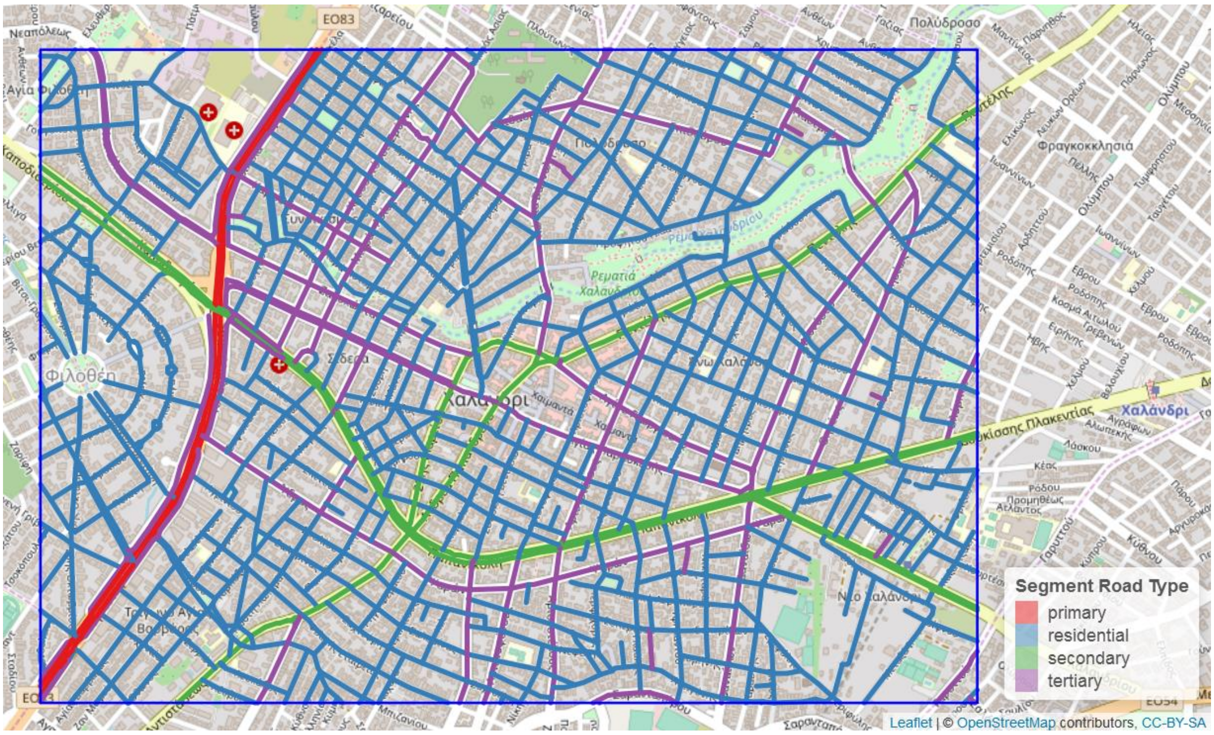


Figure 4-6: Mapping of road segment types in Chalandri area

4.1.3 Large-scale naturalistic driving data exploration

Following the determination and measurements of the various considered geometric characteristics, the naturalistic trip data are examined after being obtained from the OSeven application.

The provided dataset corresponded to a period of two months; specifically from 01-10-2019 to 29-11-2019. During that period, 3294 trips were provided from 230 individual drivers. As previously explained, driver data are completely anonymized, therefore other driver-specific information such as gender, age, aggressiveness or crash history is completely unknown. These trips were not necessarily confined in the Chalandri training area; some had origins and/or destinations on road segments outside the borders depicted on Figure 4-1. However, they were all cropped so that only the length of each trip that fell into the training area was considered – the remaining information was discarded.

Before processing, the provided trips had an average duration of 1410 seconds (or 23.50 minutes); some trips reached more than 10,000 seconds (or 2.7 hours). As each second of the recorded trip is represented by a row in the respective file, this resulted in a trip file with 4,648,555 rows. This size cannot be easily loaded and manipulated by conventional software (e.g. Microsoft Excel) and can be considered to lie towards the 'big data' classification. After cropping the trips, trip duration was reduced to a mean of 304 seconds, for a file of 1,000,273 entries. This is expected from trips within areas comparable to urban municipality. The histogram of trip durations is shown on Figure 4-7. While there are some trips with increased duration, these are few and can be attributed to heavy traffic in the area combined with a more cyclical route.

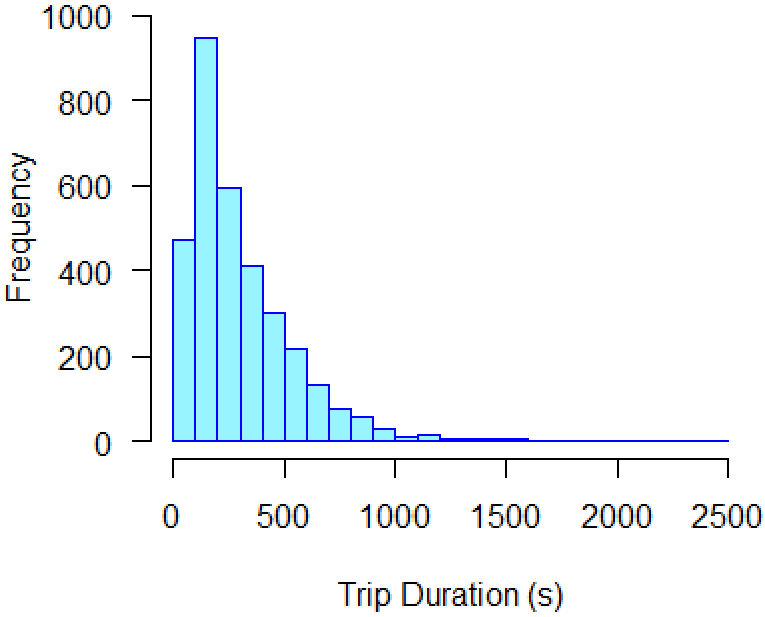


Figure 4-7: Histogram of trip durations in Chalandri area

In these trips, a number of harsh events have occurred and were recorded alongside normal driving conditions, consisting of 1348 harsh braking events and 921 harsh acceleration events. OSeven classifies harsh events to three categories of intensity, 1 – low, 2 – medium and 3 – high. As per the aforementioned, for the purposes of this dissertation, these events are considered as point-data in space (i.e. without considering the length in which they occur). Furthermore, the analyses are made on an aggregated level – events are examined uniformly regardless of intensity. In this dissertation, intensity categories are renamed as 1 – mild, 2 – moderate and 3 – severe. The numbers of harsh events per intensity category appear on **Table 4-4**.

Table 4-4: Harsh events per intensity category in Chalandri area

Event intensity category	Harsh events			
	Harsh brakings		Harsh accelerations	
1 – mild	778	57.72 %	524	56.89 %
2 – modest	409	30.34 %	291	31.60 %
3 – severe	161	11.94 %	106	11.51 %
Total	1348	100.00 %	921	100.00 %

Similar to geometric characteristics, harsh events can be depicted on the map of the training area, as shown on Figure 4-8 for harsh brakings (hb) and on Figure 4-9 for harsh accelerations (ha).

As evident from the maps, the majority of events tend to occur concentrated on roads of higher categories, which feature longer segments with more lanes and heavier traffic (namely primary and secondary roads in OSM classification).

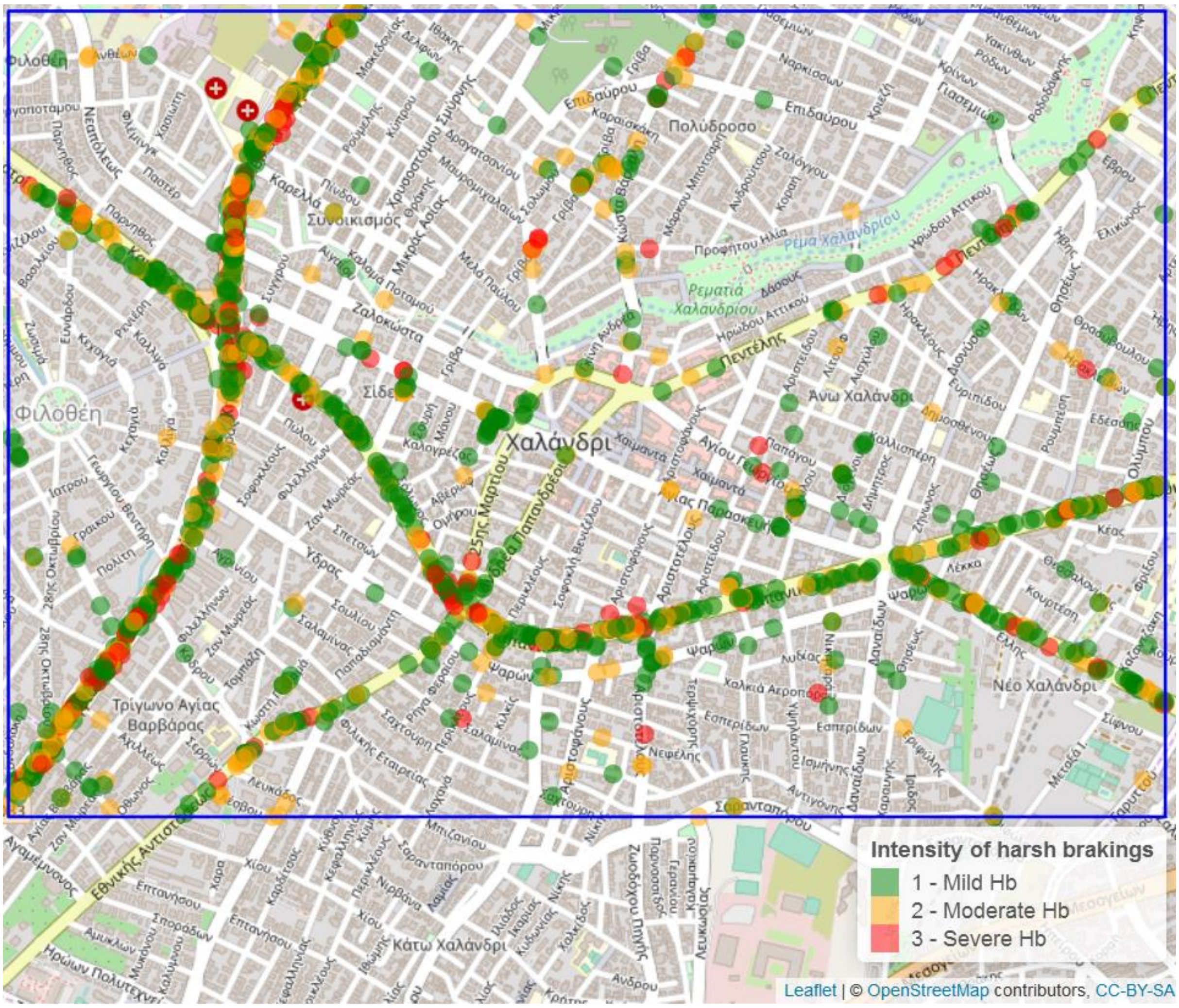


Figure 4-8: Harsh braking events in Chalandri area

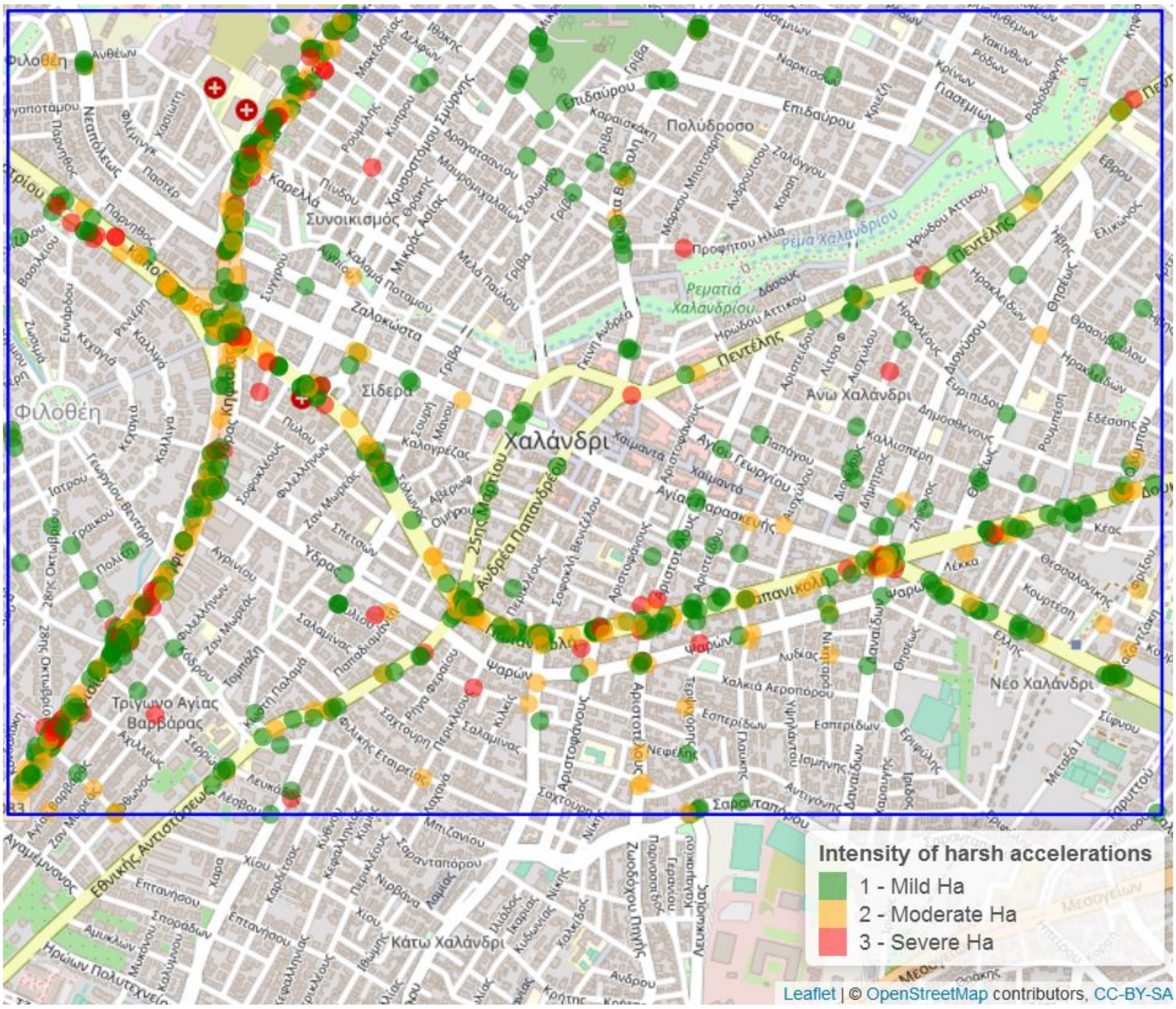


Figure 4-9: Harsh acceleration events in Chalandri area

These busier segments are the ones that would be typically selected by a driver to arrive or depart from the area; therefore they are likely to be at least part of many trips. This form of event distribution provides additional incentive to include road segment length and trip number as exposure parameters in the models. Several events have occurred outside the primary and secondary roads, however, and these might serve as an indicator of unsafe segments or at least a guide for model calibration.

While GPS accuracy is always an issue, it does not pose an insurmountable limitation for the allocation of point-like phenomena, such as harsh events. This is because in case they occur in large categories of roads in clusters, any position inaccuracies are expected to largely cancel out in the overall distributions due to the overall number of events. When harsh events occur in tertiary or residential road segments, the map layout is such where small inconsistencies in meters are absorbed by the map-matching process.

4.1.4 Map-matching results

The geometric and naturalistic data were successfully imported, subjected to a quality check and yielded the previous descriptive statistics. The next step was the implementation of the map-matching algorithm for the training area, as described in Section 3.4.3. This process was conducted for all 1,000,273 trip-seconds in the naturalistic driving big dataset, assigning each trip-second to the nearest OSM road segment. The runtime of the map-matching for Chalandri area was 7 hours and 35 minutes on a server-level computer. Descriptive statistics for the obtained parameters are shown on **Table 4-5**. Parameters with an asterisk (*) are reported only for segments that had non-zero trips, since they are calculated with either the adjusted pass count (trips per segment) or pass seconds per segment. The segments with non-zero trips are 782 out of the total 869, with 87 segments receiving no trips from the 230 drivers during the data collection period.

Table 4-5: Descriptive statistics for road segments in Chalandri area after map-matching

Segment characteristics from naturalistic driving	**Descriptive statistics**						
	Average	Min	Median	Max	St. Dev.	Skewness	Kurtosis
Pass count per segment (assigned)	191.2911	0	51	5068	451.4869	5.7548	42.6651
Adjusted pass count per segment	76.5121	0	22	933	132.4259	2.9083	9.8008
Harsh brakings per segment	1.5512	0	0	51	4.6302	5.6063	40.7254
Harsh braking rate per segment *	0.0002	0.0000	0.0000	0.0118	0.0007	9.1712	112.9984
Harsh accelerations per segment	1.0587	0	0	34	2.9703	6.1818	52.2270
Harsh acceleration rate per segment *	0.0008	0.0000	0.0000	0.3758	0.0137	26.2986	711.2173
Pass seconds per segment	1150.0035	0	245	25628	2589.2332	4.1790	21.8646
Mobile use seconds per segment	32.9275	0	6	810	75.3562	4.6195	28.2723
Mobile use percentage per segment *	3.99 %	0.00 %	2.10 %	100.00 %	7.75 %	6.7435	65.8524
Speeding seconds per segment	20.3751	0	0	1051	85.9273	7.4023	66.9560
Speeding percentage per segment *	1.82 %	0.00 %	0.00 %	100.00 %	7.50 %	9.8151	113.9056
Average driver speed per segment	21.3180	0.0000	19.8081	195.8333	16.1035	5.2435	44.6536

The descriptive statistics obtained from map-matching offer additional initial insights to the spatial examination of harsh event frequencies in the training area. The obtained values from map-matching are all positive real numbers, which is expected since they represent frequency counts and their respective rates or percentages.

The majority of road segments in the training area were assigned at least one trip, namely 782 out of 869 or 89.99% of the total. Conversely, 87 segments did not have any trips, amounting to 10.01% of the total. This indicates a good spatial coverage of the training area, although it is not uniform, as evidenced by the high standard deviation of the (adjusted) pass count per segment. The adjusted pass count assignment is depicted in the heatmap of Figure 4-10.

It is evident that drivers were not overly aggressive, since the averages of harsh braking and harsh acceleration frequencies per segment are low and the respective medians are zero. Harsh event rates appear to follow highly asymmetrical distributions, which show considerable kurtosis. Therefore there are hints of several road segments with high outliers. Furthermore, drivers tend to abide by standing driving regulations in most instances, since in most trip-seconds there is no record of mobile phone use or speeding occurring. This is a generalized statement from a frequentist perspective, however. The respective heatmaps are shown on Figure 4-11 and Figure 4-12.

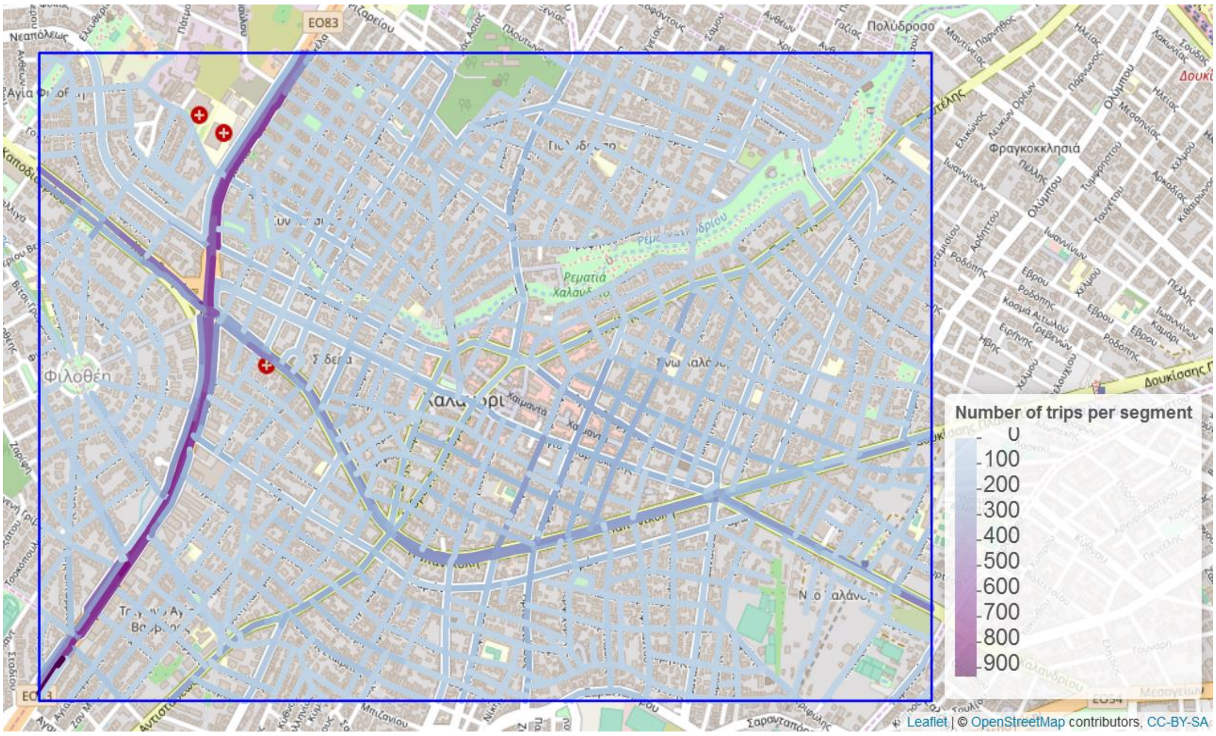

Number of trips per segment
0
100
200
300
400
500
600
700
800
900
Leaflet | © OpenStreetMap contributors, CC-BY-SA

Figure 4-10: Heatmap of adjusted pass counts of segments in Chalandri area

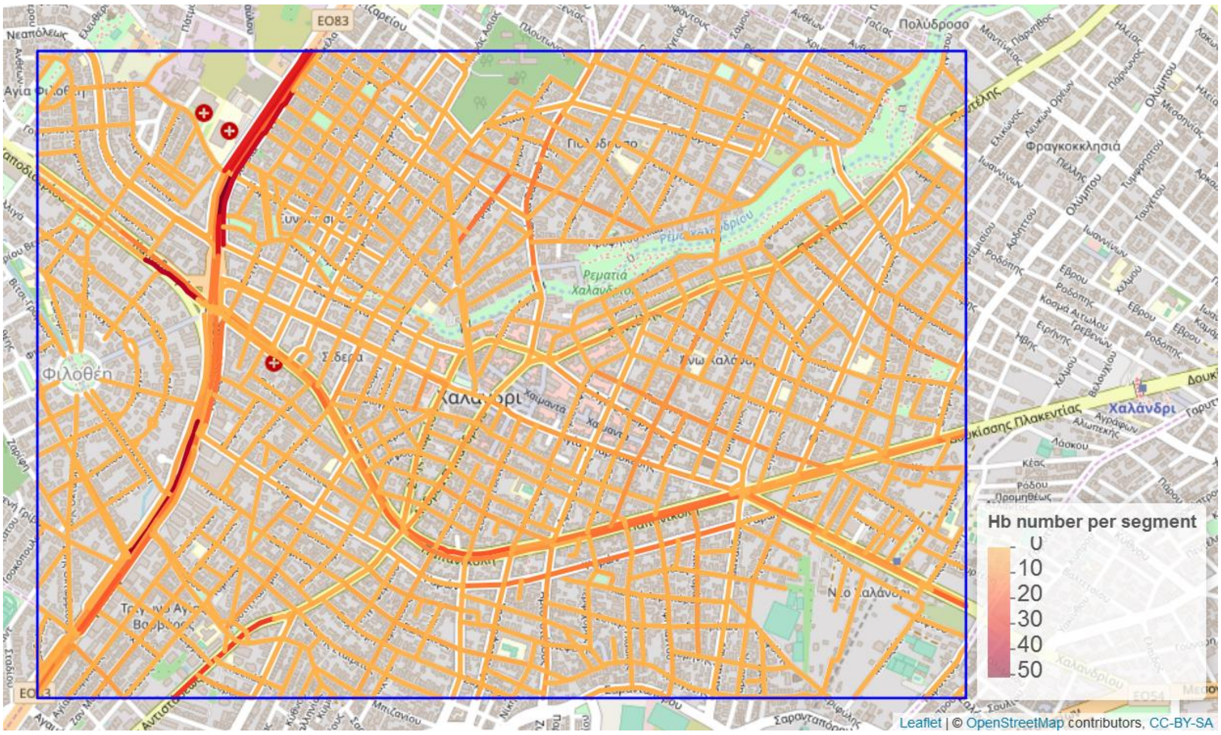


Figure 4-11: Heatmap of harsh braking frequencies of road segments in Chalandri area

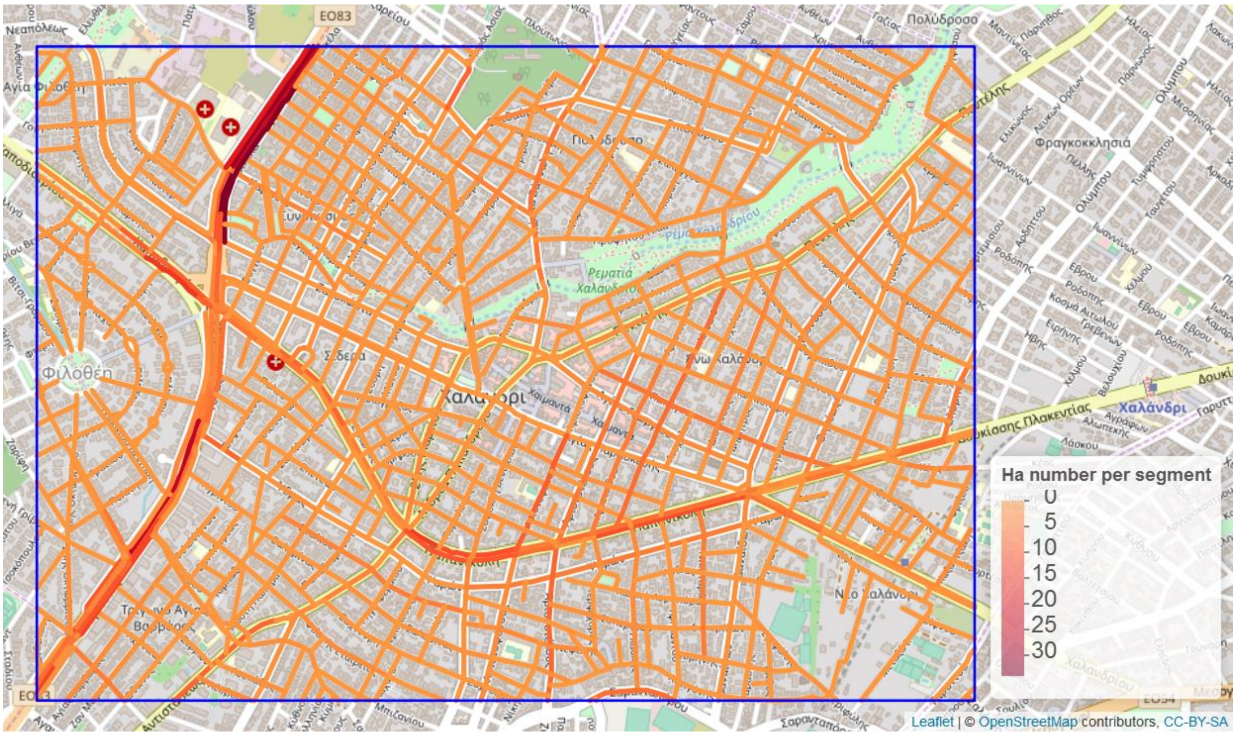


Figure 4-12: Heatmap of harsh acceleration frequencies of road segments in Chalandri area

There is positive skewness in all variable distributions, which reveals asymmetrical distributions with longer right tails. The kurtosis of all calculated variables is considerably higher than the value for normal distributions, which is 3, and significantly 'heavy-tailed' (leptokurtic) distributions for all variables. The high kurtosis values indicate the presence of infrequent sizeable deviations present for each variable in the dataset; this can be attributed to the simultaneous examination of several road categories. High kurtosis is especially pronounced in the distribution of speeding seconds and percentages per segment, in other words, drivers tend to speed infrequently but considerably – this is also indicated by the median of zero speeding percentage per segment.

From the obtained descriptive statistics, it can be discerned that the adjustment of pass count per segment was a necessity in the algorithm since initial results showed considerable double-counting in segments where several road axes travel in parallel. This led to considerable double-counting in trip sections, inflating the calculated pass count unrealistically.

Lastly, is worth noting that driver speed, speeding and mobile use parameters are calculated as a proof-of-concept at this stage and was not inserted in the models for urban networks, as explained previously.

4.2 Test area – Omonoia

4.2.1 Initial study area examination

The overall process for the test area mirrors the one followed for the training area, in order to eliminate any inconsistencies between the two datasets. The OSM data export was conducted via the same API by defining a different area on the map. As before, an initial visual exploratory check was conducted to determine any discrepancies between the map image and the raw OSM data import; no discrepancies were detected. The map with the axes of the imported segments (in green) is shown on Figure 4-13 Figure 4-13 for the selected Omonoia area. Walkways and similar footpaths have been removed from the segments, hence they will not be appearing on the processed maps apart from the baseline.

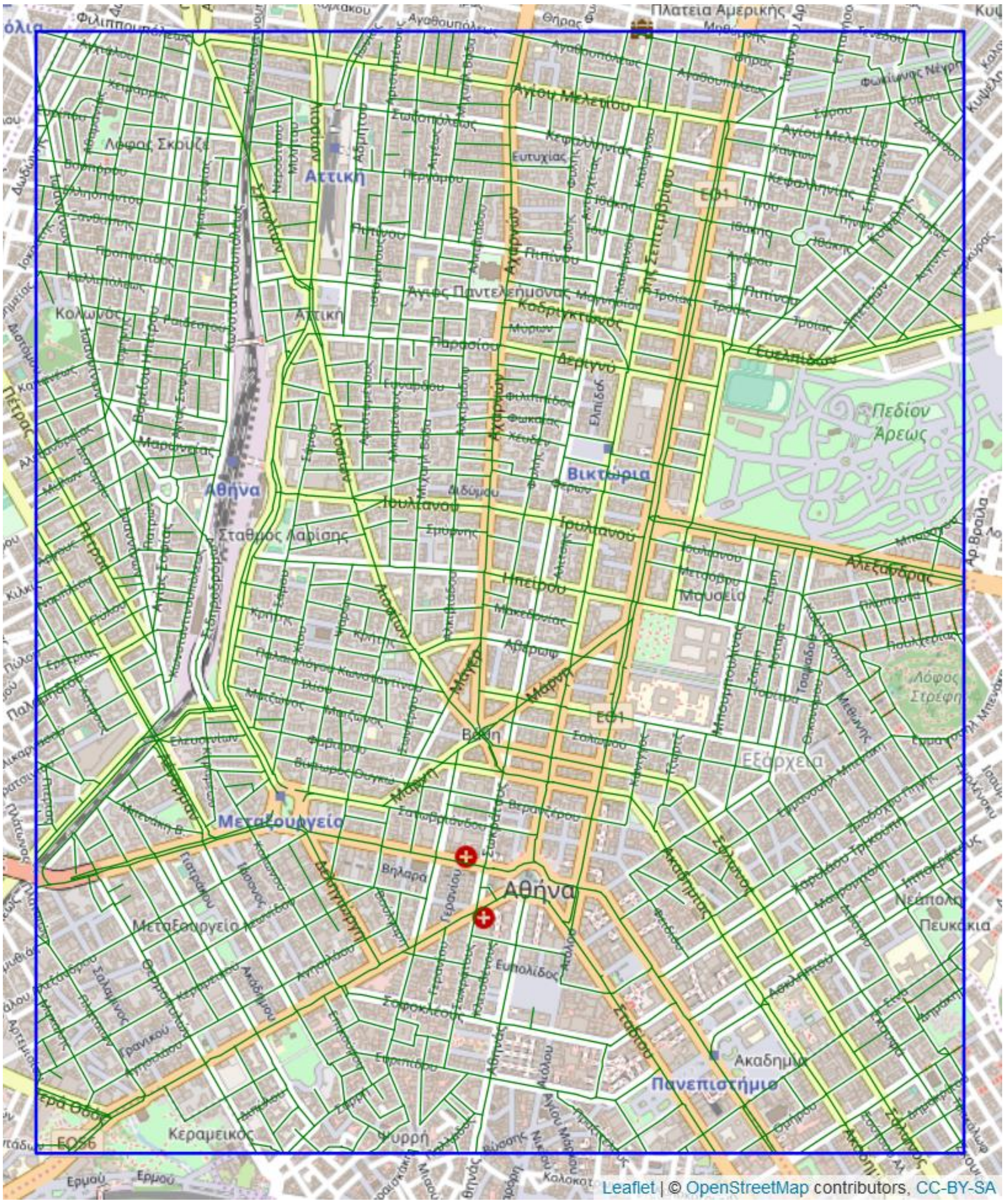


Figure 4-13: Omonoia road segments following import from OSM and removal of footways

Having exported the raw OSM data, and after enhancing them with SRTM data, the processing phase was ready to begin for the test area.

4.2.2 Road geometry characteristics

The initial raw data for Omonoia show that in the examined area 1315 segments (ways in OSM terms) are initially included which are consisted of 6115 nodes; relation data were again not exported since they are not needed in the analyses in order to reduce processing times. The exclusion of 78 footway segments followed, leading to 1237 remaining road segments. This is a higher number of roads compared to the training area, though the two areas have comparable surfaces (5.247 km^2 for Chalandri and 5.754 km^2 for Omonoia). The oldest and more central district of Omonoia features somewhat smaller roads, more densely packed in a cellular fashion.

Footways are also more numerous in the Omonoia area; this is expected due to the fact that the region is historically significant for Athens, and there have been no significant changes from its present form in decades. There were no other exclusions necessary, for instance due to missing data. All remaining road segments are utilized. The various road type frequencies of the segments per road direction and lane number appear on **Table 4-6**.

Table 4-6: Road type frequencies in Omonoia area

Road type	**Road direction**									
	One-way segments				Two-way segments				**Total**	**Total [%]**
	Lanes: 1	Lanes: 2	Lanes: 3	Lanes: 4	Lanes: 1	Lanes: 2	Lanes: 3	Lanes: 4		
Primary	5	66	54	13	0	11	0	14	163	13.18 %
Secondary	37	105	2	16	3	11	0	7	181	14.63 %
Tertiary	177	58	0	1	40	6	0	0	282	22.80 %
Residential	576	6	0	0	29	0	0	0	611	49.39 %
Total	795	235	56	30	72	28	0	21	**1237**	**100.00 %**

Once again, in the test area, residential one-way, one-lane roads comprise the majority of the segment sample. There is a relative increase in primary and tertiary roads from the training area, again attributed to the importance of the city center.

Descriptive values for the obtained geometric and road network characteristics appear on **Table 4-7**. As a reminder, gradient and neighborhood complexity are dimensionless quantities, and negative gradient values refer to downhill slopes.

Table 4-7: Descriptive statistics for the obtained geometric characteristics for road segments in Omonoia area

Geometric characteristics	**Descriptive statistics**						
	Average	Min	Median	Max	St. Dev.	Skewness	Kurtosis
Segment Length [m]	121.1647	1.5579	82.538	1085.0278	120.5802	2.4037	8.7001
Curvature [m^{-1}]	0.0024	0.0000	0.0000	0.0924	0.0080	5.3844	36.0892
Gradient [–]	0.0014	-0.4399	0.0012	0.2618	0.0646	-0.1310	3.1892
Neighborhood Complexity [–]	4.9633	2.8331	4.9970	5.6779	0.3341	-1.1335	3.9069

All geometric characteristics exhibit significant fluctuations and dispersion of values. The geometric results are comparable to the ones in the training area in orders of magnitude, but do have different values.

Additional interesting trends that have also been observed in the training area is that the standard deviation of segment length is very close to the average denoting borderline high dispersion of the variable. Furthermore, the skewness of gradient denotes a slightly more asymmetrical distribution than in the training area. The skewness of neighborhood complexity remains expectedly negative.

There is higher kurtosis in the segment length and curvature variables, suggesting heavy-tailed distributions more frequent outliers with higher values in these variables; also compared to the training area. Gradient and neighborhood complexity have kurtosis values somewhat closer to 3 featured by univariate normal distributions.

The presence of traffic lights and pedestrian crossings detected from tags in the OSM data is provided on **Table 4-8** for the test area. In total, 319 segments featuring traffic lights and 317 segments featuring pedestrian crossings were detected. It is possible that segments feature both network characteristics.

Table 4-8: Fixed network characteristics in Omonoia area

Road type	**Network characteristics**							
	Presence of traffic lights				**Presence of pedestrian crossing**			
	No		Yes		No		Yes	
Primary	133	10.75 %	30	2.43 %	136	10.99 %	27	2.18 %
Secondary	141	11.40 %	40	3.23 %	141	11.40 %	40	3.23 %
Tertiary	207	16.73 %	75	6.06 %	215	17.38 %	67	5.42 %
Residential	437	35.33 %	174	14.07 %	428	34.60 %	183	14.79 %
Total	918	74.21 %	319	25.79 %	920	74.37 %	317	25.63 %
Grand Total			**1237**	**100.00 %**			**1237**	**100.00 %**

Several maps and heatmaps can be produced from the above characteristics, similarly to the training area. Figure 4-14 to 4-16 provide an intuitive presentation of network features mirroring those extracted for the training area indicatively.

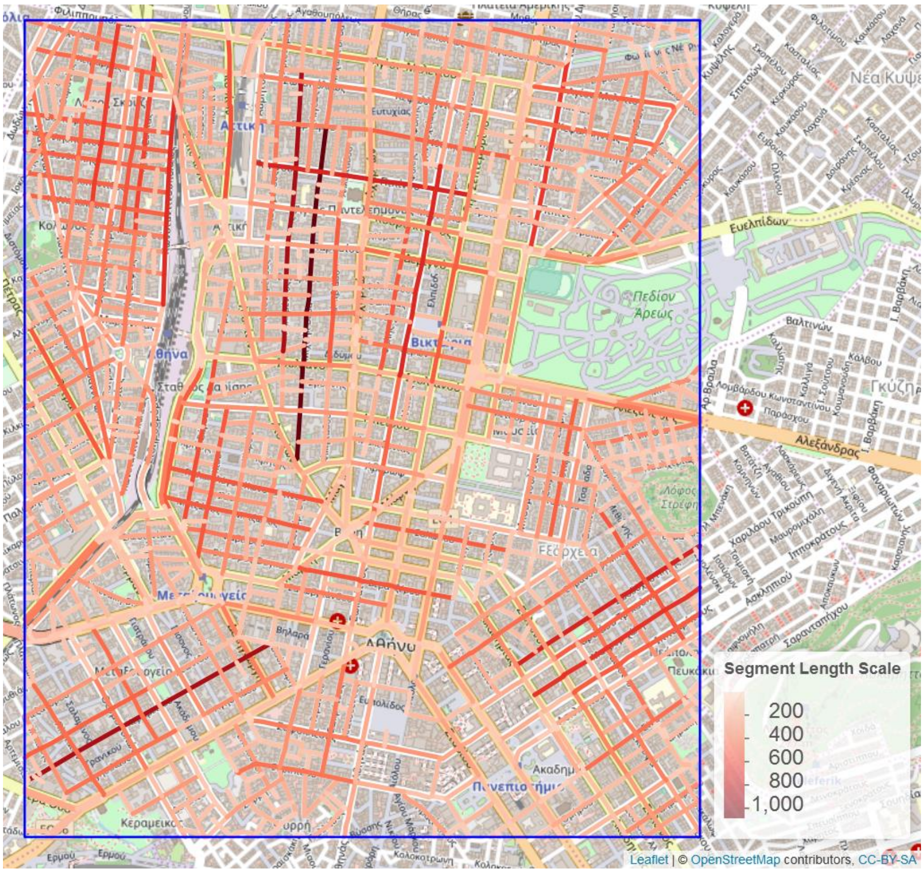


Figure 4-14: Heatmap of road segment lengths in Omonoia area

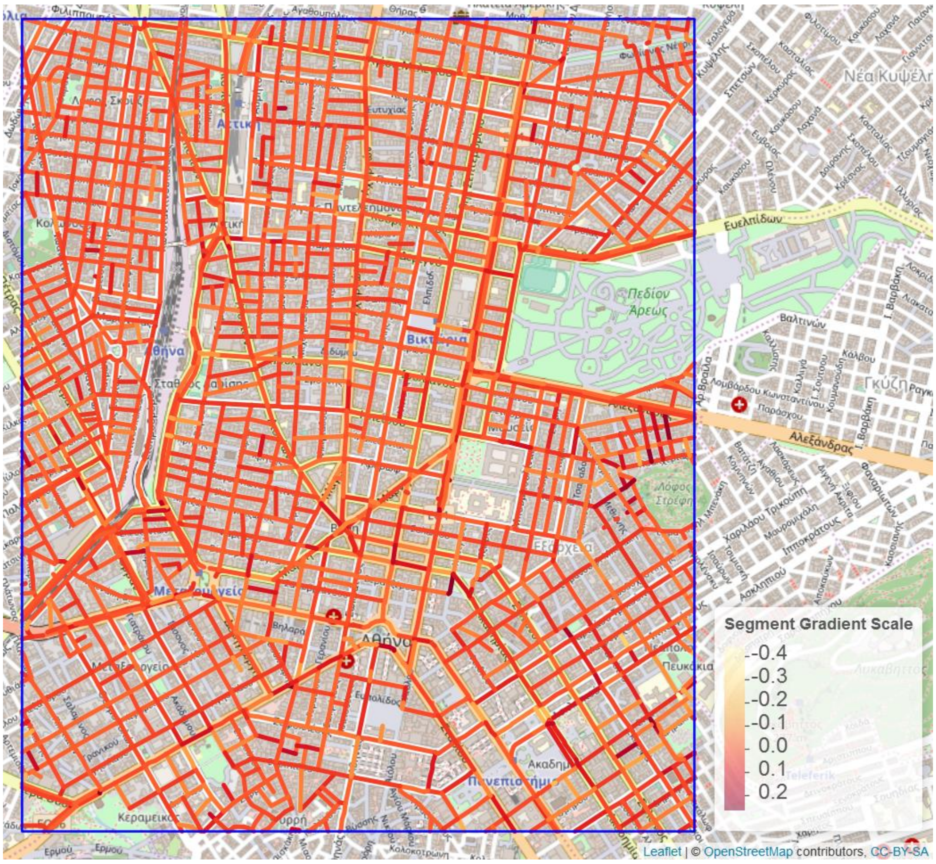


Figure 4-15: Heatmap of road segment gradients in Omonoia area

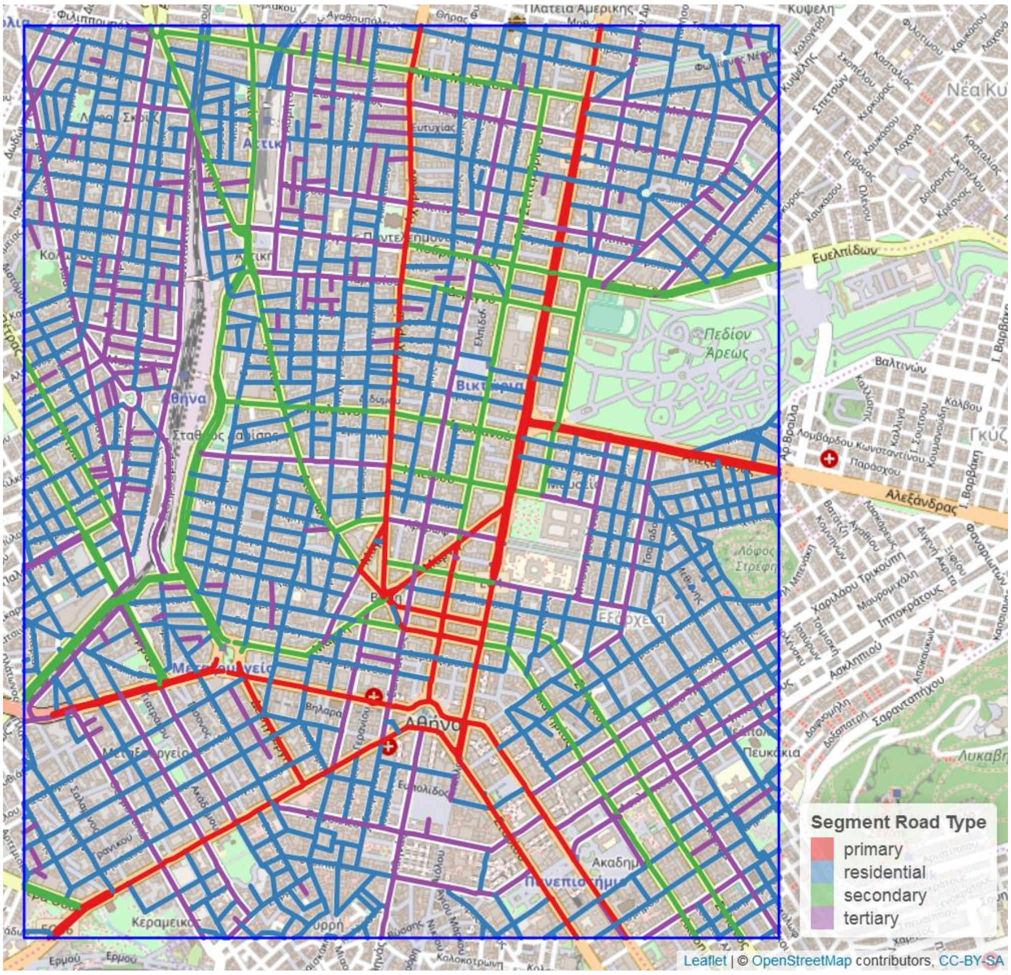


Figure 4-16: Mapping of road segment types in Omonoia area

4.2.3 Large-scale naturalistic driving data exploration

The dataset of naturalistic trip data is examined after being obtained from the OSeven application for the test area. Data corresponding to a period of two months matching with Chalandri area were provided, namely from 01-10-2019 to 29-11-2019. During that period, 2615 trips were provided from 257 individual drivers in an anonymous format. These trips were not necessarily confined in the Omonoia test area; some had origins and/or destinations on road segments outside the borders depicted on Figure 4-13. However, they were all cropped so that only the length of each trip that fell into the test area was considered – the remaining information was discarded.

Before processing, the trips had an average duration of 1354 seconds (or 22.56 minutes). This resulted in another big data trip file with 3,542,131 rows. After cropping the trips, trip duration was reduced to a mean of 369 seconds, for a file of 964,693 entries, again expected from trips within a portion of the city center. The histogram of trip durations is shown on Figure 4-17.

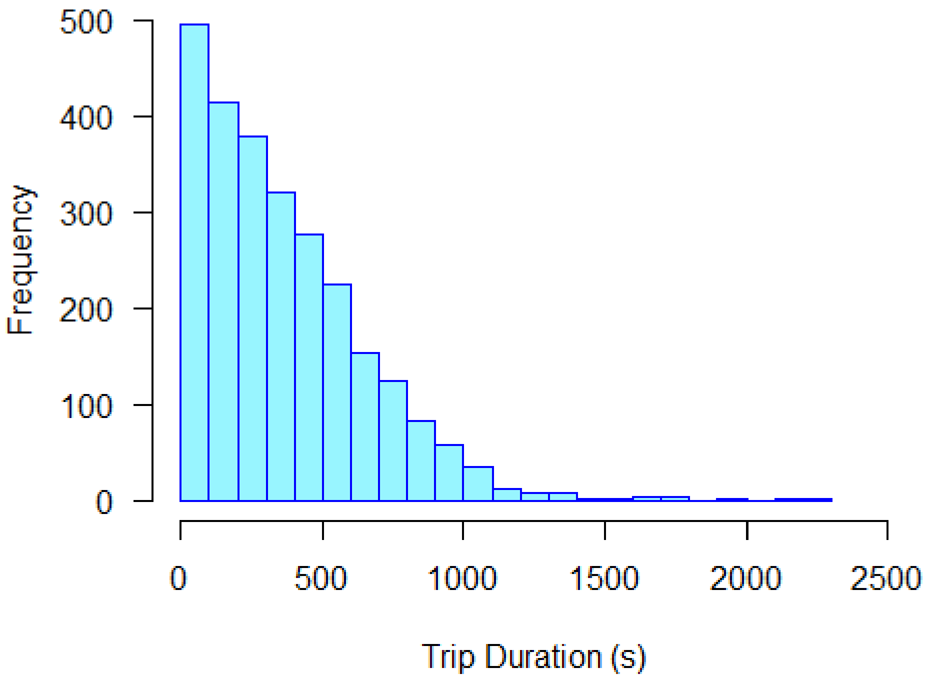


Figure 4-17: Histogram of trip durations in Omonoia area

There seems to be an increase in trips of the lowest duration compared to the Chalandri area. This can be attributed to the fact that Omonoia is often an intermediate section of trips passing through the city center. Smaller intermediate portions of these trips, especially at the edge of the area, may have been captured by the process. They feature smaller duration compared to the origin/destination trips – which include searching for parking – found in the training dataset.

In these trips, a number of harsh events have occurred and were recorded alongside normal driving conditions, consisting of 1036 harsh braking events and 938 harsh acceleration events. As mentioned previously, events are examined uniformly regardless of intensity. So far, the produced metrics and quantities for the training and test datasets appear to be very similar, which is a desirable intermediate step.

The numbers of harsh events per intensity category appear on **Table 4-9**.

Table 4-9: Harsh events per intensity category in Omonoia area

Event intensity category	Harsh events			
	Harsh brakings		Harsh accelerations	
1 – mild	528	50.97%	438	46.70%
2 – modest	350	33.78%	279	29.74%
3 – severe	158	15.25%	221	23.56%
Total	1036	100.00 %	938	100.00 %

Harsh events are then projected on the map of the test area, as shown on Figure 4-18 for harsh brakings (hb) and on Figure 4-19 for harsh accelerations (ha).

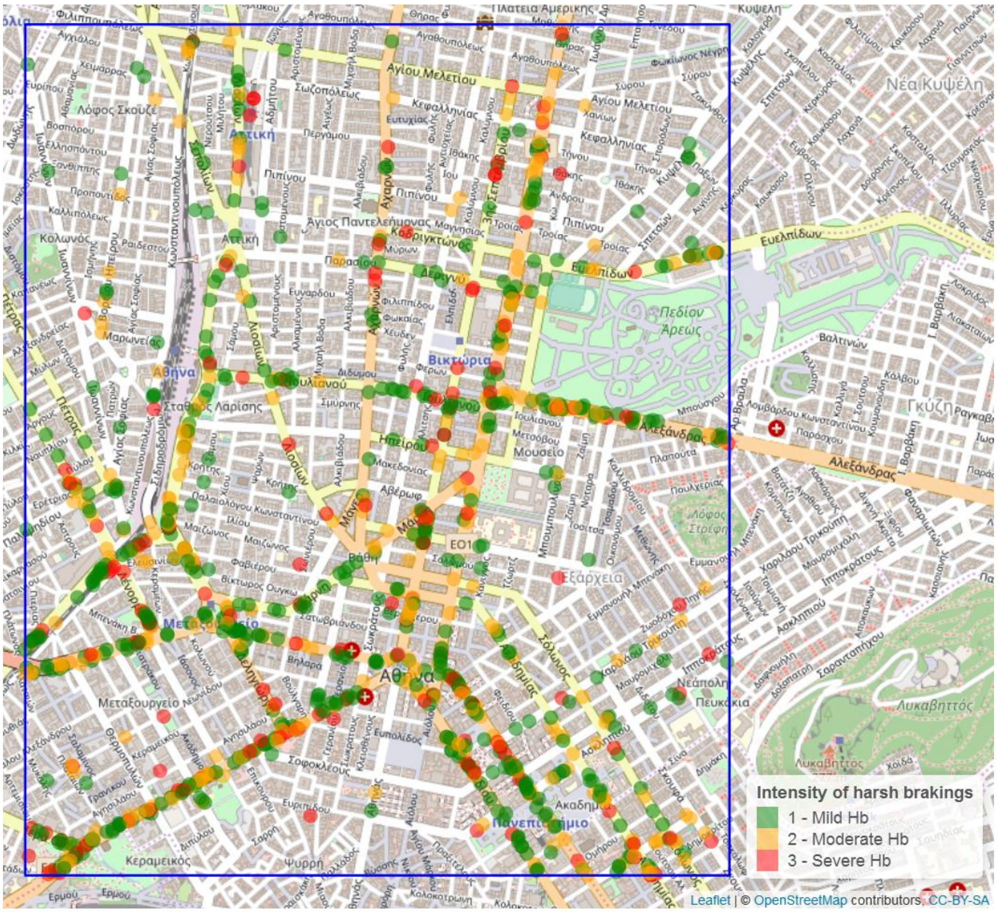


Figure 4-18: Harsh braking events in Omonoia area

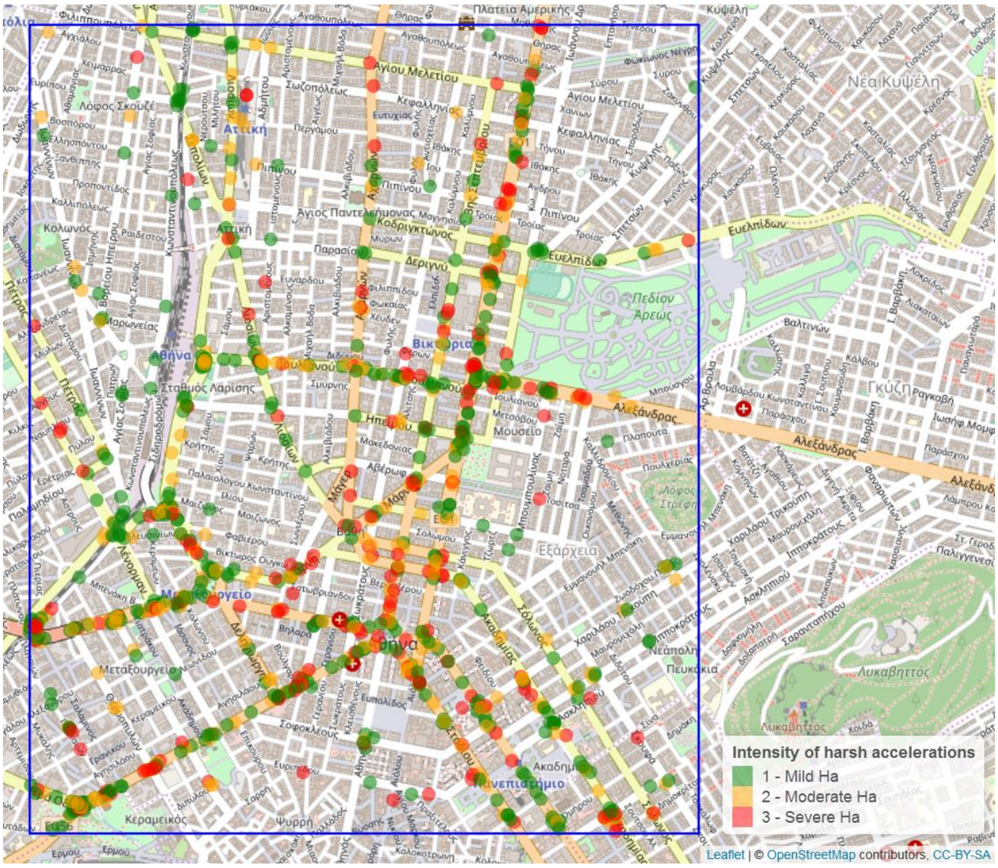


Figure 4-19: Harsh acceleration events in Omonoia area

Similar to the training area, harsh events tend to occur on more significant roads, namely primary and secondary OSM road categories. The Omonoia area features a more uniform distribution of road categories throughout its entire surface. They do appear to be some roads of smaller categories with a higher event concentration, in the southeast part of the testing area in particular.

4.2.4 Map-matching results

Map-matching was conducted this time for 964,693 trip-seconds in the test area naturalistic driving big dataset, assigning each trip-second to the nearest OSM road segment. The runtime of the map-matching for Omonoia area was 6 hours and 52 minutes on a server-level computer. Descriptive statistics for the obtained parameters are shown on **Table 4-10**. Parameters with an asterisk (*) are reported only for segments that had non-zero trips, since they are calculated with the adjusted pass count (trips per segment) or pass seconds per segment. The segments with non-zero trips are 1066 out of the total 1237, with 171 segments receiving no trips from the 257 drivers during the data collection period.

The descriptive statistics provide glimpses in the parameter values of the test area. The values are in similar magnitudes to the ones in the training dataset, and consist of frequency counts, rates, and percentages.

Table 4-10: Descriptive statistics for road segments in Omonoia area after map-matching

Segment characteristics from naturalistic driving	**Descriptive statistics**						
	Average	Min	Median	Max	St. Dev.	Skewness	Kurtosis
Pass count per segment (assigned)	91.3888	0	39	1866	145.4430	4.4864	37.0040
Adjusted pass count per segment	45.7939	0	18	437	65.8621	2.1767	5.1484
Harsh brakings per segment	0.8375	0	0	23	2.2041	4.3907	24.6982
Harsh braking rate per segment *	0.0002	0.0000	0.0000	0.0282	0.0012	16.3138	319.7128
Harsh accelerations per segment	0.7583	0	0	17	1.7973	3.9487	20.4729
Harsh acceleration rate per segment *	0.0002	0.0000	0.0000	0.03250	0.0012	19.5639	496.5927
Pass seconds per segment	779.4082	0	172	12976	1437.1352	3.2628	13.8098
Mobile use seconds per segment	29.9313	0	6	884	60.5026	4.7988	41.3488
Mobile use percentage per segment *	4.98 %	0.00 %	2.94 %	100.00 %	8.52 %	5.5713	44.6966
Speeding seconds per segment	3.9660	0	0	307	19.3752	9.8063	121.7070
Speeding percentage per segment *	0.60 %	0.00 %	0.00 %	100.00 %	3.44 %	21.8807	603.0301
Average driver speed per segment	16.0257	0.0000	14.8939	71.7500	9.3012	0.7492	1.5123

The majority of road segments in the training area were assigned at least one trip, namely 1066 out of 1237 or 86.18% of the total. Conversely, 171 segments did not have any trips, amounting to 13.82% of the total. Continuing the trend from the training area, there is good spatial coverage of the training area, although it is not uniform, as evidenced by the high standard deviation of the (adjusted) pass count per segment. The adjusted pass count assignment is depicted in the heatmap of Figure 4-20.

Driver aggressiveness remains an infrequent phenomenon overall, as averages of harsh event frequencies are low, and their means are zero. Once again, harsh event rates appear to follow highly asymmetrical distributions, which show considerable kurtosis. Therefore there are hints of several road segments with high outliers of driver behavior. The respective heatmaps of event frequencies are shown on Figure 4-21 for harsh brakings and Figure 4-22 for harsh accelerations.

Kurtosis values remain large overall for almost all variables apart from average driver speed, suggesting more outliers with more diverging values in the parameters. The differences in the distribution of average speed of the users are quite notable. Firstly, the distribution of variable values across the minimum – average – maximum range are lower, and secondly, the skewness and kurtosis values are significantly lower than in the training area. The descriptive statistics disclose a test area with different speed profiles. In the test area average driving speeds are limited, and more symmetrically spread as noted by the smaller skewness. Speeding outliers are fewer and much less pronounced, as the distribution is 'light-tailed' (platykurtic).

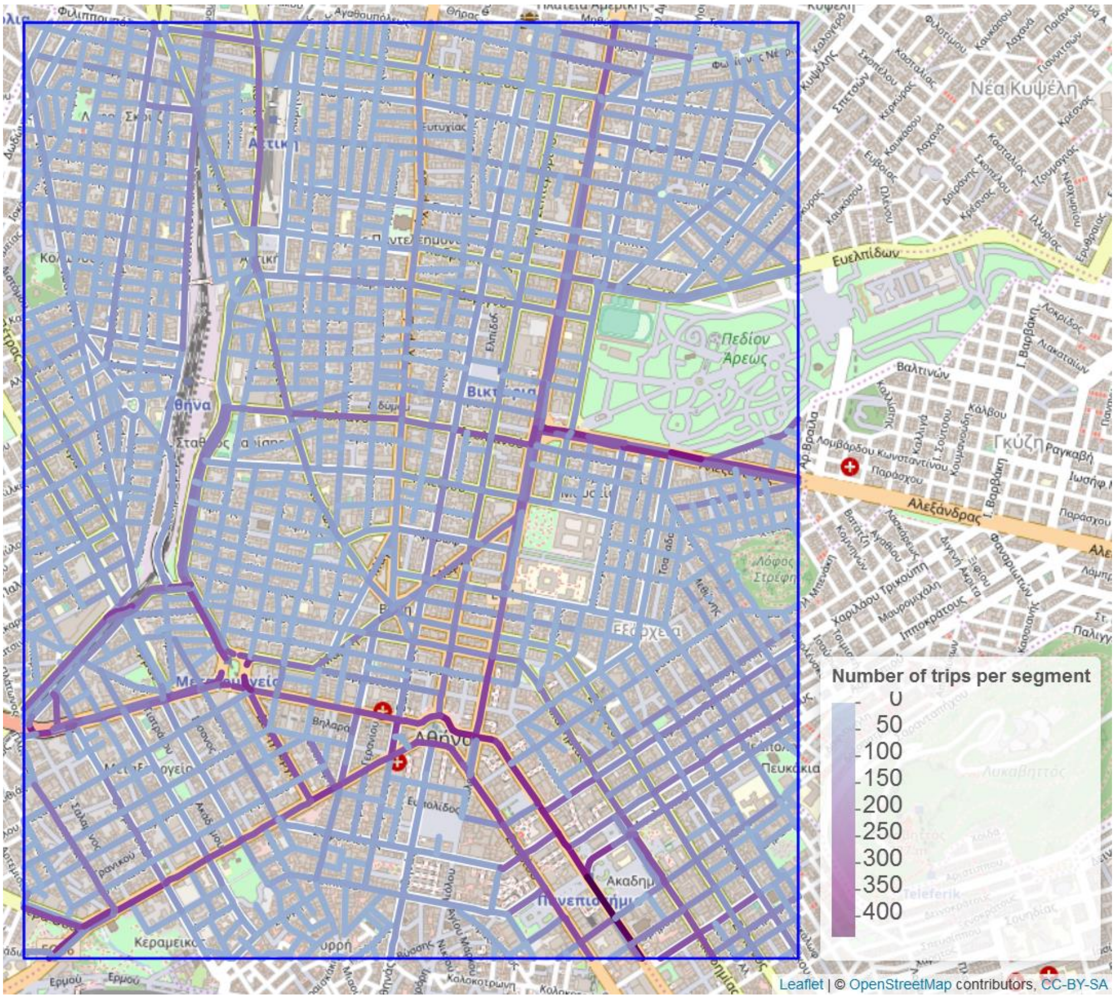

Number of trips per segment
0
50
100
150
200
250
300
350
400
Leaflet | © OpenStreetMap contributors, CC-BY-SA

Figure 4-20: Heatmap of adjusted pass counts of segments in Omonoia area

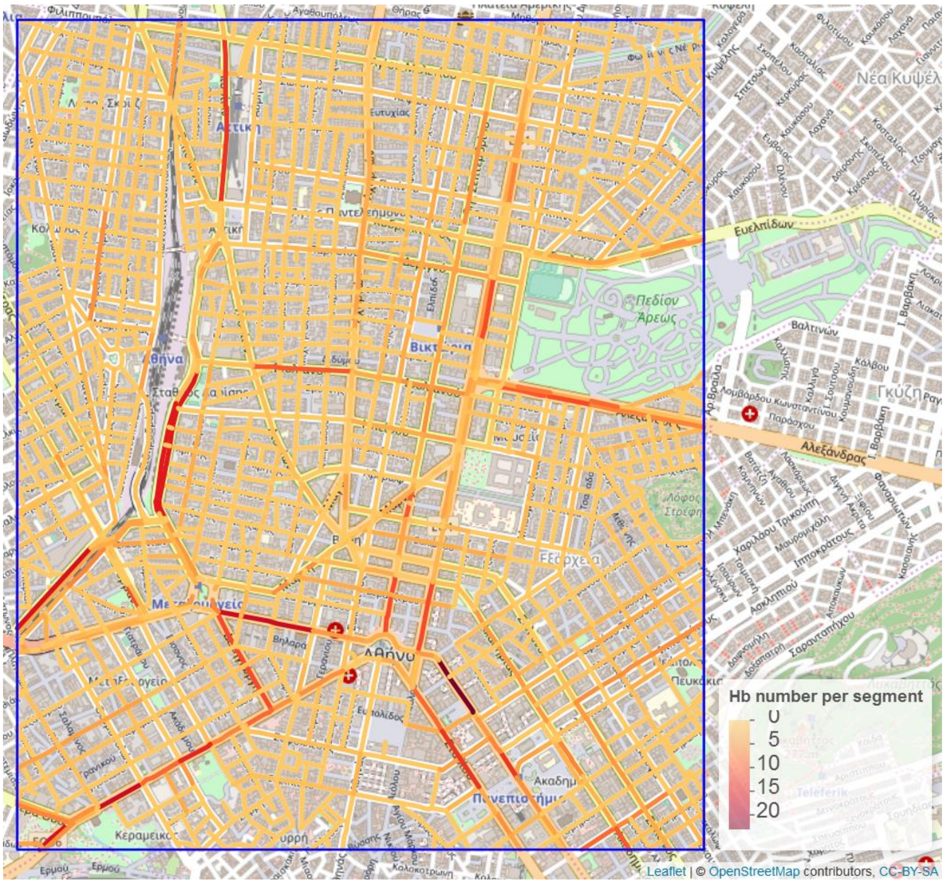


Figure 4-21: Heatmap of harsh braking frequencies of road segments in Omonoia area

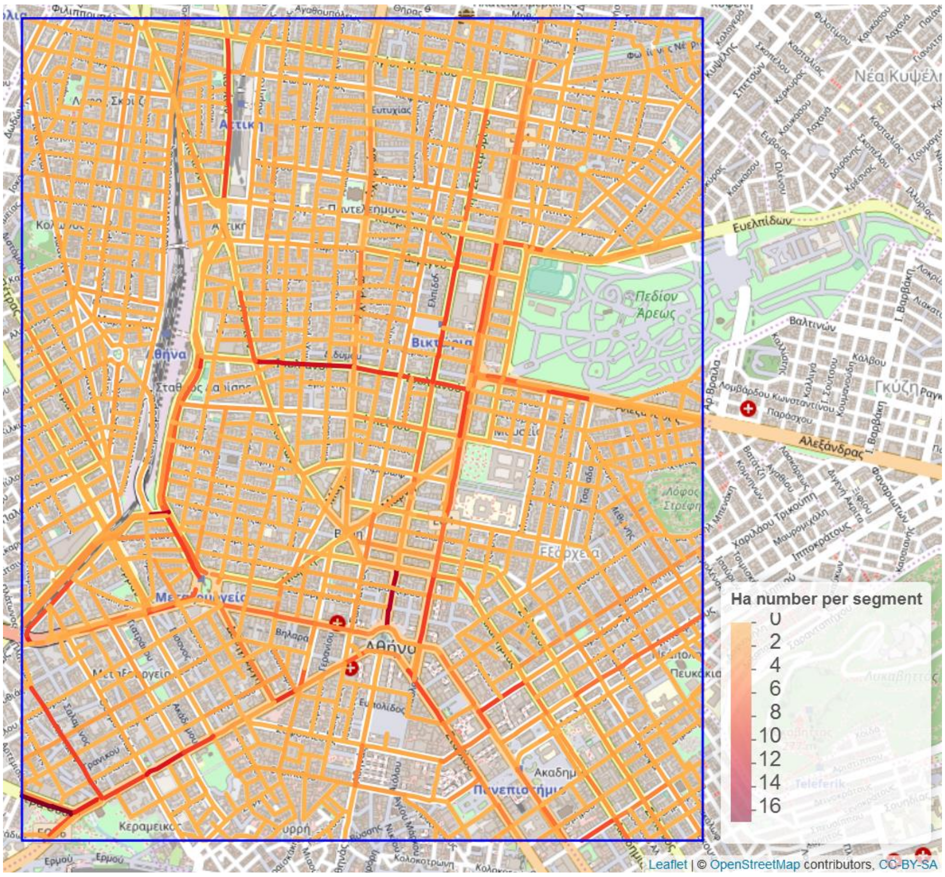


Figure 4-22: Heatmap of harsh acceleration frequencies of road segments in Omonoia area

4.3 Spatial data frame samples

In geocoding and geospatial processing terms, the objects used to represent road segments in this dissertation are composite data structures comprising (i) geometrical design features and (ii) corresponding data frames where additional information not directly referring to geometry is stored, such as harsh event number, for instance. These objects can be transformed to a variety of forms, depending largely on the applications/packages used to manipulate them, such as OpenGIS Simple Features Reference Objects (known as OGR objects), Geospatial Data Abstraction Library Objects (known as GDAL objects) or Spatial Lines Data Frame Objects (or S4 objects in R-studio).

Apart from the intuitive presentation in maps, the data frames themselves were also perused for the extraction of descriptive statistics and in preparation for the main analyses. The structure of the data frames is important for performing spatial analyses and interpreting the results, therefore a sample of each data frame is presented here in order to showcase that structure and to add more context to the descriptive statistics already provided.

A sample of 6 rows from the data frame component of each of the training and test areas is provided below to convey a general overview of the data structure after the phases of data input, geometric characteristic derivation, map-matching and processing have been complete. Following the convention of this doctoral dissertation, each row represents a different road segment based on OSM segmentation. **Table 4-11** is extracted from the training area and **Table 4-12** is extracted from the test area.

In essence, all of the collected data are combined and allocated to parameters on the 869 training data frame rows and test 1237 data frame rows representing road segments. This is a considerable merit and convenience of the adopted approach: the computationally demanding stage is data pre-processing and processing. After datasets with up to millions of lines have been fused and allocated to more manageable scales at hundreds of lines, model training (calibration) is a faster process and can allow for the exploration of several model configurations.

Table 4-11: Spatial data frame sample from Chalandri area

OSM Segment id	**Spatial data frame attributes**																					
	Type (OSM)	Road type	Road direction	Lanes	Lat. Nom.	Lon. Nom.	Segment Lengths	Curvature	Gradient	Neighb. Complex.	Traffic lights	Ped. crossing	Ha. No.	Hb. No.	Trip No.	Pass count	Mob. Use. Secs	Speed. Secs	Avg. driver speed	Ha rate	Hb rate	Mob. use perc.	Speed. perc.
106931615	way	tertiary	One-way	1	38.0230	23.8111	103.0574	0.0000	0.0129	4.5747	0	0	1	0	7	25	0	0	21.16	0.0014	0.0000	0.000	0.000
106931603	way	tertiary	One-way	1	38.0231	23.8151	43.64057	0.0000	-0.0198	3.7136	1	0	0	2	61	251	4	0	35.59	0.0000	0.0008	0.016	0.000
753871425	way	residential	One-way	1	38.0221	23.8153	21.8534	0.0000	-0.0083	4.0775	0	0	0	0	31	205	1	0	7.14	0.0000	0.0000	0.005	0.000
106931607	way	residential	One-way	1	38.0246	23.8105	269.5592	0.0002	0.0017	4.6444	0	1	0	2	25	704	20	0	14.86	0.0000	0.0003	0.028	0.000
32118803	way	residential	One-way	1	38.0209	23.8135	83.2370	0.0000	0.0221	4.6052	0	0	0	0	2	51	0	0	28.18	0.0000	0.0000	0.000	0.000
98782026	way	residential	One-way	1	38.0211	23.8130	115.4324	0.0338	0.0393	4.7095	0	0	0	0	3	9	0	0	21.87	0.0000	0.0000	0.000	0.000

Table 4-12: Spatial data frame sample from Omonoia area

OSM Segment id	Spatial data frame attributes																				
	Type (OSM)	Road type	Road direction	Lanes	Lat. Nom.	Lon. Nom.	Segment Lengths	Curvature	Gradient	Neighb. Complex.	Traffic lights	Ped. crossing	Ha. No.	Hb. No.	Trip No.	Pass count	Mob. Use. Secs	Speed. Secs	Avg. driver speed	Ha rate	Hb rate	Mob. use perc.	Speed. perc.
5656493	way	tertiary	One-way	1	37.9793	23.7338	90.2450	0.0212	0.0670	4.5218	0	1	2	0	42	881	59	28	29.96	0.0005	0.0000	0.032	0.067
10741722	way	tertiary	One-way	2	37.9854	23.7258	158.1550	0.0001	0.0128	5.5568	1	1	0	0	43	351	12	0	9.245	0.0000	0.0000	0.000	0.034
10741726	way	residential	One-way	1	37.9863	23.7233	328.2705	0.0000	-0.0209	5.0562	0	0	0	0	8	241	12	1	12.36	0.0000	0.0000	0.004	0.050
10741731	way	residential	One-way	1	37.9852	23.7270	367.5570	0.0000	0.0170	5.3982	0	1	1	0	75	1599	61	1	11.69	0.0000	0.0000	0.001	0.038
10741733	way	primary	One-way	3	37.9855	23.7284	193.7320	0.0087	-0.0397	5.4205	1	0	14	11	130	4067	300	2	15.2	0.0006	0.0004	0.000	0.074
10741734	way	primary	One-way	2	37.9872	23.7267	8.1501	0.0000	0.0222	5.5413	1	0	0	0	56	67	4	0	21.07	0.0000	0.0000	0.000	0.060

5 Urban road network segment analyses

In this section, spatial analysis results are presented for urban networks. Initially, exploratory spatial analyses are conducted, in the form of global and local Moran's *I* coefficient calculations, empirical variogram plotting and theoretical variogram fitting. Subsequently, fitted Geographically Weighted Poisson Regression (GWPR) models, Conditional Autoregressive Prior (CAR) models and Extreme Gradient Boosting (XGBoost) machine learning methods with and without spatial cross-validation are presented and their results are elaborated upon. All processes are conducted both for harsh braking and for harsh acceleration event frequencies, and refer to road segments as units of analysis. Analysis results are based on the final road segment datasets for urban networks, samples of which appear in Section 4.3.

5.1 Exploratory spatial analysis

5.1.1 Global Moran's *I*

As explained in Section 3.2.2.1, Moran's *I* coefficient is the most widely used exploratory metric for the detection of spatial dependence in the data. Following Bivand et al. (2008), global Moran's *I* calculations are conducted for harsh braking and harsh acceleration frequencies in urban networks. Apart from presentation of the results, the current section also serves as an exploration of the malleability and flexibility of the value of Moran's *I* coefficients depending on the weighting system used.

5.1.1.1 Distance-based weighting

As a first step, the entire training area is considered globally. For each road segment, weights of all the other segments are assigned based on the distance of their centroids from the examined segment centroid. Afterwards, weights are row-standardized so that their sum equals to 1 for each segment. The resulting weighting scheme is used to calculate global Moran's *I*; results appear on **Table 5-1** for Chalandri area.

Table 5-1: Global Moran's I in Chalandri area with distance-based weighting

Global Moran's I	Training area			
	Coefficient value	Expectation	Variance	p-value
Harsh brakings	-0.0043	-0.0012	0.0000	$6.6 * 10^{-6}$
Harsh accelerations	-0.0071	-0.0012	0.0000	$< 2.2 * 10^{-16}$

Initially, it would appear that there is little overall spatial autocorrelation in harsh event frequencies when the entire area is considered. The coefficient values denote very close to zero spatial autocorrelation for both harsh braking and harsh acceleration frequencies in the training area, in other words, a random spatial distribution of events. In fact, the expected values were slightly higher, indicating that slightly more clustering was expected a priori from events in the training area than the outcome. Both coefficient values are statistically significant. As a note, the result for p-values of harsh accelerations is '< 2.2e-16' in R-studio, which is scientific notation denoting a number of $2.2 * 10^{-16}$, which is very close to zero for all practical and statistical applications.

Although events in the test area were primarily used for model accuracy assessment validation, it is fruitful to measure their spatial autocorrelation as well, for comparative purposes and as a verification of data quality. Results appear on **Table 5-2** for Omonoia area.

Table 5-2: Global Moran's I in Omonoia area with distance-based weighting

Global Moran's I	Training area			
	Coefficient value	Expectation	Variance	p-value
Harsh brakings	-0.0083	-0.0008	0.0000	$< 2.2 * 10^{-16}$
Harsh accelerations	-0.0092	-0.0008	0.0000	$< 2.2 * 10^{-16}$

The results are not noticeably different than those of the training area. Once again, the coefficient values denote overall spatial autocorrelation very close to zero for both harsh braking and harsh acceleration frequencies. Relatively to the expectation values, spatial autocorrelation is slightly negative, compared to the higher (more clustered) expectation values.

5.1.1.2 Nearest-neighbors weighting

The previous values of Moran's I show a largely random spatial distribution of event frequencies on segments. However, as per the aforementioned, Moran's I is heavily influenced by the weighting scheme which should be compatible with the underlying phenomenon under examination (Tiefelsdorf & Boots, 1997).

It is apparent from Figure 4-8, Figure 4-9, Figure 4-18 and Figure 4-19 that harsh event frequencies do form clusters to a certain degree, mainly on primary and secondary roads. Therefore, it can be argued that not all segments in an area contribute to harsh event frequencies for a specific segment. It appears reasonable to look for areas of clustering in the data by using the k most important neighbors for each segment.

The obvious question that arises is: what would be a proper value for k? In other words, which closest neighbors should be considered for each road segment centroid? As Harris (2013) explains, by examining the correlations of the values of the dependent variable(s) of a road segment with each of its nearest neighbors, it is possible to make a choice of the optimal number of neighbors based on a threshold of the correlation value. The maximum value of k is one-third of the total spatial observations. In other words, for each road segment the candidate neighbors can rise to up to one-third of the total segments in the urban network area (namely 290 road segments).

The correlation values are plotted on Figure 5-1 for harsh brakings to better visualize the effects of each neighbor. A simple trend line fitted with locally-weighted polynomial regression is also provided.

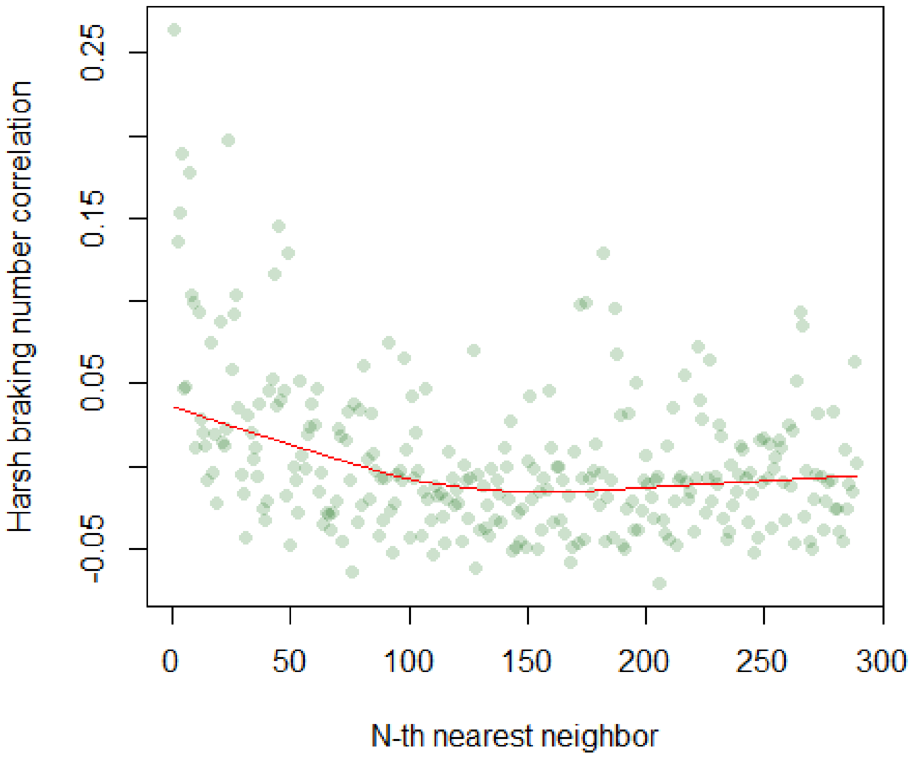


Figure 5-1: Harsh braking correlation values for N-nearest neighbors in Chalandri area

By examining the scatterplot for harsh brakings, it appears that the correlation values firstly drop below 0.1 after the 5th nearest neighbor, while they firstly drop below 0 after the 15th nearest neighbor. The respective scatterplot for harsh accelerations is shown on Figure 5-2.

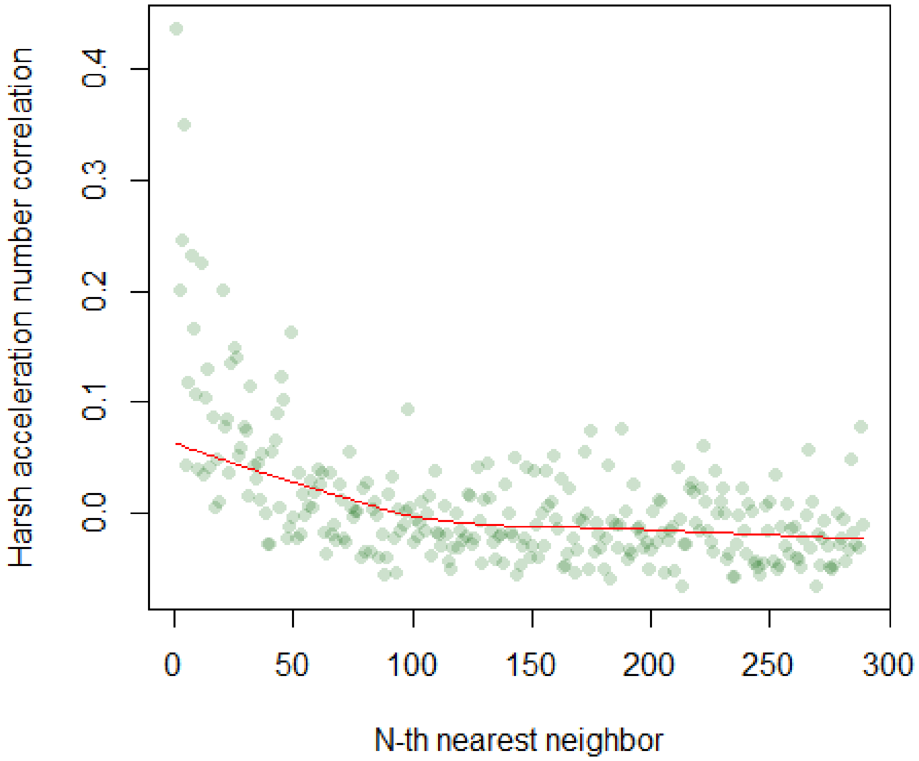


Figure 5-2: Harsh acceleration correlation values for N-nearest neighbors in Chalandri area

For harsh accelerations, correlation values firstly drop below 0.1 after the 5th nearest neighbor, while they firstly drop below 0 after the 39th nearest neighbor.

Much like p-value thresholds in traditional statistics, the choice of nearest neighbors remains largely subjective. The calculated values for global Moran's I are provided on **Table 5-3** for the training area.

Table 5-3: Global Moran's I in Chalandri area with nearest-neighbor calculations

Global Moran's I	Training area					
	Correlation threshold	k	Coefficient value	Expectation	Variance	p-value
Harsh brakings	0.0	15	0.0806	-0.0012	0.0001	$7.9 * 10^{-13}$
Harsh accelerations		39	0.0945	-0.0012	0.0000	$< 2.2 * 10^{-16}$
Harsh brakings	0.1	5	0.1421	-0.0012	0.0003	$1.7 * 10^{-13}$
Harsh accelerations		5	0.2206	-0.0012	0.0003	$< 2.2 * 10^{-16}$

The difference with the results of **Table 5-1** is considerable. While the examination of Moran's I is still global, meaning that the coefficients refer to the spatial autocorrelation of the entirety of the training area, some of the results are dramatically changed by taking the contributions of only the k-nearest neighbors into account. This time, Moran's I coefficients indicate more clustering than anticipated across all values. With the implementation of the stricter correlation threshold of 0.1, harsh accelerations start to approach positive spatial autocorrelation, denoting increased clustering in the data.

The calculated values for global Moran's I and k are provided on **Table 5-4** for the test area.

Table 5-4: Global Moran's I in Omonoia area with nearest-neighbor calculations

Global Moran's I	Test area					
	Correlation threshold	k	Coefficient value	Expectation	Variance	p-value
Harsh brakings	0.0	29	0.0889	-0.0008	0.0000	$< 2.2 * 10^{-16}$
Harsh accelerations		92	0.0672	-0.0008	0.0000	$< 2.2 * 10^{-16}$
Harsh brakings	0.1	2	0.1388	-0.0008	0.0007	$< 2.2 * 10^{-16}$
Harsh accelerations		5	0.1597	-0.0008	0.0003	$< 2.2 * 10^{-16}$

The correlations for Omonoia are denser and have lower values than Chalandri overall, dropping rapidly from 0.15 towards 0.05 and then towards 0, hence the smaller values of k when the correlation threshold is increased, and the larger numbers of nearest neighbors are required to reach 0 correlation.

The observed trends continue here as well; the contributions of only the k nearest neighbors influence Moran's I from the negative values towards the more positive values denoting slight clustering. Comparatively, the patterns remain between the two areas: Fewer nearest neighbors are required for brakings, and Moran's I coefficient values increase as the threshold increases.

The weighting scheme of contiguity, or adjacency, which is typically used for polygons in geographical analyses is not preferred in the current approach for two reasons. Firstly, the conversion of road segment axis lines to polygons would introduce an unknown amount of polygon overlap, with respective biases or errors, to the entire calculation. Secondly, relations of adjacency are not a completely accurate depiction of road networks. Adjacency cannot adequately describe the collectively higher local curvature of certain road clusters, such as those in Filothei area. Additionally, harsh braking events can occur on a road due to various effects from non-adjacent segments, for instance spillover effects or disruptive high beam lights during the night from more distant segments.

5.1.2 Local Moran's *I*

Global Moran's *I* can be disaggregated to create a localized measure of spatial autocorrelation (Anselin, 1995). Following Bivand et al. (2019), local Moran's *I* values are calculated based on the approaches of distance-based weighting and nearest-neighbors weighting.

5.1.2.1 Distance-based weighting

For the training area, local Moran's *I* results calculated based on the approach of distance-based weighting appear on **Table 5-5**.

Table 5-5: Local Moran's I in Chalandri area with distance-based weighting

Local Moran's I		Training area			
		Coefficient value	Expectation	Variance	p-value
Harsh brakings	Average	-0.0043	-0.0012	0.0003	–
	Min	-0.5805	-0.0012	0.0004	$< 2.2 * 10^{-16}$
	Median	0.0009	-0.0012	0.0002	0.8857
	Max	0.1040	-0.0012	0.0003	$1.1 * 10^{-11}$
	St. Dev.	0.0465	0.0000	–	–
Harsh accelerations	Average	-0.0071	-0.0012	0.0003	–
	Min	-0.7246	-0.0012	0.0004	$< 2.2 * 10^{-16}$
	Median	0.0004	-0.0012	0.0003	0.9335
	Max	0.1099	-0.0012	0.0003	$< 2.2 * 10^{-16}$
	St. Dev.	0.0566	0.0000	–	–

As can be observed, the average local Moran's *I* values correspond to the respective global ones of **Table 5-1** (i.e. those of the distance-based calculation). Overall, local Moran's *I* values vary considerably, denoting the occurrence of both – some – positive autocorrelation (clustering) or negative autocorrelation (dispersion) of events across road segments.

It is worth noting that there are several instances of values that are not statistically significant; perhaps due to low event observations on the segment under consideration and/or neighboring segments. Furthermore, certain segments on the edge of the study area might lack strong contributing contiguous segments due to reduced directions from which information from proximal segments is available.

Similar to network characteristics, distance-based (DB) local Moran's *I* values can be displayed in maps, as shown in Figure 5-3 and Figure 5-4.

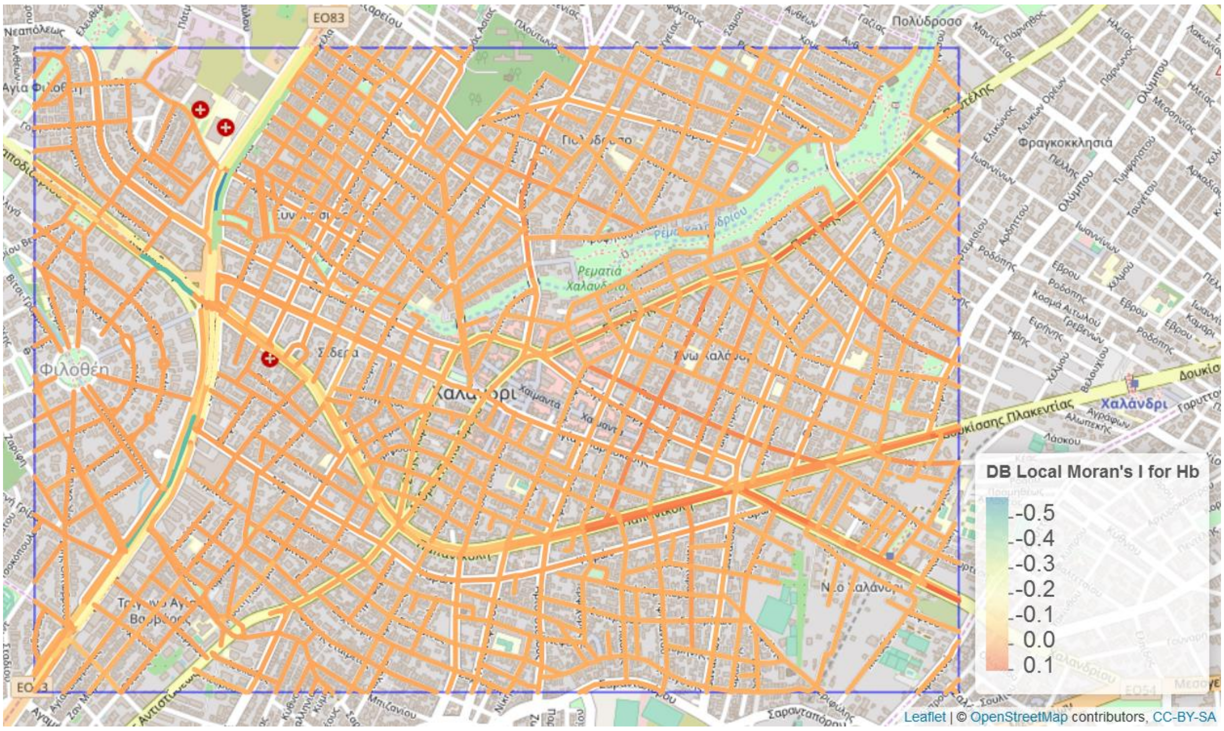

DB Local Moran's I for Hb
-0.5
-0.4
-0.3
-0.2
-0.1
0.0
0.1
Leaflet | © OpenStreetMap contributors, CC-BY-SA

Figure 5-3: Local Moran's I values in Chalandri area based on distance-based weighting for harsh braking events

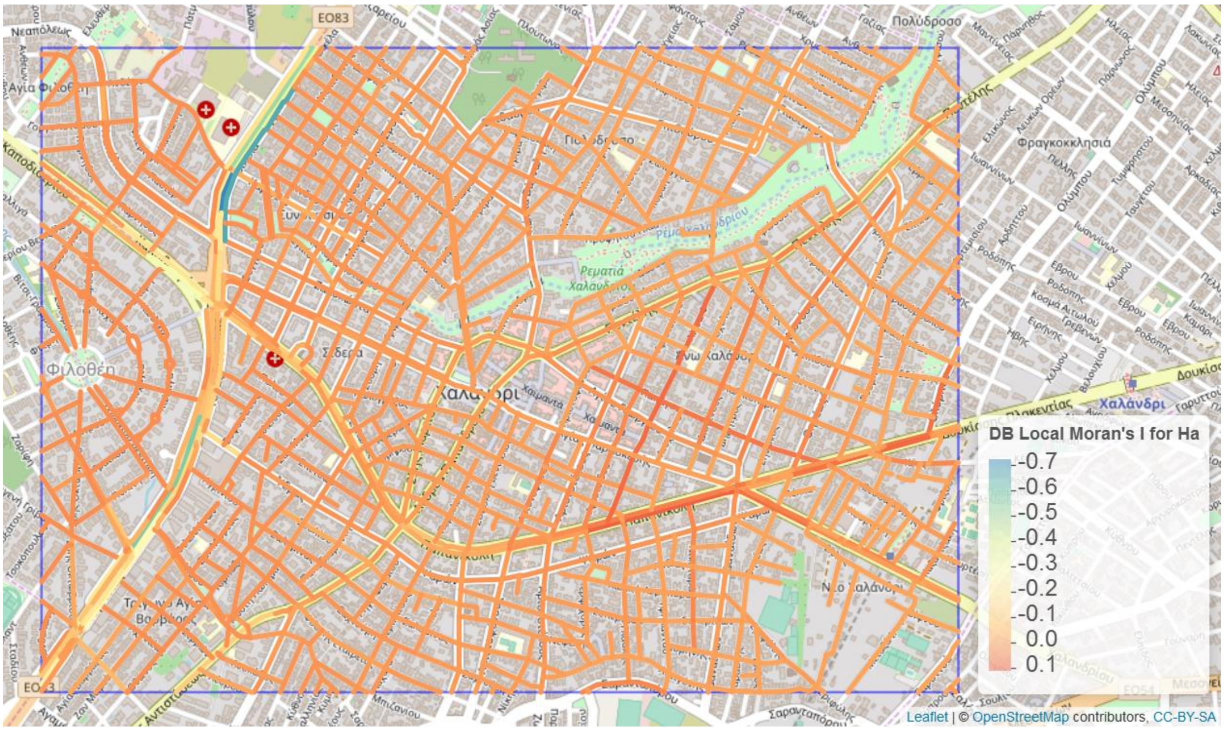

DB Local Moran's I for Ha
-0.7
-0.6
-0.5
-0.4
-0.3
-0.2
-0.1
0.0
0.1
Leaflet | © OpenStreetMap contributors, CC-BY-SA

Figure 5-4: Local Moran's I values in Chalandri area based on distance-based weighting for harsh acceleration events

An interesting finding from the previous figures is that segments with harsh event dispersion –denoted by the red side of the spectrum – belong to the larger road categories (segments belonging on primary and secondary roads). The values which slightly hint towards event clustering – denoted by the bluer side – also appear on the same roads for harsh brakings and harsh accelerations.

5.1.2.2 Nearest-neighbors weighting

Following a parallel reasoning to the one for global Moran's I, nearest-neighbor calculations can be conducted. This enables local Moran's I calculations by only taking into account only the neighboring segments, until correlations of harsh events drop lower than the specified correlation threshold.

Results calculated based on the approach of nearest-neighbors weighting with a correlation threshold of 0.0 appear on **Table 5-6**. The number of nearest-neighbors k were the same (15 and 39 respectively) for dropping below the correlation threshold, as the data remains unchanged. Once again, the average local Moran's I values correspond to the respective global ones of **Table 5-3** (i.e. those of the nearest-neighbors weighting calculation).

Table 5-6: Local Moran's I in Chalandri area with nearest-neighbors weighting

Local Moran's I		Training area			
		Coefficient value	Expectation	Variance	p-value
Harsh brakings [k=15]	Average	0.0806	-0.0012	0.0624	–
	Min	-0.9748	-0.0012	0.0624	0.0001
	Median	0.0399	-0.0012	0.0624	0.8694
	Max	10.1190	-0.0012	0.0624	$< 2.2 * 10^{-16}$
	St. Dev.	0.5919	0.0000	–	–
Harsh accelerations [k=39]	Average	0.0945	-0.0012	0.0231	–
	Min	-0.5063	-0.0012	0.0231	$< 2.2 * 10^{-16}$
	Median	0.0614	-0.0012	0.0231	0.7972
	Max	7.2223	-0.0012	0.0231	$< 2.2 * 10^{-16}$
	St. Dev.	0.5021	0.0000	–	–

Towards the maximum range, the values of local Moran's I are considerably larger than the conventional upper bound of 1. Several studies (e.g. Anselin, 1995; Waller and Gotway, 2004; Bivand et al., 2008) do report similarly ranging values. Anselin (1995) suggests the examination of the coefficient values and subsequent comparison with the mean and two-sigma rule, similar to outliers of a normal distribution. This comparison appears on Figure 5-5 for harsh brakings and on Figure 5-6 for harsh accelerations in the training area; the mean is denoted with a blue line, while the two-sigma limit is denoted towards the left with a red dotted line.

In these particular cases, it can be seen that most local Moran's I values are within the two-sigma rule. The remaining values gradually deviate from it at first, instead of single spikes. In addition to the previous, Anselin (1995) also cautions that this visual inspection does not constitute a test for outlier exclusion; indeed, global Moran's I values would drop by eliminating the segments with heightened local Moran's I values, and its statistical significance would drop as well. Therefore, segments with high local Moran's I values were not excluded on an outlier (two-sigma) basis. Rather, the results are considered to be an indication of strong spatial autocorrelations in specific segments, which are further incentive for the use of spatial models to study the phenomena of harsh events. Comparable results appear for local Moran's I for the test area, which are not shown here for brevity.

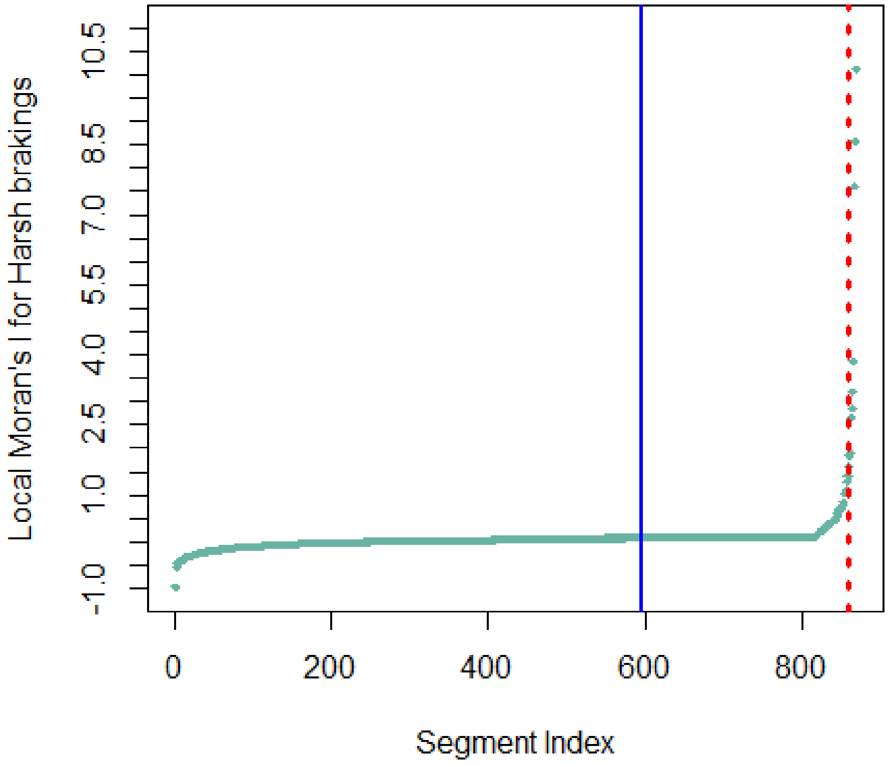


Figure 5-5: Local Moran's I values in Chalandri area based on nearest-neighbors weighting for harsh braking events

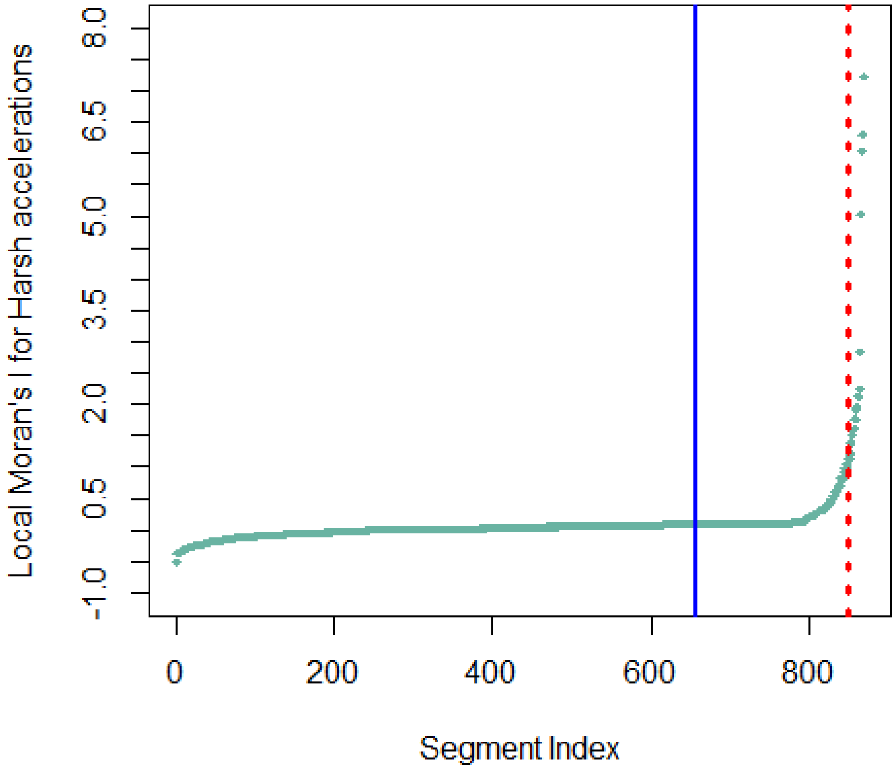


Figure 5-6: Local Moran's I values in Chalandri area based on nearest-neighbors weighting for harsh acceleration events

As per the previous, k nearest-neighbors (kNN) based local Moran's I values can be displayed in maps, as shown in Figure 5-7 and Figure 5-8.

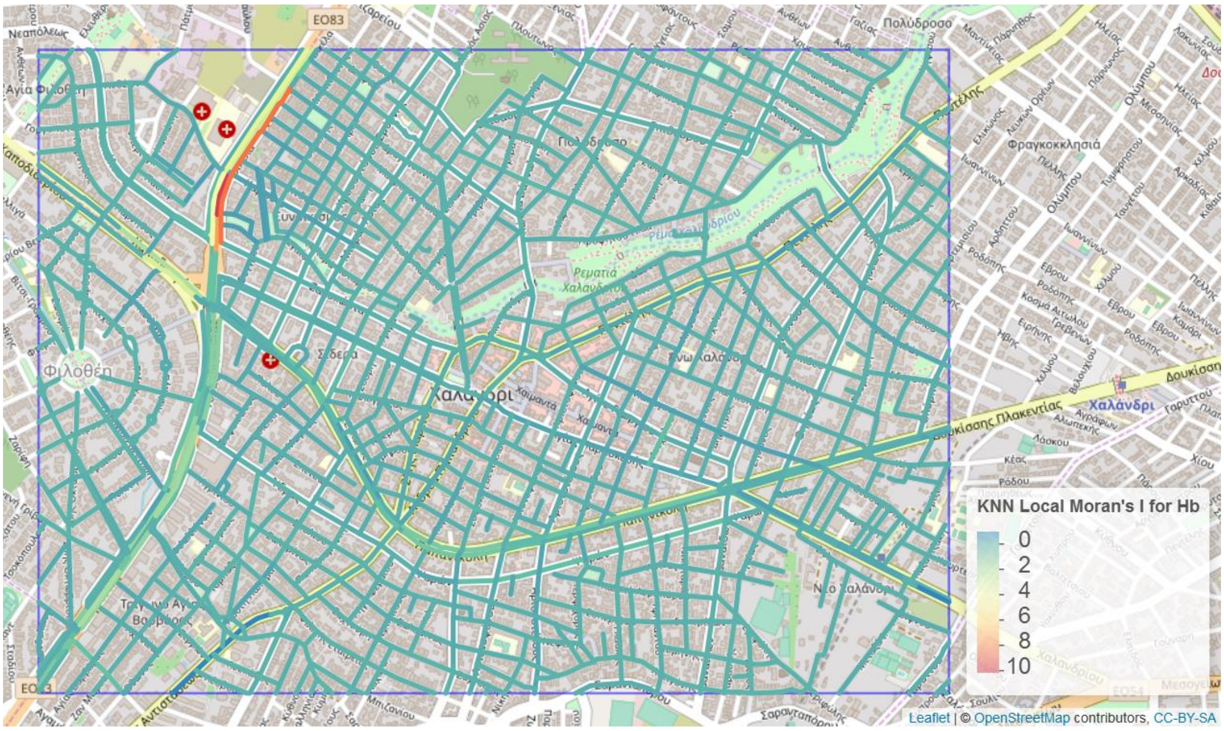


Figure 5-7: Local Moran's I values in Chalandri area based on kNN-based weighting for harsh braking events

From the related tables and figures, it is obvious that kNN-based local Moran's I values are completely different from distance-based Moran's I values. The smaller extend of the considered locations per segment in the kNN weighting approach lead to lower denominators for Moran's I values. Furthermore, the sign of the average values is reversed, leading to signs of opposite values per segment. In other words, there is a trend reversal combined with the magnitude reversal when transitioning form distance-based weighting to k-nearest neighbors weighting due to the underlying mathematical structure of the coefficient calculations.

The large discrepancies in Moran's I values highlight the sensitivities in the specification of Moran's I. If a choice between the two approaches is considered, k-nearest neighbors weighting has a more direct and more sensible physical interpretation: In the context of spatial autocorrelation, a road segment is more likely to be mostly affected by its direct neighbors rather than the entire area that it is located in. Therefore there is large positive local spatial autocorrelation of harsh brakings and harsh accelerations in the northwestern primary road segments of Chalandri, as shown in Figure 5-7 and Figure 5-8.

Overall, Moran's I coefficient is adequate for initial exploratory analysis and for the confirmation of the presence of spatial autocorrelation of harsh events in the considered road segments. However, beyond that level, the coefficient is highly volatile. These results confirm the need for more in-depth spatial statistical analysis of data, such as the ones described in the following.

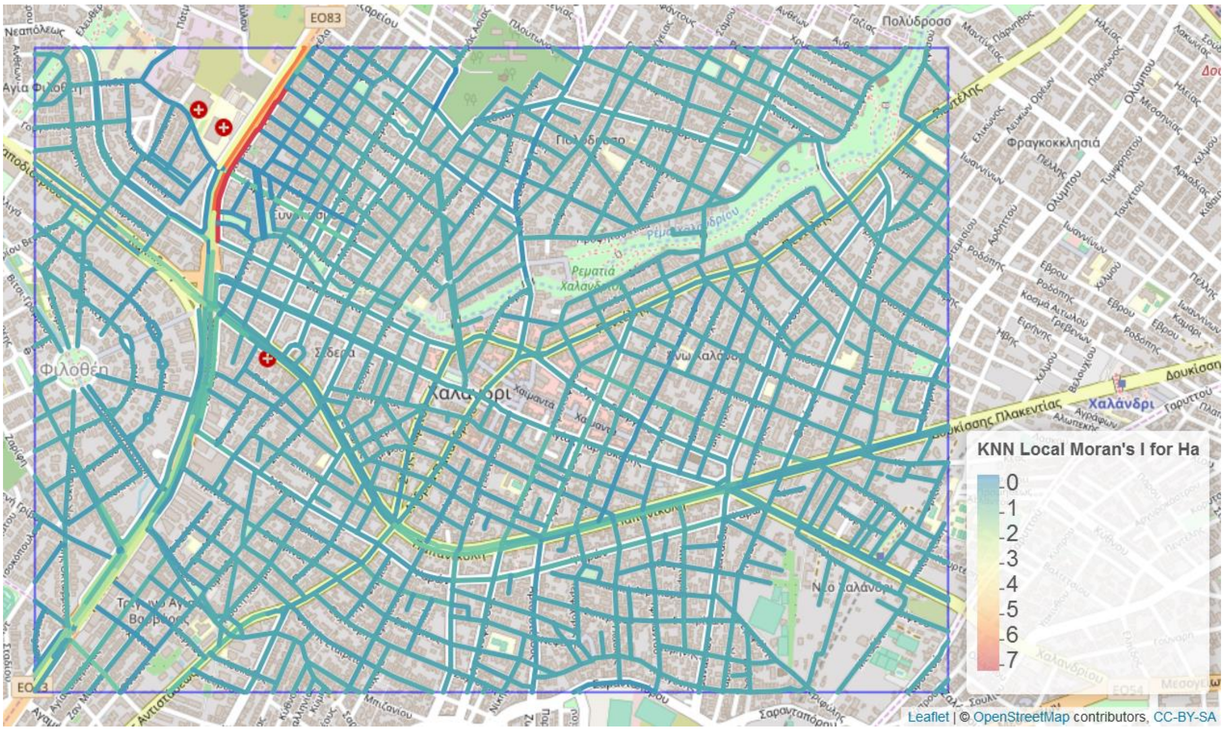


Figure 5-8: Local Moran's I values in Chalandri area based on kNN-based weighting for harsh acceleration events

5.1.3 Harsh event variograms

Empirical variograms are plotted and their respective theoretical models are also fitted for harsh event frequencies per road segment in Chalandri area. The variograms are plotted and fitted each time by considering the event frequencies as single predictors with a constant mean following Pebesma & Graeler (2013). After tests of various theoretical modelling forms, it was found that the spherical variogram with a non-zero nugget fits the data by minimizing error distance. An initial variogram for harsh brakings of the training area appears on Figure 5-9.

Before commenting, it is worth noting that variograms are created by merging spatial points by distance, not direction, which is ignored. This can lead to slightly misleading interpretations. To keep a measure of point direction, variograms can be created for each heading based on compass degrees (Bivand et al., 2008): 0^{o} – North, 90^{o} – East, 180^{o} – South and 270^{o} – West. Any point between 45^{o} and 135^{o} would be assigned to the East variogram, for instance. Thus distance now represents removal from the central point of the study area. Direction-based variograms are shown on Figure 5-10 for harsh braking events in the training area. Distance is measured in km from each road segment centroid.

The partial sill of the spherical harsh braking variogram is 10.8175, with a range of 0.1890 km, while the nugget is 12.4828. The full sill (or maximum semivariance) after stabilization of the variogram is 23.3003. In practice, this indicates that on average, about 190 m from each road segment centroid there is no observable spatial autocorrelation for harsh braking events.

Furthermore, the semivariance can give an idea of data dispersion. In theoretical large road segment samples, the observations of harsh braking frequencies can be expected to be, on average, within the square root of the maximum semivariance from the mean, namely 4.83 harsh brakings. Most of the observations can be expected to lie within the range of two times that value, namely 9.65, based on the two-sigma rule.

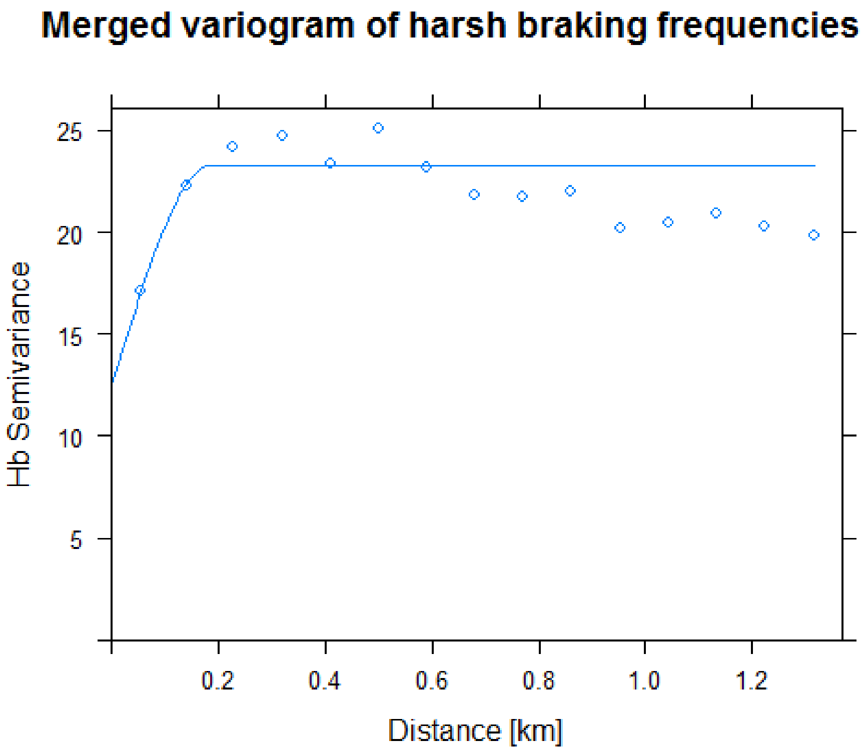


Figure 5-9: Merged direction empirical and theoretical variogram for harsh braking events in Chalandri area

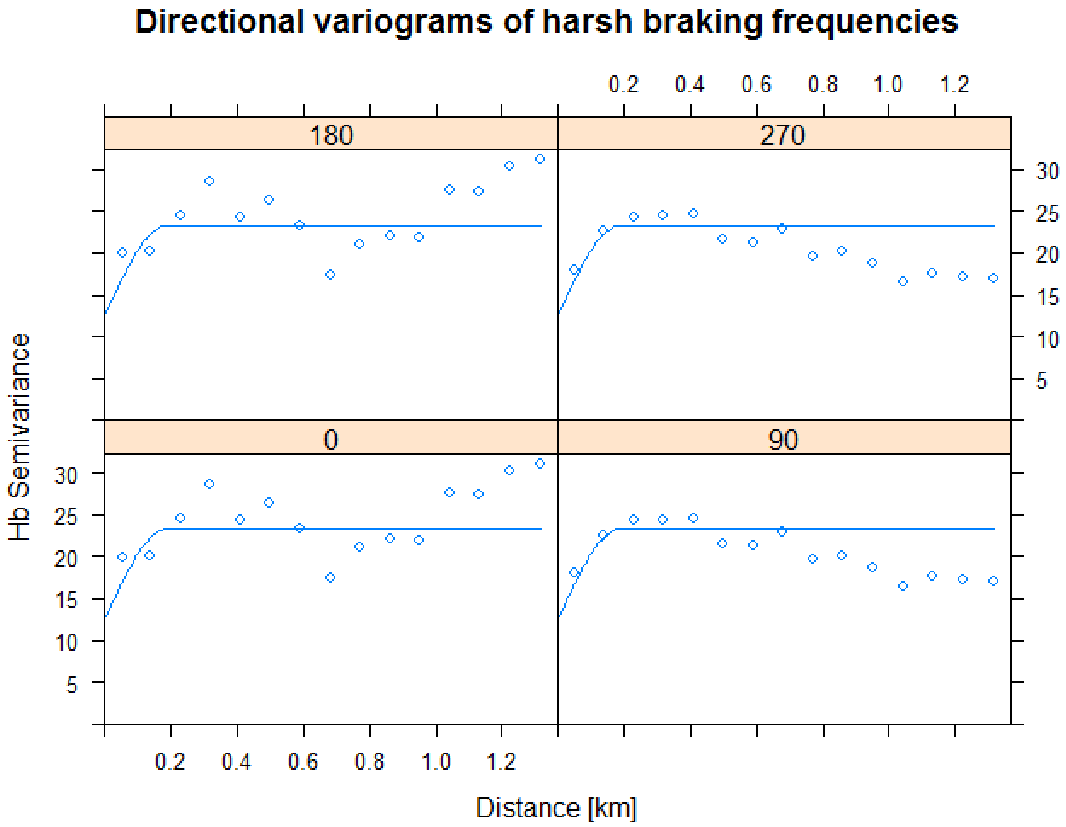


Figure 5-10: Directional empirical and theoretical variograms for harsh braking events in Chalandri area

Respectively, direction-based variograms are shown on Figure 5-11 for harsh acceleration events in the training area.

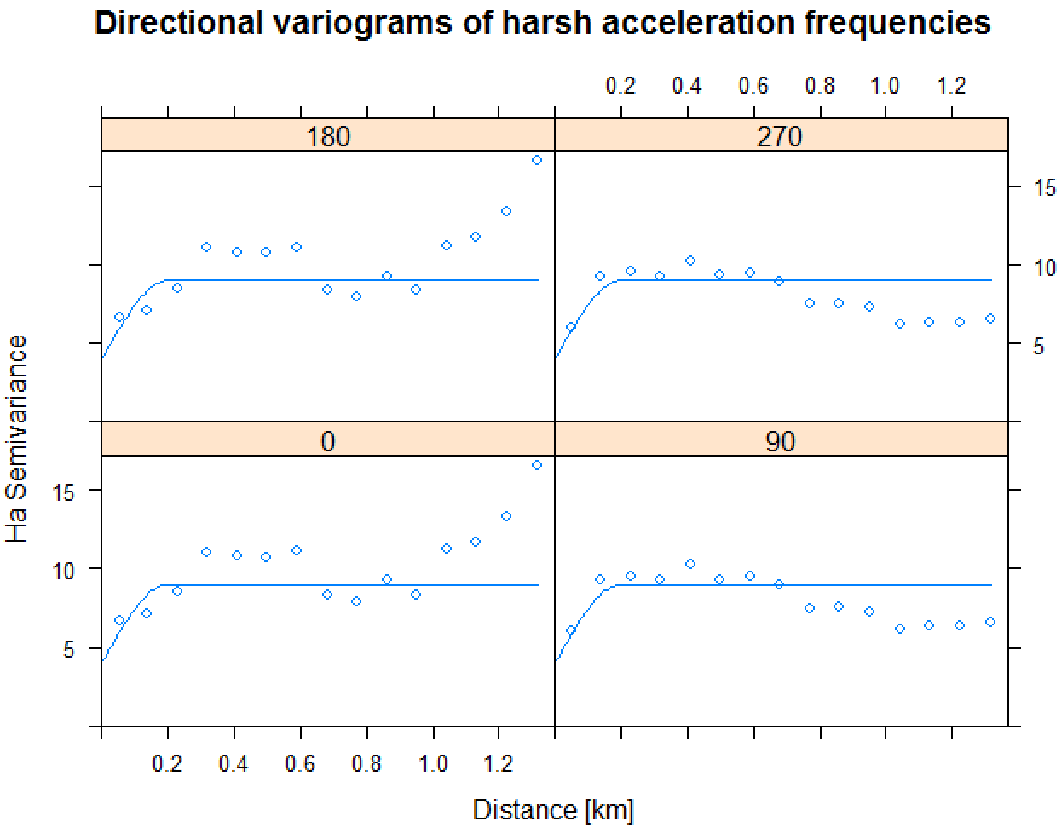


Figure 5-11: Directional empirical and theoretical variograms for harsh acceleration events in Chalandri area

The partial sill of the spherical harsh acceleration variogram is 5.0260, with a range of 0.2006 km, while the nugget is 3.9660. The full sill (or maximum semivariance) after stabilization of the variogram is 8.992. In practice, this indicates that on average, about 200 m from each road segment centroid there is no observable spatial autocorrelation for harsh acceleration events.

Furthermore, in theoretical large road segment samples, the observations of harsh acceleration frequencies can be expected to be, on average, within the square root of the maximum semivariance from the mean, namely 3.00 harsh accelerations. Most of the observations can be expected to lie within the range of two times that value, namely 6.00, based on the two-sigma rule.

Another noteworthy point is that for the directions of North and South, increased fluctuations on semivariance are observed which are not observed in the other directions. This is indicative of geographic anisotropy. Some spatial cyclicity is also observed in the North-South axis for both harsh braking and harsh acceleration frequencies, which is constitutes a wave-repetition pattern in the variogram, observed in other sciences, such as geology (Gringarten and Deutsch, 2001). These findings constitute further incentive for the utilization of spatial statistical models for harsh event analysis.

5.2 Geographically Weighted Poisson Regression results

In this section, Geographically Weighted Poisson Regression (GWPR) models are presented after calibration on the training area dataset for harsh braking and harsh acceleration frequencies. The respective coefficients and various model metrics are interpreted. Furthermore, predictions are conducted for the respective harsh event frequency values in the test area, and their performance is assessed. As a reminder, models are trained in the training area (Chalandri) and their predictions are assessed against the dataset of the test area (Omonoia).

5.2.1 Model selection criteria

Initially all independent variables were inserted in the models, and then they were eliminated one by one, following the method of backward elimination. In other words, variables were removed starting with the ones with the lowest statistical significance every time, corresponding to the highest p-values. This process is preferred to the alternatives (such as forward selection or block-wise selection) because the underlying phenomena are not documented well enough to allow for educated guesses of the correct variable mix. Backward elimination is preferred as it provides a better overview of variable importance before the removal of any independent variables. The widely accepted threshold of 95% probability for statistical significance is observed.

As per standard practices, of all the tested models, the models considered to describe the data optimally were the ones with the lowest corrected Akaike Information Criterion (AICc) and highest MacFadden pseudo-R^2 goodness-of-fit measures (MacFadden, 1977). As a note, MacFadden's pseudo-R^2 (also known as MacFadden's rho) is considered to typically display lower values than linear R^2 coefficients, with values of 0.2 – 0.4 indicating a 'very good model fit' as mentioned by MacFadden in Hensher and Stopher (1979).

In addition, the lowest RMSE/MAE/RMSLE metrics were computed and sought after for the training dataset, without considering the test dataset at this stage yet. A quality check was also conducted for the coefficient values, and especially their signs – positive or negative – to ensure that no irrational relationships are described by the model. For instance, pass counts are always expected to contribute positively to harsh event frequencies. Custom accuracy was also calculated. Last but not least, variable significance was always checked to ensure that included variables continue to contribute in every model iteration. It should be mentioned that categorical variables with many categories (such as lane number) are not removed under the condition that at least one of the categories is indicated as statistically significant.

5.2.2 Harsh braking model

GWPR is conducted for all the road segments in the training area that are traversed by vehicles. Excluding segments with no trips would limit calibration data-points for GWPR, and, more importantly, would introduce bias by changing the geometrical layout of the area that is analyzed during spatial modelling.

5.2.2.1 Cross-validation: Bandwidth selection

Selecting an appropriate kernel bandwidth is an important step when conducting GWR/GWPR. Bivand (2017) mentions that GWR bandwidth choice is potentially demanding, as each step requires the fitting of a number of regressions equal to the local area dataset. An advantage of the adopted approach of this doctoral dissertation is that the naturalistic big data from the driver trips, which might make such an approach otherwise unfeasible, were integrated in a much more manageable number of road segments during the data pre-processing stage.

Following Bivand et al. (2017) and Lu et al. (2013), bandwidth values were tested and their respective cross-validation (CV) score was calculated. The calculations are performed in an iterative process until convergence, namely until there is no significant differentiation in the CV scores. Indicative results appear on **Table 5-7** – bandwidths are shown in km.

Table 5-7: Indicative bandwidth selection iterations for GWPR on harsh brakings

Iteration number	Bandwidth value [km]	CV score
1	1.5617	10649.47
5	1.0326	9934.99
10	0.8491	9812.04
15	0.8497	9812.04
Optimal bandwidth:	**0.8497**	**9812.04**

The bandwidth of 0.84969 km (~ 850 m) was selected for yielding optimal results in the training dataset by providing the minimum CV score. A series of GWPR regressions with different variable sets and subsequent backward elimination were conducted with the optimal bandwidth.

5.2.2.2 Model presentation

The resulting final GWPR model for harsh brakings in urban road networks appears on **Table 5-8**. The p-values of statistically significant continuous variables and categorical variable categories (p-value ≤ 0.05) are shown in bold.

Table 5-8: GWPR model results for harsh brakings in urban road networks

Independent variables	Coefficients			
	Estimate	Std. Error	z-value	p-value
Intercept	0.4636	0.4057	1.143	0.253
Gradient	-2.4864	0.6330	-3.928	**0.000**
Neighborhood complexity	-0.2919	0.0765	-3.815	**0.000**
Segment length	0.0039	0.0001	28.412	**0.000**
Pass count	0.0040	0.0002	21.383	**0.000**
Traffic lights: Yes [Ref.: Traffic lights: No]	0.2563	0.1034	2.479	**0.013**
Pedestrian crossing: Yes [Ref.: Pedestrian crossing: No]	-0.1463	0.0881	-1.661	0.097
Lanes: 2 [Ref.: Lanes: 1]	-0.2435	0.1132	-2.151	**0.031**
Lanes: 3 [Ref.: Lanes: 1]	0.3669	0.1415	2.593	**0.010**
Lanes: 4 [Ref.: Lanes: 1]	0.3578	0.2572	1.391	0.164
Road type: secondary [Ref.: Road type: primary]	1.0520	0.1173	8.969	**0.000**
Road type: tertiary [Ref.: Road type: primary]	-0.0070	0.1537	-0.045	0.964
Road type: residential [Ref.: Road type: primary]	-1.0084	0.1845	-5.467	**0.000**

As stated in the theoretical background, the GWPR/GWR model family incorporates spatial effects in the model coefficients by conducting micro-regressions. Therefore, in addition to the previous overall results, descriptive statistics are provided for these coefficients on **Table 5-9**, to showcase their spatial variation.

Table 5-9: Coefficient estimates of GWPR model for harsh brakings in urban road networks

Independent variables	Coefficient estimates					
	Average	Min.	1st Quadrant	Median	3rd Quadrant	Max.
Intercept	0.4636	0.4621	0.4634	0.4639	0.4646	0.4660
Gradient	-2.4864	-2.4872	-2.4867	-2.4865	-2.4863	-2.4860
Neighborhood complexity	-0.2920	-0.2925	-0.2922	-0.2920	-0.2919	-0.2916
Segment length	0.0039	0.0039	0.0039	0.0039	0.0039	0.0039
Pass count	0.0040	0.0040	0.0040	0.0040	0.0040	0.0040
Traffic lights: Yes [Ref.: Traffic lights: No]	0.2563	0.2562	0.2563	0.2563	0.2563	0.2564
Pedestrian crossing: Yes [Ref.: Pedestrian crossing: No]	-0.1463	-0.1465	-0.1464	-0.1463	-0.1462	-0.1461
Lanes: 2 [Ref.: Lanes: 1]	-0.2435	-0.2439	-0.2437	-0.2436	-0.2435	-0.2433
Lanes: 3 [Ref.: Lanes: 1]	0.3669	0.3666	0.3668	0.3669	0.3669	0.3670
Lanes: 4 [Ref.: Lanes: 1]	0.3578	0.3568	0.3573	0.3577	0.3579	0.3583
Road type: secondary [Ref.: Road type: primary]	1.0520	1.0519	1.0520	1.0520	1.0520	1.0520
Road type: tertiary [Ref.: Road type: primary]	-0.0070	-0.0073	-0.0071	-0.0070	-0.0069	-0.0067
Road type: residential [Ref.: Road type: primary]	-1.0084	-1.0086	-1.0085	-1.0084	-1.0084	-1.0082

Model evaluation metrics are shown on **Table 5-10**:

Table 5-10: Evaluation metrics for the training of the GWPR model for harsh brakings

Metric	Value	Metric	Value
Data-points	869	RMSE	3.2954
AIC	1836.991	MAE	1.3048
AICc	1837.417	RMSLE	0.5569
McFadden pseudo-R^2	0.631	CA	80.90%

5.2.2.3 Discussion of results

Initially, the influence of each significant independent variable is examined. The variable of gradient has a negative sign, denoting that harsh braking frequencies decrease in positive (uphill) slopes, while they increase in negative (downhill) slopes – a quite intuitive outcome, as gravity helps the drivers decelerate their vehicles in the first case while it hinders them in the second case.

Interestingly, the effect of neighborhood complexity is negative, meaning that in more dense/complex areas, fewer harsh brakings tend to occur. This is explained as the manifestation of compensatory effects in driving behavior. In more complex environments, where driver attention is required to be split in several directions, and drivers may have to examine additional signage or accesses, they tend to be more cautious. This results in lower aggressiveness on the part of drivers, possibly by lowering their speed, and fewer harsh brakings.

Segment length and pass count are both exposure variables, and their influences are, as expected, positive. Marginal Effects at the Means (MEM) are calculated following Washington et al. (2010) considering the mean data points of these variables. For the segment length average of 144.3 m, an increase of 1 meter leads to an increase of $MEM_{Seg\_Length} = 0.0067$ harsh brakings. Respectively, for the pass count average of 191 passes, an increase of 1 pass leads to an increase of $MEM_{Pass\ count} = 0.0087$ harsh brakings. Therefore it seems that pass counts, which are more related to traffic variables and route choice lead to more harsh brakings per unit compared to segment lengths, which is a fixed geometrical segment characteristic.

The presence of traffic lights was also found to increase harsh braking frequencies per road segment. This indicates that more unexpected events tend to happen in roads with traffic lights, forcing drivers to brake forcefully. One explanation is that drivers tend to behave more aggressively near traffic lights in order to rush through the junction, and traffic conflicts ensue with nearby vehicles. There is also a possibility that pedestrians may also be involved in junctions, however it is noteworthy that pedestrian crossings do not seem to affect harsh braking occurrence statistically significantly as a variable except from the more lenient 90% level. Its removal, however, worsened the metrics of the GWPR model (mainly AICc, RMSE and MAE), so it was retained in its final form.

Both discrete variables with many categories seem to show partial statistical significance. Road segments with two lanes show an increased harsh braking frequency, compared to road segments with one lane only. This trend is reversed with three lanes, which display a decreasing influence in frequency. The explanation here is not very straightforward, and ought to be visualized within a context of traffic and speed in road segments with different lanes. It can be thought that two-lane segments provide some ways of avoiding harsh events, when drivers are driving with reasonable controlled speeds. In three-lane segments, it is probable that the speeds are less restrained and conflicts and the resulting harsh events can occur from more directions. This is enough to reverse the previous trend, and event frequencies increase. Four-lane segments do not seem to be significantly differentiated from the baseline by that feature alone.

Compared to primary road segments, secondary road segments show increased harsh event frequencies. Again, this is interpreted within context; roads with secondary class have less space in which to maneuver but not reduced enough speeds, leading to more harsh braking events in comparison. Tertiary road segments do not seem to be significantly differentiated from the baseline by that feature alone, while residential roads show reduced influence on harsh event frequencies.

Neither curvature nor road direction (one-way vs. two-way segments) were found to be statistically significant for harsh braking explanations. Lastly, the intercept (constant term) is not statistically significant either. This hints that no significant unobserved effects remain after unobserved parameters are integrated in the spatial fluctuation of the estimated coefficients, as explained in the following.

The spatial fluctuation of the estimated coefficients as shown on **Table 5-9** is low, but manifests in all geometric and network characteristic variables apart from two notable exceptions: the exposure variables of segment length and pass count. In practice, this means that the influence of segment length and pass count is interpreted as stable across the training area for harsh brakings. In contrast, the influence of the rest of the geometric and network variables varies slightly across the training area, as spatial effects include unobserved parameters across the network. These unobserved effects could be related to a number of reasons: unobserved additional fixed network characteristics, for instance local signage, obstacles or roadworks on the road, or unobserved flow characteristics, for instance entry/exit points providing additional flows such as multi-floor parking garages. This apparent spatial stability is also explained by their high value range compared to the other variables; spatial fluctuations are in negligible orders of magnitude.

It can be considered that the calibrated GWPR models are an example of high heterogeneity, as every observation is different and there is a different parameter β_{ik} for each variable k of each observation i. This is a consequence and also liberty of big data approaches, as proposed by Anselin et al. (2014).

As in standard Poisson Models, the McFadden pseudo-R^2 for the GLM component is at a very satisfactory level at 0.63, given its typical lower values than linear R^2 coefficients. The other model evaluation metrics shown on **Table 5-10** are calculated for the training dataset only initially. The RMSE value suggests that the average magnitude of the error is about 3.3 harsh braking counts, while MAE is considerably lower at 1.3 harsh braking counts. This indicates some isolated modest discrepancies in the predictions of the GWPR model, which increase RMSE, however overall performance is very good with a low MAE. RMSLE is also considerably lower than both metrics, as it has logarithmic properties. Lastly, the custom accuracy (CA) value for the training dataset indicates that the GWPR model correctly predicts harsh braking frequencies in the training area with a tolerance of ± 1 harsh braking per segment 81% of the times. These metrics indicate a very good model fit.

Due to the unique configuration of GWR/GWPR, maps can be created for the localized coefficient values of every variable in the model for the training area. Figure 5-12 features the mapping of the coefficient of gradient, indicatively. It should be highlighted that the graphical scale is significantly exaggerated compared to the low spatial fluctuations of the coefficient. This low spatial variability could be attributed to the fact that most included variables were able to capture harsh braking frequencies and adapt well on the global regression scale, leaving only a small amount of residual variance to be explained by local regressions.

Nonetheless, there is a clear visible trend: Gradient appears to contribute to more harsh brakings in segments located in the northwest side of the map compared to segments in the southeast, with the middle sector serving as a smooth middle ground transition for the coefficient.

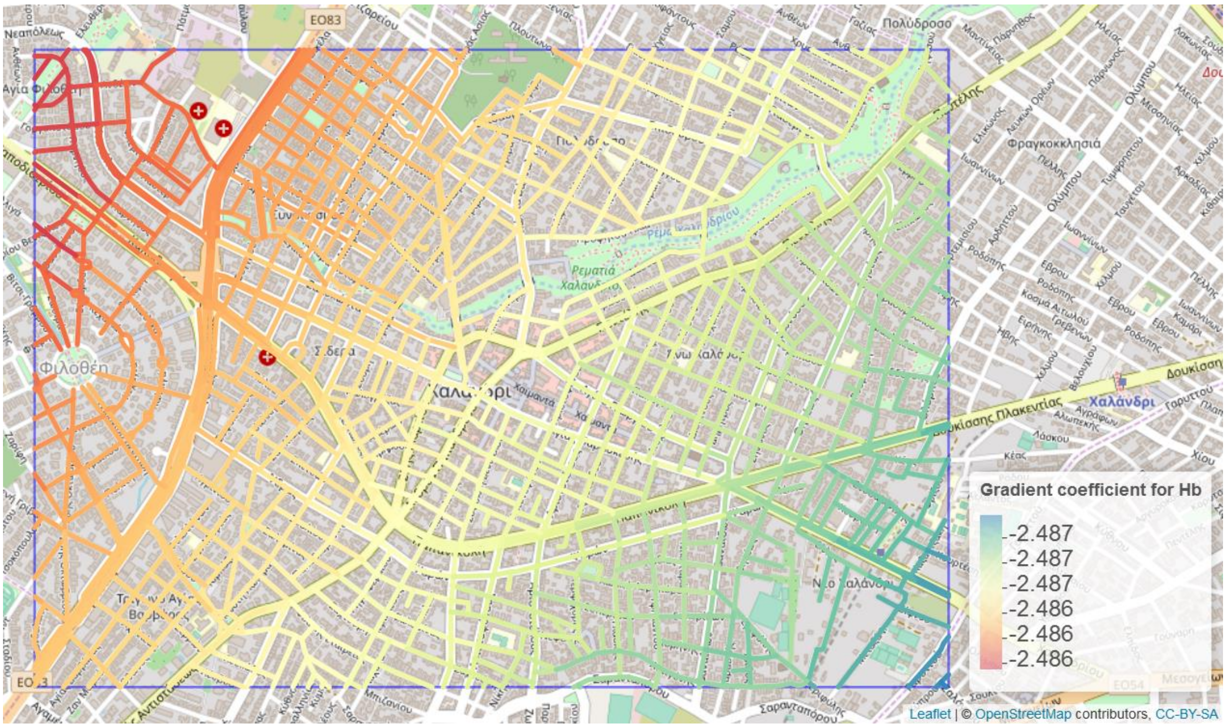

Gradient coefficient for Hb
-2.487
-2.487
-2.487
-2.486
-2.486
-2.486
Leaflet | © OpenStreetMap contributors, CC-BY-SA

Figure 5-12: GWPR gradient coefficients of harsh brakings in Chalandri area

5.2.2.4 Prediction and transferability capabilities

As stated in the methodology section, for GWR/GWPR the transferability of spatial effects is typically limited, because it requires prior knowledge of the distribution of the dependent variable in the test area. While in this particular case this knowledge exists, in a typical road safety problem it does not exist, especially in crash/event forecasting situations. In other words, to gain knowledge of the spatial effects in the test area, GWPR models would have to be trained in the full dataset of the test area, which is not its intended purpose.

Therefore it was decided to eschew the spatial effects of GWPR and conduct predictions with the non-spatial generalized linear model (GLM) part of GWPR, also termed 'global regression' for lack of its spatial effects. This is a Poisson-lognormal model with values corresponding to the average coefficient estimates found in **Table 5-8**. True values in the test area dataset and the respective predictions are plotted on Figure 5-13; there are higher concentrations of lower frequencies noted with a bolder color from observation overlap. Predictions are conducted only for the 1066 road segments with non-zero trips, as pass count is required as input.

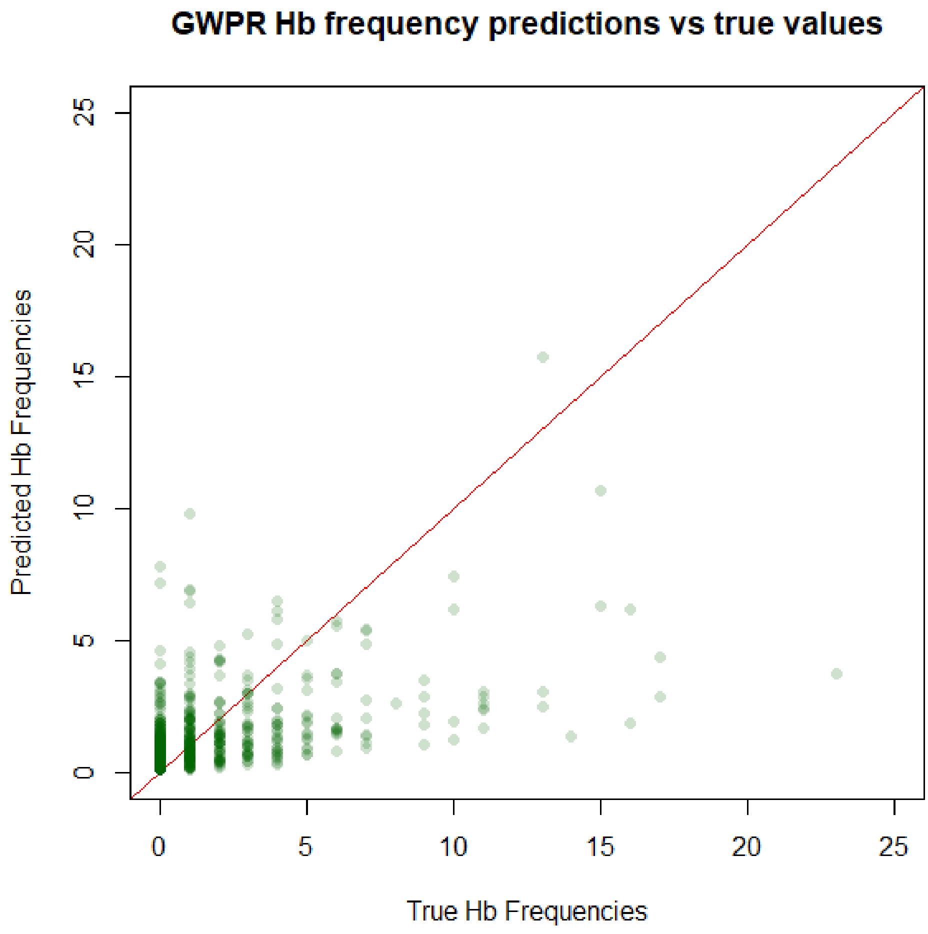


Figure 5-13: True and frequentist GLM predicted frequencies of harsh brakings in Omonoia area

The respective metrics of the predictions of the test set values from the model trained in the training set values are shown on **Table 5-11**.

Table 5-11: Evaluation metrics for predictions of the GLM part of the GWPR model for harsh brakings

Metrics	Value
RMSE	1.9792
MAE	1.0265
RMSLE	0.5508
CA	82.64%

All three error metrics have reduced values from their counterparts of **Table 5-10**. The reduction of RMSE is the most pronounced, and it also hints at fewer large errors in the predictions for the test area. This is an indicator of good GLM/GWPR model predictive capabilities and transferability to another comparable area.

The increased value of CA is very interesting, and mainly attributed to effective cross-validation in the GWPR training process. It should be mentioned, however, that part of this increase might be circumstantial from the test dataset as well, favored by the tolerance built in the CA calculation.

Harsh braking maps can be created for the predictions of the GLM/GWPR model predictions in any of the two areas – Figure 5-14 shows the predictions in the test area.

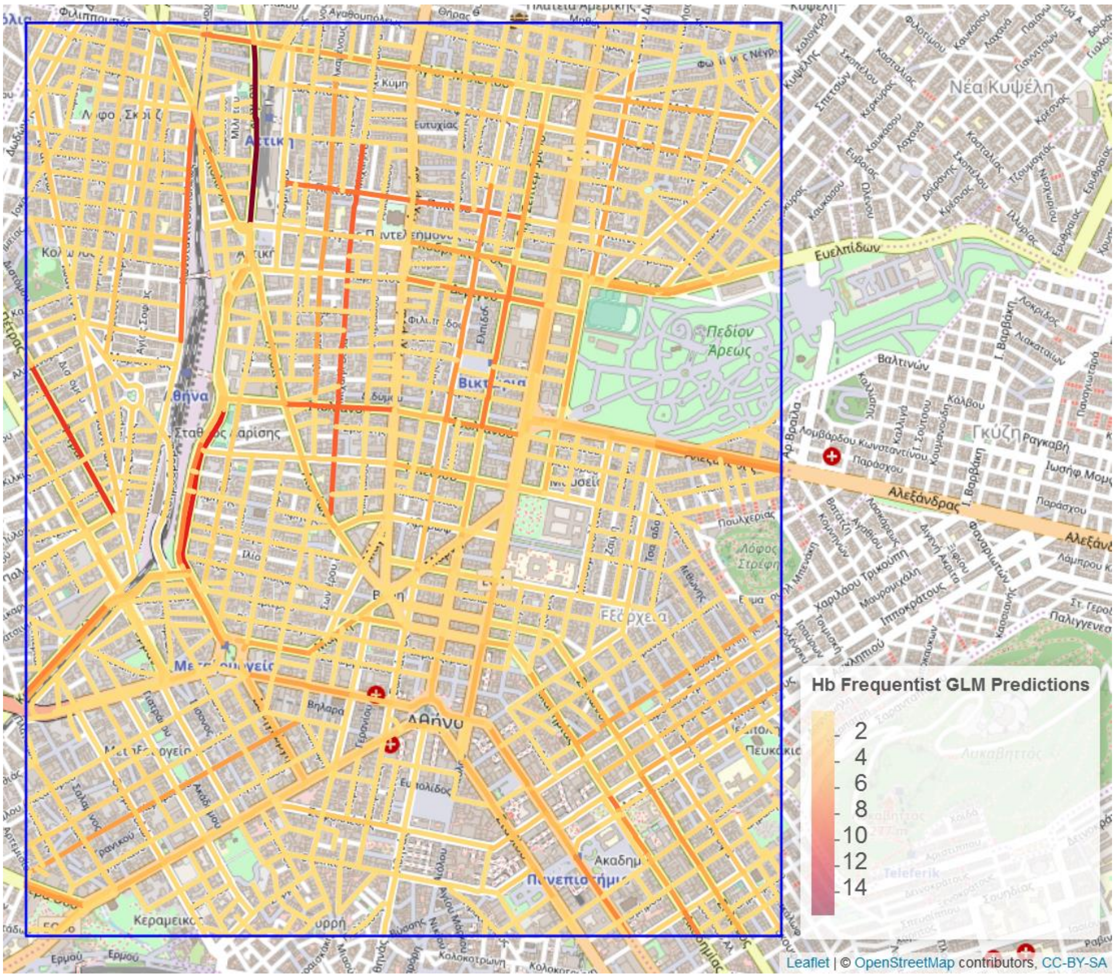

Hb Frequentist GLM Predictions
2
4
6
8
10
12
14
Leaflet | © OpenStreetMap contributors, CC-BY-SA

Figure 5-14: Frequentist GLM predicted harsh braking frequencies in Omonoia area

5.2.3 Harsh acceleration model

An equivalent process is followed for modelling harsh acceleration frequencies using GWPR.

5.2.3.1 Cross-validation: Bandwidth selection

For the bandwidth of harsh accelerations, indicative results appear on **Table 5-12** – bandwidths are shown in km.

Table 5-12: Indicative bandwidth selection iterations for GWPR on harsh accelerations

Iteration number	Bandwidth value [km]	CV score
1	1.5617	4402.00
5	0.3718	3502.67
10	0.3513	3501.04
15	0.3598	3497.06
16	0.3599	3497.06
Optimal bandwidth:	**0.3599**	**3497.06**

The bandwidth of 0.359860 km (~ 360 m) was selected for yielding optimal results in the training dataset by providing the minimum CV score. A series of GWPR regressions with different variable sets and subsequent backward elimination were conducted with the optimal bandwidth.

5.2.3.2 Model presentation

The resulting final GWPR model for harsh accelerations in urban road networks appears on **Table 5-13**. The p-values of statistically significant continuous variables and categorical variable categories (p-value ≤ 0.05) are shown in bold.

Table 5-13: GWPR model results for harsh accelerations in urban road networks

Independent variables	Coefficients			
	Estimate	Std. Error	z-value	p-value
Intercept	-1.4230	0.2135	-6.667	**0.000**
Curvature	9.0471	2.7282	3.316	**0.001**
Segment length	0.0030	0.0002	17.486	**0.000**
Pass count	0.0042	0.0002	19.818	**0.000**
Traffic lights: Yes [Ref.: Traffic lights: No]	0.3791	0.1176	3.222	**0.001**
Lanes: 2 [Ref.: Lanes: 1]	0.0794	0.1306	0.608	0.543
Lanes: 3 [Ref.: Lanes: 1]	0.4741	0.1786	2.655	**0.008**
Lanes: 4 [Ref.: Lanes: 1]	0.3828	0.3197	1.198	0.231
Road type: secondary [Ref.: Road type: primary]	0.7323	0.1473	4.973	**0.000**
Road type: tertiary [Ref.: Road type: primary]	0.3720	0.1847	2.014	**0.044**
Road type: residential [Ref.: Road type: primary]	-0.6642	0.2216	-2.997	**0.003**

In addition to the previous overall results, descriptive statistics are provided for the variable regression coefficients on **Table 5-14**, to showcase their spatial variation.

Table 5-14: Coefficient estimates of GWPR model for harsh accelerations in urban road networks

Independent variables	Coefficient estimates					
	Average	Min.	1st Quadrant	Median	3rd Quadrant	Max.
Intercept	-1.4230	-1.4246	-1.4235	-1.4229	-1.4225	-1.4216
Curvature	9.0471	8.9979	9.0248	9.0469	9.0613	9.0795
Segment length	0.0030	0.0030	0.0030	0.0030	0.0030	0.0030
Pass count	0.0042	0.0042	0.0042	0.0042	0.0042	0.0042
Traffic lights: Yes [Ref.: Traffic lights: No]	0.3791	0.3785	0.3790	0.3793	0.3796	0.3800
Lanes: 2 [Ref.: Lanes: 1]	0.0794	0.0783	0.0788	0.0790	0.0794	0.0800
Lanes: 3 [Ref.: Lanes: 1]	0.4741	0.4727	0.4733	0.4736	0.4738	0.4743
Lanes: 4 [Ref.: Lanes: 1]	0.3828	0.3802	0.3814	0.3819	0.3827	0.3841
Road type: secondary [Ref.: Road type: primary]	0.7324	0.7319	0.7324	0.7326	0.7329	0.7336
Road type: tertiary [Ref.: Road type: primary]	0.3720	0.3715	0.3717	0.3719	0.3720	0.3723
Road type: residential [Ref.: Road type: primary]	-0.6642	-0.6658	-0.6647	-0.6642	-0.6637	-0.6624

Model evaluation metrics are shown on **Table 5-15**:

Table 5-15: Evaluation metrics for the training of the GWPR model for harsh accelerations

Metric	Value	Metric	Value
Data-points	869	RMSE	2.0861
AIC	1245.987	MAE	0.9125
AICc	1246.297	RMSLE	0.4704
McFadden pseudo-R^2	0.606	CA	84.69%

5.2.3.3 Discussion of results

Similarly with the previous models, the influence of each significant independent variable is examined. Interestingly, curvature seems to have a positive effect on harsh acceleration frequencies. This may be attributed to drivers rushing ahead to exploit an open headway due to obstacles in the tighter side of the curves. In larger road segments, this can also indicate sensation-seeking ('joyride'), related to more aggressive driving.

Segment length and pass count are both exposure variables, and their influences are expectedly positive, and very close to the harsh braking model. Marginal Effects at the Means (MEM) are calculated following Washington et al. (2010) considering the mean data points of these variables. For the segment length average of 144.3 m, an increase of 1 meter leads to an increase of $MEM_{Seg\_Length} = 0.0046$ harsh accelerations. Respectively, for the pass count average of 191 passes, an increase of 1 pass leads to an increase of $MEM_{Pass\ count} = 0.0093$ harsh accelerations. Therefore it seems that pass counts lead to more events per unit compared to segment lengths for harsh accelerations as well.

The presence of traffic lights was also found to increase harsh acceleration frequencies per road segment. This pinpoints the common occurrence of drivers rushing through the junction, leading to more harsh accelerations compared to segments without traffic lights. Pedestrian crossings do not seem to affect harsh acceleration frequencies statistically significantly at any level.

Regarding lane numbers, only road segments with three lanes show increased harsh acceleration frequencies compared to one-lane segments. This hints at some inherent differentiation in three-lane segments. It is possible that in three-lane segments, drivers accelerate harshly to avoid obstacles or fill gaps in traffic. Another explanation is the rush of drivers to switch lanes, especially for exclusive lane turning positions. In four-lane segments, the trend disappears into statistical non-significance, possibly because of enough room to maneuver between obstacles or lane changing without the necessity of harsh acceleration.

All road types have been found to significantly affect harsh acceleration frequencies. Secondary and tertiary road categories display increased harsh accelerations compared to primary roads, possibly for similar reasons with lane number: driver aggressiveness and the urge to exploit large headways while lacking adequate space to do so without harshly accelerating. In residential roads, the trend is reversed, and fewer harsh accelerations are observed.

Neither gradient nor road direction seem to affect harsh acceleration frequencies in a statistically significant manner. However, the intercept is statistically significant, indicating additional unobserved effects that are not described by the included variable mix.

As expected, the variances in frequencies of harsh brakings and harsh accelerations are explained by different variable combinations. The two phenomena can be thought as different, if not opposite, action or reaction mechanisms from drivers. Harsh accelerations can be thought to be more closely related to drivers choosing a more aggressive behavior to navigate in the road environment, due to haste or aggressiveness, and may or may not involve interaction with other road users or obstacles. Conversely, harsh brakings can be thought to be more directly in reaction to avoid a collision or conflict with other road users or obstacles.

Similarly with harsh brakings, the spatial fluctuation of GWPR coefficients for harsh accelerations as shown on **Table 5-14** is low, and manifests in all geometric and network characteristic variables except the exposure variables of segment length and pass count. Once again, the influence of segment length and pass count is interpreted as stable across the training area for harsh accelerations. In contrast, the influence of the rest of the geometric and network variables varies slightly across the training area, as spatial effects include unobserved parameters across the network.

As in standard Poisson Models, the McFadden pseudo-R^2 for the GLM component is at a very satisfactory level at 0.61, given its typical lower values than linear R^2 coefficients. The other model evaluation metrics shown on **Table 5-15** are calculated for the training dataset only initially. The RMSE value suggests that the average magnitude of the error is about 2.1 harsh acceleration counts, while MAE is considerably lower at 0.9 harsh acceleration counts. This indicates fewer isolated discrepancies in the predictions of the GWPR model compared to harsh brakings; overall performance is very good with a low MAE. RMSLE is also considerably lower than both metrics, as it has logarithmic properties. Lastly, the custom accuracy (CA) value for the training dataset indicates that the GWPR model correctly predicts harsh acceleration frequencies in the training area with a tolerance of ± 1 harsh acceleration per segment 85% of the times. These metrics indicate – again – a very good model fit.

Figure 5-15 features the mapping of the coefficient of curvature, indicatively for harsh accelerations.

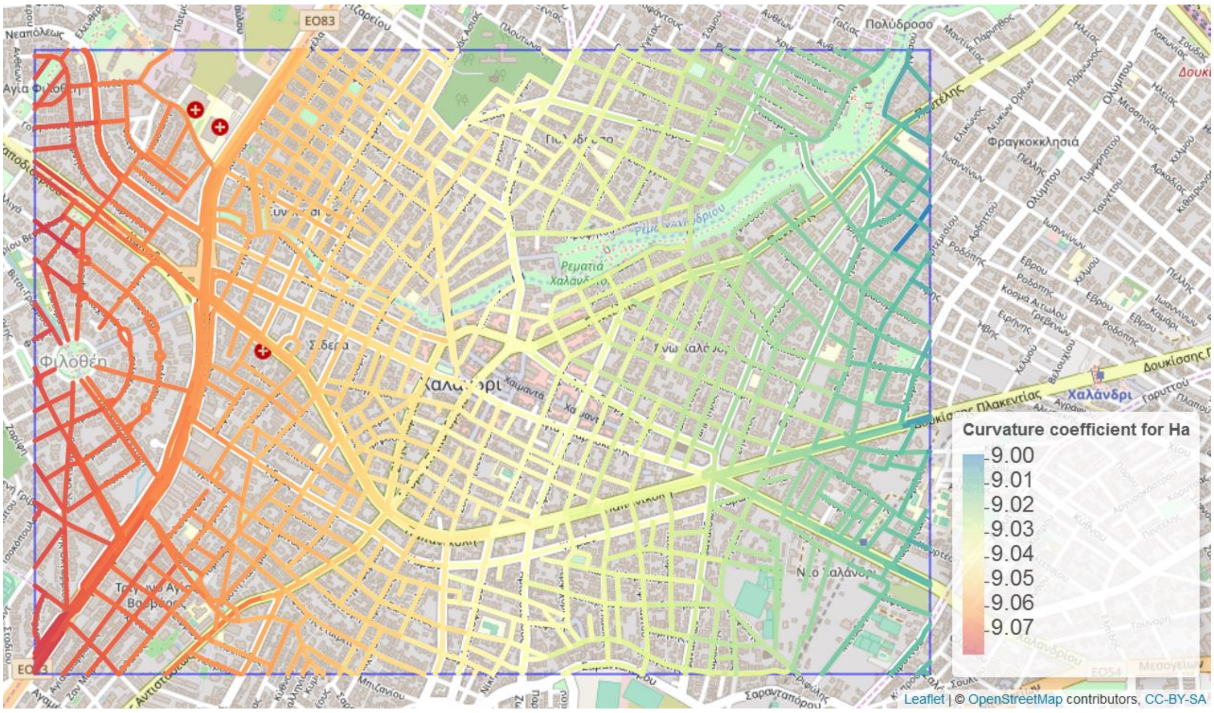


Figure 5-15: GWPR curvature coefficients of harsh accelerations in Chalandri area

The pattern of the spatial fluctuation of the coefficient of curvature for harsh accelerations is similar with the one of gradient for harsh brakings. Curvature appears to contribute to more harsh accelerations in segments located in the west side of the map compared to segments in the east, with the middle sector serving as a smooth middle ground transition for the coefficient. The magnitude of the fluctuations is more pronounced in this instance.

5.2.3.4 Prediction and transferability capabilities

As with harsh brakings, the respective GLM with values corresponding to the average coefficient estimates found in **Table 5-13** is used for prediction. True values in the test area dataset and the respective predictions are plotted on Figure 5-16; there are higher concentrations of lower frequencies noted with a bolder color from observation overlap. Predictions are conducted only for the 1066 road segments with non-zero trips, as pass count is required as input.

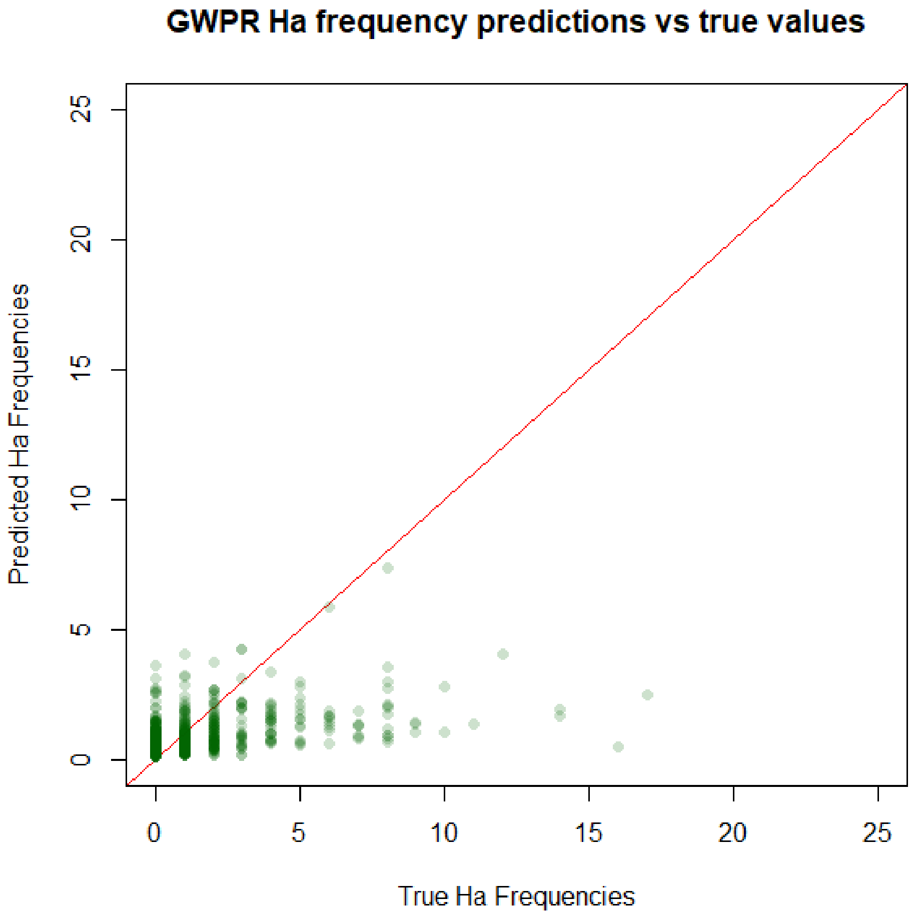


Figure 5-16: True and frequentist GLM predicted frequencies of harsh accelerations in Omonoia area

The respective metrics of the predictions of the test set values from the model trained in the training set values are shown on **Table 5-16**.

Table 5-16: Evaluation metrics for predictions of the GLM part of the GWPR model for harsh accelerations

Metrics	Value
RMSE	1.6836
MAE	0.8721
RMSLE	0.5082
CA	87.71%

RMSE, MAE and CA have reduced values from their counterparts of **Table 5-15**. RMSLE features a slight increase. This is explained by slight local variations between the training and test datasets at low values (namely 0 or 1) which are influenced by the built-in addition of +1 inherent in the RMSLE calculation. CA is again slightly elevated in the test dataset, indicating adequate cross-validation by the selection of an appropriate GWPR bandwidth. All metric values are indicators of overall good GLM/GWPR model predictive capabilities and transferability to another comparable area.

GLM/GWPR model predictions in the test area are shown on Figure 5-17.

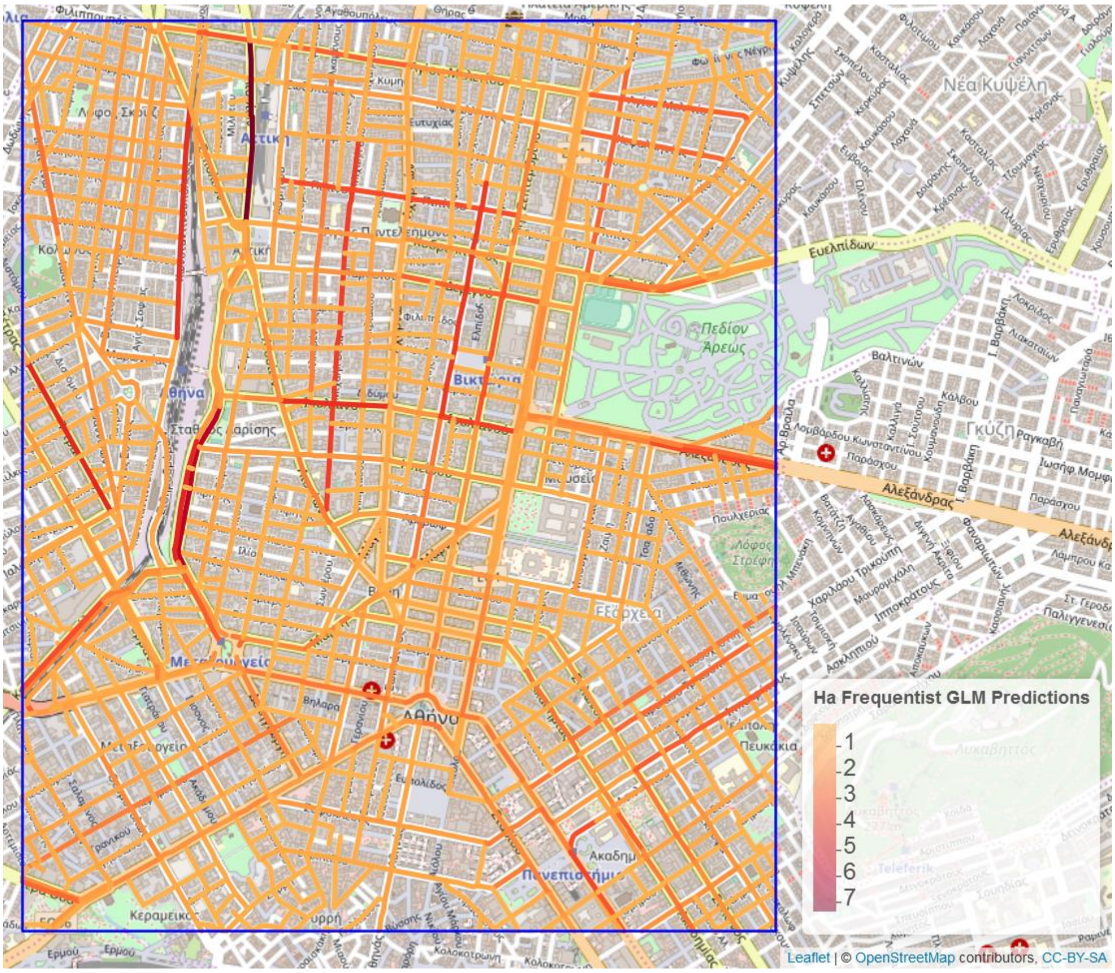


Figure 5-17: Frequentist GLM predicted harsh acceleration frequencies in Omonoia area

5.2.4 Issues with Linear Geographically Weighted Regression

Apart from the previous approach, linear GWR was also explored as an option for harsh event frequencies. The reasoning was to integrate the exposure parameters into harsh event frequencies, thus obtaining harsh event rates: namely harsh braking/acceleration frequencies per pass count per meter. The linear analysis was then conducted with harsh braking/acceleration rates hb_rate_w and ha_rate_w as the dependent variables. However, a number of issues arose during the process.

In all model configuration attempts, no significant correlations were found between harsh braking/acceleration rates and any of the independent variables previously calculated and utilized. The developed linear GWR models had R^2 values that were close to zero, and the statistical significance of all variables was practically non-existent (p-values > 0.5).

In addition, in each model there were predictions that were negative, which is impossible for event rates; this is another sign of very poor model performance. These are indicators that the phenomena of harsh brakings/acceleration frequencies are not adequately described by linear relationships.

Additional consideration was given to circumventing these obstacles by using the logarithm of event rates instead. This proved quickly futile as there are many zeros in the event rate dataset, representing roads without events for the investigation period. Any effort to assign a very low logarithmic values to zero rates led to similarly very skewed results and was abandoned.

Furthermore, while the rate approach entails a standardization which is intuitive, the process inherently forces exposure variable coefficients to 1 in the rate denominator. Therefore, their influence is integrated in the remaining model coefficients. This presents two inherent limitations. Firstly, it binds exposure variable coefficients together and limits the flexibility of the model that would explain some of the variance via these coefficients. Secondly, any interpretation of the remaining model coefficients becomes slightly more perplexed than the more straightforward Poisson approach.

All in all, the exploration of linear GWR models has revealed that linear and log-linear relationships and the harsh event rate approach are inadequate to explain harsh event frequencies. This finding is expected – the vast majority of the literature adopts count-based GLM approaches for spatial analysis of crashes, as evident from **Table 2-1** to **Table 2-4**. The same trend manifests with harsh events, which are point-data road safety observations as well. The count approach enables more direct comparison with Bayesian Poisson-lognormal CAR models as well.

5.3 Conditional Autoregressive Prior models

In this section, Bayesian Poisson lognormal models with conditional autoregressive priors (CAR models) are presented after calibration on the training area dataset for harsh braking and harsh acceleration frequencies. The respective coefficients and various model metrics are interpreted. Furthermore, predictions are conducted for the respective harsh event frequency values in the test area, and their performance is assessed.

5.3.1 Model selection criteria

Parts of the CAR model selection process mirror that of GWPR described in Section 5.2.1. The backward elimination process was observed, and variable significance was determined by examining the respective Bayesian Credible Intervals (BCI). The widely accepted threshold of 95% probability for statistical significance is observed. Therefore variable coefficients with posterior samples (i.e. distribution values) with the same sign at the 2.5% and 97.5% distribution percentiles – which represent the middle 95% – were noted as statistically significant.

The models considered to describe the data optimally were the ones with the lowest Deviance Information Criterion (DIC), lowest Watanabe's modified Akaike Information Criterion (WAIC) and highest Log Marginal Predictive Likelihood (LMPL) based on Spiegelhalter et al. (2002) and Lee (2013). In the case of two very closely competing models a voting system of 'best of three' was implemented, meaning that the model with two or three more desirable values was retained.

In addition, the lowest RMSE/MAE/RMSLE metrics were computed and sought after for the training dataset, without considering the test dataset at this stage yet. A quality check was also conducted for the coefficient values, and especially their signs – positive or negative – to ensure that no irrational relationships are described by the model. Custom accuracy was also calculated. Last but not least, variable significance was always checked to ensure that included variables continue to contribute in every model iteration. It should be mentioned that categorical variables with many categories (such as lane number) are not removed under the condition that at least one of the categories is indicated as statistically significant.

5.3.2 Harsh braking model

Similar to GWPR, Bayesian analysis with CAR models is conducted for all the road segments in the training area that are traversed by vehicles.

5.3.2.1 Model preparation

The preparation phase of the CAR models initially involved constructing the spatial weighting matrix of adjacent road segments. Several spatial alternatives were examined: (i) binary neighbor weighting, (ii) row-standardized inverse distance weights and (iii) row-standardized inverse squared distance weights, mirroring the techniques described in Section 3.2.2.1. For (i), weights are given on a rook-polygon analogy by creating Thiessen (voronoi) polygons for road segment centroids. The closest one-third of the total road segments are considered for each road segment, which aids computational times – in this case, the training area featured 869 road segments, therefore the closest 289 roads were considered for each one.

For the independent variables, Bayesian inference was conducted by initially giving non-informative priors to variable distributions, which are assumed to be Gaussian (normal) with a mean of 0 and a variance of 10^6. The equivalent priors of tau^2 and ${\sigma_{\theta i}}^2$, which represent spatially structured and unstructured effects respectively, are assumed to be Gaussian (normal) with a mean of 1 and a variance of 10^6. As a reminder, the CAR priors are assumed to follow the Besag-York-Mollie (or BYM) specification which has been used in many road safety past studies to spatially model crash frequencies since its inception (e.g. Huang et al., 2016; Cai et al., 2018; Zhai et al., 2018; Wen et al., 2019).

Bayesian inference is then conducted using Markov Chain Monte Carlo (MCMC) simulation. It was found that the best performing models required a large burn-in period before stabilization. After several trials, the posterior summaries for the best-fitting models were obtained by a chain with 410,000 iterations, the first 400,000 of which were discarded as the burn-in sample. The remainder 10,000 samples are thinned by 100 to reduce autocorrelation and the resulting values describe the posterior distributions. Computationally, a fixed value for the random number generation processes is also required to ensure the replicability of results.

5.3.2.2 Model presentation

CAR Models were calibrated following Lee (2013). The resulting final CAR model for harsh brakings in urban road networks appears on **Table 5-17**. The 95% BCI values are calculated at 2.5% (lower bound) to 97.5% (upper bound), and median values refer to this 95% BCI margin only. 95% BCI values of statistically significant continuous variables and categorical variable categories – which retain the same signs – are shown in bold.

Table 5-17: CAR model results for harsh brakings in urban road networks

Independent variables	Posterior values				
	Mean	St. Dev.	Median	2.5% value	97.5% value
Intercept	-1.4134	0.5538	-1.1438	**-2.2422**	**-0.5845**
Gradient	-9.7538	0.9612	-9.9088	**-11.3826**	**-8.1250**
Neighborhood Complexity	-0.1787	0.1116	-0.1919	**-0.3535**	**-0.0038**
Segment length	0.0075	0.0004	0.0073	**0.0069**	**0.0080**
Pass count	0.0086	0.0003	0.0084	**0.0081**	**0.0091**
Traffic lights: Yes [Ref.: Traffic lights: No]	-0.0902	0.0628	-0.0534	-0.1982	0.0178
Pedestrian crossing: Yes [Ref.: Ped. cross.: No]	0.3820	0.1182	0.2625	**0.1614**	**0.6025**
Lanes: 2 [Ref.: Lanes: 1]	-0.1713	0.0612	-0.1543	**-0.3055**	**-0.0371**
Lanes: 3 [Ref.: Lanes: 1]	-0.5719	0.0682	-0.5673	**-0.6885**	**-0.4552**
Lanes: 4 [Ref.: Lanes: 1]	1.9169	0.0726	1.8785	**1.7990**	**2.0348**
Road type: secondary [Ref.: Road type: primary]	-0.1094	0.1480	-0.1549	-0.3869	0.1682
Road type: tertiary [Ref.: Road type: primary]	-1.6389	0.1811	-1.6566	**-1.9854**	**-1.2924**
Road type: residential [Ref.: Road type: primary]	-2.5578	0.1358	-2.5842	**-2.8039**	**-2.3116**
Sigma-phi^2 [Spatially structured effects]	700.3172	93.2742	672.7877	**532.4443**	**868.1901**
Sigma-theta^2 [Spatially unstructured effects]	2.3455	0.2470	2.3362	**1.8810**	**2.8100**

Model evaluation metrics are shown on **Table 5-18**:

Table 5-18: Evaluation metrics for the training of the CAR model for harsh brakings

Metric	Value	Metric	Value
Data-points	869	RMSE	1.2830
DIC	1584.104	MAE	0.4115
WAIC	1589.478	RMSLE	0.1727
LMPL	-834.286	CA	96.32%

5.3.2.3 Discussion of results

Upon inspection, it is determined that the coefficient signs of statistically significant variables are mostly similar to those obtained from GWPR analysis. Specifically, road segment gradient and neighborhood complexity retain their negative signs, thus their increases contribute negatively to harsh braking occurrence. Conversely, the exposure variables of segment length and pass count retain their positive signs, contributing positively to harsh braking occurrence.

MEM can be again calculated following Washington et al. (2010), albeit without including any spatially structured or unstructured effects. For the segment length average of 144.3 m, an increase of 1 meter leads to an increase of $MEM_{Seg\_Length} = 0.0221$ harsh brakings. Respectively, for the pass count average of 191 passes, an increase of 1 pass leads to an increase of $MEM_{Pass\ count} = 0.0446$ harsh brakings. Again it seems that pass counts, which are more related to traffic variables and route choice lead to more harsh brakings per unit compared to segment lengths, which is a fixed geometrical segment characteristic.

The influence of each road type was found to be similar overall with previous results. This time there is no statistically significant contribution of secondary road segments to harsh braking frequencies, compared to the reference category of primary road segments. Tertiary and residential road segments are found to contribute less to harsh braking frequency, compared to primary segments, which is a reasonable and intuitive result.

All lane number categories show statistical significance in the CAR model. Two lanes continue to influence harsh events negatively compared to the one-lane baseline. This time, however, three lane segments also follow the negative trend, which is reversed at four lane segments, which appear to be statistically significant for the first time. This discrepancy is explained partly by different components in the Bayesian equation (in other words, varying variable mixes and different way of spatial effect integration) and partly by different modelling approaches (Bayesian – MCMC versus frequentist – MLE).

Pedestrian crossings were found to contribute positively to harsh braking occurrence in the CAR model. This finding indicates increased traffic conflicts in the proximity of pedestrian crossings, possibly with pedestrians. This in turn suggests that the study area crossings may feature poor visibility or functionality and pedestrians may appear to drivers unexpectedly, causing harsh braking events. It is interesting to note that pedestrian crossing presence has substituted that of traffic lights in the GWPR model as a significant variable. Parallel to GWPR, the removal of traffic lights from the CAR model led to much worse overall fits, in regards of DIC, WAIC, LMPL and error metrics.

The term ${\sigma_{\varphi i}}^2$, representing the standard deviation of the spatially structured effect distribution, also merits elaboration. While the value is considerably high, it is important to remember that this is not a regression coefficient. Rather, it represents a significant amount of variance fitted and explained by road segment-specific spatial effects of the CAR model. The variation is also more pronounced than the localized fluctuations of GWPR. This also ties with the very high CA of the model. In comparison, the term ${\sigma_{\theta i}}^2$, representing the variance of unstructured spatial effects is lower by two orders of magnitude, indicating that the effect of globally unobserved factors is significantly smaller for harsh braking occurrence. The intercept is likely to have absorbed several such effects as well.

CAR model outputs for the three spatial weighting alternatives were compared, and results were close in terms of coefficient values, variable significance and DIC values. The optimal model was found to involve binary neighbor weighting, in which for each segment all directly nearest neighbors are given a weight of 1 and the rest are given a weight of 0, while the closest one-third of the total study area is considered.

Performance metrics are calculated initially for the training dataset. The RMSE value suggests that the average magnitude of the error is about 1.3 harsh braking counts, while MAE is considerably lower at 0.4 harsh braking counts. RMSLE is also considerably lower than both metrics, as it has logarithmic properties. The custom accuracy (CA) value for the training dataset indicates that the CAR model correctly predicts harsh braking frequencies in the training area with a tolerance of ± 1 harsh braking per segment 96.3% of the times, which is a remarkable performance. This is attributed heavily to the meticulous calibration of spatial effects in the training area during the significant number of MCMC repetitions and subsequent Bayesian inference, which updates the uninformative posteriors based on the real data. Overall, the CAR harsh braking model displays considerably good goodness-of-fit metrics and low error related metrics.

5.3.2.4 Prediction and transferability capabilities

As stated in the methodology section, for CAR, as GWPR, the transferability of spatial effects is typically limited, because it requires prior knowledge of the distribution of the dependent variable in the test area. While in this particular case this knowledge exists, in a typical road safety problem it does not exist, especially in crash/event forecasting situations. In other words, to gain knowledge of the spatial effects in the test area, CAR models would have to be trained in the full dataset of the test area, which is not its intended purpose.

Moreover, CAR models do not contain a baseline (global) regression by default as is the case with GWPR models. Thus predictions are conducted by calibrating a Bayesian Poisson-lognormal model without spatial effects with the same specifications (410,000 iterations, 400,000 of which were burn-in and remainder 10,000 are thinned by 100 to reduce autocorrelation, and a fixed random generator value for replicability). The output appears on **Table 5-19**:

Table 5-19: Baseline Bayesian Poisson-lognormal model results for harsh braking predictions in urban road networks

Independent variables	Posterior values				
	Mean	St. Dev.	Median	2.5% value	97.5% value
Intercept	0.4820	0.3916	0.5177	1.3272	-0.3309
Gradient	-2.5163	0.6529	-2.4913	**-1.3484**	**-3.8121**
Neighborhood Complexity	-0.2890	0.0688	-0.2988	**-0.1559**	**-0.4021**
Segment length	0.0039	0.0001	0.0039	**0.0041**	**0.0036**
Pass count	0.0040	0.0002	0.0040	**0.0044**	**0.0037**
Traffic lights: Yes [Ref.: Traffic lights: No]	0.2415	0.1022	0.2493	**0.4198**	**0.0445**
Pedestrian crossing: Yes [Ref.: Ped. cross.: No]	-0.1471	0.0868	-0.1513	0.0241	-0.2752
Lanes: 2 [Ref.: Lanes: 1]	-0.2584	0.1063	-0.2597	**-0.0456**	**-0.4775**
Lanes: 3 [Ref.: Lanes: 1]	0.3394	0.1392	0.3335	**0.6155**	**0.0080**
Lanes: 4 [Ref.: Lanes: 1]	0.2780	0.2354	0.2857	0.7392	-0.2153
Road type: secondary [Ref.: Road type: primary]	1.0328	0.1139	1.0405	**1.2424**	**0.7668**
Road type: tertiary [Ref.: Road type: primary]	-0.0372	0.1513	-0.0297	0.3135	-0.3662
Road type: residential [Ref.: Road type: primary]	-1.0514	0.1774	-1.0657	**-0.6824**	**-1.4457**

As expected, without spatially structured and unstructured random effects, the baseline Bayesian Poisson-lognormal model results revert to being much closer to the baseline frequentist Poisson-lognormal model results shown on **Table 5-8** compared to CAR models in terms of coefficient values and variable significance. Predictions can now be conducted for the test area with the baseline Bayesian Poisson-lognormal model. The respective metrics of the predictions of the test set values from the model trained in the training set values are shown on **Table 5-20**.

Table 5-20: Evaluation metrics for predictions of the Bayesian Poisson-lognormal model for harsh brakings

Metrics	Value
RMSE	1.9804
MAE	1.0290
RMSLE	0.5520
CA	82.74%

True values in the test area dataset and the respective predictions are plotted on Figure 5-18; there are higher concentrations of lower frequencies noted with a bolder color from observation overlap. Predictions are conducted only for the 1066 road segments with non-zero trips, as pass count is required as input.

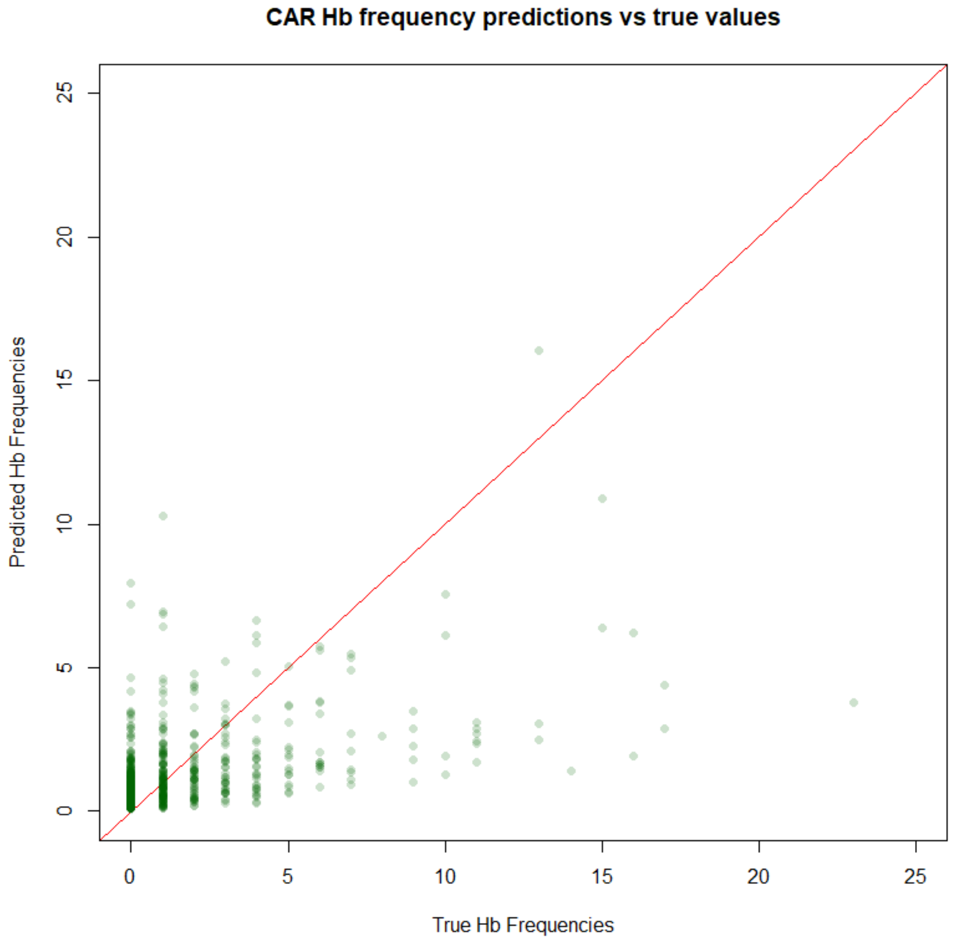


Figure 5-18: True and Bayesian GLM predicted frequencies of harsh brakings in Omonoia area

It appears that the removal of spatially structured and unstructured random effects has led to decreased error and CA metrics compared to the training area, another finding which is expected. A more interesting comparison is the one with the values are shown on **Table 5-11**: It appears that the Bayesian and frequentist Poisson-lognormal models are performing almost identically. On one hand, this is an expected finding, and a good quality check for these similar model predictions; on the other, this means that the two methods contribute similar values and are not expected to round out different discrepancies.

Harsh braking maps can be created for the predictions of the Bayesian GLM/CAR model in any of the two areas – Figure 5-19 shows the predictions in the test area.

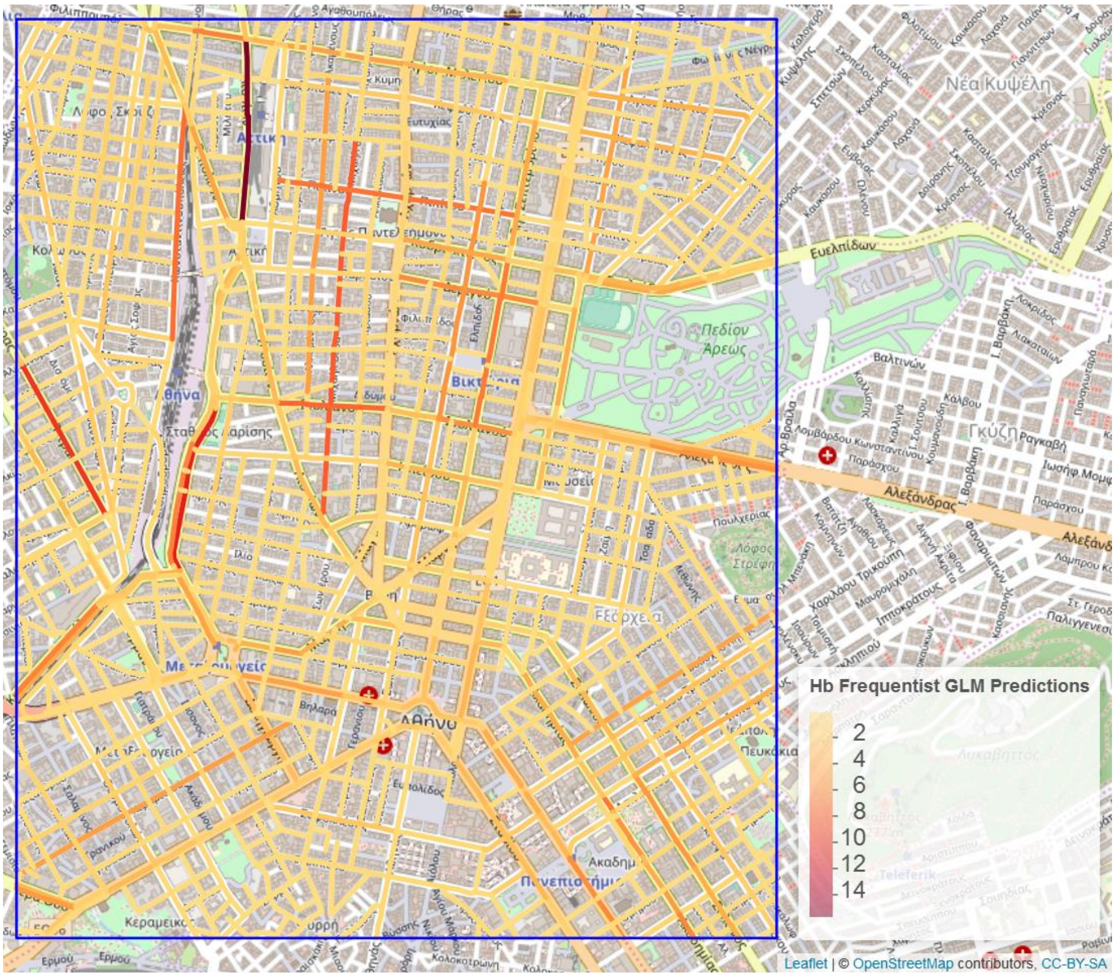

Hb Frequentist GLM Predictions
2
4
6
8
10
12
14
Leaflet | © OpenStreetMap contributors, CC-BY-SA

Figure 5-19: Bayesian GLM predicted harsh braking frequencies in Omonoia area

5.3.3 Harsh acceleration model

An equivalent process is followed for modelling harsh acceleration frequencies using CAR models.

5.3.3.1 Model preparation

The same weighting schemes as in the harsh braking model were examined, and the uninformative priors had equal values. Bayesian inference is then conducted using Markov Chain Monte Carlo (MCMC) simulation. It was found that the best performing models required a large burn-in period before stabilization, albeit smaller than the one for harsh brakings. After several trials, the posterior summaries for the best-fitting models were obtained by a chain with 110,000 iterations, the first 100,000 of which were discarded as the burn-in sample. The remainder 10,000 samples are thinned by 100 to reduce autocorrelation and the resulting values describe the posterior distributions. Computationally, a fixed value for the random number generation processes is also required to ensure the replicability of results.

5.3.3.2 Model presentation

CAR Models were calibrated following Lee (2013). The resulting final CAR model for harsh accelerations in urban road networks appears on **Table 5-21**. The 95% BCI values are calculated at 2.5% (lower bound) to 97.5% (upper bound), and median values refer to this 95% BCI margin only. 95% BCI values of statistically significant continuous variables and categorical variable categories – which retain the same signs – are shown in bold.

Table 5-21: CAR model results for harsh accelerations in urban road networks

Independent variables	**Posterior values**				
	Mean	St. Dev.	Median	2.5% value	97.5% value
Intercept	-1.2399	0.3158	-1.1892	**-1.7769**	**-0.7537**
Curvature	6.3926	2.9976	6.0723	**2.5188**	**10.5868**
Neighborhood Complexity	-0.2308	0.0650	-0.2594	**-0.3179**	**-0.1150**
Segment length	0.0038	0.0002	0.0038	**0.0035**	**0.0041**
Pass count	0.0071	0.0002	0.0072	**0.0068**	**0.0073**
Traffic lights: Yes [Ref.: Traffic lights: No]	0.1147	0.0469	0.4400	**0.4317**	**0.4946**
Pedestrian crossing: Yes [Ref.: Ped. cross.: No]	0.4554	0.0225	0.0978	**0.0543**	**0.1919**
Lanes: 2 [Ref.: Lanes: 1]	-0.0134	0.0653	-0.0476	-0.0800	0.0874
Lanes: 3 [Ref.: Lanes: 1]	-0.1702	0.0284	-0.1512	**-0.2134**	**-0.1459**
Lanes: 4 [Ref.: Lanes: 1]	0.4380	0.0608	0.4521	**0.3466**	**0.5154**
Road type: secondary [Ref.: Road type: primary]	0.7202	0.1480	0.7506	**0.5656**	**0.8443**
Road type: tertiary [Ref.: Road type: primary]	0.3610	0.0938	0.4303	**0.1287**	**0.5241**
Road type: residential [Ref.: Road type: primary]	-0.6715	0.1513	-0.6408	**-0.9129**	**-0.4609**
Sigma-phi^2 [Spatially structured effects]	255.3276	35.1953	259.2946	**203.6138**	**303.0744**
Sigma-theta^2 [Spatially unstructured effects]	0.2827	0.1259	0.2100	**0.1666**	**0.4714**

Model evaluation metrics are shown on **Table 5-22**:

Table 5-22: Evaluation metrics for the training of the CAR model for harsh accelerations

Metric	**Value**	**Metric**	**Value**
Data-points	869	RMSE	0.7961
DIC	1512.394	MAE	0.4111
WAIC	1544.994	RMSLE	0.2512
LMPL	-754.853	CA	95.74%

5.3.3.3 Discussion of results

The most obvious observation is that more variables are statistically significant in the CAR model for harsh accelerations compared to the GWPR model. Furthermore, it is determined that the coefficient signs of statistically significant variables are retained in variables that are common between the two models for the most part. Specifically, curvature, traffic lights and the exposure variables of segment length and pass count all retain their positive contribution towards harsh event frequency.

MEM can be again calculated following Washington et al. (2010), albeit without including any spatially structured or unstructured effects. For the segment length average of 144.3 m, an increase of 1 meter leads to an increase of $MEM_{Seg\_Length} = 0.0066$ harsh accelerations. Respectively, for the pass count average of 191 passes, an increase of 1 pass leads to an increase of $MEM_{Pass\ count} = 0.0276$ harsh accelerations. Again it seems that pass counts, which are more related to traffic variables and route choice lead to more harsh accelerations per unit compared to segment lengths.

The influence of road type categories are similar: Secondary and tertiary road categories display increased harsh accelerations compared to primary roads, while in residential roads the trend is reversed, and fewer harsh accelerations are observed.

The signs of lane numbers are different, with the CAR model output suggesting that three-lane segments show reduced harsh acceleration frequencies compared to one-lane segments, while on the other hand four-lane segments show increased harsh acceleration frequencies. The similarity with GWPR is that two-lane segments do not contribute any statistically significant effects. The most likely explanation for this discrepancy is the integration of part of the effect of lane number influence in the spatial effects term in the CAR model, which is very localized. Therefore, this finding suggests that the influence of lane number on harsh acceleration frequencies is unclear overall based on the current approach.

For the first time for harsh accelerations, pedestrian crossings were found to contribute positively to harsh acceleration occurrence. This could be explained by aggressive drivers accelerating to avoid being boxed in, in enclosed spaces created by pedestrians, obstacles and nearby traffic. Similarly, decreased neighborhood complexity seemed to provide less distraction sources to drivers, allowing them the temporal and focus margin to harshly accelerate.

As in the harsh braking CAR model, the term ${\sigma_{\varphi i}}^2$, representing the standard deviation of the spatially structured effect distribution shows a considerably high value. Therefore a significant amount of variance fitted and explained by road segment-specific spatial effects of the CAR model. The term ${\sigma_{\theta i}}^2$, representing the variance of unstructured spatial effects is lower by three orders of magnitude, indicates that the effect of globally unobserved factors is significantly smaller for harsh braking occurrence. The intercept is likely to have absorbed several such effects as well.

CAR model outputs for the three spatial weighting alternatives were compared, and results were close in terms of coefficient values, variable significance and DIC values. Once again, the optimal model was found to involve binary neighbor weighting, in which for each segment all directly nearest neighbors are given a weight of 1 and the rest are given a weight of 0, while the closest one-third of the total study area is considered.

Performance metrics are calculated initially for the training dataset. The RMSE value suggests that the average magnitude of the error is about 0.8 harsh acceleration counts, while MAE is considerably lower at 0.4 harsh acceleration counts. RMSLE is also considerably lower than both metrics, as it has logarithmic properties. The custom accuracy (CA) value for the training dataset indicates that the CAR model correctly predicts harsh acceleration frequencies in the training area with a tolerance of ± 1 harsh acceleration per segment 95.7% of the times. This is once again a remarkable performance, due to the chain convergence from several MCMC iterations which updates the uninformative posteriors based on the real data. Overall, the CAR acceleration braking model displays considerably good goodness-of-fit metrics and low error related metrics.

5.3.3.4 Prediction and transferability capabilities

Mirroring the process for harsh braking models, a Bayesian Poisson-lognormal model without spatial effects with the same specifications is calibrated (110,000 iterations, 100,000 of which were burn-in and remainder 10,000 are thinned by 100 to reduce autocorrelation, and a fixed random generator value for replicability). The output appears on **Table 5-23**:

Table 5-23: Baseline Bayesian Poisson-lognormal model results for harsh acceleration predictions in urban road networks

Independent variables	**Posterior values**				
	Mean	St. Dev.	Median	2.5% value	97.5% value
Intercept	-0.8460	0.4912	-1.7905	**-0.8409**	**0.1402**
Curvature	9.1757	2.6633	3.9392	**9.2673**	**14.3167**
Neighborhood Complexity	-0.1196	0.0938	-0.2939	-0.1254	0.0656
Segment length	0.0030	0.0002	0.0027	**0.0030**	**0.0033**
Pass count	0.0042	0.0002	0.0038	**0.0042**	**0.0046**
Traffic lights: Yes [Ref.: Traffic lights: No]	0.3729	0.1169	0.1320	**0.3762**	**0.5931**
Pedestrian crossing: Yes [Ref.: Ped. cross.: No]	-0.0190	0.1039	-0.2302	-0.0171	0.1741
Lanes: 2 [Ref.: Lanes: 1]	0.0703	0.1245	-0.1595	**0.0698**	**0.3110**
Lanes: 3 [Ref.: Lanes: 1]	0.4692	0.1719	0.1426	**0.4707**	**0.7930**
Lanes: 4 [Ref.: Lanes: 1]	0.3560	0.3177	-0.2781	**0.3728**	**0.9669**
Road type: secondary [Ref.: Road type: primary]	0.7465	0.1444	0.4637	**0.7465**	**1.0377**
Road type: tertiary [Ref.: Road type: primary]	0.3689	0.1755	0.0148	**0.3607**	**0.7321**
Road type: residential [Ref.: Road type: primary]	-0.6798	0.2117	-1.1078	**-0.6718**	**-0.2690**

As expected, without spatially structured and unstructured random effects, the baseline Bayesian Poisson-lognormal model results revert to being closer to the baseline frequentist Poisson-lognormal model results shown on **Table 5-13** compared to CAR models. This applies in terms of coefficient values and variable significance, i.e. neighborhood complexity and pedestrian crossings devolve again to not statistically significant variables. Predictions can now be conducted for the test area with the baseline Bayesian Poisson-lognormal model. The respective metrics of the predictions of the test set values from the model trained in the training set values are shown on **Table 5-24**:

Table 5-24: Evaluation metrics for predictions of the Bayesian Poisson-lognormal model for harsh accelerations

Metrics	**Value**
RMSE	1.6841
MAE	0.8700
RMSLE	0.5071
CA	87.62%

True values in the test area dataset and the respective predictions are plotted on Figure 5-20; there are higher concentrations of lower frequencies noted with a bolder color from observation overlap. Predictions are conducted only for the 1066 road segments with non-zero trips, as pass count is required as input.

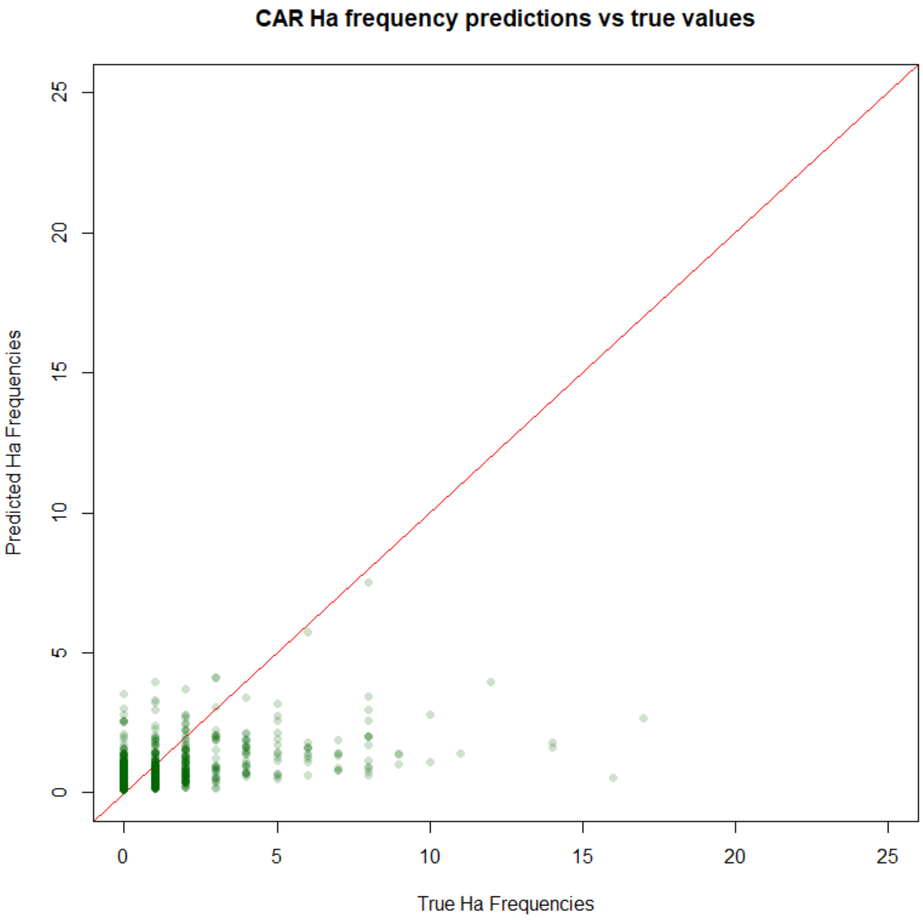


Figure 5-20: True and Bayesian GLM predicted frequencies of harsh accelerations in Omonoia area

It appears that the removal of spatially structured and unstructured random effects has led to decreased error and CA metrics compared to the training area, another finding which is expected. Similar to harsh braking models, and after comparison with the values of **Table 5-16**, it appears that once again the Bayesian and frequentist Poisson-lognormal models are performing almost identically. On one hand, this is an expected finding, and a good quality check for these similar model predictions; on the other, this means that the two methods contribute similar values and are not expected to round out different discrepancies.

Harsh acceleration maps can be created for the predictions of the Bayesian GLM/CAR model in any of the two areas – Figure 5-21 shows the predictions in the test area.

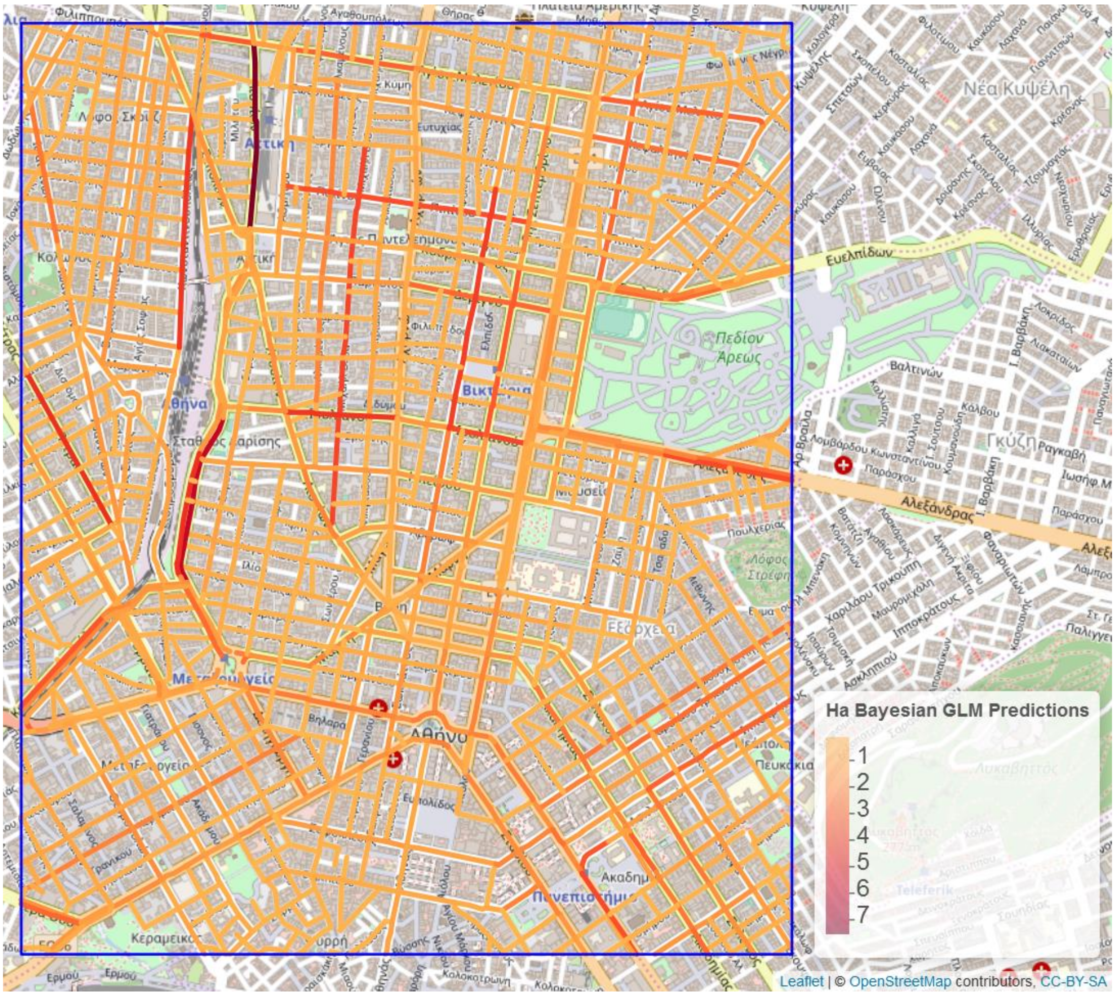


Figure 5-21: Bayesian GLM predicted harsh acceleration frequencies in Omonoia area

5.4 XGBoost algorithms

5.4.1 XGBoost selection process

In the rapidly advancing Machine Learning (ML) world, there are several algorithms and predictive techniques available for a variety of scientific problems. Several of them have been implemented in a road safety context (for instance, Lin et al., 2015; Katrakazas et al. 2017; Schratz et al. 2018). A reasonable question then arises: Why was XGBoost selected?

The main reason leading to XGBoost selection lies with the problem at hand and its frequency (or count-based) dependent variable. The road safety literature has long determined that generalized linear models (GLMs) such as Poisson and negative binomial models are appropriate for modelling the structure of such variables (Lord & Mannering, 2010).

However, to the knowledge of the author, and from the methodological research and literature review conducted for the present doctoral dissertation, very few ML algorithms that support count-based modelling are available to this date. Therefore, before exploring more sophisticated tools, this left the option of exploring popular ML algorithms with continuous variable structure, in other words, solving ML regression problems. Since the efficiency of all ML algorithms is directly dependent on the examined dataset, this exploratory phase had to be conducted after all the data collection and processing processes described in Section 4 had been completed – thus it is described at this stage.

Support Vector Machine (SVM) algorithms and Random Forest (RF) algorithms were considered. SVM and RF algorithms were both tested with the exact training and test datasets used as input for the previous GWPR and CAR models. The dependent variables that were considered was both harsh event frequencies – as would be inserted in a GLM – and harsh event rates, as described in Section 3.4.3.5.

The outcomes were very subpar, and similar to those described in Section 5.2.4: Almost no significant correlations were found between harsh event rates/frequencies and the explanatory independent variables. The predictive power of these algorithms was very weak (large error values and CA < 40%). More alarmingly, each algorithm made numerous predictions that were negative, which is impossible for event rates. As per the aforementioned, this is a sign of very poor model performance. Furthermore, logarithmic transformation that would counter these discrepancies was not possible due to several zero values of the dependent variables in the test and train datasets.

Overall, these results were not considered worthy of presentation or discussion in the context of this dissertation and were thus omitted, as was the theoretical background for SVM and RF algorithms. This process lead to the conclusion that SVM and RF algorithms are inadequate to analyze harsh events spatially per road segment in the present approach. As described in Section 2.1, these algorithms were used for classification tasks of spatial analyses in road safety, such as RF classification of hotspots (Jiang et al., 2016) and SVM crash injury severity prediction (Effati et al., 2015), and not for count-based modelling.

The next step was resorting to one of the few algorithms that have the capability of supporting proper GLM – Poisson regression. Algorithm efficiency and swiftness of calculation was another desired quality. In addition, the existence of spatial effects in harsh event frequencies was previously proven by GWPR and CAR model results, as well as global and local Moran's I examinations. Therefore it is fruitful to explore and compare results by implementing spatial cross-validation as described in Section 3.2.7.3, and in the relevant literature (Schratz et al. 2018; Lovelace et al., 2019).

Extreme Gradient Boosting (XGBoost) is a fast and efficient algorithm that fits the desired criteria. It has been shown to outclass other machine learning techniques, such as traditional classification and regression trees, both in road safety (Ting et al., 2020) and in other fields (Nielsen, 2016). Furthermore, XGBoost has been included in the popular 'mlr' package for R-studio (Bischl et al., 2016). This means that it could be easily modified to operate in a spatial analysis concept by allowing spatial cross-validation. For all these reasons, XGBoost was selected to augment GWPR/CAR model results.

The following sections present the calibration process of XGBoost algorithms and the respective results that these algorithms yield for harsh brakings and harsh accelerations. Initially, XGBoost algorithms are trained using traditional random cross-validation (RCV), followed by algorithms of spatial cross-validation (SPCV). All algorithms concern count-based modelling of harsh event frequencies, and were thus conducted with the Poisson cost function as described by Equation (67) in Section 3.2.7.2.

5.4.2 Harsh braking RCV XGBoost implementation

The present section presents the calibration process of the XGBoost algorithm and the respective results for harsh brakings using random cross-validation (RCV).

5.4.2.1 Hyperparameter tuning

The hyperparameter tuning process involved the determination of optimal parameter values for the training dataset. The optimized hyperparameters are those previously mentioned in Section 3.2.7.2: (i) learning rate, (ii) Gamma, (iii) maximum tree depth, (iv) evaluation metric, (v) number of rounds for cost function convergence. Results from hyperparameter optimization appear on **Table 5-25**:

Table 5-25: Hyperparameter optimization results for RCV XGBoost for harsh brakings

Hyperparameter	Examined range	Optimized Value
Learning rate	0.000 – 1.000	0.590
Gamma	0 – 100	0
Maximum tree depth	1 – 50	6
Evaluation metric	RMSE | RMSLE | MAE | Logloss | poisson-nloglik	RMSE
Number of rounds	1 – 1000	100

The number of k-folds for each cross-validation task was also investigated. Due to the limited – at least for machine learning standards – number of data-points (or rows of the dataset) which represent road segments in the training and test areas, it quickly became apparent that large values of k would be unrealistic and would lead to underfitting models. A decision was thus made to set $k = 5$ and thus conduct 5-fold RCV. The number of 5 folds was also retained for SPCV to allow for more straightforward comparisons between the models.

5.4.2.2 Result presentation

A 5-fold RCV XGBoost with the Poisson cost function was then trained on the training dataset following Bischl et al. (2016). The resulting feature (or independent variable) importance parameters found by executing RCV XGBoost after hyperparameter optimization for harsh braking frequencies are shown on **Table 5-26**:

Table 5-26: Feature importance of RCV XGBoost for harsh brakings

Independent variables – Features	Gain	Cover	Frequency
Pass count	0.6788	0.4090	0.2366
Segment length	0.1436	0.2031	0.2252
Gradient	0.0806	0.1079	0.2061
Curvature	0.0444	0.0913	0.1412
Neighborhood complexity	0.0344	0.1403	0.1298
Lane number	0.0072	0.0275	0.0191
Road type	0.0049	0.0028	0.0191
Traffic lights	0.0037	0.0075	0.0153
Pedestrian crossing	0.0024	0.0107	0.0076

Algorithm evaluation metrics are shown on **Table 5-27** for the training dataset.

Table 5-27: Evaluation metrics for the training of RCV XGBoost for harsh brakings

Metric	Value
RMSE	1.4215
MAE	0.4971
RMSLE	0.3140
CA	90.56%

5.4.2.3 Discussion of results

Similarly with the previous statistical approaches, XGBoost draws a significant amount of information gain from the two exposure variables, namely pass count and segment length. The calculation of feature importance showcases that pass count is significantly more informative than road segment length. Most of the other included independent variables also contribute to the creation of the rules of the XGBoost tree ensemble, with the exception of road direction. Overall, these results are consistent with the outputs of the CAR and GWPR statistical methods.

However, the limited interpretability of results is obvious, as it is not feasible to investigate the isolated effect of parameters or the manner in which they split the ensemble of XGBoost trees. Furthermore, there is no specific inclusion or otherwise investigation of spatial effects affecting the data at this stage, and no conclusions can be drawn for any spatial dependence present solely from the algorithm.

Regarding the three error metrics and custom accuracy (CA), XGBoost features a good performance on the training dataset, ranking better than GWPR but lower than Bayesian CAR models. These metrics indicate a good model fit.

5.4.2.4 Prediction and transferability capabilities

The trained RCV XGBoost algorithm for harsh brakings is readily applied to the test dataset without any modification or recalibration, thus showcasing a strength of ML methods. True values in the test area dataset and the respective predictions are plotted on Figure 5-22; there are higher concentrations of lower frequencies noted with a bolder color from observation overlap. Predictions are conducted only for the 1066 road segments with non-zero trips, as pass count is required as input.

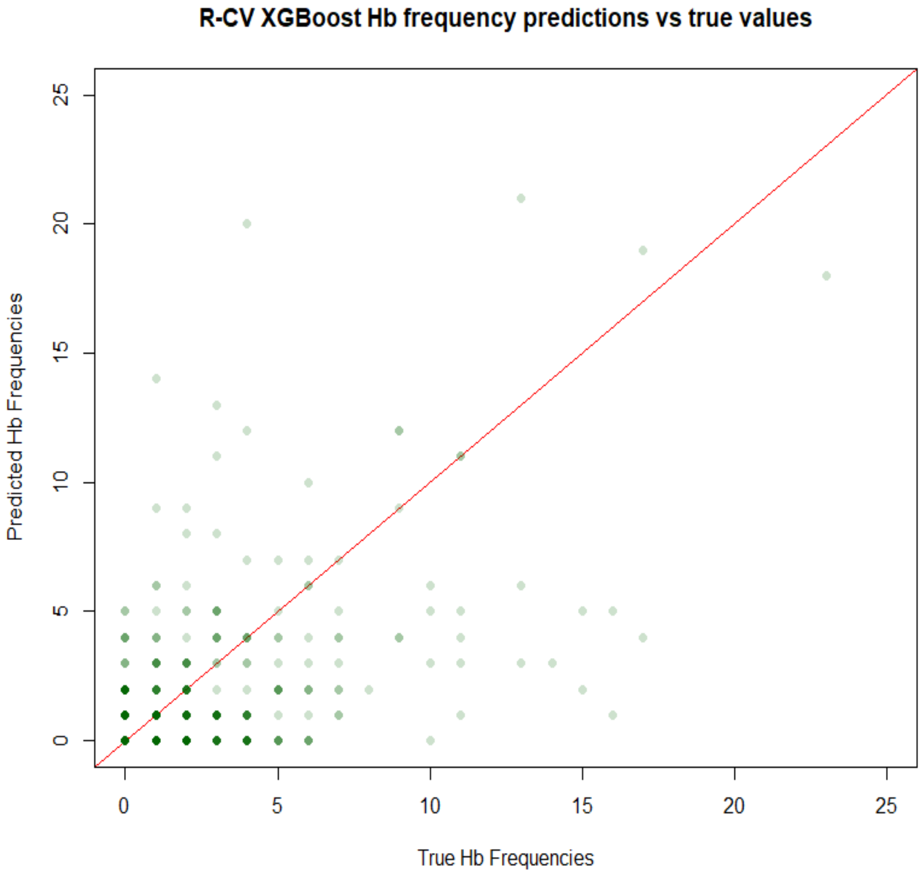


Figure 5-22: True and RCV XGBoost predicted frequencies of harsh brakings in Omonoia area

Evaluation metric results appear on **Table 5-28** for the test dataset.

Table 5-28: Evaluation metrics for predictions of RCV XGBoost for harsh brakings

Metric	Value
RMSE	1.9834
MAE	0.8415
RMSLE	0.5484
CA	83.40%

For the test dataset, error metrics are slightly elevated compared to those of the training dataset, and there is a slight drop in CA, as expected. A more interesting comparison is the one with the metrics of **Table 5-11** and **Table 5-20**: Compared to frequentist and Bayesian Poisson-lognormal models, RCV XGBoost yields about 18% lower MAE and slightly better CA for harsh braking frequency prediction. RMSE and RMSLE values are almost identical between the three methods. A similar visual comparison of Figure 5-22 with Figure 5-13 and Figure 5-18 shows that RCV XGBoost predicted harsh braking values have more symmetric dispersion when plotted against the true harsh braking values.

Harsh braking maps can be created for the predictions of RCV XGBoost in any of the two areas – Figure 5-23 shows the predictions in the test area.

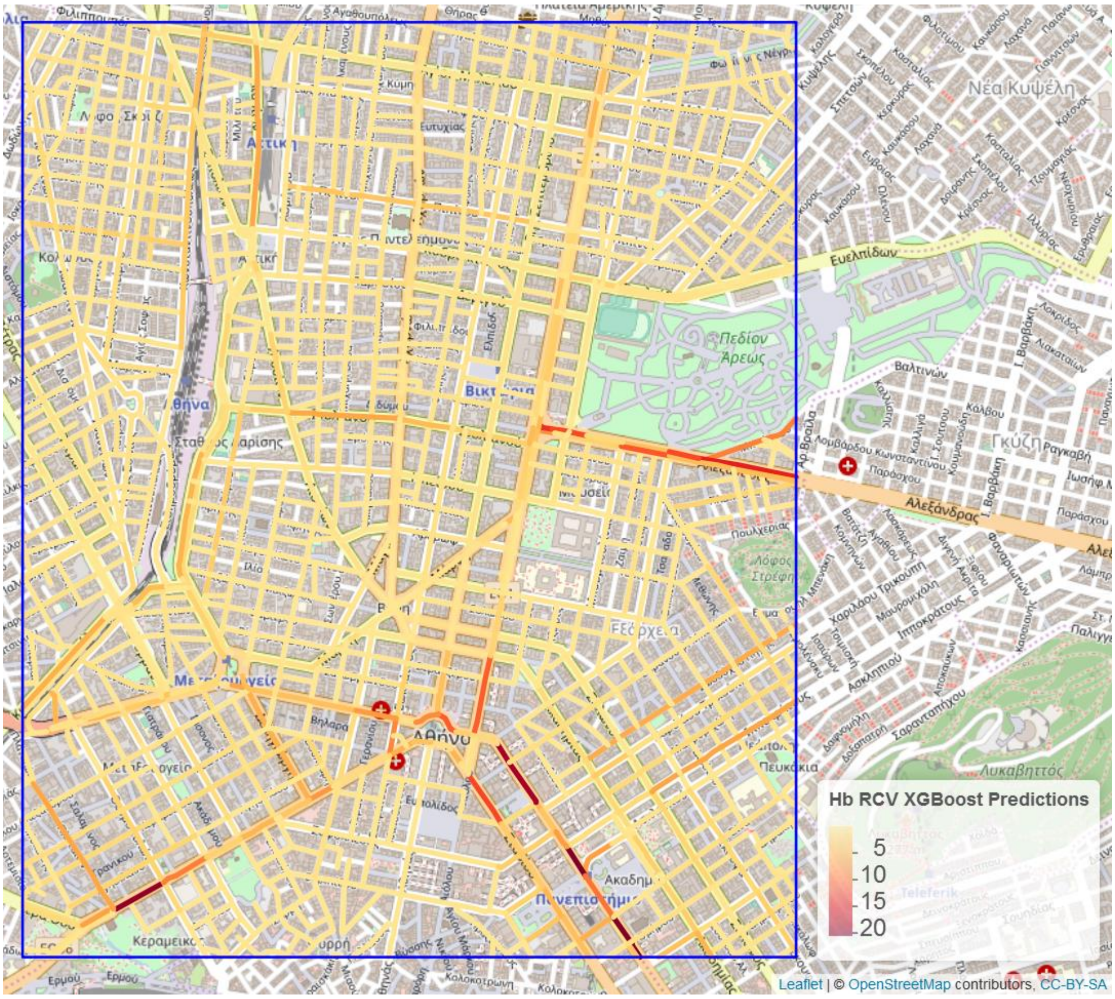


Figure 5-23: RCV XGBoost predicted harsh braking frequencies in Omonoia area

5.4.3 Harsh acceleration RCV XGBoost implementation

The present section presents the calibration process of the XGBoost algorithm and the respective results for harsh acceleration using RCV.

5.4.3.1 Hyperparameter tuning

The hyperparameter tuning process involved the determination of optimal parameter values for the training dataset. The optimized hyperparameters are those previously mentioned in Section 3.2.7.2: (i) learning rate, (ii) Gamma, (iii) maximum tree depth, (iv) evaluation metric, (v) number of rounds for cost function convergence. Results from hyperparameter optimization appear on **Table 5-29**:

Table 5-29: Hyperparameter optimization results for RCV XGBoost for harsh accelerations

Hyperparameter	Examined range	Optimized Value
Learning rate	0.000 – 1.000	0.400
Gamma	0 – 100	0
Maximum tree depth	1 – 50	6
Evaluation metric	RMSE | RMSLE | MAE | Logloss | poisson-nloglik	RMSE
Number of rounds	1 – 1000	100

The number of k-folds was retained to 5 for consistency across all RCV and SPCV models for harsh accelerations as well.

5.4.3.2 Result presentation

A 5-fold RCV XGBoost with the Poisson cost function was then trained on the training dataset following Bischl et al. (2016). The resulting feature (or independent variable) importance parameters found by executing RCV XGBoost after hyperparameter optimization for harsh acceleration frequencies are shown on **Table 5-30**:

Table 5-30: Feature importance of RCV XGBoost for harsh accelerations

Independent variables – Features	Gain	Cover	Frequency
Pass count	0.7184	0.4344	0.1946
Segment length	0.1058	0.2354	0.2865
Gradient	0.0588	0.1271	0.1784
Neighborhood complexity	0.0532	0.0912	0.1541
Curvature	0.0323	0.0752	0.1108
Road type	0.0109	0.0119	0.0189
Traffic lights	0.0069	0.0059	0.0189
Road direction	0.0060	0.0046	0.0162
Pedestrian crossing	0.0045	0.0066	0.0108
Lane Number	0.0033	0.0077	0.0108

Algorithm evaluation metrics are shown on **Table 5-31** for the training dataset.

Table 5-31: Evaluation metrics for the training of RCV XGBoost for harsh accelerations

Metric	Value
RMSE	0.9128
MAE	0.3728
RMSLE	0.3000
CA	93.32%

5.4.3.3 Discussion of results

As with harsh brakings, XGBoost draws a significant amount of information gain for harsh accelerations from the two exposure variables, namely pass count and segment length, and again the calculation of feature importance showcases that pass count is significantly more informative than road segment length. For the first time in all urban network analyses, road direction offers some – comparably small – amount of information on the explanation of harsh event frequencies. Therefore all independent variables are used as informative to an extent. This discrepancy is explained by differences in XGBoost training, which aims to minimize the cost function, with Bayesian (MCMC) and frequentist (MLE) approaches.

As previously explained, the limited interpretability of results is obvious, as it is not feasible to investigate the isolated effect of parameters or the manner in which they split the ensemble of XGBoost trees. Furthermore, there is no specific inclusion or otherwise investigation of spatial effects affecting the data at this stage, and no conclusions can be drawn for any spatial dependence present solely from the algorithm.

A noteworthy observation is that the trend in error metrics and CA persists for harsh accelerations as well. By comparison of XGBoost results with **Table 5-15** and **Table 5-22**, it appears that XGBoost features a good performance on the training dataset, ranking better than GWPR but lower than Bayesian CAR models overall. Overall, the error values are quite low and these metrics indicate a good model fit.

5.4.3.4 Prediction and transferability capabilities

The trained RCV XGBoost algorithm for harsh accelerations is readily applied to the test dataset without any modification or recalibration. True values in the test area dataset and the respective predictions are plotted on Figure 5-24; there are higher concentrations of lower frequencies noted with a bolder color from observation overlap. Predictions are conducted only for the 1066 road segments with non-zero trips, as pass count is required as input.

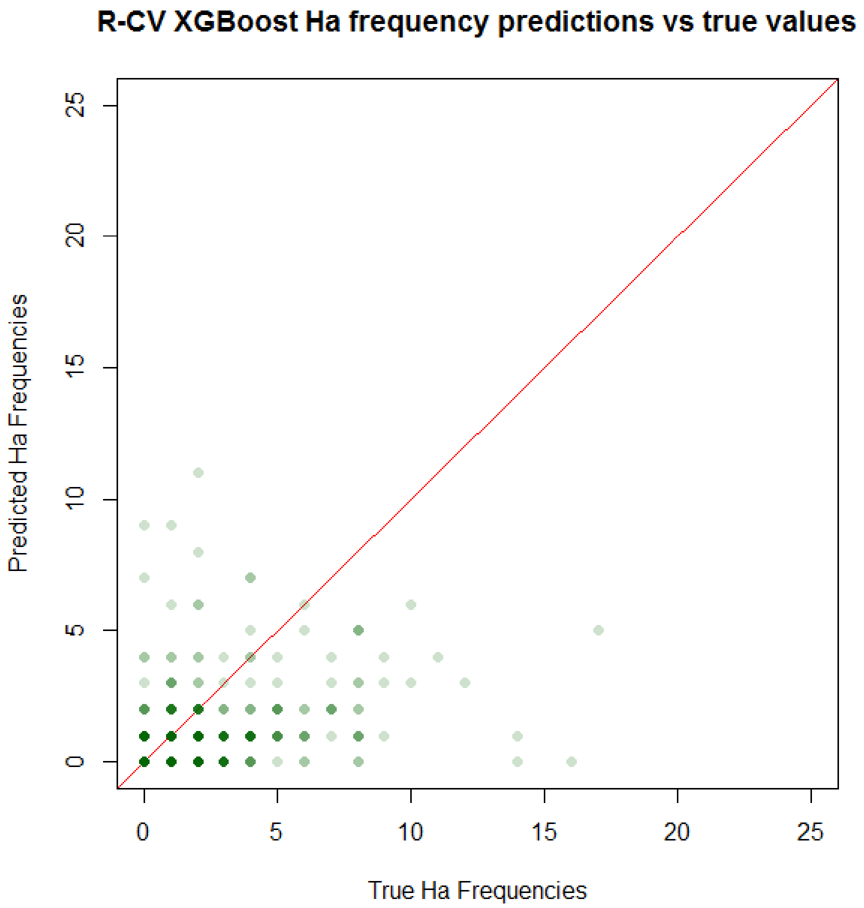
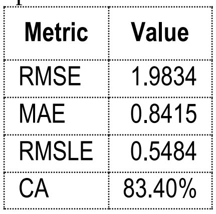


Figure 5-24: True and RCV XGBoost predicted frequencies of harsh accelerations in Omonoia area

Evaluation metric results appear on **Table 5-32** for the test dataset.

Table 5-32: Evaluation metrics for predictions of RCV XGBoost for harsh accelerations



Metric	Value
RMSE	1.9834
MAE	0.8415
RMSLE	0.5484
CA	83.40%

For the test dataset, error metrics are elevated compared to those of the training dataset, and there is a modest drop in CA. A more interesting comparison is the one with the metrics of **Table 5-16** and **Table 5-24**. Compared to frequentist and Bayesian Poisson-lognormal models, RCV XGBoost yields about 4% lower MAE. However, all other metrics are worse, and there is a 4% reduction in RCV XGBoost CA. The outcome is the opposite of the one obtained for harsh brakings, revealing the fact that all three methods are performing comparatively and none clearly outclasses the others. Once again, RCV XGBoost yields a more symmetric dispersion of predicted vs. true harsh acceleration values, as shown in Figure 5-24, compared to the previous results of Figure 5-16 and Figure 5-20.

Harsh acceleration maps can be created for the predictions of RCV XGBoost in any of the two areas – Figure 5-25 shows the predictions in the test area.

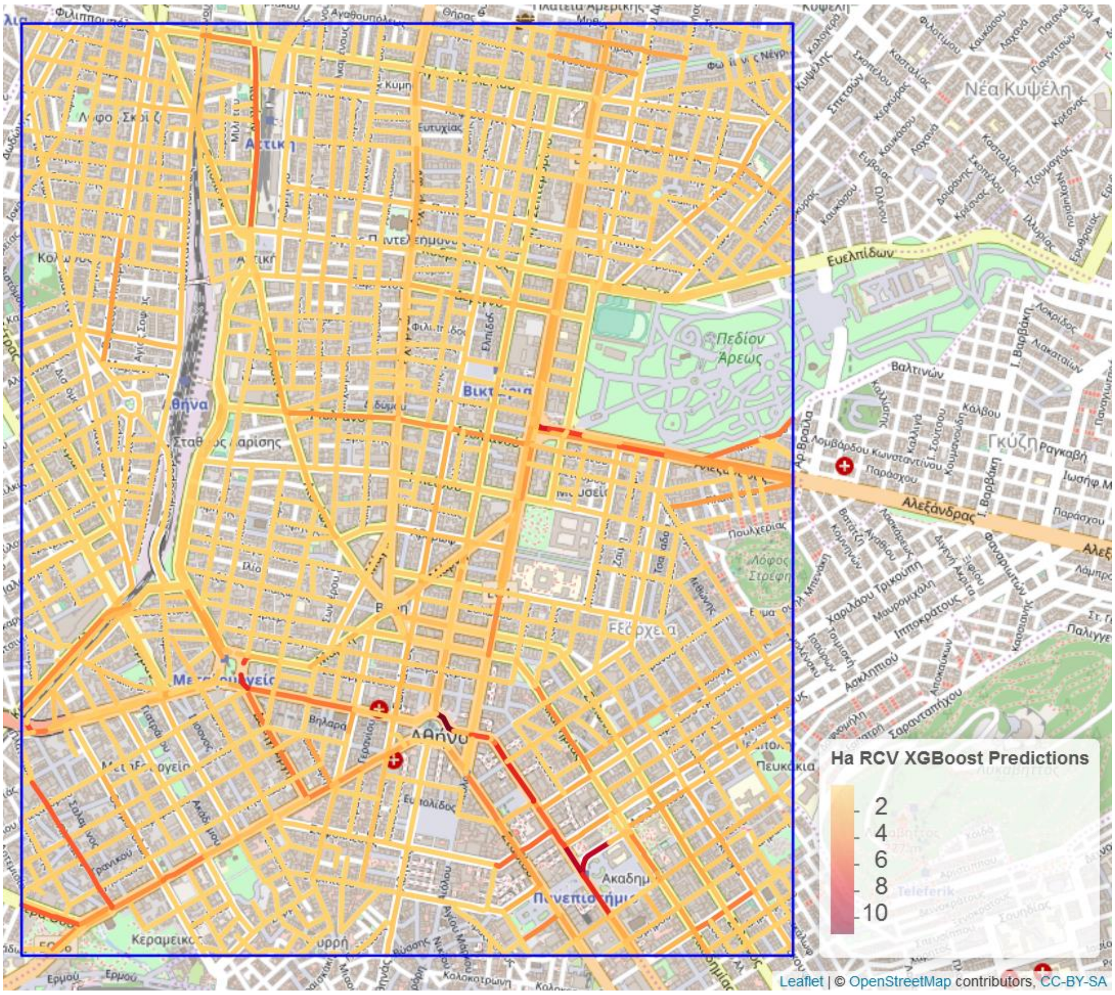


Figure 5-25: RCV XGBoost predicted harsh acceleration frequencies in Omonoia area

5.4.4 Harsh braking SPCV XGBoost implementation

XGBoost algorithms are now trained with spatial cross-validation (SPCV), as presented in Lovelace et al. (2019) for the prediction of harsh braking frequencies.

5.4.4.1 Hyperparameter tuning

The hyperparameter tuning process involved the determination of optimal parameter values for the training dataset. The optimized hyperparameters are those previously mentioned in Section 3.2.7.2: (i) learning rate, (ii) Gamma, (iii) maximum tree depth, (iv) evaluation metric, (v) number of rounds for cost function convergence. Results from hyperparameter optimization appear on **Table 5-33**:

Table 5-33: Hyperparameter optimization results for SPCV XGBoost for harsh brakings

Hyperparameter	Examined range	Optimized Value
Learning rate	0.000 – 1.000	0.300
Gamma	0 – 100	3.81
Maximum tree depth	1 – 50	8
Evaluation metric	RMSE | RMSLE | MAE | Logloss | poisson-nloglik	RMSE
Number of rounds	1 – 1000	72

As mentioned in Section 3.2.7.3, k-folds refer to spatial folds for SPCV. In other words, observations are not split randomly into k equal subsets. Rather, k equal-sized neighborhood folds are formed in order to preserve any spatial effects, representing local traits and road network particularities that are inherently expressed in the data. The number of k neighborhood folds was retained to 5 for consistency across all RCV and SPCV models for harsh brakings.

5.4.4.2 Result presentation

A 5-fold SPCV XGBoost with the Poisson cost function was then trained on the training dataset following Bischl et al. (2016) and Lovelace et al. (2019). The resulting feature (or independent variable) importance parameters found by executing SPCV XGBoost after hyperparameter optimization for harsh braking frequencies are shown on **Table 5-34**:

Table 5-34: Feature importance of SPCV XGBoost for harsh brakings

Independent variables – Features	Gain	Cover	Frequency
Pass count	0.6271	0.3813	0.2201
Segment length	0.1400	0.2222	0.2117
Gradient	0.0860	0.1257	0.1761
Neighborhood complexity	0.0684	0.1467	0.1929
Curvature	0.0626	0.0883	0.1572
Road type	0.0078	0.0048	0.0189
Lane number	0.0048	0.0245	0.0147
Pedestrian crossing	0.0024	0.0065	0.0063
Traffic lights	0.0010	0.0001	0.0021

Algorithm evaluation metrics are shown on **Table 5-35** for the training dataset.

Table 5-35: Evaluation metrics for the training of SPCV XGBoost for harsh brakings

Metric	Value
RMSE	1.8293
MAE	0.4994
RMSLE	0.2390
CA	91.71%

5.4.4.3 Discussion of results

XGBoost with spatial cross-validation yields overall outputs that are comparable with those of random cross-validation for harsh brakings. The two exposure variables are the most crucial for information gain for the training of the tree ensemble. Most of the other included independent variables also contribute to the creation of the rules of the XGBoost tree ensemble, with the exception of road direction. These results are overall consistent with the outputs all previous methods, albeit with the limited ML interpretability. A noteworthy fact is that the change of the cross-validation method led to the creation of a different tree ensemble in terms of ranking of feature importance as expressed by Gain. SPCV caused neighborhood complexity to emerge as more informative, for instance, and led to a reordering of variable rankings for all variables apart from the ones expressing exposure and gradient.

Compared to the results of RCV XGBoost for harsh brakings, higher RMSE values are observed, contrasted by lower RMSLE values, while MAE performance is unchanged. This is interpreted as SPCV XGBoost being a less conservative model outcome in the training area regarding segments with higher harsh braking counts (higher RMSE), but achieving better fit in lower values (lower RMSLE). The latter category outweighs the former, as denoted by the slightly elevated CA despite the higher RMSE.

5.4.4.4 Prediction and transferability capabilities

The trained SPCV XGBoost algorithm for harsh brakings is readily applied to the test dataset without any modification or recalibration, again showcasing a strength of ML methods. True values in the test area dataset and the respective predictions are plotted on Figure 5-26; there are higher concentrations of lower frequencies noted with a bolder color from observation overlap. Predictions are conducted only for the 1066 road segments with non-zero trips, as pass count is required as input.

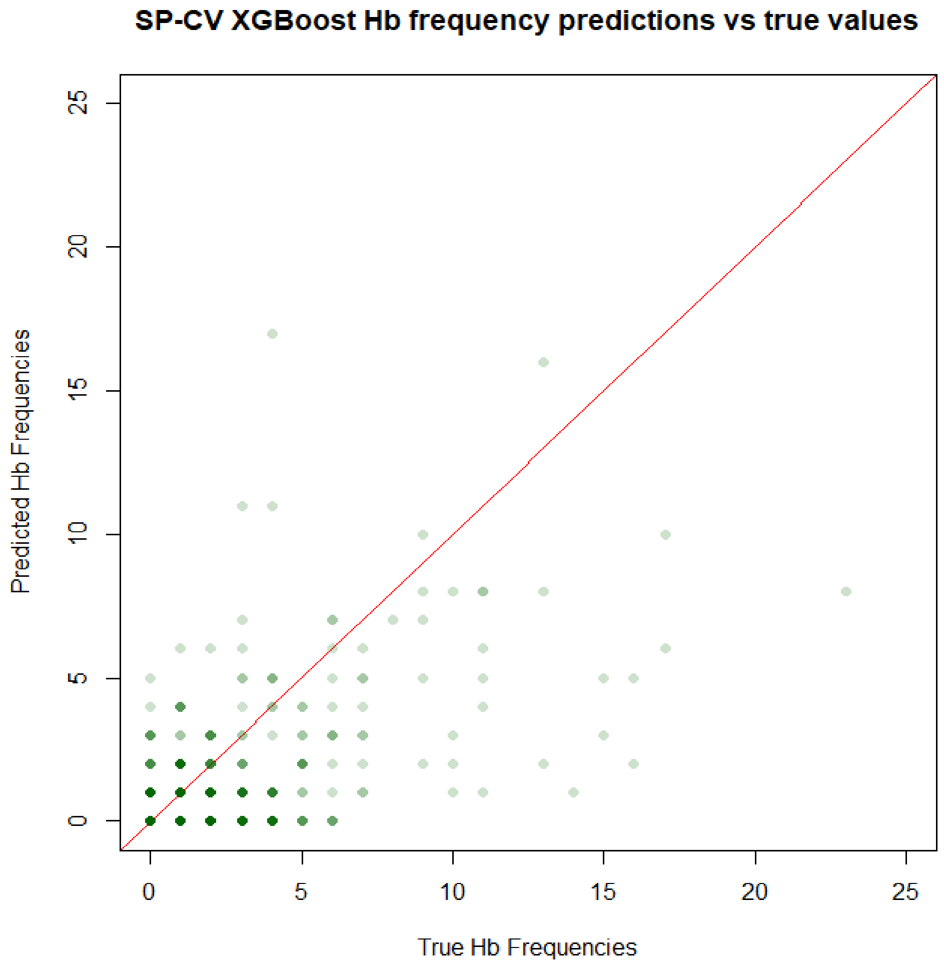


Figure 5-26: True and SPCV XGBoost predicted frequencies of harsh brakings in Omonoia area

Evaluation metric results appear on **Table 5-36** for the test dataset.

Table 5-36: Evaluation metrics for predictions of SPCV XGBoost for harsh brakings

Metric	Value
RMSE	1.8418
MAE	0.7542
RMSLE	0.5189
CA	85.27%

For the test dataset, error metrics are slightly elevated compared to those of the training dataset, and there is a modest drop in CA. A more interesting comparison is the one with the metrics of **Table 5-11**, **Table 5-20** and **Table 5-28**: Compared to frequentist and Bayesian Poisson-lognormal models, as well as RCV XGBoost, SPCV XGBoost is found to be the most accurate method yet, with lower error metrics and higher CA scores. Indicatively, SPCV XGBoost MAE is found to be about 10% lower compared to the second-lowest value belonging to RCV XGBoost. A visual comparison of Figure 5-26 with Figure 5-13, Figure 5-18 and Figure 5-22 shows that SPCV XGBoost predicted harsh braking values are slightly closer to the diagonal than RCV XGBoost while retaining the desirable increase in symmetric dispersion when plotted against the true harsh braking values.

The explanation for the improvement of the ensemble tree performance lies with the structural splits of the data employed during cross-validation and then used by the algorithm as input. By retaining geographical proximity in favor of random splits, more meaningful separations can be made in the XGBoost tree structures. While the algorithm does not 'remember' or transfer the spatial effects of a specific training area to a different test area, it allows for different spatial effects to manifest for each neighborhood by acknowledging a geographical structure – in other words, relative locations – in the area.

Harsh braking maps can be created for the predictions of SPCV XGBoost in any of the two areas – Figure 5-27 shows the predictions in the test area.

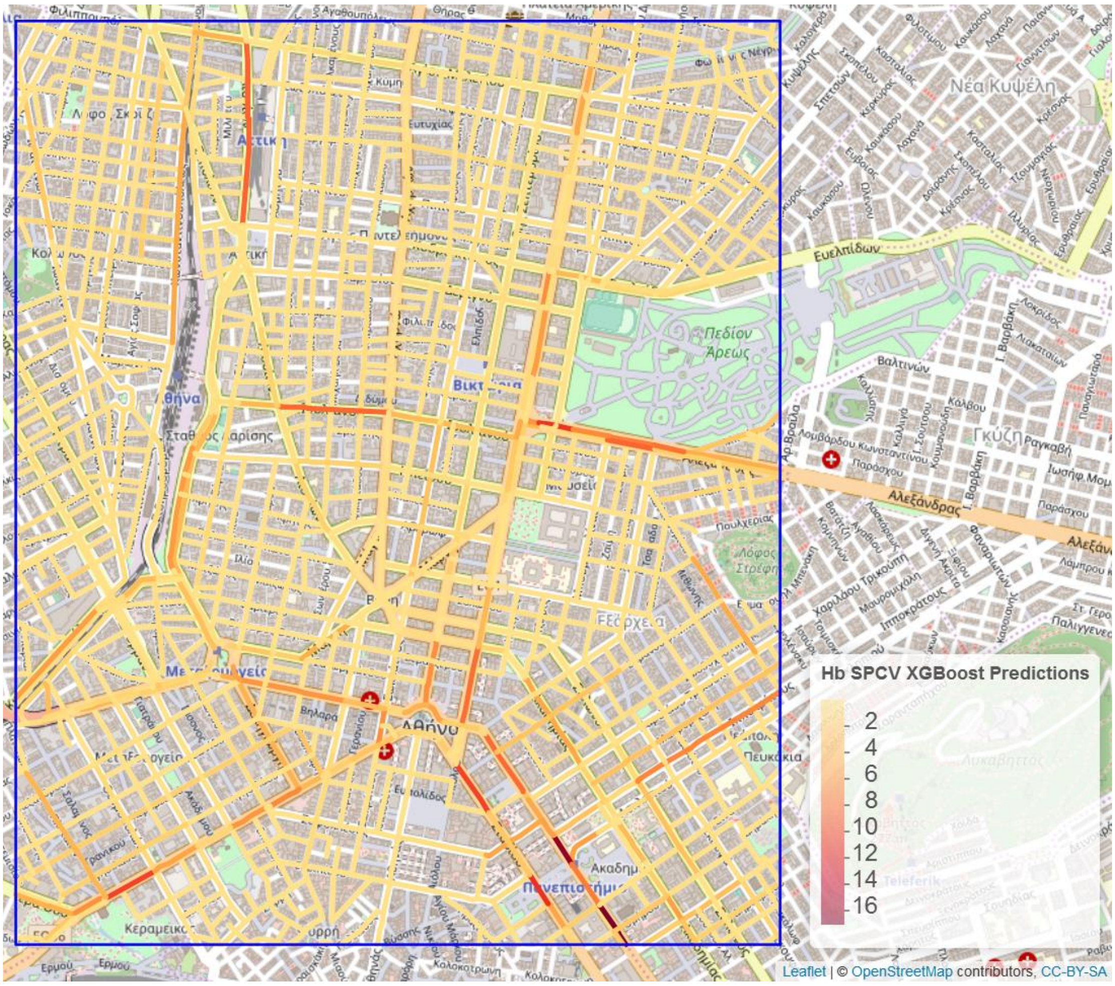


Figure 5-27: SPCV XGBoost predicted harsh braking frequencies in Omonoia area

5.4.5 Harsh acceleration SPCV XGBoost implementation

In the last analysis for urban networks, XGBoost algorithms are now trained with spatial cross-validation (SPCV), as presented in Lovelace et al. (2019) for the prediction of harsh acceleration frequencies.

5.4.5.1 Hyperparameter tuning

The hyperparameter tuning process involved the determination of optimal parameter values for the training dataset. The optimized hyperparameters are those previously mentioned in Section 3.2.7.2: (i) learning rate, (ii) Gamma, (iii) maximum tree depth, (iv) evaluation metric, (v) number of rounds for cost function convergence. Results from hyperparameter optimization appear on **Table 5-37**:

Table 5-37: Hyperparameter optimization results for SPCV XGBoost for harsh accelerations

Hyperparameter	Examined range	Optimized Value
Learning rate	0.000 – 1.000	0.200
Gamma	0 – 100	3.94
Maximum tree depth	1 – 50	4
Evaluation metric	RMSE | RMSLE | MAE | Logloss | poisson-nloglik	RMSE
Number of rounds	1 – 1000	217

The number of k neighborhood folds was retained to 5 for consistency across all RCV and SPCV models for harsh accelerations as well.

5.4.5.2 Result presentation

A 5-fold SPCV XGBoost with the Poisson cost function was then trained on the training dataset following Bischl et al. (2016) and Lovelace et al. (2019). The resulting feature (or independent variable) importance parameters found by executing SPCV XGBoost after hyperparameter optimization for harsh acceleration frequencies are shown on **Table 5-38**:

Table 5-38: Feature importance of SPCV XGBoost for harsh accelerations

Independent variables – Features	Gain	Cover	Frequency
Pass count	0.8253	0.5050	0.2926
Segment length	0.0766	0.1869	0.2394
Neighborhood complexity	0.0355	0.0795	0.1436
Curvature	0.0309	0.0698	0.1117
Gradient	0.0189	0.1087	0.1223
Road type	0.0065	0.0110	0.0372
Lane number	0.0027	0.0327	0.0213
Traffic lights	0.0026	0.0051	0.0213
Pedestrian crossing	0.0011	0.0012	0.0106

Algorithm evaluation metrics are shown on **Table 5-39** for the training dataset.

Table 5-39: Evaluation metrics for the training of SPCV XGBoost for harsh accelerations

Metric	Value
RMSE	1.1327
MAE	0.4891
RMSLE	0.3504
CA	89.87%

5.4.5.3 Discussion of results

For harsh accelerations, SPCV XGBoost yields results that are slightly worse compared to RCV XGBoost. Consistently with all previous XGBoost models, the two exposure variables of pass count and segment length are the most crucial for information gain for the training of the tree ensemble. These results are overall consistent with the outputs all previous methods, albeit with the limited ML interpretability. Once again, SPCV led to the creation of a different tree ensemble in terms of ranking of feature importance as expressed by Gain compared to RCV. SPCV caused neighborhood complexity to emerge as more informative, and led to a reordering of variable rankings for all variables apart from the ones expressing exposure.

Compared to RCV XGBoost for harsh accelerations, there is a slight increase across all three error metrics, and a drop of about 4% in CA values. A notable difference with RCV XGBoost is that the variable of road direction does not contribute to the algorithm any statistically significant information that would explain harsh acceleration frequency variance.

5.4.5.4 Prediction and transferability capabilities

The trained SPCV XGBoost algorithm for harsh accelerations is readily applied to the test dataset without any modification or recalibration. True values in the test area dataset and the respective predictions are plotted on Figure 5-28; there are higher concentrations of lower frequencies noted with a bolder color from observation overlap. Predictions are conducted only for the 1066 road segments with non-zero trips, as pass count is required as input.

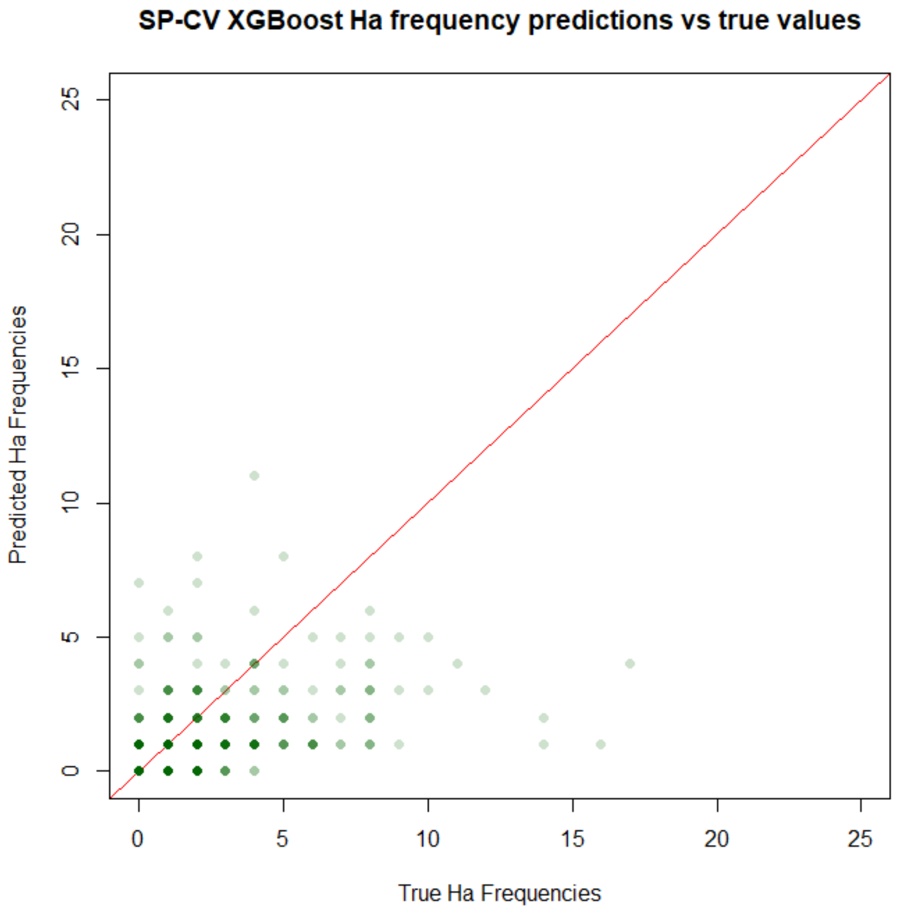
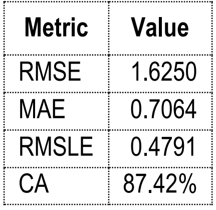


Figure 5-28: True and SPCV XGBoost predicted frequencies of harsh accelerations in Omonoia area

Evaluation metric results appear on **Table 5-40** for the test dataset.

Table 5-40: Evaluation metrics for predictions of SPCV XGBoost for harsh accelerations



Metric	Value
RMSE	1.6250
MAE	0.7064
RMSLE	0.4791
CA	87.42%

After application on the test area dataset it is found that error metrics are slightly elevated compared to those of the training dataset. However, there is only a slight drop in CA when comparing the training area with the test area. Overall, SPCV XGBoost performance is very comparable with the frequentist and Bayesian Poisson regression models and somewhat better than RCV XGBoost with a 4% improvement in CA. Compared to the metrics of **Table 5-16**, **Table 5-24** and **Table 5-32**, it is determined that SPCV XGBoost displays the lowest RMSE, MAE and RMSLE for harsh accelerations.

As shown in Figure 5-28, compared to the previous results of Figure 5-16, Figure 5-20 and Figure 5-24, SPCV XGBoost yields a more symmetric dispersion of predicted vs. true harsh acceleration values compared to the frequentist and Bayesian Poisson regression models, while it is comparable to RCV XGBoost.

Harsh acceleration maps can be created for the predictions of SPCV XGBoost in any of the two areas – Figure 5-29 shows the predictions in the test area.

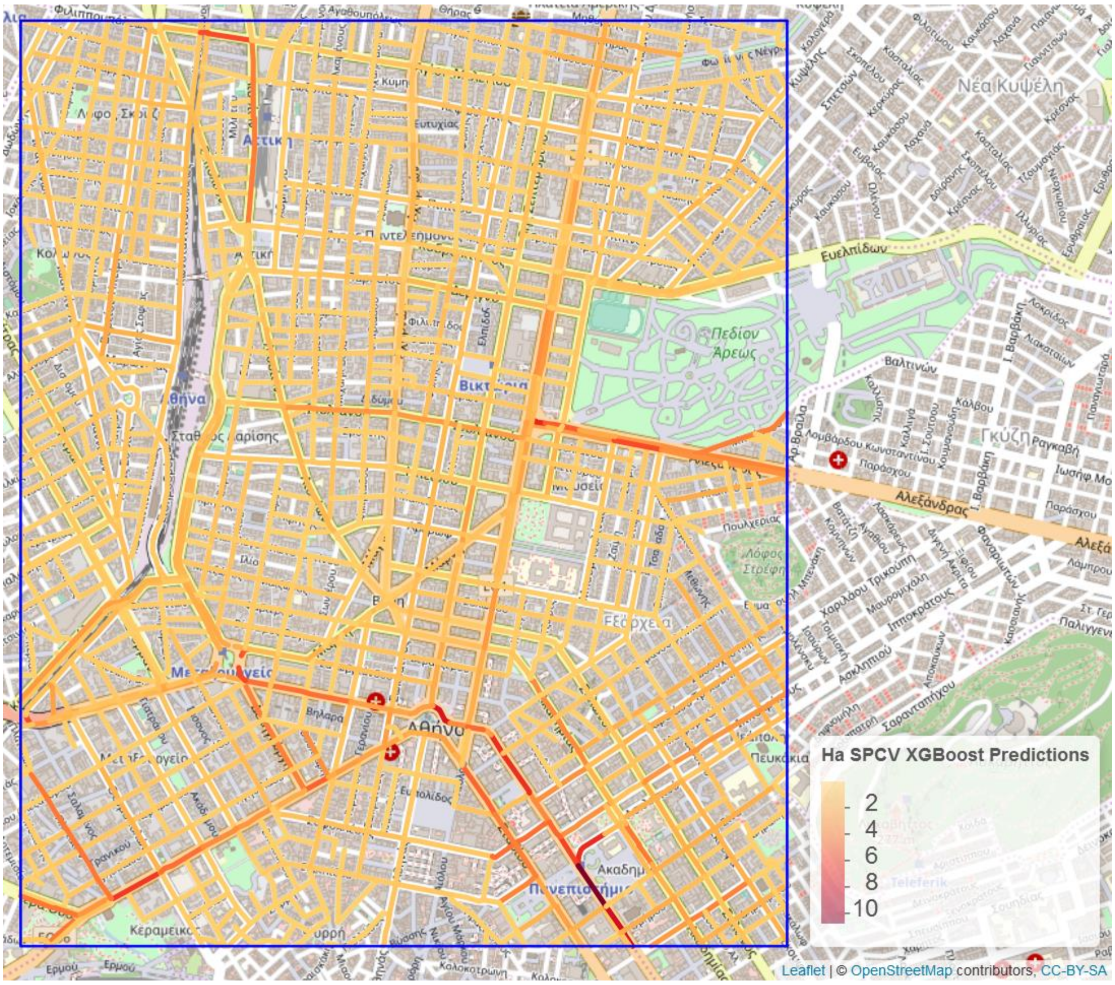

Ha SPCV XGBoost Predictions
2
4
6
8
10
Νέα Κυψέλη
Ευελπίδων
Πεδίον Άρεως
Γκύζη
Αλεξάνδρας
Αθήνα
Leaflet | © OpenStreetMap contributors, CC-BY-SA

Figure 5-29: SPCV XGBoost predicted harsh acceleration frequencies in Omonoia area

5.5 Overall urban network results

5.5.1 Combined prediction maps

Based on the overall findings of Section 5, it is evident that no method clearly outclasses the others for the prediction of harsh event frequencies per segment. Since the frequentist and Bayesian predictions are conducted with their base aspatial Poisson lognormal GLMs, and without spatial effects, it is obvious they stem from a related framework. Therefore it is decided to include the predictions of both RCV and SPCV XGBoost to have equal weighting amongst the four – turned two – predictions of harsh events when transferring in another area.

In other words, the combined predictions are obtained by averaging the predictions of the developed (i) Frequentist GLM model, (ii) Bayesian GLM model, (iii) RCV XGBoost algorithm and (iv) SPCV XGBoost algorithm for harsh brakings and harsh accelerations respectively.

The combined prediction heatmap for the test area appears on Figure 5-30 for harsh braking frequencies.

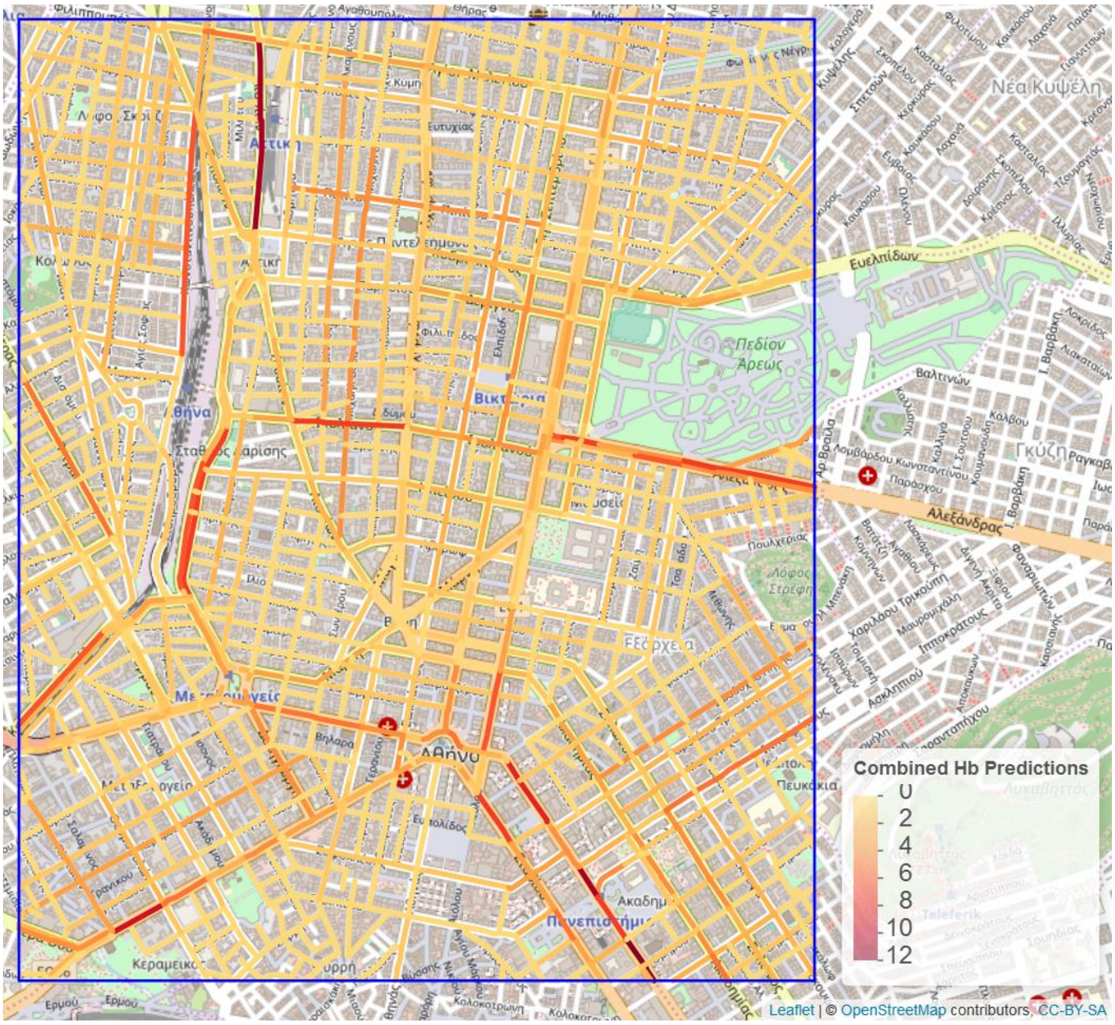


Figure 5-30: Combined prediction heatmap of harsh braking frequencies in Omonoia area

The evaluation metrics are shown on **Table 5-41** for combined predictions of harsh braking numbers.

Table 5-41: Evaluation metrics for combined model predictions for harsh brakings

Metric	Value
RMSE	1.6114
MAE	0.6645
RMSLE	0.4514
CA	87.55%

The combined prediction heatmap for the test area appears on Figure 5-31 for harsh acceleration frequencies.

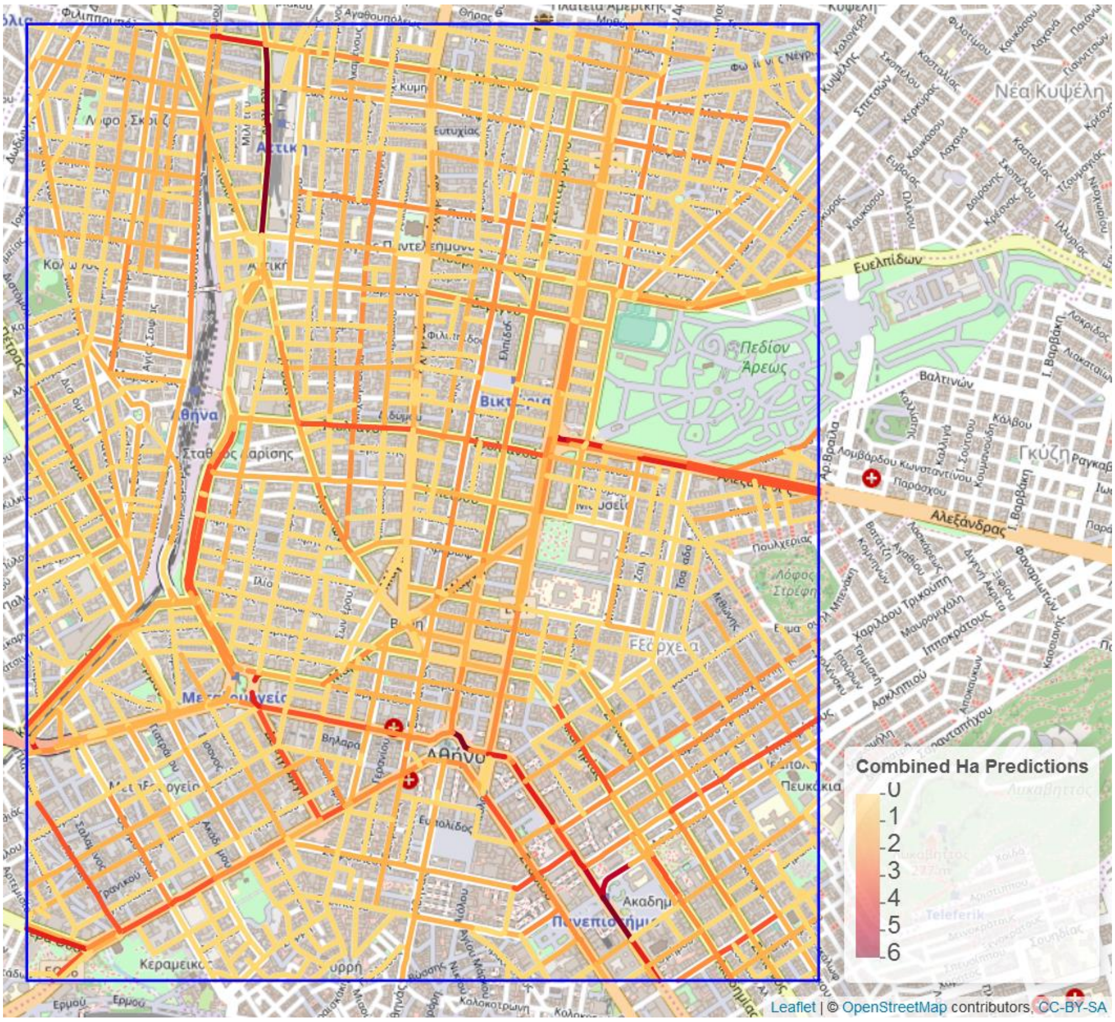


Figure 5-31: Combined predictions of harsh acceleration frequencies in Omonoia area

The evaluation metrics are shown on **Table 5-42** for combined predictions of harsh acceleration numbers.

Table 5-42: Evaluation metrics for combined model predictions for harsh accelerations

Metric	Value
RMSE	1.5010
MAE	0.6903
RMSLE	0.4316
CA	89.09%

The heatmaps of Figure 5-30 & Figure 5-31 and the evaluation metrics of **Table 5-41** & **Table 5-42** are compared with the respective prediction heatmaps and evaluation metrics of all individual models of Section 5. Another equally critical comparison is with the real harsh event data of Figure 4-18 & Figure 4-19 (point-data) and Figure 4-21 & Figure 4-22 (events per road segment).

These comparisons reveal that the combinations of Frequentist GLM, Bayesian GLM, RCV XGBoost and SPCV XGBoost model predictions yield the optimal results that best approach the real data of the test area. Combined predictions display better RMSE, MAE, RMSLE metrics compared to individual models horizontally and in all instances. Moreover, there is a gain of more than 2% in CA compared to the second best performing individual models.

Having discarded explicit spatial effects for prediction models, the underlying Poisson-based prediction mechanisms highlight different road segments as problematic locations – hotspots. However, through averaging, the models cover the shortcomings of each other and result to a unified prediction that is much closer to reality. By examining the maps, it is very interesting to note that the combination is fruitful by focusing on the hotspot segments: frequentist and Bayesian GLMs highlighted segments in the northwest part of the test area as more dangerous, while RCV XGBoost and SPCV XGBoost pinpointed segments in the southeast part. In a way, the combination functions like overlapping focusing lenses or color layers in photography, leading to a much more precise picture.

These results show considerable promise of transferability of the followed methodological approach to other urban road network areas without naturalistic driving data availability. Furthermore, they provide increased incentive to conduct more spatial analysis with additional model types in the future. Their contributions can be included in the combined predictions in order to examine if they further improve predictive performance. Naturally, larger naturalistic driving datasets that include more events per road segment are expected to yield even more precise results as well.

5.5.2 Discussion of main findings

The present section conducted a series of analyses involving GWPR, CAR RCV XGBoost and SPCV XGBoost models for harsh braking and harsh acceleration frequencies across road segments of urban networks. The main findings can be summarized in the following points:

1. Large scale high resolution naturalistic big data obtained from smartphone sensors can be meaningfully combined with geographical primary and secondary characteristics and road infrastructure parameters to form informative spatial datasets. Appropriate custom-made data cleaning, geometric characteristic derivation, map-matching algorithms with vote-counting systems for pass count adjustment were required. These datasets allow detailed spatial analysis of harsh event frequencies on a road segment basis for urban network areas.

2. 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.

3. Harsh event frequencies can be spatially analyzed as counts of events across road segments which have geographical neighborhood structures. 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. The combination of significant variables is different in each model.

4. 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 braking frequencies.

5. 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.

6. 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 can only 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.

7. 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.

8. CAR models can fit on a specific study area extremely well for harsh event frequencies (CA > 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.

9. 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.

10. 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%).

11. 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.

12. Non-count based modelling methods, including linear spatial methods such as GWR, and regression ML methods such as SVM and RF proved inappropriate to analyze harsh event frequencies either as count variables or as harsh event rates. The harsh event phenomena are highly non-linear, leading to poor model fits, poor CA and large error metrics. Additionally, road segment datasets contain zeros which do not allow for log-linear methods. Furthermore, harsh event rates lead to loss of information by forcing exposure variables to have coefficients bound to 1.

6 Urban arterial data collection and processing

This section provides technical information on the process of data collection, descriptive statistics, exploratory parameters and various additional information for the data describing urban arterial segments in the study area of Kifisias Avenue. The structure is largely equivalent to that of Section 4.

6.1 Study area – Kifisias Avenue

6.1.1 Initial study area examination

The relevant map data exports were again conducted with the OSM overpass-turbo API (https://overpass-turbo.eu/). An initial visual exploratory check was conducted to determine any discrepancies between the map image and the raw OSM data import; no discrepancies were detected. As a reminder, a span of 7.90 km was considered, due to the presence of conductive loop detectors and the respective traffic data availability there (as explained in Section 3.3.4.2).

For the following analyses, only the urban arterial segments were retained – i.e. no other connected road segments, walkways or cyclist paths were considered. The map with the axes of the imported segments (in green) is shown on Figure 6-1 for the selected length of Kifisias Avenue, together with the northbound and southbound traffic measurement locations.

Having exported the raw OSM data, and after enhancing them with SRTM data for the training area, the processing phase was ready to begin. The stages outlined in Sections 3.4.2 – 3.4.5 were followed consecutively.

6.1.2 Road geometry characteristics

There examined urban arterial length includes 152 road segments – ways in OSM terms – and 658 nodes. All segments featured complete data with no missing information and were thus utilized. The distribution of segments based on lane number per road bearing (Northbound or Southbound) is shown on **Table 6-1**.

Table 6-1: Lane numbers of Kifisias Avenue segments per road bearing

Road segment bearing	Lane number			
	1	2	3	4
Northbound	9	13	50	3
Southbound	6	11	52	8

As expected, it is evident that the majority of urban arterial segments feature a large number of lanes. One and two-lane segments usually refer to a ramp or exit lane merging with or breaching off the main road and as such coincide.

Descriptive values for the obtained geometric and road network characteristics appear on **Table 6-2**. As a reminder, gradient and neighborhood complexity are dimensionless quantities, and negative gradient values refer to downhill slopes.

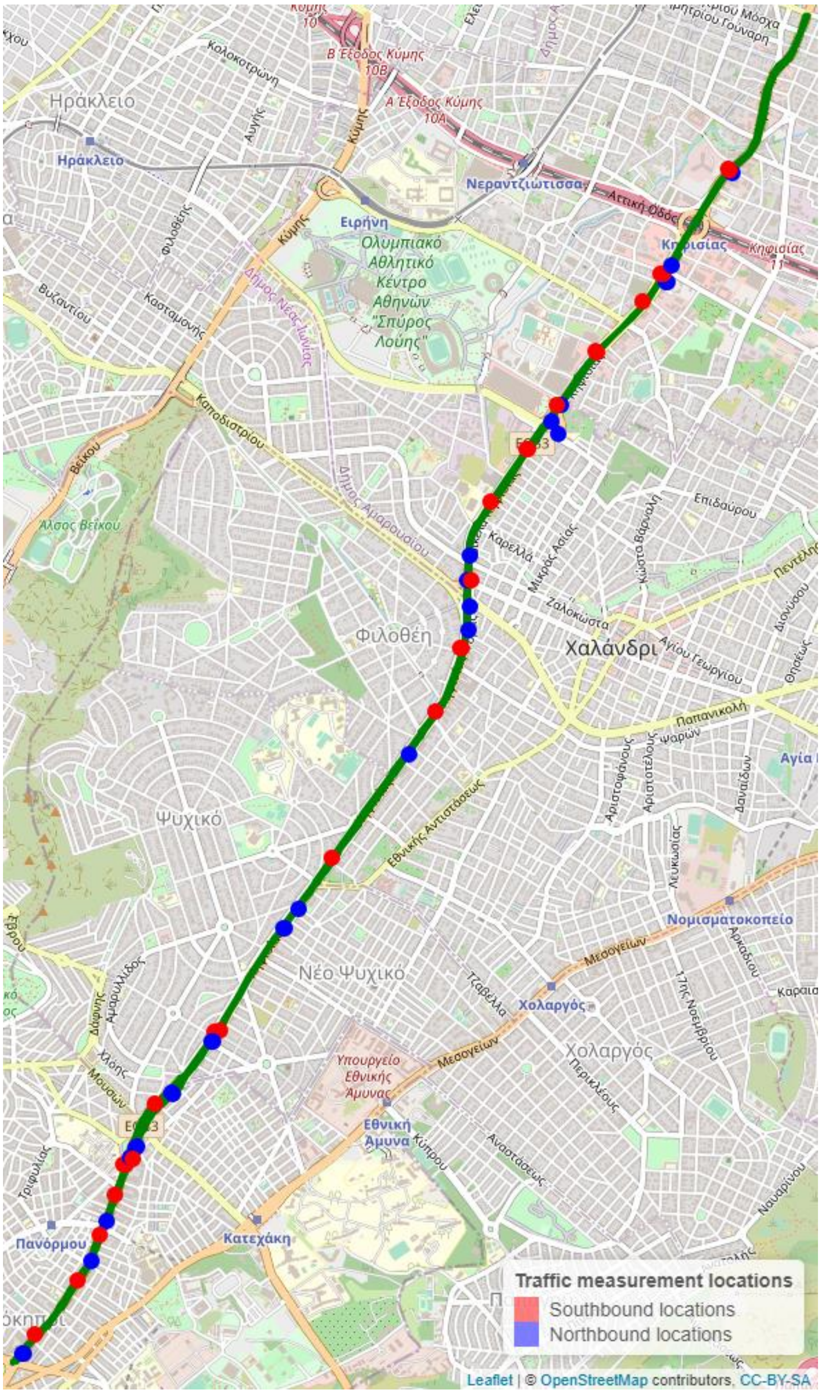


Figure 6-1: Kifisias road segments following import from OSM

The geometric characteristics of urban arterial road segments appear to feature some fluctuations and show values which are overall comparable to urban road networks (**Table 4-2** and **Table 4-7**). The standard deviations of geometric characteristics are lower across all metrics, as expected from segments of a specific avenue axis which constitute a more homogenous sample; values towards the extremes are more localized.

The skewness of segment length, curvature and gradient is positive, which reveals asymmetrical distributions with longer right tails. As per the aforementioned, neighborhood complexity is a logarithm of the number of proximal nodes of each segment, negative skewness is expected.

Table 6-2: Descriptive statistics for the obtained geometric characteristics for urban arterial segments in Kifisias Avenue

Geometric characteristics	**Descriptive statistics**						
	Average	Min	Median	Max	St. Dev.	Skewness	Kurtosis
Segment Length [m]	122.6286	3.5194	98.3146	446.5301	100.5204	1.1451	0.8666
Curvature [m^{-1}]	0.0011	0.0000	0.0002	0.0156	0.0021	4.0644	21.3629
Gradient [–]	0.0033	-0.1605	-0.0050	0.3905	0.0635	1.4937	8.2379
Neighborhood Complexity [–]	4.9523	4.1902	5.0148	5.4384	0.2981	-0.7020	-0.2162

The kurtosis of each of the geometric characteristics signify a departure from normal distributions. There is positive kurtosis in gradient and considerably more in segment curvature, signifying 'heavy-tailed' (leptokurtic) distributions for these variables which feature more and more frequent outliers. On the other hand, segment length and neighborhood complexity show low and negative kurtosis values, respectively, signifying a platykurtic distribution. This denotes lower numbers of outliers which are less frequent for these variables.

The presence of traffic lights and pedestrian crossings was again detected from tags in the OSM data: from the 152 total urban arterial road segments, 15 road segments feature traffic lights while 21 road segments feature pedestrian crossings. Once again, heatmaps can be produced from the previous characteristics. Figure 6-2 provides an intuitive presentation of road segment lengths of the study area.

6.1.3 Large-scale naturalistic driving data exploration

The dataset of naturalistic trip data is examined after being obtained from the OSeven application for the area of Kifisias Avenue. Data corresponding to a period of three months were provided, specifically from 01-09-2019 to 29-11-2019. During that period, 8756 trips were provided from 314 individual drivers in an anonymous format. These trips were not confined in the study area; some had origins and/or destinations on road segments outside the borders depicted on Figure 6-1. However, they were all cropped so that only the length of each trip that fell into the urban arterial area was considered – the remaining information was discarded.

Before processing, the trips had an average duration of 1306 seconds (or 21.76 minutes). This resulted in a very large big data file with 11,435,150 rows. However, this number corresponded to trip data within the surface of a rectangle in which Kifisias Avenue corresponded roughly to the second diagonal. An initial map-matching process was executed in order to remove trips outside Kifisias Avenue and reduce computational times when matching naturalistic driving data and traffic data.

The area depicted on Figure 6-1 includes larger areas that are not roads, and thus do not contain OSM nodes or ways. This is the reason that the moving window had to be considerably larger from the urban network cases. However, a spatial-based reduction from the entire spatial dataset was again necessary. In practice, dimensions of 350*350 on the OSM grid were used, corresponding to about 280m*380m. The runtime of this initial map-matching process for Kifisias Avenue was 76 hours and 17 minutes (or 3.26 days) on a server-level computer.

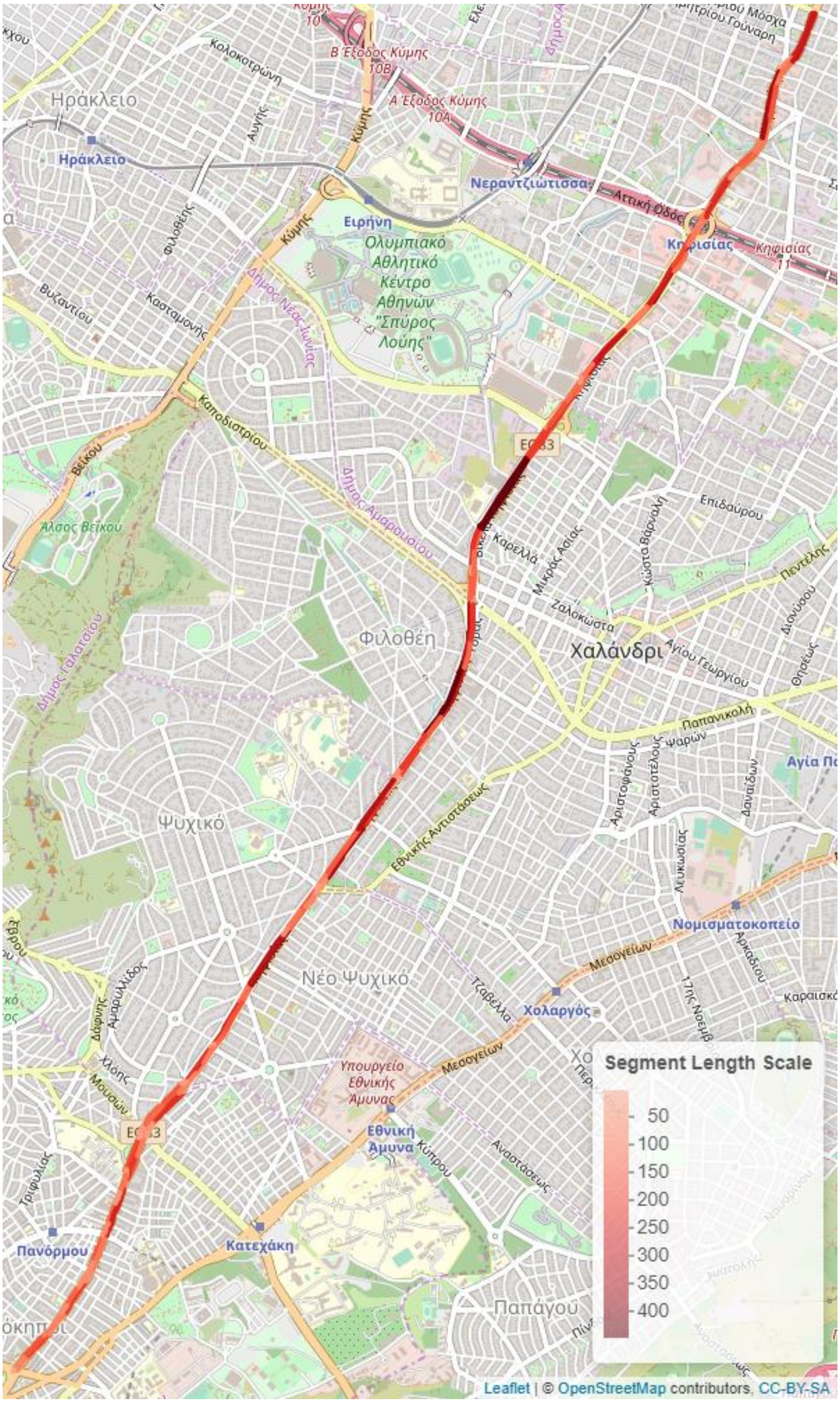


Figure 6-2: Heatmap of road segment lengths in Kifisias Avenue

As expected, the trips strictly on Kifisias Avenue segments were a fraction of the total amount. After cropping the trips, trip duration was reduced to a mean of 220.67 seconds, for a file of 930,346 entries. The histogram of trip durations is shown on Figure 6-3. It is possible that the concentration of small durations is due to drivers passing by some road segments of Kifisias Avenue as a small portion of a trip perpendicular to the main axis or a similar detour between suburbs.

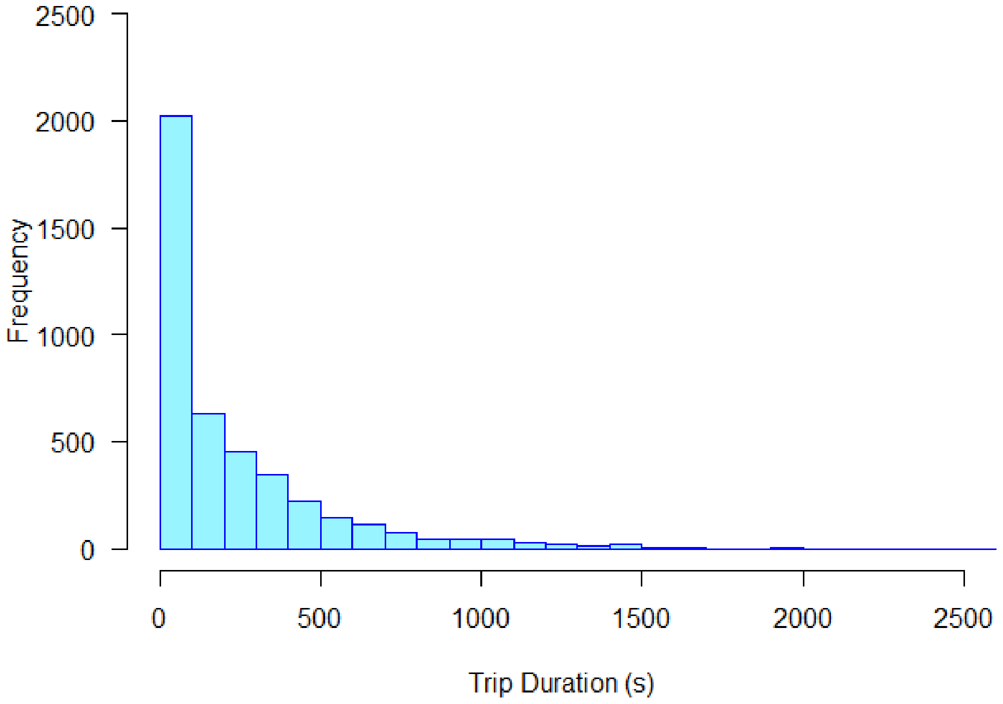


Figure 6-3: Histogram of trip durations in Kifisias Avenue

In these trips, 1543 harsh brakings and 1033 harsh accelerations occurred and were recorded alongside normal driving conditions. As per the aforementioned, for the purposes of this dissertation, these events are considered as point-data in space (i.e. without considering the length in which they occur). Furthermore, the analyses are made on an aggregated level – events are examined uniformly regardless of intensity. The numbers of harsh events per intensity category appear on **Table 6-3**.

Table 6-3: Harsh events per intensity category in Kifisias area

Event intensity category	Harsh events			
	Harsh brakings		Harsh accelerations	
1 – mild	771	49.97%	548	53.05%
2 – modest	546	35.39%	310	30.01%
3 – severe	226	14.65%	175	16.94%
Total	1543	100.00%	1033	100.00%

Similar to geometric characteristics, harsh events can be depicted on the map of the study area as point-data, as shown on Figure 6-4 for harsh brakings (hb) and on Figure 6-5 for harsh accelerations (ha). From the two figures it is obvious upon initial visual examination that there is an evenly spread distribution of events across the length of the Avenue, with no segment standing out in particular. There was a slight exception regarding a thinning of harsh accelerations at the junction with Amarisias Artemidos roughly at 75% of the Avenue length towards the North.

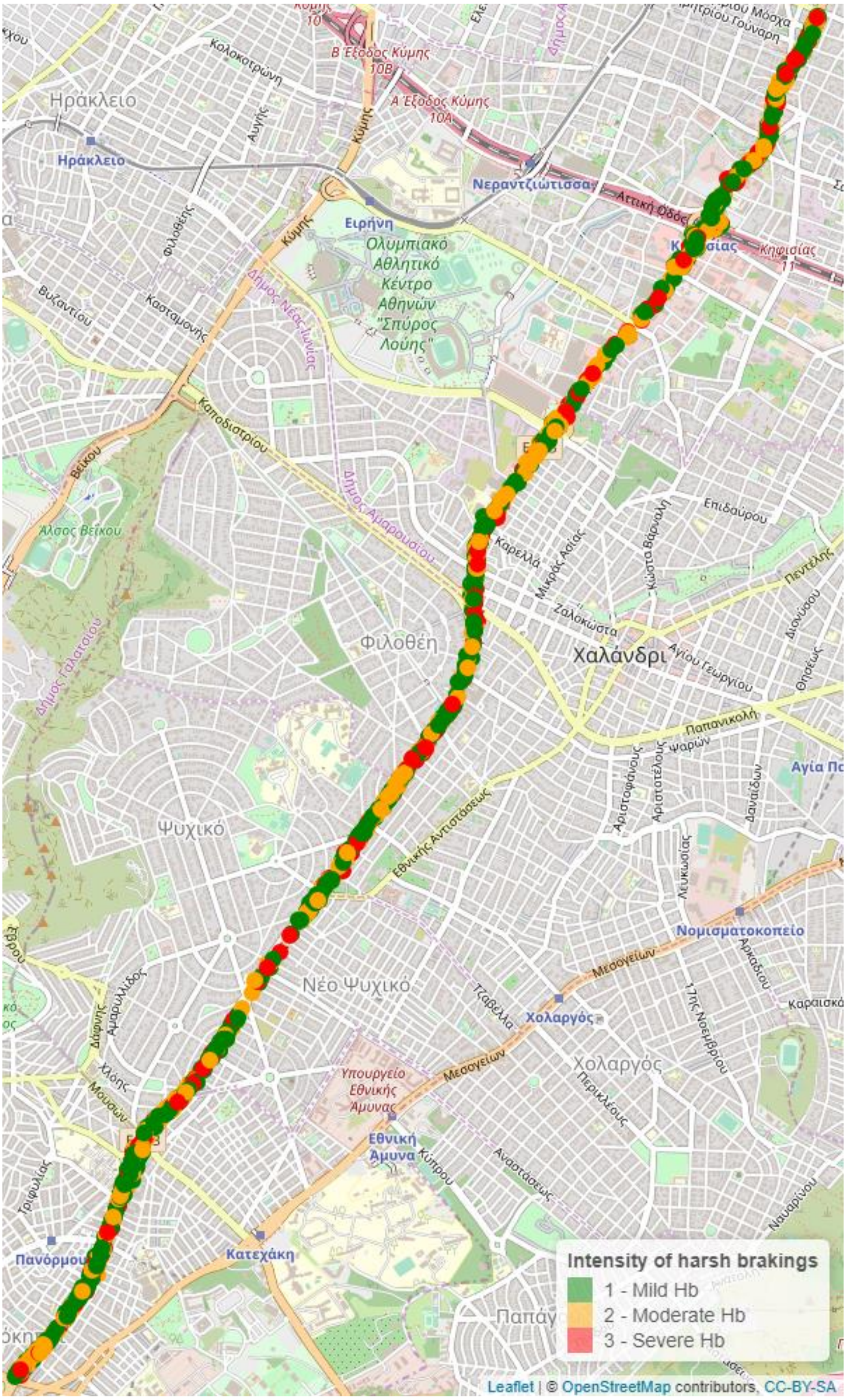


Figure 6-4: Total harsh braking events in Kifisias Avenue

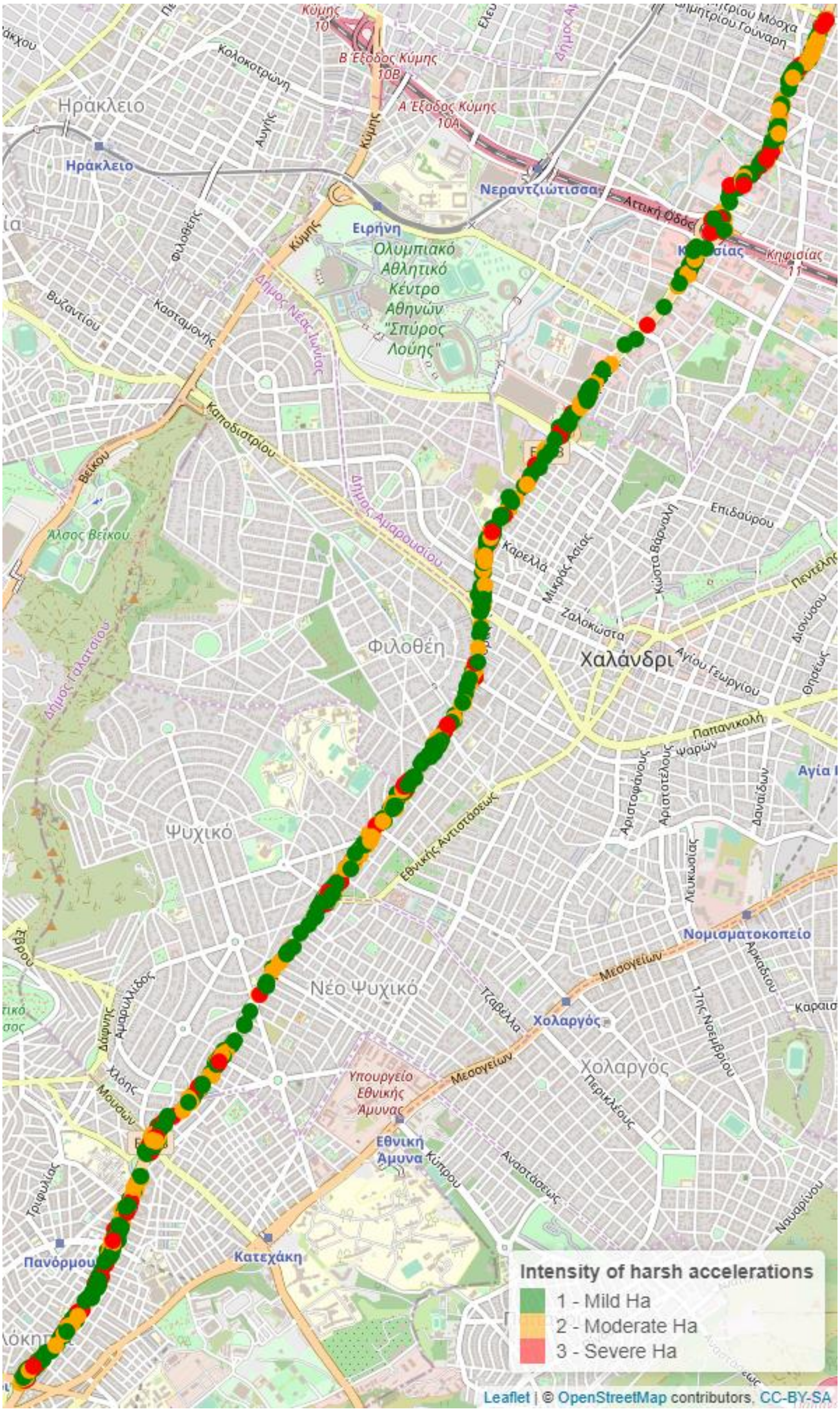


Figure 6-5: Total harsh acceleration events in Kifisias Avenue

6.1.4 Integration of traffic parameters for urban arterial segments

Following the extraction and compilation of the geometric dataset regarding the infrastructure of Kifisias Avenue, and the acquisition of the naturalistic driving big dataset for that area, the next step concerned the acquisition of traffic data for the corresponding period of analysis, namely from 01-09-2019 to 29-11-2019. The process described on Section 3.4.5 was followed, and data was collected from the Athens TMC for the 54 main and not-main traffic measurement locations corresponding with Kifisias Avenue for the same time period.

A sample of TMC measurements of traffic volume (flow), speed and occupancy for the measurement location MS268 is shown on **Table 6-4**. The measurements have a temporal resolution of 90 s and were taken during 05/09/2019; the gradual onset of congestion after 08:22:30 can be observed as speed drops and occupancy increases.

Table 6-4: Sample TMC measurements of traffic volume, speed and occupancy

MS code	Timestamp	Volume [veh/h]	Occupancy [%]	Speed [km/h]	MS bearing	MS type
MS268	05.09.2019 08:04:30	2040	9	50	Northbound	Main_road
MS268	05.09.2019 08:06:00	2640	8	61	Northbound	Main_road
MS268	05.09.2019 08:07:30	3120	9	57	Northbound	Main_road
MS268	05.09.2019 08:09:00	2120	6	62	Northbound	Main_road
MS268	05.09.2019 08:10:30	2560	8	55	Northbound	Main_road
MS268	05.09.2019 08:12:00	2840	10	54	Northbound	Main_road
MS268	05.09.2019 08:13:30	3400	11	55	Northbound	Main_road
MS268	05.09.2019 08:15:00	2160	7	59	Northbound	Main_road
MS268	05.09.2019 08:16:30	2080	6	56	Northbound	Main_road
MS268	05.09.2019 08:18:00	2640	8	53	Northbound	Main_road
MS268	05.09.2019 08:19:30	2680	9	53	Northbound	Main_road
MS268	05.09.2019 08:21:00	2520	8	55	Northbound	Main_road
MS268	05.09.2019 08:22:30	2720	17	34	Northbound	Main_road
MS268	05.09.2019 08:24:00	2560	38	21	Northbound	Main_road
MS268	05.09.2019 08:25:30	2400	27	23	Northbound	Main_road
MS268	05.09.2019 08:27:00	3000	27	28	Northbound	Main_road
MS268	05.09.2019 08:28:30	1960	21	18	Northbound	Main_road
MS268	05.09.2019 08:30:00	2920	41	21	Northbound	Main_road
MS268	05.09.2019 08:31:30	1560	45	16	Northbound	Main_road
MS268	05.09.2019 08:33:00	2960	44	16	Northbound	Main_road
MS268	05.09.2019 08:34:30	2400	35	16	Northbound	Main_road
MS268	05.09.2019 08:36:00	1760	48	13	Northbound	Main_road
MS268	05.09.2019 08:37:30	1960	35	12	Northbound	Main_road
MS268	05.09.2019 08:39:00	2920	40	16	Northbound	Main_road
MS268	05.09.2019 08:40:30	3080	37	16	Northbound	Main_road
MS268	05.09.2019 08:42:00	2040	39	18	Northbound	Main_road
MS268	05.09.2019 08:43:30	2040	34	16	Northbound	Main_road
MS268	05.09.2019 08:45:00	2760	47	14	Northbound	Main_road
MS268	05.09.2019 08:46:30	1600	45	11	Northbound	Main_road

The term timestamp denotes a character string variable containing the information of the precise date and time of the measurements. As previously stated, TMC detection systems are subject to constant physical strain and can fail temporarily. This can lead to the recording of unrealistic values, such as occupancy > 100 %, speed > 200 km/h or speed > 0 along with traffic volume = 0, which were discarded and not considered in the database.

Having collected TMC measurements for the entirety of Kifisias Avenue, the fundamental traffic volume diagrams can be produced as defined by Greenshields et al. (1935). Figure 6-6 shows the empirical traffic volume-speed diagram produced from traffic data.

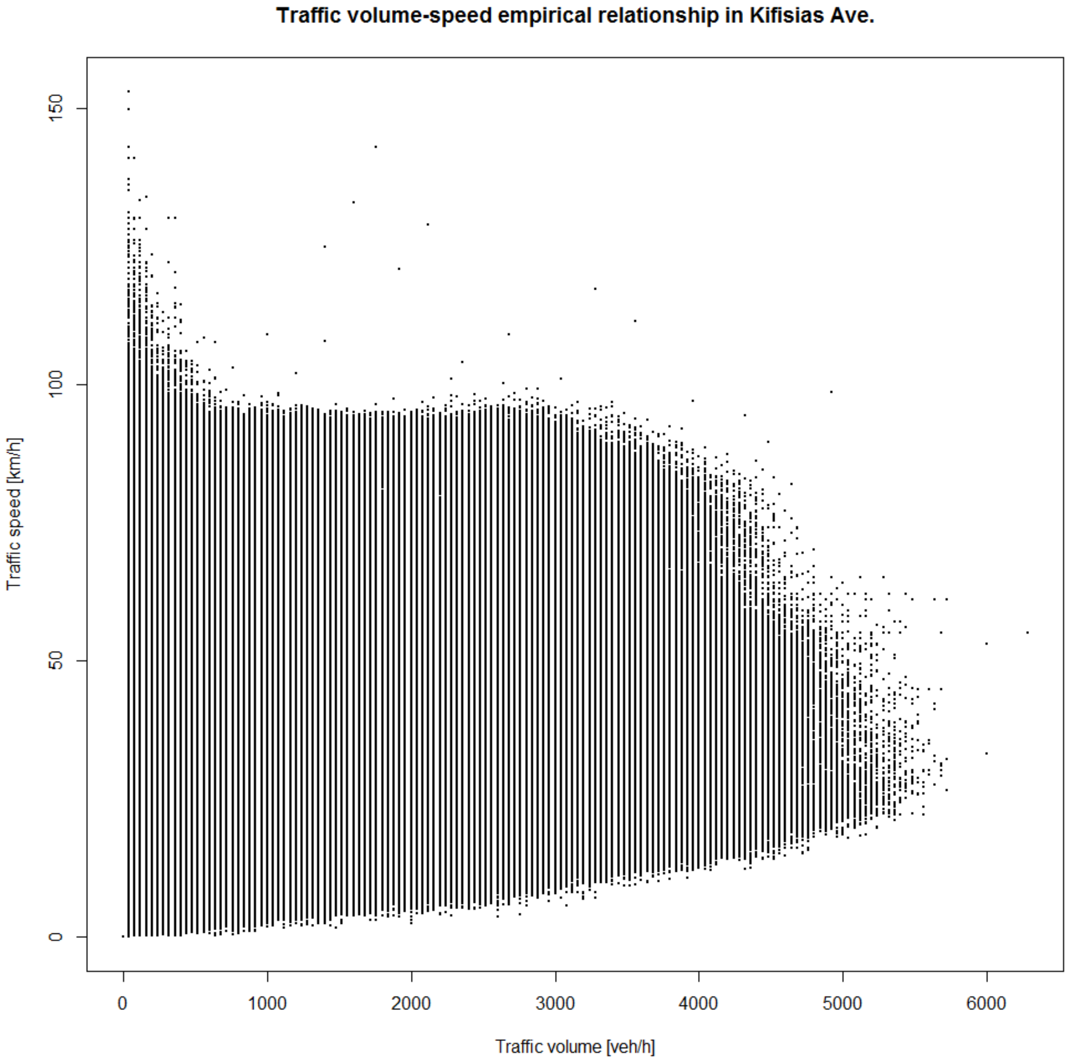


Figure 6-6: Empirical diagram of traffic volume and speed in Kifisias Avenue

It is evident that apart from the enveloping theoretical curve, in practice the inner area is also populated with all possible intermediate conditions. Similarly, Figure 6-7 shows the empirical occupancy and traffic volume diagram produced from traffic data.

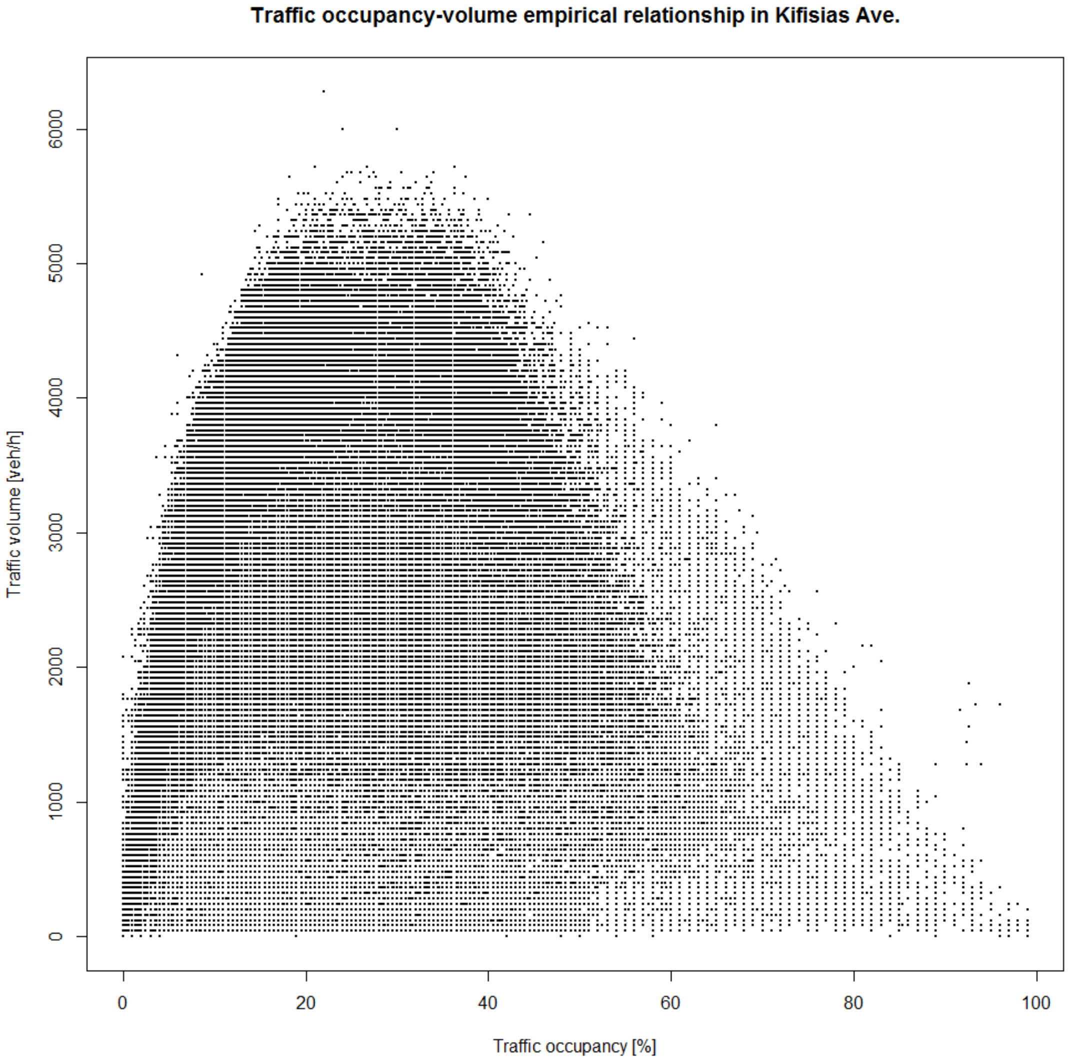


Figure 6-7: Empirical diagram of occupancy and traffic volume in Kifisias Avenue

6.1.5 Dataset merging and map-matching results

The traffic management dataset was then matched to the naturalistic driving dataset. The merging of traffic and naturalistic driving data was a time-consuming process even for the reduced dataset containing trip-seconds only on Kifisias Avenue. The runtime of this matching process was 74 hours and 49 minutes (or 3.18 days) on a server-level computer.

In the enhanced dataset, the variables of current traffic flow, traffic state and speed difference were calculated, as described in Section 3.4.5.2. Afterwards, traffic volume was transformed to vehicles per cycle per lane. This provided a common framework and the calculation of traffic states as established by Vlahogianni et al. (2008) and shown in Figure 3-28. This transformed traffic volume describes a snapshot of nearby vehicles per cycle per lane and as such it was considered a meaningful quantity to retain in the urban arterial analyses.

Traffic state was used as a filter label variable to obtain urban arterial subsets for free, synchronized and congested flow conditions, as described in the following. There was an overall class imbalance between the three traffic flow regimes that were determined. Specifically, from the 930,346 trip-seconds, 661322 were classified as free flow trip-seconds, 241361 seconds were classified as synchronized flow trip-seconds and 27663 were classified as congested flow trip-seconds.

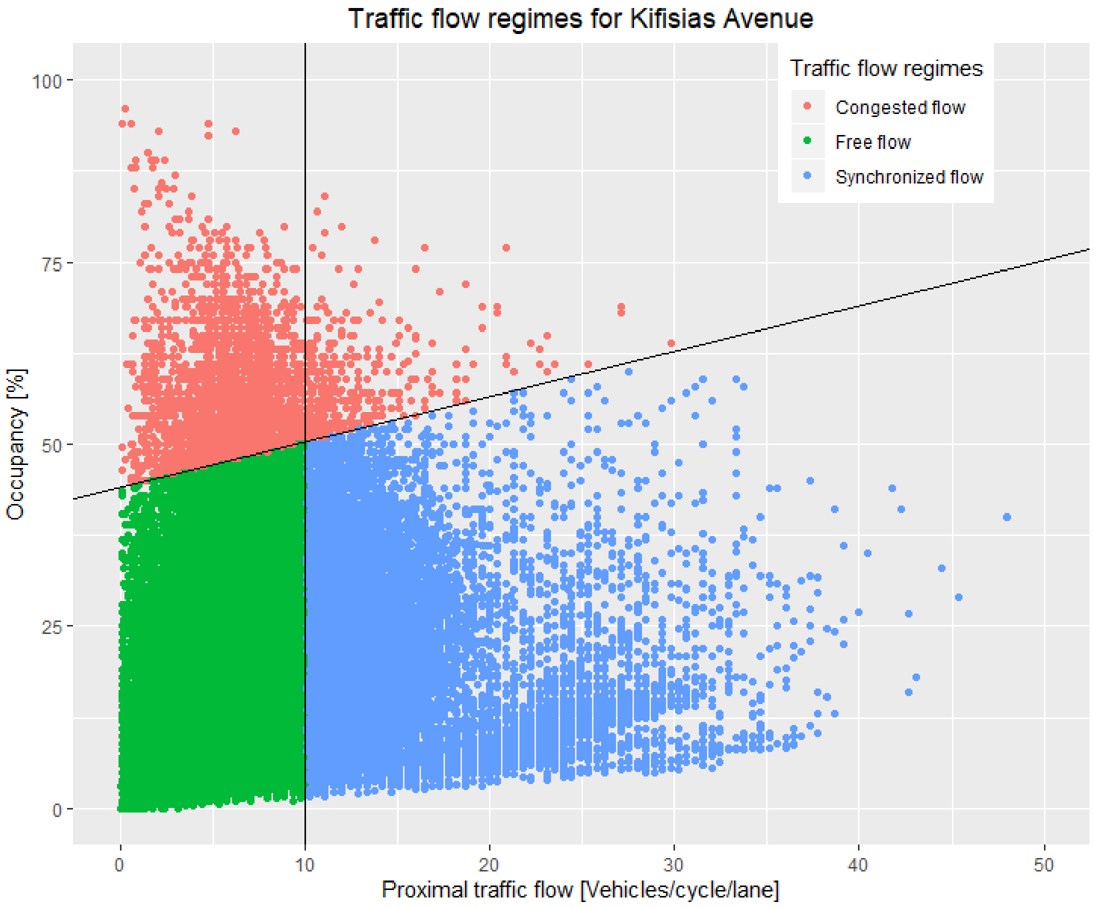


Figure 6-8: Traffic flow regimes in Kifisias Avenue

Subsequently, the spatial map-matching of the enhanced traffic and naturalistic driving subsets was conducted for each traffic state. Trip-seconds were analyzed and their parameters were attributed to road segments of Kifisias Avenue as described in Section 3.4.3. The runtime of the map-matching for free-flow conditions was about 8 hours and 40 minutes on a server-level computer; for the other regimes a fraction of that runtime was needed. Descriptive statistics for the obtained parameters are shown on **Table 6-5** for free flow, on **Table 6-6** for synchronized flow and on **Table 6-7** for congested flow conditions. Parameters with an asterisk (*) are reported only for segments that had non-zero trips, since they are calculated with the adjusted pass count (trips per segment), naturalistic driver speed or pass seconds per segment.

The descriptive statistics obtained from map-matching offer additional initial insights to the spatial examination of harsh event frequencies in Kifisias Avenue. The obtained values are all positive real numbers (with the exception of speed difference), which is expected since they represent frequency or traffic counts and their respective rates or percentages. Since Kifisias Avenue is a busy arterial, the vast majority of road segments were assigned at least one trip. Specifically for free flow all segments (100% of the total), for synchronized flow 144 out of 152 segments (94.74% of the total) and for congested flow 145 out of 152 segments (95.39% of the total) featured at least one trip. This indicates a good spatial coverage of the urban arterial segments.

Table 6-5: Descriptive statistics for free flow trip-seconds in Kifisias Avenue road segments after map-matching

Segment characteristics from naturalistic driving and traffic data	**Descriptive statistics**						
	Average	Min	Median	Max	St. Dev.	Skewness	Kurtosis
Adjusted pass count per segment	225.461	10	212	559	114.228	0.4260	-0.3626
Harsh brakings per segment	3.704	0	2	19	4.104	1.5236	2.0808
Harsh braking rate per segment *	0.0002	0.0000	0.0000	0.0001	0.0002	1.6969	2.5615
Harsh accelerations per segment	2.388	0	1	20	3.104	2.4571	8.0259
Harsh acceleration rate per segment *	0.0002	0.0000	0.0001	0.0030	0.0004	4.6284	26.4649
Pass seconds per segment	4350.809	24	3311	27386	4112.611	1.8954	5.9099
Mobile use seconds per segment	82.230	0	54	511	74.293	2.1880	7.8220
Mobile use percentage per segment *	1.89 %	0.00 %	1.73 %	6.67 %	0.83 %	1.5361	6.5007
Speeding seconds per segment	128.784	0	53	997	149.564	2.6103	9.1215
Speeding percentage per segment *	2.96 %	0.00 %	1.64 %	26.13 %	3.66 %	2.9256	11.8500
Average driver speed per segment [km/h] *	30.991	6.382	29.957	66.764	11.197	0.3321	0.0940
Average traffic volume per segment [veh/h] *	1295.967	177	1436	2722	611.464	-0.1181	-1.0784
Avg. std. traffic volume /cycle/lane per segment *	5.244	1.246	5.926	8.262	1.902	-0.6179	-0.9050
Average occupancy per segment [%] *	13.02 %	2.73 %	12.49 %	27.80 %	5.27 %	0.200	-0.7155
Average traffic speed per segment [km/h] *	41.929	18.915	41.857	65.705	10.167	0.3218	-0.5647
Speed Difference per segment [km/h] *	10.938	-14.415	10.446	53.279	11.837	0.7995	1.2738

Regarding driver aggressiveness, it appears that free-flow conditions favor the generation of more numbers of harsh events per segment. Another interesting observation is that, in free and synchronized flow conditions, individual drivers with the OSeven application tend to display lower, more conservative speeds compared to the average traffic. This trend is reversed for congested flow conditions, where they attain slightly higher speeds, which are nonetheless comparatively low, as expected for congested flow.

Table 6-6: Descriptive statistics for synchronized flow trip-seconds in Kifisias Avenue road segments after map-matching

Segment characteristics from naturalistic driving and traffic data	**Descriptive statistics**						
	Average	Min	Median	Max	St. Dev.	Skewness	Kurtosis
Adjusted pass count per segment	90.691	0	36	433	108.499	1.2467	0.5932
Harsh brakings per segment	1.414	0	0	16	2.871	2.8316	8.3284
Harsh braking rate per segment *	0.0001	0.0000	0.0000	0.0025	0.0003	5.2572	29.8478
Harsh accelerations per segment	0.934	0	0	12	1.8794	3.0261	11.0442
Harsh acceleration rate per segment *	0.0001	0.0000	0.0000	0.0031	0.0004	6.0645	41.0953
Pass seconds per segment	1587.908	2	246	13930	2749.729	2.2648	4.7652
Mobile use seconds per segment	22.263	0	3	198	38.302	2.1549	4.3051
Mobile use percentage per segment *	1.64 %	0.00 %	1.19 %	25.00 %	2.79 %	5.3931	37.5301
Speeding seconds per segment	24.730	0	3	454	59.787	4.1418	20.9217
Speeding percentage per segment *	2.49 %	0.00 %	0.63 %	25.09 %	4.63 %	2.7868	7.8758
Average driver speed per segment [km/h] *	26.892	3.133	26.509	80.754	13.618	0.4965	1.4714
Average traffic volume per segment [veh/h] *	2879.142	1052	3010	3931	599.980	-0.7708	0.4670
Avg. std. traffic volume /cycle/lane per segment *	12.462	10.111	11.692	24.200	2.778	2.8139	7.7015
Average occupancy per segment [%] *	21.24 %	7.00 %	20.73 %	42.62 %	6.74 %	0.3837	-0.2695
Average traffic speed per segment [km/h] *	43.058	15.486	41.475	66.432	11.729	0.5699	0.6300
Speed Difference per segment [km/h] *	18.410	0.000	15.119	71.231	13.326	1.1553	1.3005

Most variable distributions display positive skewness, denoting asymmetrical distributions with longer right tails. An exception to this observation is traffic flow (and standardized traffic flow per cycle per lane for free flow), which features longer left-tail distributions. In addition, the kurtosis of most behavioral parameters exceeds the threshold of 3 for normal distribution, indicating the presence of more and more frequent outliers in the data. On the other hand, most traffic-related parameters have kurtoses considerably lower than 3, indicating distributions with flatter peaks and lighter tails, as well as the absence of many and frequent outliers. Nonetheless, there are exceptions to this trend, such as the standardized traffic volume and speed difference for congested flow.

Table 6-7: Descriptive statistics for congested flow trip-seconds in Kifisias Avenue road segments after map-matching

Segment characteristics from naturalistic driving and traffic data	**Descriptive statistics**						
	Average	Min	Median	Max	St. Dev.	Skewness	Kurtosis
Adjusted pass count per segment	11.947	0	6	80	15.827	2.2116	4.5510
Harsh brakings per segment	0.066	0	0	2	0.274	4.3841	20.4293
Harsh braking rate per segment *	0.0000	0.0000	0.0000	0.0013	0.0001	7.1710	56.9838
Harsh accelerations per segment	0.026	0	0	1	0.1606	5.8601	32.5546
Harsh acceleration rate per segment *	0.0000	0.0000	0.0000	0.015	0.0013	11.7790	137.808
Pass seconds per segment	182	0	45	1900	338.480	3.2682	11.4992
Mobile use seconds per segment	3.112	0	0	46	6.630	3.4056	14.4991
Mobile use percentage per segment *	2.28 %	0.00 %	0.00 %	33.33 %	5.02 %	3.2815	12.5914
Speeding seconds per segment	1.237	0	0	23	3.374	4.3690	21.2111
Speeding percentage per segment *	0.78 %	0.00 %	0.00 %	12.59 %	1.92 %	1.9296	14.4875
Average driver speed per segment [km/h] *	18.639	0.738	17.250	52.598	10.434	0.6430	-0.0524
Average traffic volume per segment [veh/h] *	1441.143	40	1544	2522	579.797	-0.2546	-0.8918
Avg. std. traffic volume /cycle/lane per segment *	6.182	0.444	6.061	19.931	2.793	1.2268	3.6730
Average occupancy per segment [%] *	56.03 %	46.67 %	55.59 %	74.00 %	5.45 %	1.0642	1.6704
Average traffic speed per segment [km/h] *	13.016	0.277	12.575	28.419	5.390	0.2799	-0.2708
Speed Difference per segment [km/h] *	5. 212	0.176	3.579	26.002	5.352	2.1358	5.0225

Despite the very satisfactory spatial coverage of naturalistic data, the integration of traffic data led to a loss of information from the available trip-seconds and harsh events. This loss occurred due to the demands of both spatial and temporal proximity for traffic information to complement the naturalistic dataset. For a portion of the dataset, no such information was available. Only these trip seconds with information across all variables, also known as complete cases, were used for spatial modelling of harsh events. Thus, the remaining harsh events after merging and map-matching are a subset of those shown on **Table 6-3** and appear on **Table 6-8**.

Table 6-8: Complete case harsh events per intensity category in Kifisias area

Event intensity category	Harsh events											
	Free flow				Synchronized flow				Congested flow			
	HB		HA		HB		HA		HB		HA	
1 – mild	254	45%	182	50%	134	62%	78	55%	5	52%	2	50%
2 – modest	229	41%	113	31%	49	23%	44	31%	4	38%	2	50%
3 – severe	80	14%	68	19%	32	15%	20	14%	1	10%	0	0%
Total	563	100%	363	100%	215	100%	142	100%	10	100%	4	100%

From both **Table 6-7** and **Table 6-8**, it is obvious that harsh events do not typically occur under congested conditions, as their numbers are far too scarce for spatial, and even conventional, modelling techniques. This is a reasonable outcome, since the reduced spatial and temporal headways of congested flow leave very little margin to develop the speed required for harsh braking or harsh acceleration manoeuvers. Thus it was decided to proceed with modelling of harsh event frequencies for free flow and synchronized flow only. Indicative heatmaps of harsh events per road segment are shown on Figure 6-9 and Figure 6-10 for free flow state and on Figure 6-11 and Figure 6-12 for synchronized flow state.

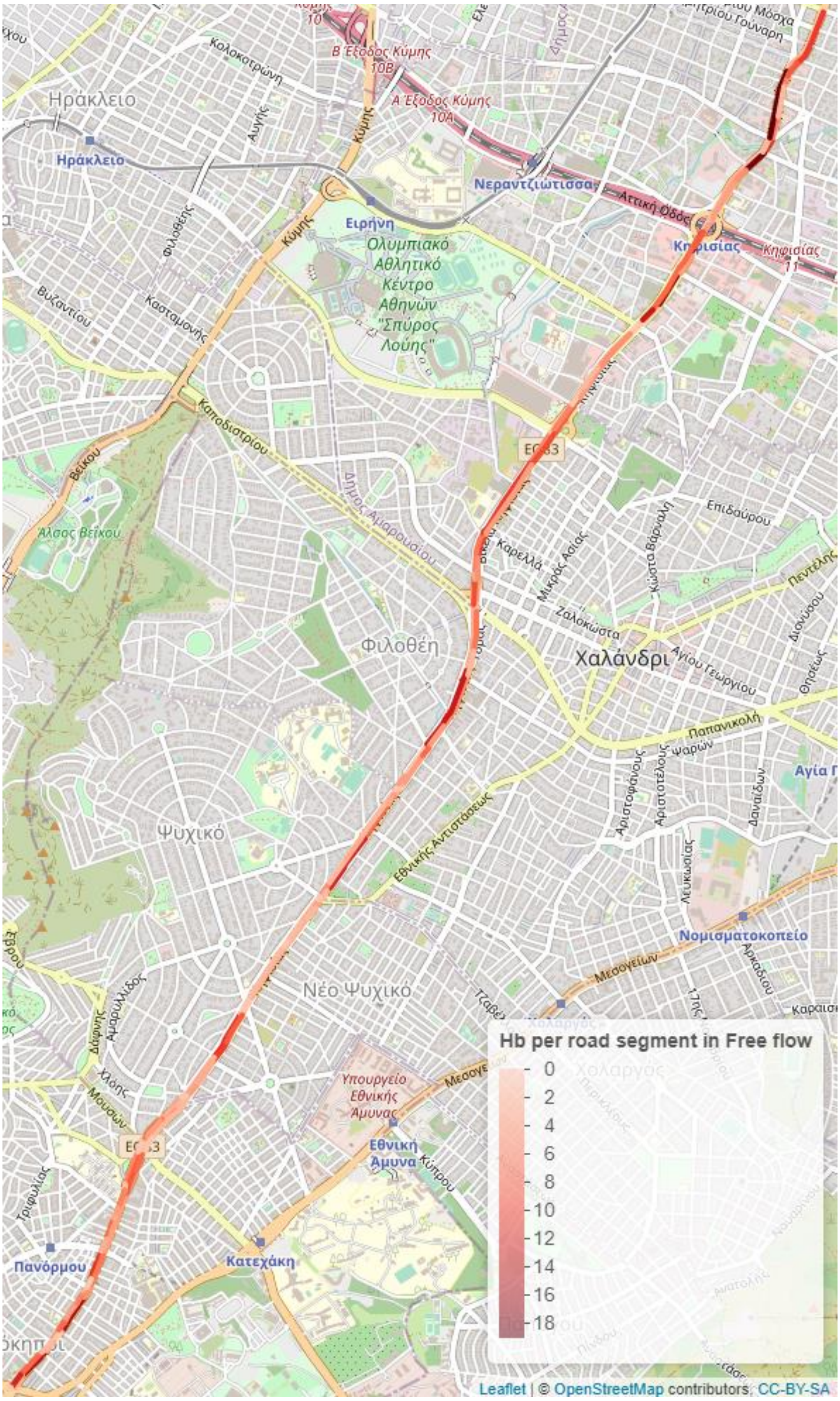


Figure 6-9: Heatmap of harsh braking frequencies of road segments in Kifisias Avenue under free flow state

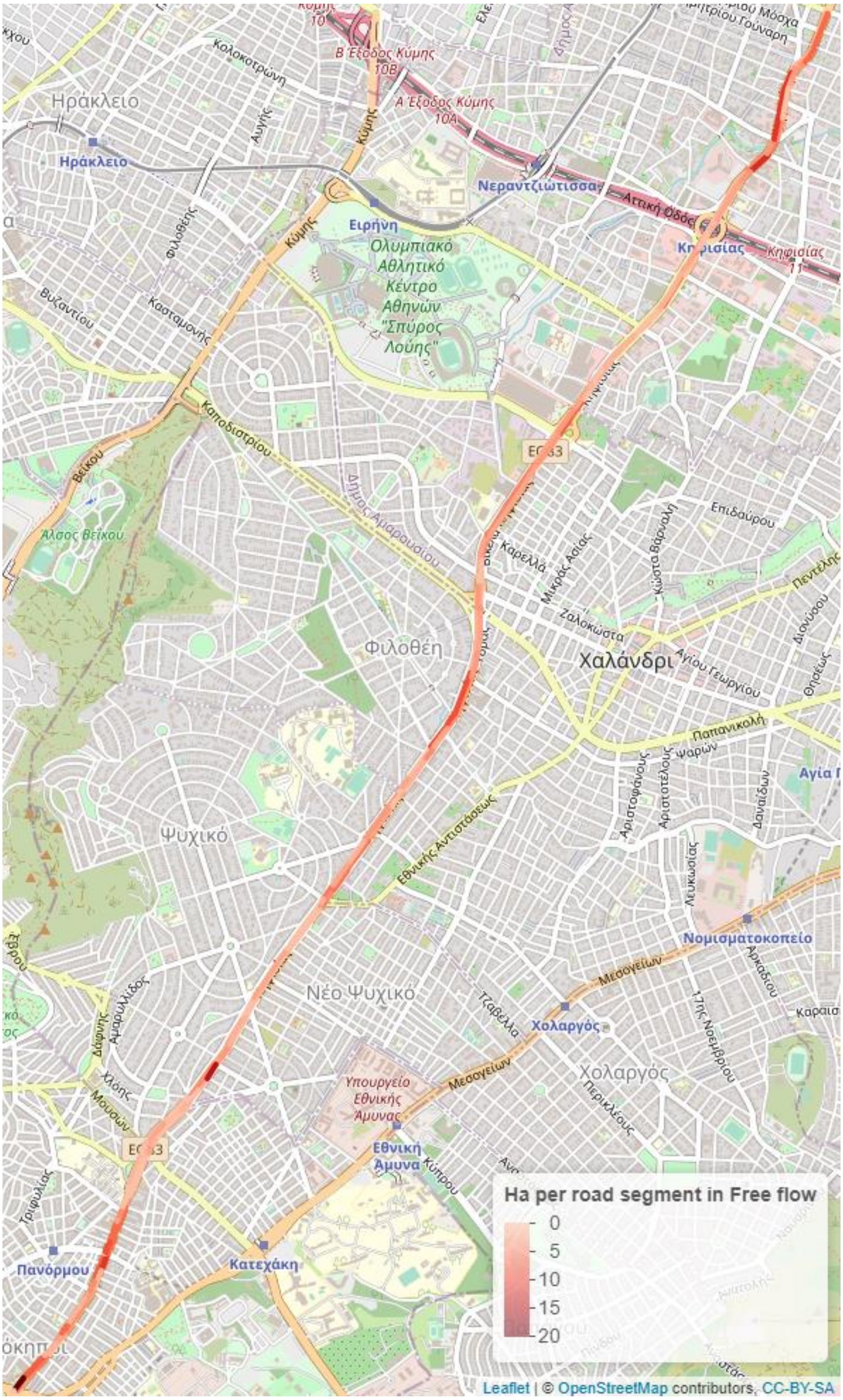


Figure 6-10: Heatmap of harsh acceleration frequencies of road segments in Kifisias Avenue under free flow state

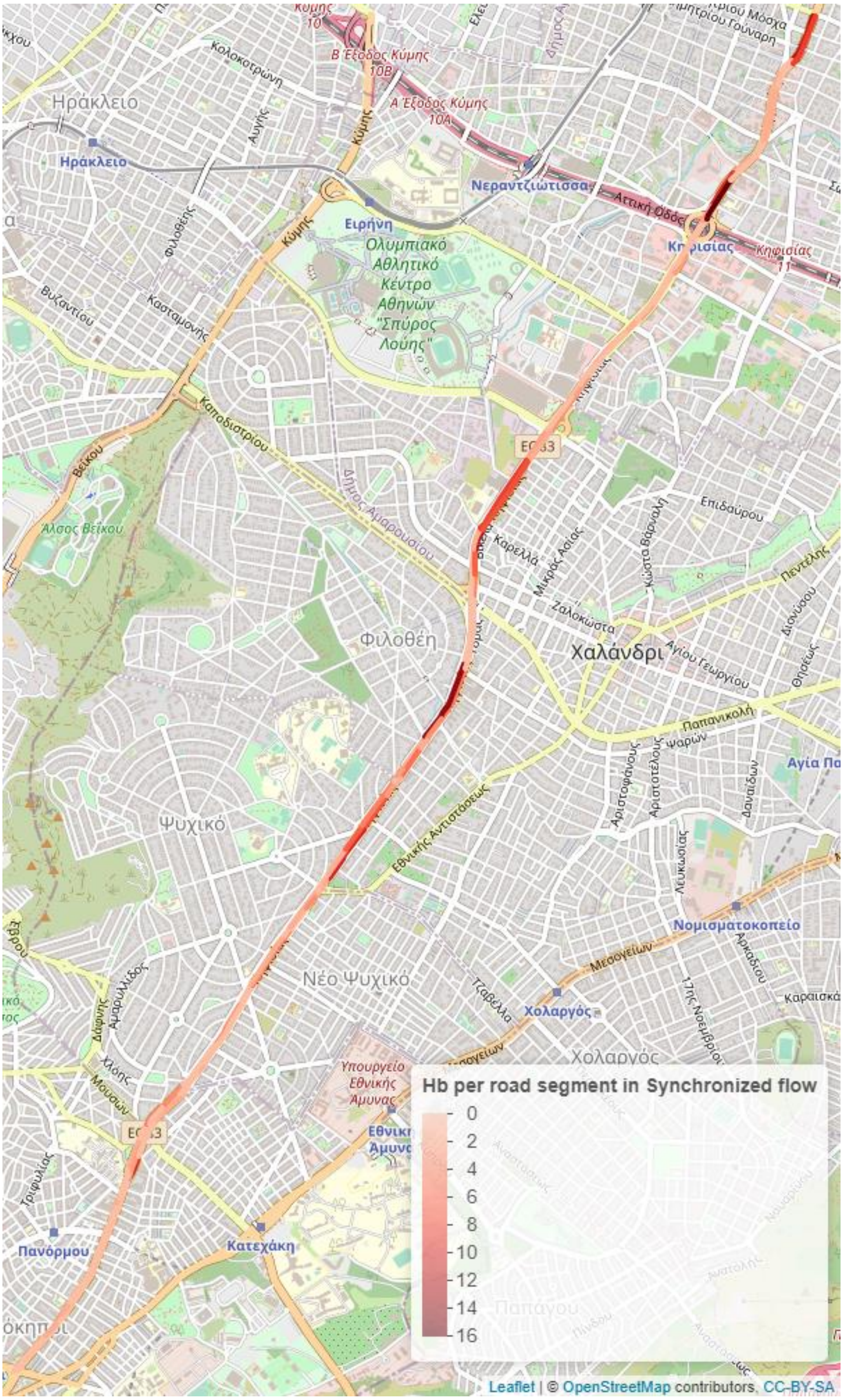


Figure 6-11: Heatmap of harsh braking frequencies of road segments in Kifisias Avenue under synchronized flow state

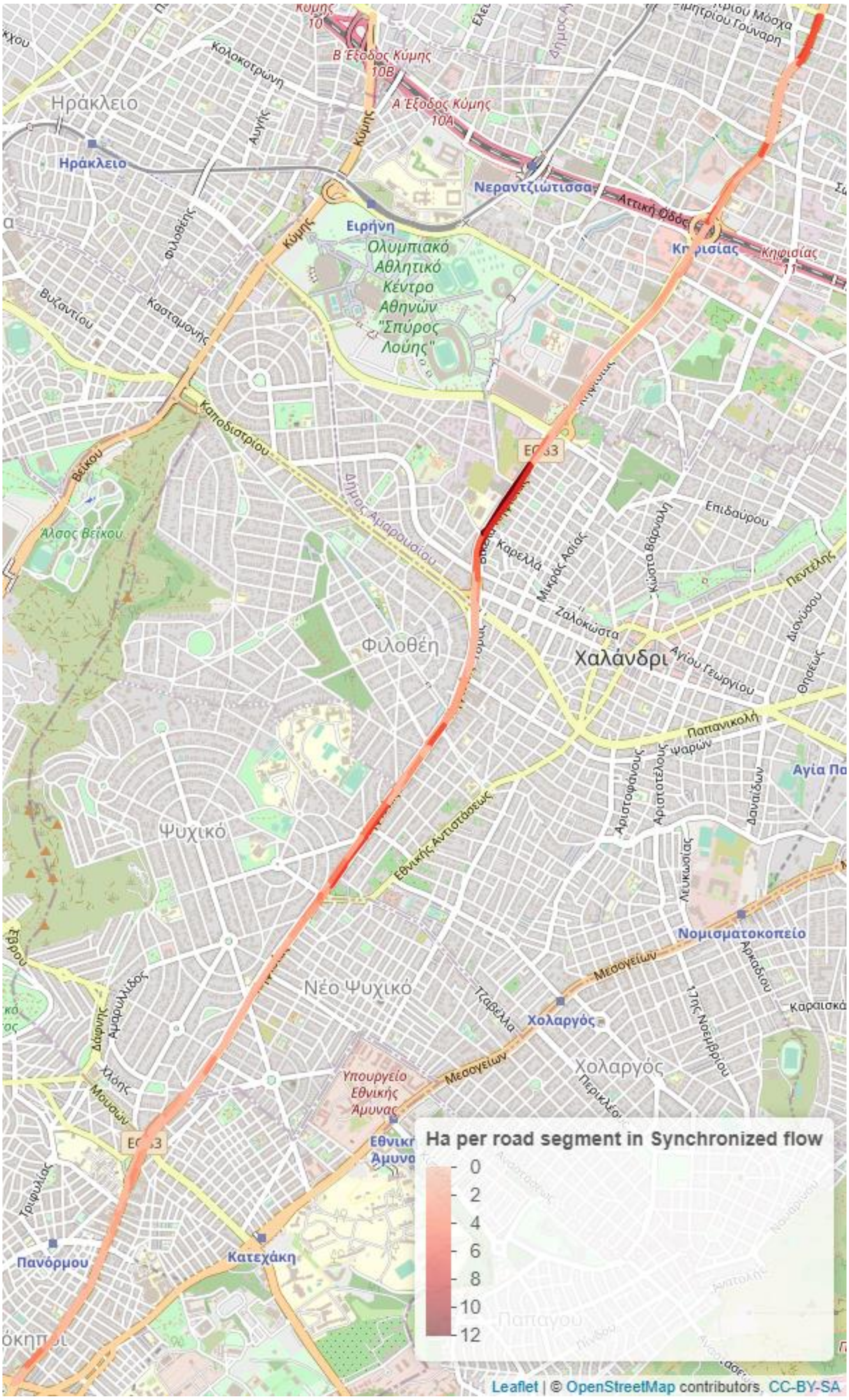


Figure 6-12: Heatmap of harsh acceleration frequencies of road segments in Kifisias Avenue under synchronized flow state

6.2 Spatial data frame sample

After the phases of data input, geometric characteristic derivation, traffic data integration, map-matching and processing have been complete, one spatial data frame is obtained for each examined traffic state, namely free flow and synchronized flow. As previously stated, the structure of the spatial data frames is important for performing spatial analyses and interpreting the results. Therefore, samples of the spatial data frames corresponding to Kifisias Avenue are presented here to showcase their structure and to add more context to the descriptive statistics already provided.

The spatial data frame sample is provided on **Table 6-9** for free flow state and on **Table 6-10** for synchronized flow state. Following the convention of this doctoral dissertation, each row represents a different road segment based on OSM segmentation. Obviously, the variables representing geometric and fixed components remain the same across the two data frames.

OSM Segment id	Spatial data frame attributes									
	Bearing	Lanes	Lat. Nom.	Lon. Nom.	Seg. Length	Curv/re	Gradient	Neighb. Comp.	Traffic Lights	Ped. Cross.
5168798	NB	1	38.0425	23.8039	17.8	0.0000	0.0000	4.3	1	1
5168803	NB	3	37.9877	23.7627	192.1	0.0006	0.0468	4.8	0	1
25117418	SB	3	37.9981	23.7692	101.4	0.0028	0.0068	5.3	0	1
25117419	NB	2	37.9989	23.7701	238.5	0.0015	0.0169	5.2	1	1
25117422	NB	2	37.9980	23.7693	99.2	0.0029	-0.0100	5.3	1	1
25117423	SB	3	37.9962	23.7682	200.2	0.0006	-0.0190	5.2	1	0

OSM Segment id	Spatial data frame attributes									
	HA No	HB No	Trip No	Pass sec	Speeding sec	Mob. use sec	HA rate	HB rate	Speeding %	Mob. use%
5168798	1	0	120	594	45	3	0.0005	0.0000	0.0758	0.0051
5168803	6	15	377	16644	50	356	0.0001	0.0002	0.0030	0.0214
25117418	3	3	306	3466	223	37	0.0001	0.0001	0.0643	0.0107
25117419	0	4	167	3346	176	94	0.0000	0.0001	0.0526	0.0281
25117422	1	2	286	4436	190	46	0.0000	0.0001	0.0428	0.0104
25117423	2	4	238	3228	123	53	0.0000	0.0001	0.0381	0.0164

OSM Segment id	Spatial data frame attributes					
	Avg. Driver Speed	Avg. Traffic Flow	Avg. Curr. Std. Traffic Flow	Avg. Occup.	Avg. Traffic Speed	Avg. Speed Diff.
5168798	34.2792	518.8041	5.7645	6.4012	47.1193	12.8402
5168803	22.3148	503.9870	1.8666	6.9158	29.6288	7.3140
25117418	33.4752	1442.7839	5.3436	12.5282	42.4920	9.0168
25117419	33.8288	895.4150	4.9745	14.6926	35.4671	1.6384
25117422	26.1111	237.6761	1.3204	9.7450	25.2650	-0.8461
25117423	26.8537	1712.7750	6.3436	16.0830	36.4070	9.5533

Table 6-9: Data frame sample from Kifisias Avenue segments for free flow conditions

OSM Segment id	Spatial data frame attributes									
	Bearing	Lanes	Lat. Nom.	Lon. Nom.	Seg. Length	Curv/re	Gradient	Neighb. Comp.	Traffic Lights	Ped. Cross.
5168798	NB	1	38.0425	23.8039	17.8	0.0000	0.0000	4.3	1	1
5168803	NB	3	37.9877	23.7627	192.1	0.0006	0.0468	4.8	0	1
25117418	SB	3	37.9981	23.7692	101.4	0.0028	0.0068	5.3	0	1
25117419	NB	2	37.9989	23.7701	238.5	0.0015	0.0169	5.2	1	1
25117422	NB	2	37.9980	23.7693	99.2	0.0029	-0.0100	5.3	1	1
25117423	SB	3	37.9962	23.7682	200.2	0.0006	-0.0190	5.2	1	0

OSM Segment id	Spatial data frame attributes									
	HA No	HB No	Trip No	Pass sec	Speeding sec	Mob. use sec	HA rate	HB rate	Speeding %	Mob. use%
5168798	4	1	275	2364	46	15	0.0008	0.0002	0.0195	0.0063
5168803	0	0	1	2	0	0	0.0000	0.0000	0.0000	0.0000
25117418	0	0	14	142	3	0	0.0000	0.0000	0.0211	0.0000
25117419	0	3	159	2774	105	24	0.0000	0.0001	0.0379	0.0087
25117422	0	1	13	230	2	0	0.0000	0.0008	0.0087	0.0000
25117423	0	0	37	270	2	1	0.0000	0.0000	0.0074	0.0037

OSM Segment id	Spatial data frame attributes					
	Avg. Driver Speed	Avg. Traffic Flow	Avg. Curr. Std. Traffic Flow	Avg. Occup.	Avg. Traffic Speed	Avg. Speed Diff.
5168798	25.6179	1701.4774	18.9053	24.7398	32.8814	7.2635
5168803	25.4500	2760.0000	10.2222	26.0000	30.9109	5.4609
25117418	12.3353	2974.0016	11.0148	28.3566	32.2021	19.8669
25117419	29.7266	2221.8844	12.3438	20.1428	41.1587	11.4321
25117422	11.6914	2287.7012	12.7095	31.2541	28.5670	16.8755
25117423	14.8664	2970.1611	11.0006	24.1813	34.9666	20.1002

Table 6-10: Data frame sample from Kifisias Avenue segments for synchronized flow conditions

7 Urban arterial segment analyses

In this section, spatial analysis results are presented for urban arterial segments. Initially, exploratory spatial analyses are conducted, in the form of global and local Moran's *I* coefficient calculations, empirical variogram plotting and theoretical variogram fitting. Subsequently, fitted Geographically Weighted Poisson Regression (GWPR) models, Conditional Autoregressive Prior (CAR) models and Extreme Gradient Boosting (XGBoost) machine learning methods with and without spatial cross-validation are presented and their results are elaborated upon.

All processes are conducted both for harsh braking and for harsh acceleration event frequencies, and refer to road segments as units of analysis. Exploratory spatial analysis are conducted in the total harsh event point-data. Modelling results are based on the final road segment datasets for urban arterials per traffic state, a sample of which appears in Section 6.2. The structure of the present section largely mirrors that of Section 5, however no predictions are conducted using the particular models.

Having departed from urban network analyses, a note on retained model variables is necessary. Several variables utilized in urban network analyses were not as meaningful in Kifisias Avenue, which is a homogenous road environment. Examples are road type, which is consistently 'primary' and road direction, which is consistently 'one-way' for the urban arterial segments.

However, a new group of variables became available and meaningful after the integration of traffic data and the shift to models that aim to investigate – instead of predicting – harsh event frequencies. These variables refer to the behavior of individual road user in the road segments (i.e. driver speed, seconds or percentages of speeding and mobile phone use), to nearby traffic in the area (i.e. traffic volume, traffic speed and occupancy), and a combination of the two (i.e. speed difference).

7.1 Exploratory spatial analysis

7.1.1 Global Moran's *I*

As in the previous cases, global Moran's *I* calculations are conducted for harsh braking and harsh acceleration frequencies in urban arterials following Bivand et al. (2008). Since Moran's *I* coefficients refer to point-data, they were calculated for all trip-seconds collectively regardless of traffic state. Therefore the complete set of harsh events of **Table 6-3** will be used for coefficient calculations.

7.1.1.1 Distance-based weighting

For each road segment, weights of all the other segments are assigned based on the distance of their centroids from the examined segment centroid. Afterwards, weights are row-standardized so that their sum equals to 1 for each segment. The resulting weighting scheme is used to calculate global Moran's *I*; results appear on **Table 7-1** for Kifisias Avenue.

Table 7-1: Global Moran's I in Kifisias Avenue with distance-based weighting

Global Moran's I	Training area			
	Coefficient value	Expectation	Variance	p-value
Harsh brakings	-0.0269	-0.0066	0.0000	0.0004
Harsh accelerations	-0.0071	-0.0066	0.0000	0.9266

Similarly with urban networks, it seems that there is almost zero spatial autocorrelation in harsh event frequencies when the entire urban arterial area is considered. In other words, these values indicate a random spatial distribution of events. The expected values were slightly higher, indicating that slightly more clustering was expected a priori from events in the study area than the outcome. The coefficient value for harsh brakings is statistically significant, while the value for harsh accelerations is not, indicating a possibility of considerably different – but unknown – true effects. This is an indication that once again, the consideration of the entire study area with distance-based weighting may not the optimal approach for the determination of Moran's *I* and therefore for the true degree of spatial autocorrelation in the data.

7.1.1.2 Nearest-neighbors weighting

As per the previous steps followed in Section 5.1.1, the correlation values are plotted on Figure 7-1 to better visualize the effects of each neighbor. A simple trend line fitted with locally-weighted polynomial regression is also provided. The maximum value of k is one-third of the total urban arterial segments (namely 50 road segments).

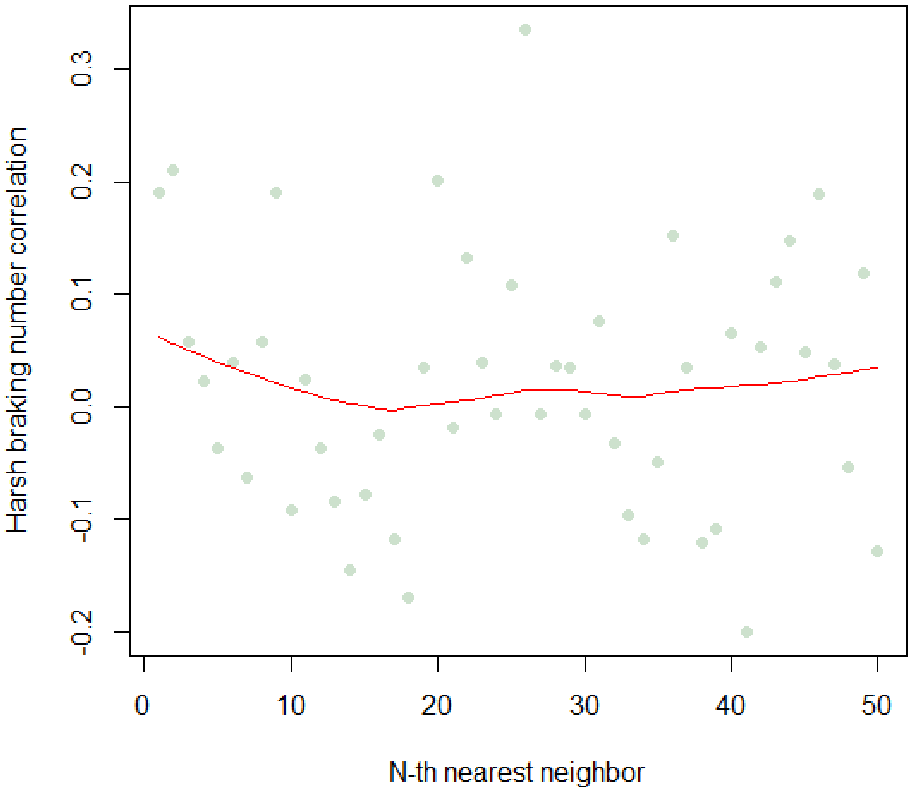


Figure 7-1: Harsh braking correlation values for N-nearest neighbors in Kifisias Avenue

The respective correlations are shown on Figure 7-2 for harsh accelerations; a steeper drop in correlation values is observed this time.

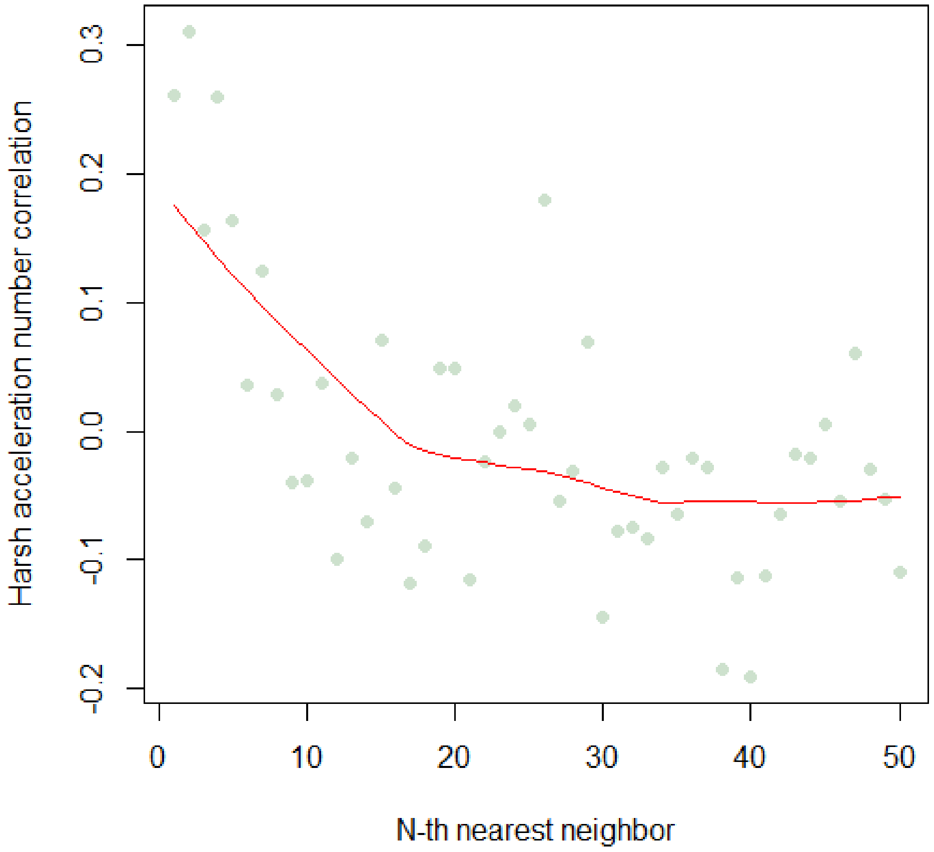


Figure 7-2: Harsh acceleration correlation values for N-nearest neighbors in Kifisias Avenue

In the case of urban arterial segments, there is reduced dimensionality of the area: instead of a two-dimensional urban network surface, there is a mostly one-dimensional – or linear – configuration. This transformation reduces the number of neighbors for each segment, and thus the correlations drop more rapidly. Thus, for nearest-neighbors weighting, the correlation threshold was set to 0.0 from the start. The calculated values for global Moran's *I* are provided on **Table 7-2** for Kifisias Avenue.

Table 7-2: Global Moran's I in Kifisias Avenue with nearest-neighbor calculations

Global Moran's I	**Training area**					
	Correlation threshold	k	Coefficient value	Expectation	Variance	p-value
Harsh brakings	0.0	5	0.0913	-0.0066	0.0023	0.0389
Harsh accelerations		9	0.1261	-0.0066	0.0012	0.0002

The trend that was observed in urban networks is retained in urban arterials: the difference with the results of **Table 7-1** is again considerable. The results are reversed by taking the contributions of only the k-nearest neighbors into account. With nearest-neighbor weighting, Moran's I coefficients indicate more clustering than anticipated for both harsh brakings and harsh accelerations.

7.1.2 Local Moran's *I*

Global Moran's *I* can be disaggregated to create a localized measure of spatial autocorrelation (Anselin, 1995). Following Bivand et al. (2019), local Moran's *I* values are calculated based on the approaches of distance-based weighting and nearest-neighbors weighting.

7.1.2.1 Distance-based weighting

For Kifisias Avenue, local Moran's *I* results calculated based on the approach of distance-based weighting appear on **Table 7-3**.

Table 7-3: Local Moran's I in Chalandri area with distance-based weighting

Local Moran's I		Training area			
		Coefficient value	Expectation	Variance	p-value
Harsh brakings	Average	-0.0269	-0.0066	0.0132	–
	Min	-0.3314	-0.0066	0.0024	0.0000
	Median	-0.0224	-0.0066	0.0002	0.7569
	Max	0.3449	-0.0066	0.0050	0.0000
	St. Dev.	0.1148	0.0000	–	–
Harsh accelerations	Average	-0.0071	-0.0066	0.0015	–
	Min	-0.0790	-0.0066	0.0020	0.1086
	Median	-0.0037	-0.0066	0.0003	0.9520
	Max	0.2186	-0.0066	0.0043	0.0006
	St. Dev.	0.0388	0.0000	–	–

As can be observed, the average local Moran's *I* values correspond to the respective global ones of **Table 7-1** (i.e. those of the distance-based calculation). Overall, local Moran's *I* values vary considerably, denoting the occurrence of both – some – positive autocorrelation (clustering) or negative autocorrelation (dispersion) of events across road segments.

It is worth noting that there are several instances of values that are not statistically significant; perhaps due to low event observations on the segment under consideration and/or neighboring segments. Furthermore, certain segments on the edge of the study area might lack strong contributing contiguous segments due to reduced directions from which information from proximal segments is available.

Distance-based (DB) local Moran's *I* values can be displayed in maps, as shown in Figure 7-3 and Figure 7-4. An interesting finding from these figures is that the maximum and minimum values of spatial autocorrelation are found in completely different road segments for harsh brakings and harsh accelerations, due to the different nature of the phenomena.

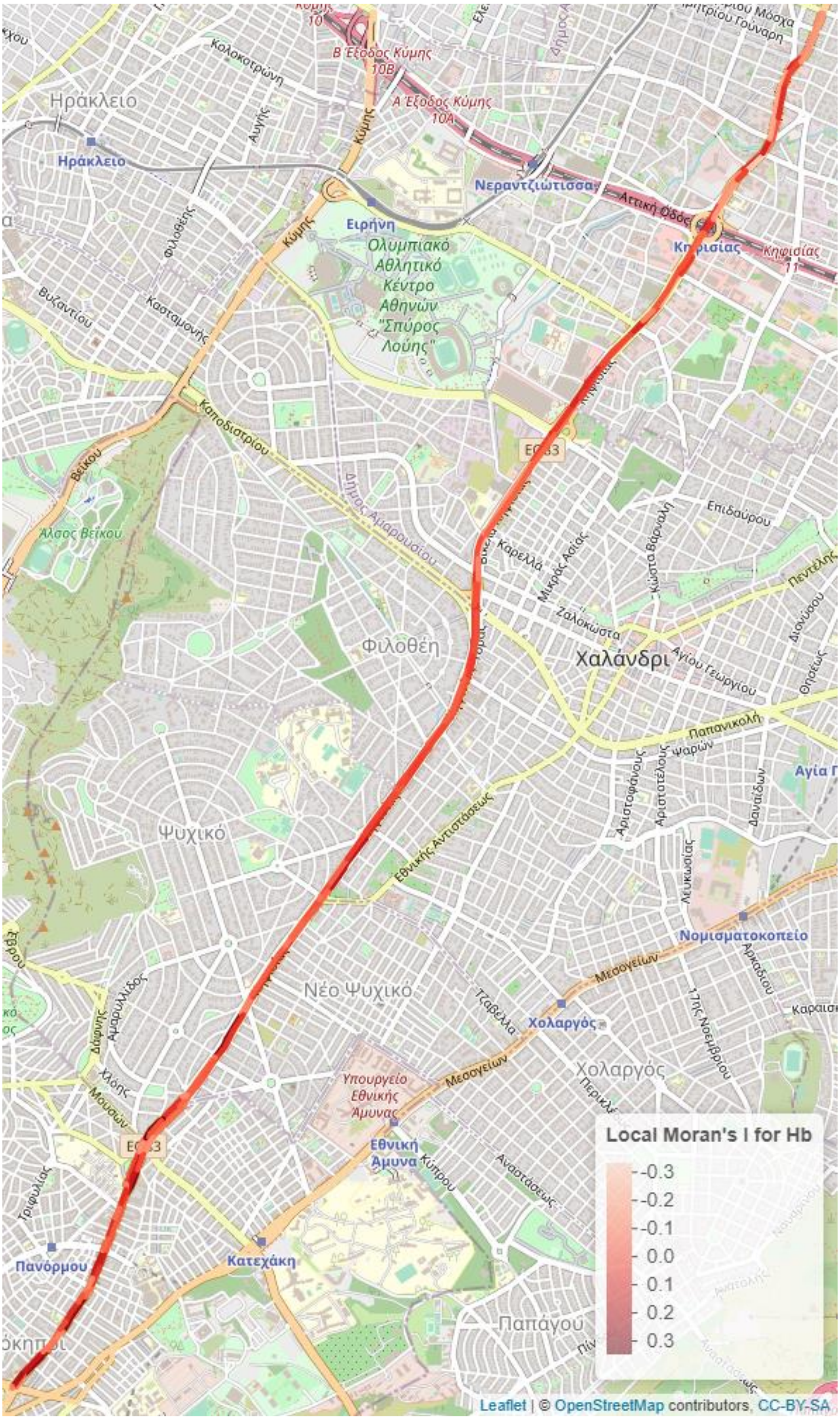


Figure 7-3: Local Moran's I values in Kifisias Avenue based on distance-based weighting for harsh braking events

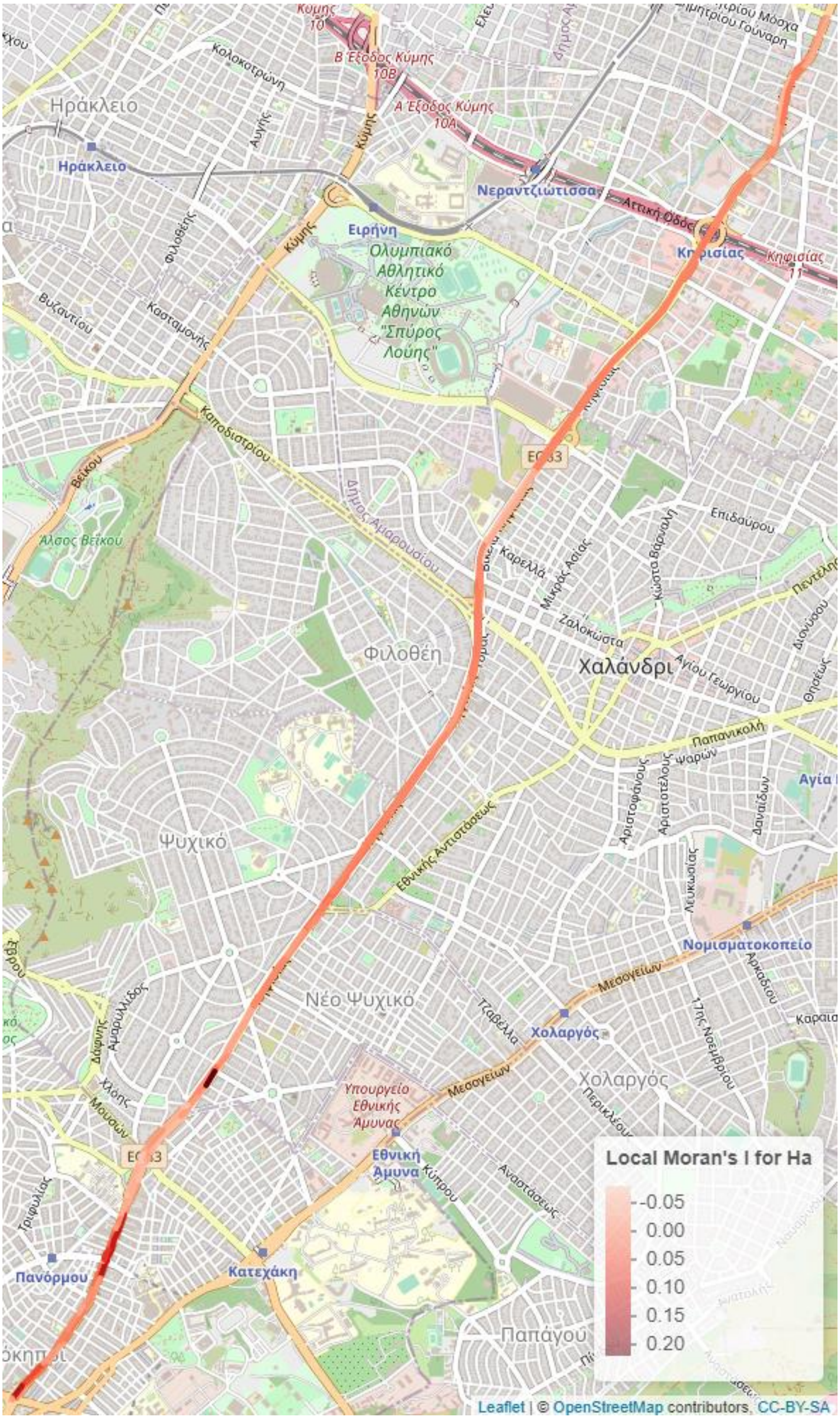


Figure 7-4: Local Moran's I values in Kifisias Avenue based on distance-based weighting for harsh acceleration events

7.1.2.2 Nearest-neighbors weighting

Results calculated based on the approach of nearest-neighbors weighting with a correlation threshold of 0.0 appear on **Table 7-4**. The number of nearest-neighbors k were the same (5 and 9 respectively) for dropping below the correlation threshold, as the data remains unchanged. The average local Moran's I values correspond to the respective global ones of **Table 7-2** (i.e. those of the nearest-neighbors weighting calculation).

Table 7-4: Local Moran's I in Kifisias Avenue with nearest-neighbors weighting

Local Moran's I		**Training area**			
		Coefficient value	Expectation	Variance	p-value
Harsh brakings [k=5]	Average	0.0914	-0.0066	0.2839	–
	Min	-0.9920	-0.0066	0.2839	0.0024
	Median	0.0156	-0.0066	0.2839	0.9610
	Max	3.6040	-0.0066	0.2839	0.0000
	St. Dev.	0.5329	0.0000	–	–
Harsh accelerations [k=9]	Average	0.2130	-0.0066	0.3037	–
	Min	-1.1600	-0.0066	0.3037	0.0077
	Median	0.1485	-0.0066	0.3037	0.7205
	Max	3.9605	-0.0066	0.3037	0.0000
	St. Dev.	0.5021	0.0000	–	–

Once again, towards the maximum range, the values of local Moran's I exceed the conventional upper bound of 1. The comparison of coefficient values and subsequent comparison with the mean and two-sigma rule, as per Anselin (1995), is shown on Figure 7-5 for harsh brakings and on Figure 7-6 for harsh accelerations. The mean is denoted with a blue line, while the two-sigma limit is denoted towards the left with a red dotted line.

By observing the figures, it is determined that most local Moran's I values are within the two-sigma rule. The remaining values gradually deviate from it at first, instead of single spikes. There are three outliers that appear as considerable deviations. However, segments with high local Moran's I values are not excluded on an outlier (two-sigma) basis. Rather, the results are considered to be an indication of strong spatial autocorrelations in specific segments, which are further incentive for the use of spatial models to study the phenomena of harsh events. Maps displaying the values of k nearest-neighbors (kNN) based local Moran's I are shown in Figure 7-7 and Figure 7-8.

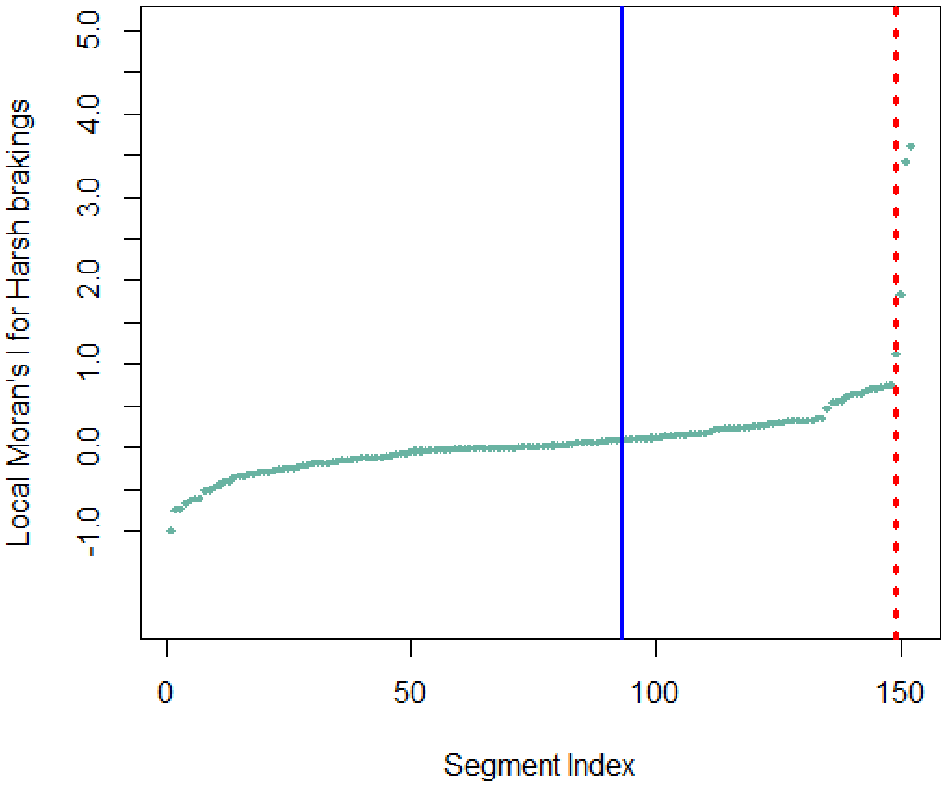


Figure 7-5: Local Moran's I values in Kifisias Avenue based on nearest-neighbors weighting for harsh braking events

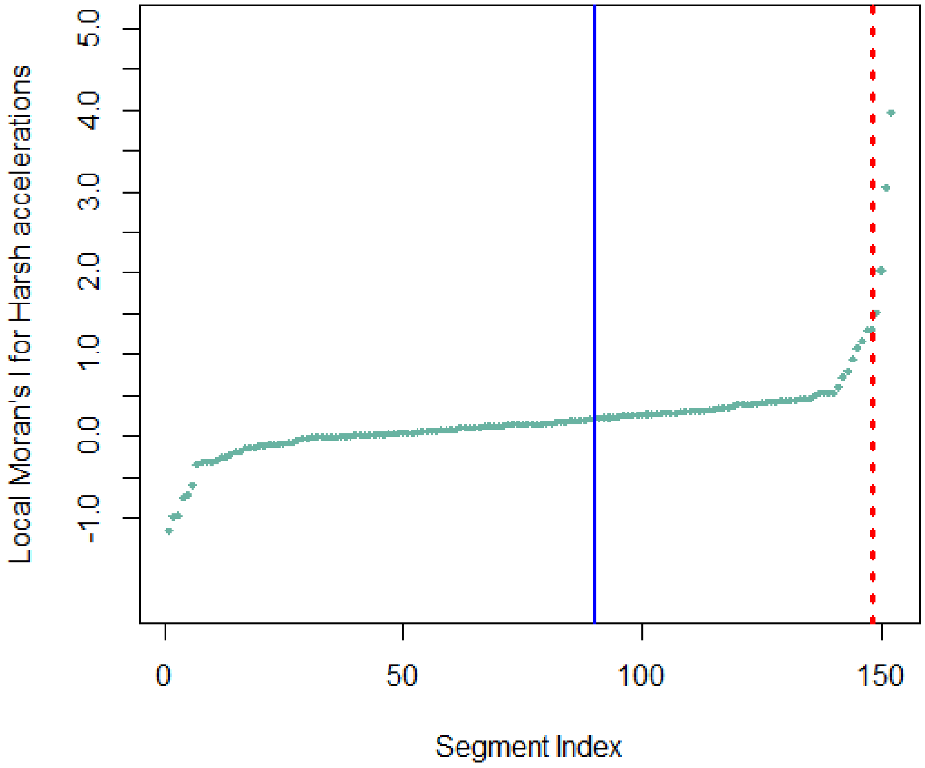


Figure 7-6: Local Moran's I values in Kifisias Avenue based on nearest-neighbors weighting for harsh acceleration events

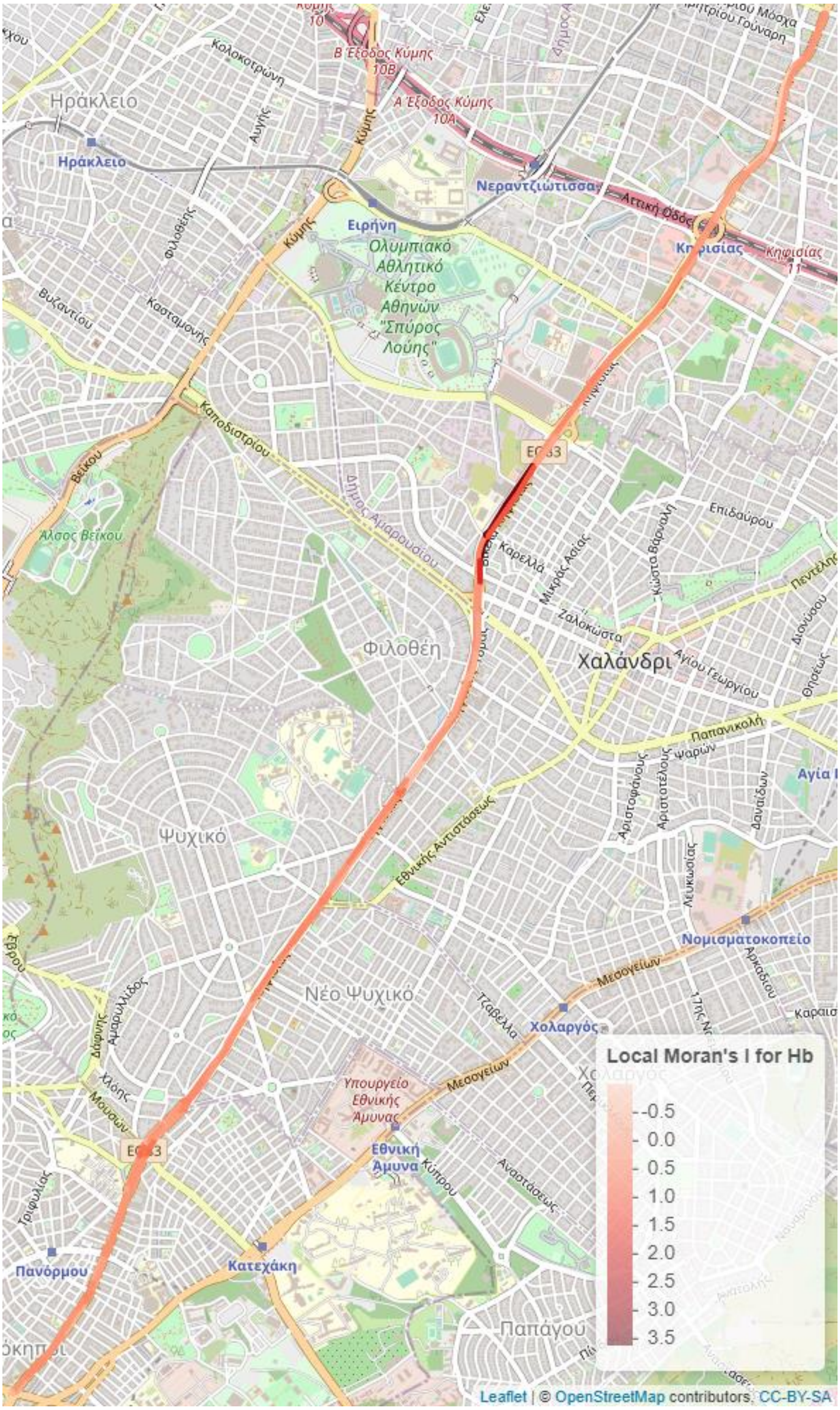


Figure 7-7: Local Moran's I values in Kifisias Avenue based on kNN-based weighting for harsh braking events

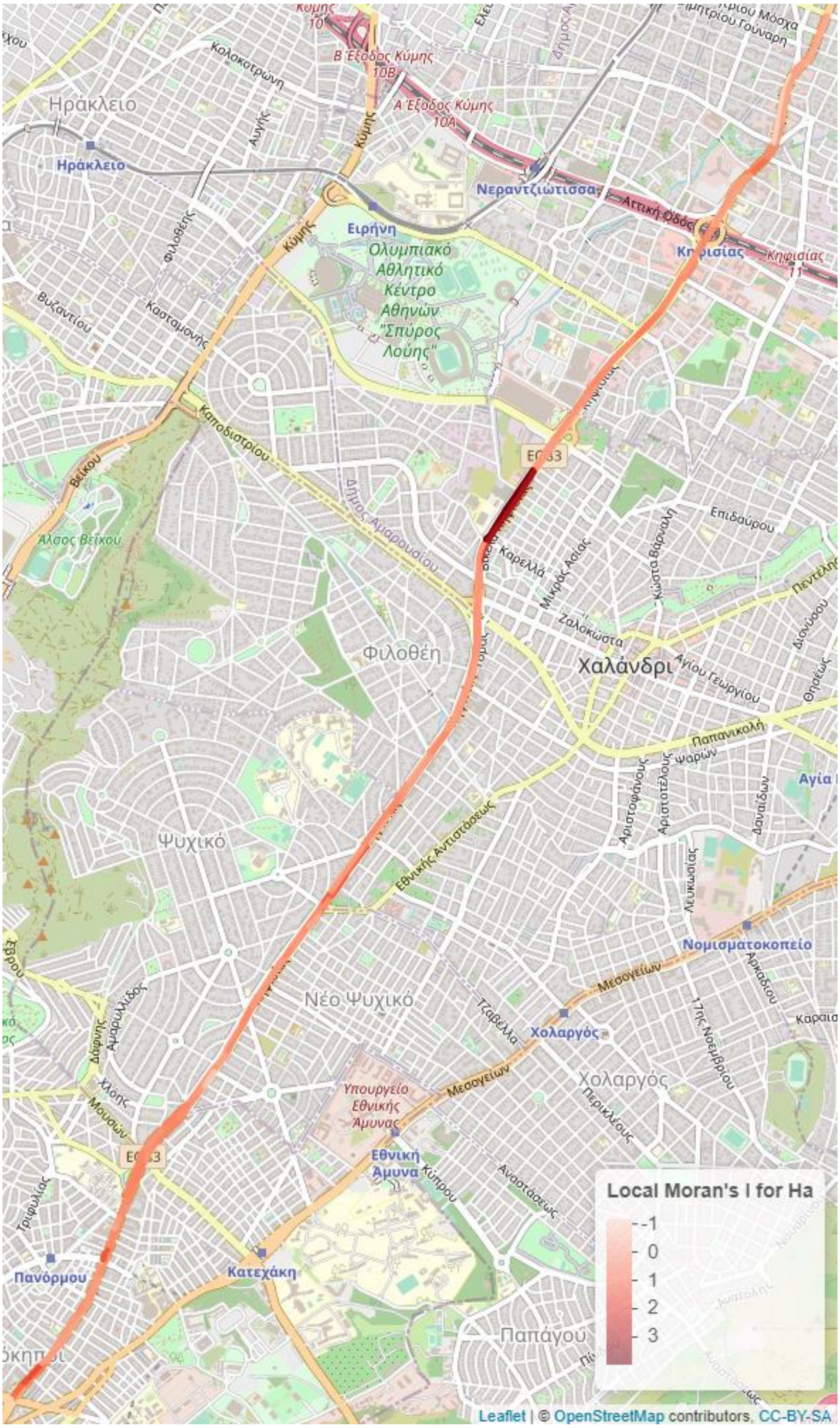


Figure 7-8: Local Moran's I values in Kifisias Avenue based on kNN-based weighting for harsh acceleration events

As in urban networks, it is obvious that kNN-based local Moran's I values are completely different from distance-based Moran's I values. The trend reversal combined with the change of magnitude both remain when transitioning form distance-based weighting to k-nearest neighbors weighting due to the underlying mathematical structure of the coefficient calculations. The sensitivities in the specification of Moran's I are once again underlined.

In the context of spatial autocorrelation, a road segment is more likely to be mostly affected by its direct neighbors rather than the entire area that it is located in. Therefore there is large positive local spatial autocorrelation of harsh brakings and harsh accelerations in certain middle road segments as highlighted on the maps of Figure 7-7 and Figure 7-8.

7.1.3 Harsh event variograms

Empirical variograms are plotted and their respective theoretical models are fitted for harsh event frequencies per road segment in Kifisias Avenue. As with Moran’s I, the entire harsh event dataset is considered Once again, the present configuration considers event frequencies as single predictors with a constant mean (Pebesma & Graeler, 2013).

Due to the one-dimensional nature of the study area, (single axis configuration), it is evident that directional variograms will not provide any added value for meaningful physical interpretation of harsh event distributions. Therefore the simple merged theoretical and empirical variograms were fitted and calculated. After tests of various theoretical modelling forms, it was found that the exponential variogram with a non-zero nugget fits the data by minimizing error distance.

The merged empirical and theoretical variograms are shown on Figure 7-9 for harsh braking events in the study area. Distance is measured in km from each road segment centroid. The partial sill of the exponential harsh braking variogram is 24.6935, with a range of 0.31059 km, while the nugget is 73.1283. The full sill (or maximum semivariance) after stabilization of the variogram is 97.8218. In practice, this indicates that about 310 m from each road segment centroid there is no observable spatial autocorrelation for harsh braking events on average.

In theoretical large road segment samples, the observations of harsh braking frequencies can be expected to be, on average, within the square root of the maximum semivariance from the mean, namely 9.89 harsh brakings. Most of the observations can be expected to lie within the range of two times that value, namely 19.78, based on the two-sigma rule.

Respectively, the empirical and theoretical variograms are shown on Figure 7-10 for harsh acceleration events in the study area. The overall magnitude of values is very similar with the harsh braking variograms. The partial sill of the exponential harsh acceleration variogram is 36.5925, with a range of 0.3218 km, while the nugget is 25.4606. The full sill (or maximum semivariance) after stabilization of the variogram is 62.0531. In practice, this indicates that about 320 m from each road segment centroid there is no observable spatial autocorrelation for harsh acceleration events on average.

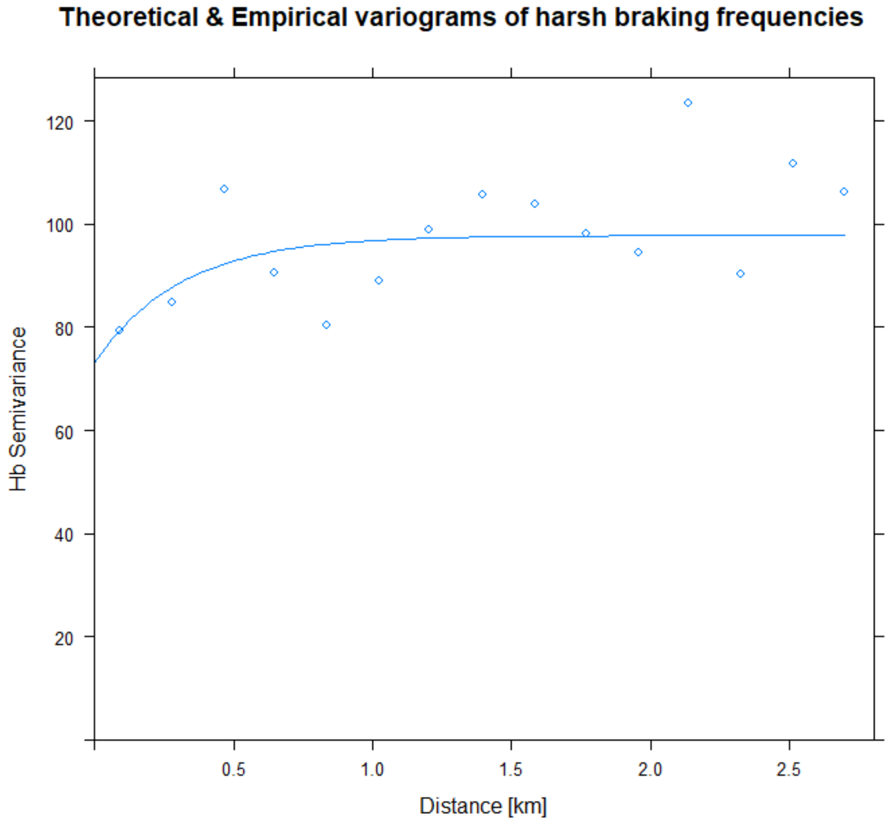


Figure 7-9: Merged empirical & theoretical variograms for harsh brakings in Kifisias Avenue

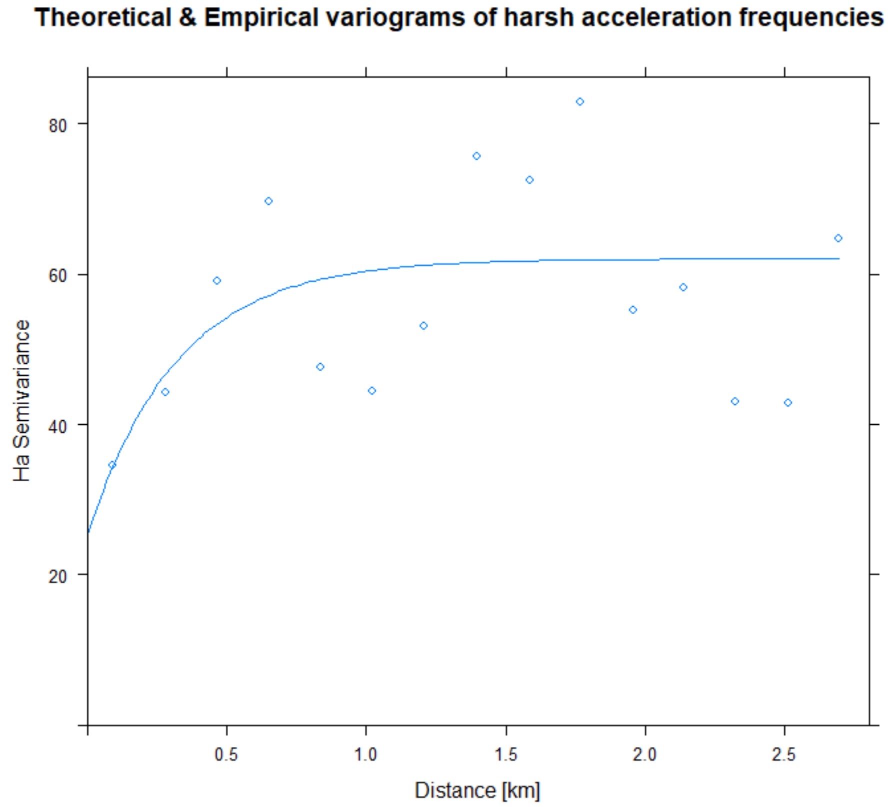


Figure 7-10: Merged empirical & theoretical variograms for harsh accelerations in Kifisias Avenue

Furthermore, in theoretical large road segment samples, the observations of harsh acceleration frequencies can be expected to be, on average, within the square root of the maximum semivariance from the mean, namely 7.88 harsh accelerations. Most of the observations can be expected to lie within the range of two times that value, namely 15.75, based on the two-sigma rule.

There are some additional noteworthy observations that can be made for the variograms of Kifisias Avenue. Firstly, compared to urban road network variograms, variograms for urban arterial segments appear to be more volatile. The definition of a specific theoretical variogram trend line was not as intuitive or apparent as in urban networks. Furthermore, there is spatial cyclicity observed in the axis for both harsh braking and harsh acceleration frequencies, which constitutes a wave-repetition pattern in the variograms. In other words, there is some repetitiveness in the patterns of data (Gringarten and Deutsch, 2001). This might indicate specific points where harsh events occur regarding their locations relevant to road segment centroids.

7.2 Geographically Weighted Poisson Regression results

In this section, Geographically Weighted Poisson Regression (GWPR) models are presented for urban arterial segments for harsh braking and harsh acceleration frequencies, for free flow and synchronized flow traffic states. The respective coefficients and various model metrics are interpreted. The model selection criteria remain identical to those described in Section 5.2.1.

7.2.1 Harsh braking models

Since the process of training GWPR models has been discussed in the previous, the following sections are a more compact description of GPWR models for free flow and synchronized flow conditions in Kifisias Avenue. Results are discussed collectively afterwards.

7.2.1.1 Free flow conditions model

Following Bivand et al. (2017) and Lu et al. (2013), bandwidth values were tested and their respective cross-validation (CV) score was calculated in an iterative process until convergence. Indicative results appear on **Table 7-5** – bandwidths are shown in km.

Table 7-5: Indicative bandwidth selection iterations for GWPR on harsh brakings for free flow conditions

Iteration number	Bandwidth value [km]	CV score
1	3.2053	2127.70
5	0.7631	1795.94
10	1.1374	1575.85
14	1.1394	1575.85
Optimal bandwidth:	**1.1394**	**1575.85**

The bandwidth of 1.14 km was selected for yielding optimal results in the study area by providing the minimum CV score. A series of GWPR regressions with different variable sets and subsequent backward elimination were conducted with the optimal bandwidth.

The resulting final GWPR model for harsh brakings in urban arterial segments for free flow conditions appears on **Table 7-6**. The p-values of statistically significant continuous variables and categorical variable categories (p-value ≤ 0.05) are shown in bold.

Table 7-6: GWPR model results for harsh brakings in urban arterial segments for free flow conditions

Independent variables	Coefficients			
	Estimate	Std. Error	z-value	p-value
Intercept	-0.2544	0.2231	-1.140	0.254
Gradient	-1.1013	0.7803	-1.411	0.158
Segment length	0.0033	0.0005	5.995	**0.000**
Pass count	0.0023	0.0005	4.636	**0.000**
Mobile use seconds	0.0022	0.0006	3.744	**0.000**
Speed difference	0.0385	0.0051	7.519	**0.000**
Average std. current traffic volume	-0.1640	0.0326	-5.025	**0.000**
Average occupancy	0.0595	0.0109	5.444	**0.000**
Bearing: Southbound [Ref.: Northbound]	-0.2611	0.0985	-2.652	**0.008**

In addition to the previous overall results, descriptive statistics are provided for the spatial variation of the coefficients on **Table 7-7**:

Table 7-7: Coefficient estimates of GWPR model for harsh brakings in urban arterial segments for free flow conditions

Independent variables	Coefficient estimates					
	Average	Min.	1st Quadrant	Median	3rd Quadrant	Max.
Intercept	-0.2544	-0.2543	-0.2542	-0.2541	-0.2540	-0.2540
Gradient	-1.1013	-1.1082	-1.1049	-1.1013	-1.0990	-1.0971
Segment length	0.0033	0.0033	0.0033	0.0033	0.0033	0.0033
Pass count	0.0023	0.0023	0.0023	0.0023	0.0023	0.0023
Mobile use seconds	0.0022	0.0022	0.0022	0.0022	0.0022	0.0022
Speed difference	0.0385	0.0385	0.0385	0.0385	0.0385	0.0386
Average std. current traffic volume	-0.1640	-0.1641	-0.1640	-0.1640	-0.1639	-0.1638
Average occupancy	0.0595	0.0594	0.0595	0.0595	0.0595	0.0596
Bearing: Southbound [Ref.: Northbound]	-0.2611	-0.2616	-0.2613	-0.2609	-0.2607	-0.2605

Model evaluation metrics are shown on **Table 7-8**:

Table 7-8: Evaluation metrics for the training of the GWPR model for harsh brakings in urban arterial segments for free flow conditions

Metric	Value	Metric	Value
Data-points	152	RMSE	2.8905
AIC	323.1902	MAE	2.0705
AICc	324.4620	RMSLE	0.6046
McFadden pseudo-R^2	0.521	CA	56.58%

7.2.1.2 Synchronized flow conditions model

Similarly, for synchronized flow conditions, bandwidth selection results appear on **Table 7-9** – bandwidths are shown in km.

Table 7-9: Indicative bandwidth selection iterations for GWPR on harsh brakings for synchronized flow conditions

Iteration number	Bandwidth value [km]	CV score
1	3.2053	630.37
5	2.4506	621.51
10	2.0660	617.71
15	2.0200	617.63
18	2.0199	617.63
Optimal bandwidth:	**2.0199**	**617.63**

The bandwidth of 2.02 km was selected for yielding optimal results in the study area by providing the minimum CV score. A series of GWPR regressions with different variable sets and subsequent backward elimination were conducted with the optimal bandwidth.

The resulting final GWPR model for harsh brakings in urban arterial segments for synchronized flow conditions appears on **Table 7-10**. The p-values of statistically significant continuous variables and categorical variable categories (p-value ≤ 0.05) are shown in bold.

Table 7-10: GWPR model results for harsh brakings in urban arterial segments for synchronized flow conditions

Independent variables	Coefficients			
	Estimate	Std. Error	z-value	p-value
Intercept	-2.1012	0.4076	-5.155	**0.000**
Curvature	71.6430	41.5065	1.726	0.084
Segment length	0.0024	0.0007	3.245	**0.001**
Pass count	0.0059	0.0008	7.658	**0.000**
Mobile use seconds	0.0113	0.0018	6.427	**0.000**
Average hourly traffic volume	-0.0002	0.0001	-2.167	**0.030**
Average occupancy	0.0495	0.0136	3.632	**0.000**

In addition to the previous overall results, descriptive statistics are provided for the spatial variation of the coefficients on **Table 7-11**:

Table 7-11: Coefficient estimates of GWPR model for harsh brakings in urban arterial segments for synchronized flow conditions

Independent variables	Coefficient estimates					
	Average	Min.	1st Quadrant	Median	3rd Quadrant	Max.
Intercept	-2.1012	-2.1014	-2.1012	-2.1010	-2.1008	-2.1007
Curvature	71.6430	71.6160	71.6279	71.6416	71.6639	71.6842
Segment length	0.0024	0.0024	0.0024	0.0024	0.0024	0.0024
Pass count	0.0059	0.0059	0.0059	0.0059	0.0059	0.0059
Mobile use seconds	0.0113	0.0113	0.0113	0.0113	0.0113	0.0113
Average hourly traffic volume	-0.0002	-0.0002	-0.0002	-0.0002	-0.0002	-0.0002
Average occupancy	0.0495	0.0495	0.0495	0.0495	0.0495	0.0495

Model evaluation metrics are shown on **Table 7-12**:

Table 7-12: Evaluation metrics for the training of the GWPR model for harsh brakings in urban arterial segments for synchronized flow conditions

Metric	Value	Metric	Value
Data-points	152	RMSE	1.6733
AIC	187.7578	MAE	0.9404
AICc	188.5362	RMSLE	0.4306
McFadden pseudo-R^2	0.696	CA	83.55%

7.2.1.3 Discussion of results

GPWR analysis of harsh braking events provides numerous insights and useful results. When comparing the model outcomes for the two traffic states, both similarities and differences emerge. It is apparent that the array of significant variables is different in the two models. The variables that are important across both traffic states – namely segment length, pass count, total seconds of mobile phone use by drivers and average occupancy – retain the same influence without trend reversals, all increasing harsh braking frequencies in urban arterial road segments.

Conversely, traffic variables displayed a largely different influence across traffic states, an outcome which was a priori expected. The exception to the previous motive is average occupancy, which was found to increase harsh braking frequencies consistently in both cases, indicating perhaps a more reliable traffic predictor.

In free flow conditions, speed difference was found to increase harsh braking frequency. This is a reasonable finding, with a straightforward physical interpretation: drivers that differentiate significantly from the speed of surrounding traffic find themselves having reduced available headways that generate more harsh brakings in turn, especially when other drivers see conservative driving as an opportunity to overtake. Average standardized current traffic volume was found to reduce harsh event frequencies. In addition, a systematic spatial difference was found in the road segments from the flag variable of bearing: northbound segments were found to have significantly increased harsh braking frequencies compared to southbound segments.

In synchronized flow conditions, average hourly traffic volume was found to be significant instead, reducing harsh event frequencies. Speed difference was no longer statistically significant.

The interpretation of the effects of traffic occupancy and traffic volume is more comprehensive when these parameters are viewed jointly. Traffic occupancy can be viewed as a representation of cluttering of the road surface, and its increases lead to more possible vehicle conflicts, thus more harsh braking frequencies. Conversely, since congested flow conditions are excluded, increases in traffic volume represent an ease of movement for traffic with fewer harsh brakings.

For each model, one geometrical variable acts as a latent ‘binding agent’ – specifically gradient for free flow and curvature for synchronized flow. These variables are not statistically significant per se, but if they are removed the respective models display lower performance as measured by AICc, error metrics and custom accuracy (CA). This indicates highly complex and perhaps volatile relationships of road geometry with harsh braking frequency which cannot be captured by models operating in a generalized linear framework; it is possible that random effects or ensemble tree methods might succeed better in this task.

Similarly, a number of variables were not found to be statistically significant in any model, namely speeding duration, traffic lights or pedestrian crossings, number of lanes (except when integrated in average standardized current traffic volume).

The spatial fluctuation of the estimated coefficients as shown on **Table 7-7** and **Table 7-11** is low. It manifests in all variables in this analysis, as opposed to urban road networks, even in the two exposure variables of segment length and pass count for harsh brakings.

Marginal Effects at the Means (MEM) are calculated for the exposure variables following Washington et al. (2010) considering the mean data points of these variables. In free flow conditions, for the segment length average of 122.6 m, an increase of 1 meter leads to an increase of $MEM_{Seg\_Length} = 0.0049$ harsh brakings. For the pass count average of 225 passes, an increase of 1 pass leads to an increase of $MEM_{Pass\ count} = 0.0037$ harsh brakings.

For synchronized flow conditions, for the segment length average of 122.6 m, an increase of 1 meter leads to an increase of $MEM_{Seg\_Length} = 0.0032$ harsh brakings; the result is different due to the change of coefficient. For the pass count average of 91 passes, an increase of 1 pass leads to an increase of $MEM_{Pass\ count} = 0.0100$ harsh brakings. Therefore when examining urban arterial segments under free flow conditions, segment lengths contribute to more harsh brakings per unit compared to pass counts, a trend which does not hold for synchronized flow conditions (or urban road networks).

The McFadden pseudo-R^2 for the GLM component is satisfactory at 0.52 for free flow and very satisfactory at 0.70 for synchronized flow, given its typical lower values than linear R^2 coefficients. Harsh brakings are predicted accurately with a tolerance of ± 1 harsh braking per segment 57% of the times for free flow and 84% for synchronized flow. The error metrics and CA show that the model for synchronized conditions fits the spatial data better. It is possible that harsh event frequencies in free flow cannot be optimally described by a generalized linear framework, which is further corroborated by the fact that the intercept is not statistically significant for free flow. However, error metrics are not directly comparable for two models, since trip-seconds in synchronized segments have smaller values, namely one third of those in free flow segments.

Due to the unique configuration of GWR/GWPR, maps can be created for the localized coefficient values of every variable in the model for the study area. Figure 7-11 features the mapping of the coefficient of speed difference, indicatively for free flow, and Figure 7-12 features the mapping of the coefficient of average occupancy, indicatively for synchronized flow. It should be noted that the graphical scale is significantly exaggerated compared to the low spatial fluctuations of the coefficient.

Nonetheless, there are clear visible trends: In both instances the examined variables appear to contribute to more harsh brakings in northern road segments compared to southern segments, with the middle sector serving as a smooth middle ground transition for the coefficients.

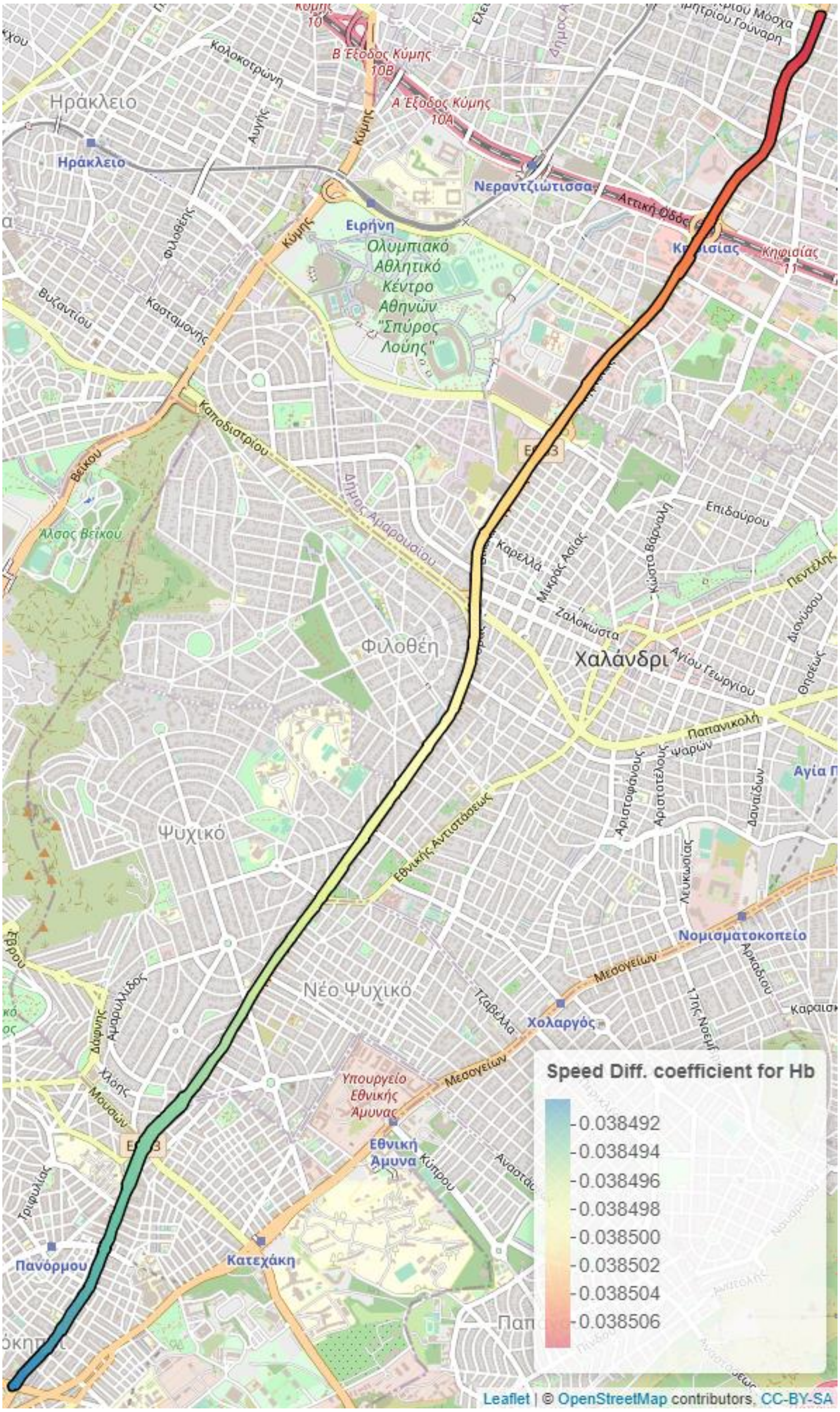


Figure 7-11: GWPR speed difference coefficients of harsh brakings in Kifisias Avenue for free flow conditions

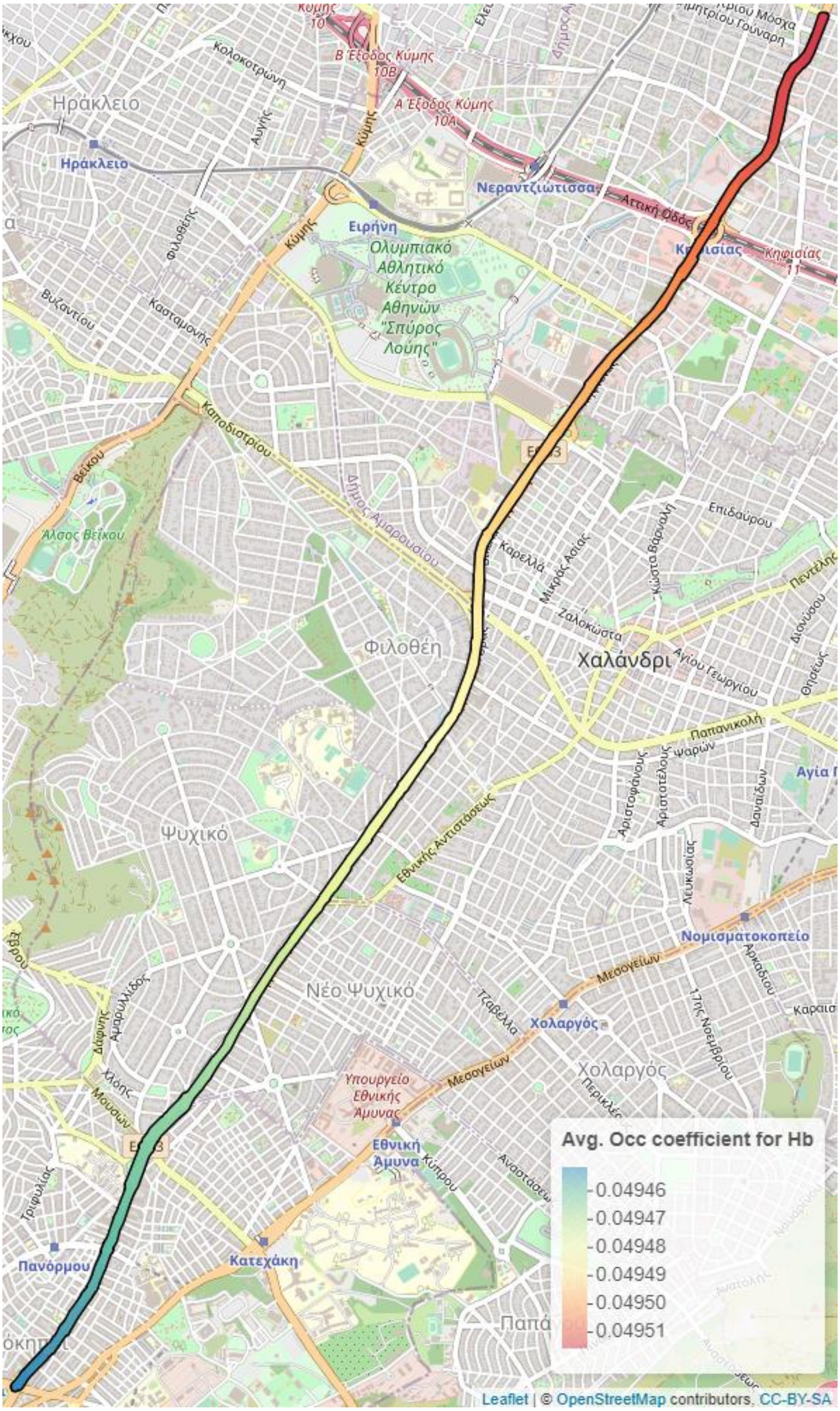


Figure 7-12: GWPR average occupancy coefficients of harsh brakings in Kifisias Avenue for synchronized flow conditions

7.2.2 Harsh acceleration models

The previous process is mirrored for harsh acceleration models in Kifisias Avenue. Results are discussed collectively afterwards.

7.2.2.1 Free flow conditions model

For free flow conditions, bandwidth selection results appear on **Table 7-13** – bandwidths are shown in km.

Table 7-13: Indicative bandwidth selection iterations for GWPR on harsh accelerations for free flow conditions

Iteration number	Bandwidth value [km]	CV score
1	3.2053	2520.33
5	4.4265	2462.58
10	3.9661	2456.60
15	3.9529	2456.60
16	3.9529	2456.60
Optimal bandwidth:	**3.9529**	**2456.60**

The bandwidth of 3.95 km was selected for yielding optimal results in the study area by providing the minimum CV score. A series of GWPR regressions with different variable sets and subsequent backward elimination were conducted with the optimal bandwidth.

The resulting final GWPR model for harsh accelerations in urban arterial segments for free flow conditions appears on **Table 7-14**. The p-values of statistically significant continuous variables and categorical variable categories (p-value ≤ 0.05) are shown in bold.

Table 7-14: GWPR model results for harsh accelerations in urban arterial segments for free flow conditions

Independent variables	Coefficients			
	Estimate	Std. Error	z-value	p-value
Intercept	-0.2237	0.4148	-0.539	0.590
Segment length	0.0017	0.0008	2.066	**0.039**
Pass count	0.0032	0.0006	4.918	**0.000**
Average traffic speed	-0.0240	0.0086	-2.778	**0.005**
Speed difference	0.0528	0.0074	7.104	**0.000**
Average occupancy	0.0258	0.0119	2.163	**0.031**
Bearing: Southbound [Ref.: Northbound]	-0.2434	0.1232	-1.977	**0.048**
Mobile use seconds	0.0027	0.0008	3.488	**0.000**
Speeding seconds	-0.0011	0.0005	-2.025	**0.043**

In addition to the previous overall results, descriptive statistics are provided for the spatial variation of the coefficients on **Table 7-15**:

Table 7-15: Coefficient estimates of GWPR model for harsh accelerations in urban arterial segments for free flow conditions

Independent variables	Coefficient estimates					
	Average	Min.	1st Quadrant	Median	3rd Quadrant	Max.
Intercept	-0.2237	-0.2240	-0.2239	-0.2238	-0.2236	-0.2234
Segment length	0.0017	0.0017	0.0017	0.0017	0.0017	0.0017
Pass count	0.0032	0.0032	0.0032	0.0032	0.0032	0.0032
Average traffic speed	-0.0240	-0.0240	-0.0240	-0.0240	-0.0240	-0.0240
Speed difference	0.0528	0.0528	0.0528	0.0528	0.0528	0.0528
Average occupancy	0.0258	0.0258	0.0258	0.0258	0.0258	0.0258
Bearing: Southbound [Ref.: Northbound]	-0.2434	-0.2435	-0.2435	-0.2434	-0.2434	-0.2434
Mobile use seconds	0.0027	0.0027	0.0027	0.0027	0.0027	0.0027
Speeding seconds	-0.0011	-0.0011	-0.0011	-0.0011	-0.0011	-0.0011

Model evaluation metrics are shown on **Table 7-16**:

Table 7-16: Evaluation metrics for the training of the GWPR model for harsh accelerations in urban arterial segments for free flow conditions

Metric	Value	Metric	Value
Data-points	152	RMSE	2.2817
AIC	296.4888	MAE	1.5816
AICc	297.7568	RMSLE	0.6305
McFadden pseudo-R^2	0.435	CA	63.16%

7.2.2.2 Synchronized flow conditions model

For synchronized flow conditions, bandwidth selection results appear on **Table 7-17** – bandwidths are shown in km.

Table 7-17: Indicative bandwidth selection iterations for GWPR on harsh accelerations for synchronized flow conditions

Iteration number	Bandwidth value [km]	CV score
1	3.2053	475.15
5	2.9454	473.93
10	2.8034	473.73
14	2.8034	473.73
Optimal bandwidth:	**2.8034**	**473.73**

The bandwidth of 2.80 km was selected for yielding optimal results in the study area by providing the minimum CV score. A series of GWPR regressions with different variable sets and subsequent backward elimination were conducted with the optimal bandwidth.

The resulting final GWPR model for harsh accelerations in urban arterial segments for synchronized flow conditions appears on **Table 7-18**. The p-values of statistically significant continuous variables and categorical variable categories (p-value ≤ 0.05) are shown in bold.

Table 7-18: GWPR model results for harsh accelerations in urban arterial segments for synchronized flow conditions

Independent variables	Coefficients			
	Estimate	Std. Error	z-value	p-value
Intercept	-1.2573	0.4381	-2.870	**0.004**
Pass count	0.0035	0.0010	3.369	**0.001**
Average traffic speed	-0.0240	0.0073	-3.299	**0.001**
Bearing: Southbound [Ref.: Northbound]	0.4721	0.1929	2.447	**0.014**
Mobile use seconds	0.0148	0.0019	8.012	**0.000**
Average traffic volume	0.0003	0.0001	2.326	**0.020**

In addition to the previous overall results, descriptive statistics are provided for the spatial variation of the coefficients on **Table 7-19**:

Table 7-19: Coefficient estimates of GWPR model for harsh accelerations in urban arterial segments for synchronized flow conditions

Independent variables	Coefficient estimates					
	Average	Min.	1st Quadrant	Median	3rd Quadrant	Max.
Intercept	-1.2573	-1.2582	-1.2579	-1.2576	-1.2571	-1.2566
Pass count	0.0035	0.0035	0.0035	0.0035	0.0035	0.0035
Average traffic speed	-0.0240	-0.0240	-0.0240	-0.0240	-0.0240	-0.0239
Bearing: Southbound [Ref.: Northbound]	0.4721	0.4721	0.4721	0.4721	0.4722	0.4722
Mobile use seconds	0.0148	0.0148	0.0148	0.0148	0.0148	0.0148
Average traffic volume	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003

Model evaluation metrics are shown on **Table 7-20**:

Table 7-20: Evaluation metrics for the training of the GWPR model for harsh accelerations in urban arterial segments for synchronized flow conditions

Metric	Value	Metric	Value
Data-points	152	RMSE	1.2978
AIC	197.9967	MAE	0.8258
AICc	198.5763	RMSLE	0.4507
McFadden pseudo-R^2	0.509	CA	86.84%

7.2.2.3 Discussion of results

When comparing the model outcomes for the two traffic states, both similarities and differences emerge once again. It is apparent that the array of significant variables is different in the two models. The naturalistic driving variables that are important across both traffic states – namely pass count and total seconds of mobile phone use by drivers and average occupancy – retain the same influence without trend reversals and increase harsh acceleration frequencies in urban arterial road segments.

Segment length also retains the same positive influence for free flow conditions, contributing to more harsh accelerations. In synchronized flow conditions, and for the first time in all GWPR analyses of the present research, segment length was not found to be a statistically significant exposure variable.

Regarding traffic variables, the first important finding is that average traffic speed reduces harsh acceleration events in both traffic conditions. This is an intuitive finding; as traffic moves faster across the arterial segments, drivers do not feel the need to abruptly accelerate their vehicles. In free flow, increases of seconds of exceeding the speed limit led to fewer harsh events. As drivers who speed move already fast, they do not have as many reasons and/or vehicle capabilities to harshly accelerate further.

For free flow conditions speed difference was found to significantly contribute to harsh accelerations as well. The interpretation is that as drivers find increased headways from their positions as surrounding traffic moves faster away from them, they may find the margin to accelerate to overtake or catch up. In synchronized flow, average hourly traffic volume was found to significantly contribute to harsh accelerations instead of speed difference. As the number of surrounding vehicles increases, it is possible that drivers feel the need to abruptly adjust and quickly cover a certain distance in order to move to more favorable positions in traffic.

The variable of segment bearing (direction) was found to be statistically significant for both traffic states, albeit with a trend reversal: In free flow more harsh accelerations occur in northbound segments, while in synchronized flow more harsh accelerations occur in southbound segments. This is an indicator of a localized systematic difference of infrastructure, which is informative for the models. Moreover, it has an important implication: it serves as a confirmation for the necessity of studying these road environments under different flow conditions, and not collectively.

It is important to note the absence of geometrical variables in the harsh acceleration models, as opposed to harsh braking models. It would appear that drivers are not affected by local geometrical characteristics when deciding to abruptly accelerate in an urban arterial. The spatial fluctuation of the estimated coefficients as shown on **Table 7-15** and **Table 7-19** is low. It manifests in almost all variables in this analysis, though for some variables in very small decimals.

Marginal Effects at the Means (MEM) are calculated for the exposure variables following Washington et al. (2010) considering the mean data points of these variables. In free flow conditions, for the segment length average of 122.6 m, an increase of 1 meter leads to an increase of $MEM_{Seg\_Length} = 0.0021$ harsh accelerations. For the pass count average of 225 passes, an increase of 1 pass leads to an increase of $MEM_{Pass\ count} = 0.0064$ harsh accelerations.

For synchronized flow conditions, for the pass count average of 91 passes, an increase of 1 pass leads to an increase of $MEM_{Pass\ count} = 0.0048$ harsh accelerations (segment lengths are not significant).

Therefore when examining urban arterial segments under free flow conditions, segment lengths contribute to fewer harsh accelerations per unit compared to pass counts.

The McFadden pseudo-R^2 for the GLM component is at satisfactory levels at 0.44 for free flow and at 0.51 for synchronized flow, given its typical lower values than linear R^2 coefficients. Harsh accelerations are predicted accurately with a tolerance of ± 1 harsh acceleration per segment 63% of the times for free flow and 87% for synchronized flow. Similarly with harsh brakings, the error metrics and CA show that the model for synchronized conditions fits the spatial data better, and the explanations are considered to be the same: free flow has more events which cannot be perfectly captured in a glm structure.

Due to the unique configuration of GWR/GWPR, maps can be created for the localized coefficient values of every variable in the model for the study area. Figure 7-13 features the mapping of the coefficient of speed difference, indicatively for free flow, and Figure 7-14 features the mapping of the coefficient of average traffic speed, indicatively for synchronized flow. It should be noted that the graphical scale is significantly exaggerated compared to the low spatial fluctuations of the coefficient.

Nonetheless, there are clear visible trends: In both instances the examined variables appear to contribute to more harsh accelerations in southern road segments compared to north segments, with the middle sector serving as a smooth middle ground transition for the coefficients.

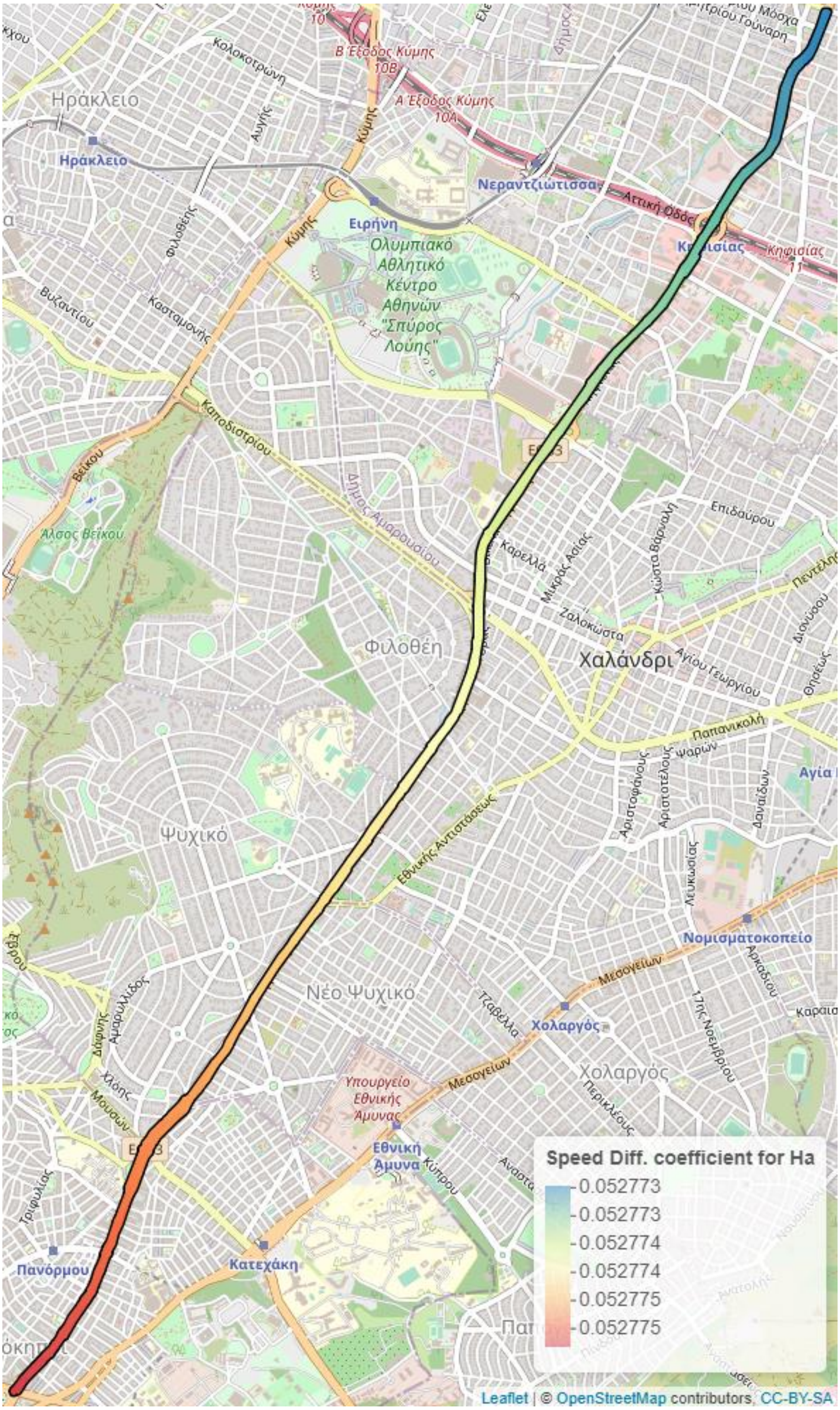


Figure 7-13: GWPR speed difference coefficients of harsh accelerations in Kifisias Avenue for free flow conditions

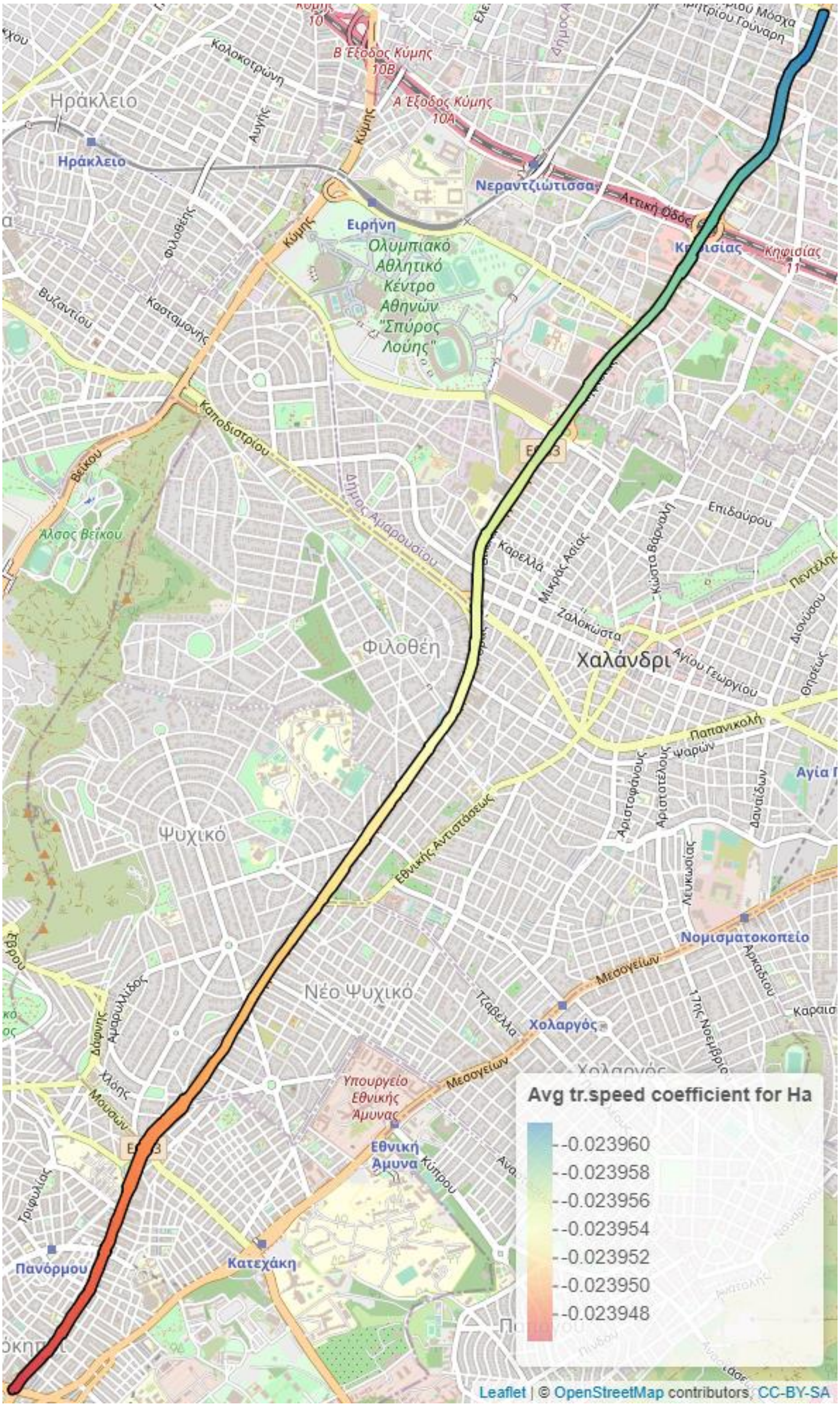


Figure 7-14: GWPR average traffic speed coefficients of harsh accelerations in Kifisias Avenue for synchronized flow conditions

7.3 Conditional Autoregressive Prior models

In this section, Bayesian Poisson lognormal models with conditional autoregressive priors (CAR models) are presented for urban arterial segments for harsh braking and harsh acceleration frequencies, for free flow and synchronized flow traffic states. The respective coefficients and various model metrics are interpreted. The model selection criteria remain identical to those described in Section 5.3.1.

7.3.1 Harsh braking models

Similar to GWPR, Bayesian analysis with CAR models is conducted for all the road segments in the training area that are traversed by vehicles. CAR Models were calibrated following Lee (2013).

7.3.1.1 Free flow conditions model

Bayesian inference is conducted using Markov Chain Monte Carlo (MCMC) simulation. It was found that the best performing models required a large burn-in period before stabilization. After several trials, the posterior summaries for the best-fitting models were obtained by a chain with 410,000 iterations, the first 400,000 of which were discarded as the burn-in sample. The remainder 10,000 samples are thinned by 100 to reduce autocorrelation and the resulting values describe the posterior distributions. A fixed value for the random number generation processes is also required to ensure the replicability of results.

The resulting final CAR model for harsh brakings in free flow conditions appears on **Table 7-21**. The 95% BCI values are calculated at 2.5% (lower bound) to 97.5% (upper bound), and median values refer to this 95% BCI margin only. 95% BCI values of statistically significant continuous variables and categorical variable categories – which retain the same signs – are shown in bold.

Table 7-21: CAR model results for harsh brakings in urban arterial segments for free flow conditions

Independent variables	**Posterior values**				
	Mean	St. Dev.	Median	2.5% value	97.5% value
Intercept	-0.4664	0.4065	-0.4508	-1.2726	0.2710
Segment length	0.0031	0.0009	0.0031	**0.0014**	**0.0050**
Pass count	0.0027	0.0010	0.0026	**0.0010**	**0.0043**
Mobile use seconds	0.0042	0.0013	0.0042	**0.0014**	**0.0068**
Bearing: Southbound [Ref.: Northbound]	-0.2746	0.1564	-0.2838	-0.5868	0.0102
Speed difference	0.0318	0.0086	0.0325	**0.0123**	**0.0462**
Average std. current traffic volume	-0.0417	0.0497	-0.0400	-0.1293	0.0458
Sigma-phi^2 [Spatially structured effects]	0.0662	0.3194	0.0120	**0.0035**	**0.5027**
Sigma-theta^2 [Spatially unstructured effects]	0.3796	0.1046	0.3760	**0.2056**	**0.6073**

Model evaluation metrics are shown on **Table 7-22**:

Table 7-22: Evaluation metrics for the training of the CAR model for harsh brakings in urban arterial segments for free flow conditions

Metric	**Value**	**Metric**	**Value**
Data-points	152	RMSE	1.1052
DIC	627.335	MAE	0.9002
WAIC	685.909	RMSLE	0.3565
LMPL	-432.384	CA	84.22%

7.3.1.2 Synchronized flow conditions model

The posterior summaries for the best-fitting models were obtained by a chain with 410,000 iterations, the first 400,000 of which were discarded as the burn-in sample. The remainder 10,000 samples are thinned by 100 to reduce autocorrelation and the resulting values describe the posterior distributions. A fixed value for the random number generation processes is also required to ensure the replicability of results.

For synchronized conditions, the resulting final CAR model for harsh brakings appears on **Table 7-23**.

Table 7-23: CAR model results for harsh brakings in urban arterial segments for synchronized flow conditions

Independent variables	Posterior values				
	Mean	St. Dev.	Median	2.5% value	97.5% value
Intercept	-2.4520	0.9653	-2.3930	**-4.5402**	**-0.8008**
Gradient	1.0782	2.3259	1.2740	-3.5162	5.0933
Curvature	6.6068	76.8309	16.5296	-187.4946	126.2128
Segment length	0.0019	0.0013	0.0019	**0.0009**	**0.0043**
Pass count	0.0057	0.0014	0.0059	**0.0030**	**0.0082**
Mobile use seconds	0.0134	0.0037	0.0130	**0.0072**	**0.0216**
Average occupancy	0.0371	0.0199	0.0387	**0.0005**	**0.0745**
Lanes: 2 [Ref.: Lanes: 1]	0.2878	0.3357	0.2608	-0.4291	0.9370
Lanes: 3 [Ref.: Lanes: 1]	-0.0207	0.3684	0.0059	-0.7539	0.7771
Lanes: 4 [Ref.: Lanes: 1]	-1.9839	0.9180	-1.9661	**-3.5933**	**-0.4680**
Speeding seconds	0.0020	0.0016	0.0020	-0.0004	0.0048
Average std. current traffic volume	-0.0195	0.0501	-0.0229	-0.1089	0.0623
Bearing: Southbound [Ref.: Northbound]	0.0119	0.2883	0.0394	-0.6320	0.4772
Sigma-phi^2 [Spatially structured effects]	0.0309	0.0469	0.0127	**0.0031**	**0.1721**
Sigma-theta^2 [Spatially unstructured effects]	0.3916	0.2210	0.3466	**0.0826**	**0.8997**

Model evaluation metrics are shown on **Table 7-24**:

Table 7-24: Evaluation metrics for the training of the CAR model for harsh brakings in urban arterial segments for synchronized flow conditions

Metric	Value	Metric	Value
Data-points	152	RMSE	0.7472
DIC	333.270	MAE	0.5206
WAIC	335.783	RMSLE	0.2971
LMPL	-172.426	CA	90.79%

7.3.1.3 Discussion of results

It is evident that the introduction of spatially structured and unstructured effects led to several differences in the models, although some similarities with previous models persist.

Speed difference was found to be positively correlated with increased harsh braking frequencies in free flow conditions. Furthermore, in synchronized flow conditions, average occupancy was found to positively contribute to harsh braking occurrence. Both of these findings are retained from the respective GWPR models. For both CAR models, average standardized current traffic volume was not statistically significant. Nonetheless, its inclusion was necessary for good model performance. Removing it lead to aberrant model behavior with explosive metrics.

In both traffic states, the exposure variables of segment length and pass count retain their positive signs, contributing positively to harsh braking occurrence. MEM can be again calculated following Washington et al. (2010), albeit without including any spatially structured or unstructured effects. In free flow conditions, for the segment length average of 122.6 m, an increase of 1 meter leads to an increase of $MEM_{Seg\_Length} = 0.0045$ harsh brakings. For the pass count average of 225 passes, an increase of 1 pass leads to an increase of $MEM_{Pass\ count} = 0.0050$ harsh brakings.

For synchronized flow conditions, for the segment length average of 122.6 m, an increase of 1 meter leads to an increase of $MEM_{Seg\_Length} = 0.0042$ harsh brakings; the result is different due to the change of coefficient. For the pass count average of 91 passes, an increase of 1 pass leads to an increase of $MEM_{Pass\ count} = 0.0010$ harsh brakings. Therefore under both traffic states, pass counts contribute to more harsh brakings per unit compared to segment lengths.

Additionally, in both traffic states, the average seconds of mobile phone use by drivers led to increased harsh braking frequencies. This finding is intuitive and in line with previous results; it also hints at the effects of driver distraction as well.

Apart from segment length, no geometric characteristics such as gradient and curvature have been found statistically significant. Similarly with average standardized current traffic volume, their inclusion was necessary for the good performance of the synchronized flow model. Since the elimination of these variables led to aberrant model behavior, it is possible that the latent information provided by gradient and curvature aids in the determination of the distribution of the spatially structured and unstructured effects.

In synchronized flow conditions, it appears that urban arterial segments with four lanes have statistically fewer harsh brakings compared to segments with one lane; this influence does not seem to extend to segments with two or three lanes. Other road characteristics, such as the presence of pedestrian crossings and traffic lights, did not retain any significant influence after network characteristics were introduced.

The value ranges of spatially structured and spatially unstructured effects, represented by ${\sigma_{\varphi i}}^2$ and ${\sigma_{\theta i}}^2$, respectively, were comparable in urban arterials, as opposed to urban networks. This indicates roughly equal magnitudes of unobserved spatial and non-spatial factors. Another noteworthy point is that the intercept is significant only in the synchronized flow model, similar to GWPR. This is a possible hint of harsh braking distributions that do not exactly conform to a generalized linear framework in free flow conditions.

The metrics of RMSE denote an average error magnitude of 1.1 and 0.7 harsh braking counts, while MAE is lower at 0.9 and 0.5 harsh braking counts. Overall, it appears that the inclusion of spatially structured and spatially unstructured effects improves model fitting performance compared to GWPR, as measured by the error metrics and CA. The improvement is particularly pronounced in free flow conditions, with a CA gain of 27.64%. On one hand, these effects allow for more precise model fitting, but on the other, they remove any possibility of transferability. Once again, the synchronized flow model fits the data better than the free flow model.

7.3.2 Harsh acceleration models

Bayesian CAR models were fitted for harsh acceleration frequencies in Kifisias Avenue as well. Results are discussed collectively afterwards.

7.3.2.1 Free flow conditions model

It was determined that the best performing models required a large burn-in period before stabilization. The posterior summaries for the best-fitting models were obtained by a chain with 210,000 iterations, the first 200,000 of which were discarded as the burn-in sample. The remainder 10,000 samples are thinned by 100 to reduce autocorrelation and the resulting values describe the posterior distributions. A fixed value for the random number generation processes is also required to ensure the replicability of results. The resulting final CAR model for harsh accelerations in free flow conditions appears on **Table 7-25**.

Table 7-25: CAR model results for harsh accelerations in urban arterial segments for free flow conditions

Independent variables	Posterior values				
	Mean	St. Dev.	Median	2.5% value	97.5% value
Intercept	-1.0912	0.3730	-1.1263	**-1.8890**	**-0.3498**
Gradient	1.2874	1.0077	1.3254	-0.8988	3.1334
Segment length	0.0011	0.0012	0.0010	-0.0010	0.0034
Pass count	0.0022	0.0010	0.0022	**0.0000**	**0.0039**
Mobile use seconds	0.0047	0.0016	0.0048	**0.0020**	**0.0075**
Speeding seconds	-0.0012	0.0007	-0.0012	**-0.0024**	**-0.0000**
Speed difference	0.0323	0.0084	0.0328	**0.0153**	**0.0479**
Average occupancy	0.0328	0.0149	0.0335	**0.0008**	**0.0599**
Bearing: Southbound [Ref.: Northbound]	-0.2327	0.1605	-0.2266	-0.5247	0.0614
Sigma-phi^{2} [Spatially structured effects]	0.5614	2.0561	0.0145	**0.0035**	**4.8643**
Sigma-theta^{2} [Spatially unstructured effects]	0.3253	0.1110	0.3137	**0.1594**	**0.5203**

Model evaluation metrics are shown on **Table 7-26**:

Table 7-26: Evaluation metrics for the training of the CAR model for harsh accelerations in urban arterial segments for free flow conditions

Metric	Value	Metric	Value
Data-points	152	RMSE	1.0912
DIC	543.354	MAE	0.8612
WAIC	549.602	RMSLE	0.4286
LMPL	-283.054	CA	86.18%

7.3.2.2 Synchronized flow conditions model

The posterior summaries for the best-fitting models were obtained by a chain with 410,000 iterations, the first 400,000 of which were discarded as the burn-in sample. The remainder 10,000 samples are thinned by 100 to reduce autocorrelation and the resulting values describe the posterior distributions. A fixed value for the random number generation processes is also required to ensure the replicability of results. The resulting final CAR model for harsh accelerations in synchronized flow conditions appears on **Table 7-27**.

Table 7-27: CAR model results for harsh accelerations in urban arterial segments for synchronized flow conditions

Independent variables	Posterior values				
	Mean	St. Dev.	Median	2.5% value	97.5% value
Intercept	-2.9399	0.8226	-2.9677	**-4.7206**	**-1.4793**
Gradient	1.7464	2.8951	1.8003	-4.5321	7.8190
Segment length	-0.0005	0.0021	-0.0003	-0.0045	0.0029
Mobile use seconds	0.0159	0.0060	0.0164	**0.0046**	**0.0264**
Average occupancy	0.0237	0.0242	0.0218	-0.0184	0.0685
Lanes: 2 [Ref.: Lanes: 1]	-1.0085	0.5069	-1.0784	**-2.0273**	**-0.0389**
Lanes: 3 [Ref.: Lanes: 1]	-1.8321	0.5409	-1.7989	**-3.0067**	**-0.7748**
Lanes: 4 [Ref.: Lanes: 1]	-2.8837	0.9416	-2.7485	**-4.8042**	**-1.3752**
Speeding seconds	-0.0039	0.0030	-0.0041	-0.0092	0.0027
Average hourly traffic volume	0.0008	0.0004	0.0008	**0.0001**	**0.0015**
Pass count	0.0065	0.0025	0.0064	**0.0020**	**0.0111**
Bearing: Southbound [Ref.: Northbound]	0.2981	0.4186	0.2363	-0.4149	1.1266
Average driver speed	-0.0224	0.0179	-0.0205	-0.0612	0.0125
Sigma-phi^2 [Spatially structured effects]	0.0699	0.1595	0.0155	**0.0032**	**0.5204**
Sigma-theta^2 [Spatially unstructured effects]	1.0935	0.4969	1.0099	**0.4153**	**2.3129**

Model evaluation metrics are shown on **Table 7-28**:

Table 7-28: Evaluation metrics for the training of the CAR model for harsh accelerations in urban arterial segments for synchronized flow conditions

Metric	Value	Metric	Value
Data-points	152	RMSE	0.5404
DIC	308.408	MAE	0.3904
WAIC	314.833	RMSLE	0.2475
LMPL	-168.141	CA	97.36%

7.3.2.3 Discussion of results

The developed CAR models reveal interesting outputs for harsh acceleration frequencies as well.

In free flow conditions, the difference of traffic and driver speeds has been found to generate more harsh accelerations. The explanation lies in the creation of spatial and temporal headways from the speed differences, which drivers then exploit to accelerate abruptly. Similarly with GPWR, increases in average occupancy lead to more harsh accelerations. It is possible that drivers feel some pressure from the increasing numbers of surrounding vehicles and seek to exploit possible windows through harsh acceleration manoeuvers.

When examining synchronized flow conditions, the variable of average hourly traffic volume is found to significantly influence harsh accelerations in a positive manner. This is another finding that persisted from GWPR models. Increased traffic volume indicates an increased amount of vehicles traversing the road segment, with higher speeds (in synchronized flow). This is an environment that can encourage more abrupt accelerating behavior on the part of drivers, thus increasing the frequency of harsh accelerations. Furthermore, it is also worth noting that the influence of traffic occupancy weakens to not being statistically significant. To continue with synchronized flow investigation, an increased number of lanes seem to provide more space to drivers, thus reducing the frequencies of harsh acceleration events.

The exposure variable of pass count was found to positively contribute to harsh acceleration frequencies across both traffic states. In free flow conditions, for the pass count average of 225 passes, an increase of 1 pass leads to an increase of $MEM_{Pass\ count} = 0.0036$ harsh accelerations. For synchronized flow conditions, for the pass count average of 91 passes, an increase of 1 pass leads to an increase of $MEM_{Pass\ count} = 0.0012$ harsh accelerations. Therefore it can be said that a single pass contributes to more harsh acceleration events (about 3 times more) when driving in synchronized flow conditions as opposed to free flow conditions.

Moreover, in both free flow and synchronized flow, increases in the average seconds of mobile use are positively correlated with more harsh accelerations in a given road segment. This finding persists from GWPR models as well. It can be attributed to driver recklessness as distraction from mobile use increases, with drivers not fully realizing they are accelerating excessively, or perhaps can be attributed to cases of urgency, as drivers also look for information on their smartphones.

In free flow and synchronized flow conditions, the inclusion of traffic and behavioral variables, as well as spatially structured and spatially unstructured effects has once again led to the deprecation of geometric variables. Interestingly, apart from gradient and curvature, the exposure variable of segment length was not statistically significant as well. However, in both cases, the inclusion of certain variables was necessary for model cohesion. This includes geometric variables (both gradient and segment length), average speeding seconds per road segment from drivers (a behavioral variable which was marginally significant for free flow conditions only) and the flag variable of road segment bearing (direction). For the CAR model of synchronized flow, average individual driver speed and average occupancy were also required. The same explanation as with harsh brakings is proposed: latent information in these variables aids the calibration of the spatially structured and unstructured effects, which are used to fit the data much better than these variables would with a coefficient.

As with harsh brakings, the value ranges of spatially structured and spatially unstructured effects, represented by ${\sigma_{\varphi i}}^2$ and ${\sigma_{\theta i}}^2$, were comparable in urban arterials. This indicates roughly equal magnitudes of unobserved spatial and non-spatial factors. This time the intercept is significant in both free and synchronized flow models, hinting at the presence of some unobserved factors.

The metrics of RMSE denote an average error magnitude of 1.1 and 0.5 harsh acceleration counts, while MAE is lower at 0.9 and 0.4 harsh acceleration counts. Once again, it appears that the inclusion of spatially structured and spatially unstructured effects improves model fitting performance compared to GWPR, as measured by the error metrics and CA. The improvement is again more pronounced in free flow conditions, with a CA gain of 23.02%. Once again, the synchronized flow model fits the data better than the free flow model. This is attributed to an increase of harsh events per segment, and to a lower representation of urban arterial segments with few events, especially when contrasted with urban networks. Nonetheless, it is very interesting to notice this pattern persisting for harsh accelerations as well as harsh brakings.

7.4 XGBoost algorithms

The following sections present the calibration process of XGBoost algorithms and the respective results that these algorithms yield for harsh brakings and harsh accelerations, and for free flow and synchronized flow conditions. Initially, XGBoost algorithms are trained using traditional random cross-validation (RCV).

The application of XGBoost algorithms utilizing spatial cross-validation (SPCV) constitutes the last spatial analyses conducted in the present doctoral dissertation. It is worth noting that the linear nature of the urban arterial study area confines the SPCV process: the study area is partitioned in k linearly continuous segment subsets (spatial folds), which could be considered continuous road sections.

All algorithms concern count-based modelling of harsh event frequencies, and were thus conducted with the Poisson cost function as described by Equation (67) in Section 3.2.7.2.

7.4.1 Harsh braking RCV XGBoost implementation

7.4.1.1 Free flow conditions

The hyperparameter tuning process involved the determination of optimal parameter values for the study area in free flow conditions. The optimized hyperparameters are those previously mentioned in Section 3.2.7.2: (i) learning rate, (ii) Gamma, (iii) maximum tree depth, (iv) evaluation metric, (v) number of rounds for cost function convergence. Results from hyperparameter optimization appear on **Table 7-29**:

Table 7-29: Hyperparameter optimization results for RCV XGBoost for harsh brakings in free flow conditions

Hyperparameter	Examined range	Optimized Value
Learning rate	0.000 – 1.000	0.811
Gamma	0 – 10	0
Maximum tree depth	1 – 50	6
Evaluation metric	RMSE | RMSLE | MAE | Logloss | poisson-nloglik	RMSE
Number of rounds	1 – 1000	100

The number of k-folds for each cross-validation task was also investigated. It was decided to retain the value of $k = 5$ and thus conduct 5-fold RCV that was used in urban road networks. A 5-fold RCV XGBoost with the Poisson cost function was then trained on the urban arterial dataset for free flow conditions following Bischl et al. (2016). The resulting feature (or independent variable) importance parameters found by executing RCV XGBoost after hyperparameter optimization for harsh braking frequencies are shown on **Table 7-30**:

Table 7-30: Feature importance of RCV XGBoost for harsh brakings in free flow conditions

Independent variables – Features	Gain	Cover	Frequency
Mobile use seconds	0.4583	0.2274	0.0927
Speeding seconds	0.1374	0.1293	0.0795
Speed difference	0.0755	0.0907	0.0795
Gradient	0.0642	0.0501	0.0861
Pass count	0.0577	0.0997	0.1391
Segment length	0.0454	0.0915	0.1656
Average driver speed	0.0387	0.0810	0.0861
Average occupancy	0.0370	0.0171	0.0530
Average std. current traffic volume	0.0328	0.0798	0.0728
Curvature	0.0208	0.0826	0.0530
Bearing	0.0195	0.0074	0.0331
Average traffic speed	0.0115	0.0202	0.0530
Lane number	0.0013	0.0233	0.0066

Algorithm evaluation metrics are shown on **Table 7-31** for the study area dataset.

Table 7-31: Evaluation metrics for the training of RCV XGBoost for harsh brakings in free flow conditions

Metric	Value
RMSE	0.4730
MAE	0.1579
RMSLE	0.0579
CA	98.03%

7.4.1.2 Synchronized flow conditions

The hyperparameter tuning process involved the determination of optimal parameter values for the study area in synchronized flow conditions. Results from hyperparameter optimization appear on **Table 7-32**:

Table 7-32: Hyperparameter optimization results for RCV XGBoost for harsh brakings in synchronized flow conditions

Hyperparameter	Examined range	Optimized Value
Learning rate	0.000 – 1.000	0.620
Gamma	0 – 10	0
Maximum tree depth	1 – 50	6
Evaluation metric	RMSE | RMSLE | MAE | Logloss | poisson-nloglik	RMSE
Number of rounds	1 – 1000	100

The number of k-folds for each cross-validation task was retained to $k = 5$. A 5-fold RCV XGBoost with the Poisson cost function was then trained on the urban arterial dataset for synchronized flow conditions following Bischl et al. (2016). The resulting feature (or independent variable) importance parameters found by executing RCV XGBoost after hyperparameter optimization for harsh braking frequencies are shown on **Table 7-33**:

Table 7-33: Feature importance of RCV XGBoost for harsh brakings in synchronized flow conditions

Independent variables – Features	Gain	Cover	Frequency
Mobile use seconds	0.6446	0.2885	0.0909
Pass count	0.0833	0.1486	0.1488
Speeding seconds	0.0810	0.1161	0.0992
Gradient	0.0491	0.1028	0.1240
Average std. current traffic volume	0.0298	0.0341	0.0909
Segment length	0.0282	0.0605	0.1405
Average driver speed	0.0219	0.0265	0.0496
Average occupancy	0.0216	0.0767	0.0744
Speed difference	0.0167	0.0549	0.0661
Average traffic speed	0.0092	0.0433	0.0496
Curvature	0.0084	0.0426	0.0496
Bearing	0.0060	0.0034	0.0083
Traffic lights	0.0000	0.0021	0.0083

Algorithm evaluation metrics are shown on **Table 7-34** for the study area dataset.

Table 7-34: Evaluation metrics for the training of RCV XGBoost for harsh brakings in synchronized flow conditions

Metric	Value
RMSE	0.2433
MAE	0.0461
RMSLE	0.0268
CA	99.34%

7.4.1.3 Discussion of results

The application of XGBoost algorithms allows for a certain amount of information to be gleaned based on the ranking and magnitude of variable importance. As an initial remark, it is apparent that the XGBoost tree ensemble draws information from the data in a different manner than frequentist or Bayesian modelling in the case of urban arterials as well. Therefore, there are similarities but also differences in the statistical significance from the previous models.

It appears that in both traffic states, driver distraction as expressed by average seconds of mobile phone use per segment is the most important variable to describe harsh braking frequencies. This variable dominates gain scores in both free and synchronized flow states. The ranking of subsequent variables is different across traffic states.

In free flow conditions, the average seconds of speeding by drivers per road segment appears to be the second most important variable, followed by speed difference. Gradient and the exposure variables of pass count and segment length follow afterwards. Certain traffic parameters, such as average driver speed, average occupancy and average standardized current traffic volume have influences of about equal magnitudes as well. The tree ensemble is crated with smaller contributions from curvature, bearing (road direction), average traffic speed and lane number.

In synchronized flow conditions, the number of passes per segment raises to be the second most important characteristic. Speeding seconds and gradient also play a high-ranking role, however the influence of speed difference ranks considerably lower this time. Segment length and the various traffic parameters occupy the middle-ranking positions, while the rest of the variables complete the ensemble. For the first time in urban arterials, traffic lights seem to display a minor influence for harsh braking frequency.

For both traffic states, these results are overall very consistent with the outputs of the CAR and GWPR statistical methods. However, the limited interpretability of results is obvious, as it is not feasible to investigate the isolated effect of parameters or the manner in which they split the ensemble of XGBoost trees. Furthermore, there is no specific inclusion or otherwise investigation of spatial effects affecting the data at this stage or any latent spatial dependence that manifests in urban arterials from the algorithm.

Regarding the three error metrics and custom accuracy (CA), XGBoost features an excellent performance on the urban arterial area, ranking better than both GWPR and Bayesian CAR models in this case. The algorithm has effectively fitted the data for free flow (which was harder for GPWR and CAR) and synchronized flow, adapting to the urban arterial dataset.

7.4.2 Harsh acceleration RCV XGBoost implementation

The present section presents the calibration process of the XGBoost algorithm and the respective results for harsh accelerations using random cross-validation (RCV).

7.4.2.1 Free flow conditions

For free flow conditions, results from hyperparameter optimization appear on **Table 7-35**:

Table 7-35: Hyperparameter optimization results for RCV XGBoost for harsh accelerations in free flow conditions

Hyperparameter	Examined range	Optimized Value
Learning rate	0.000 – 1.000	0.755
Gamma	0 – 10	0
Maximum tree depth	1 – 50	6
Evaluation metric	RMSE | RMSLE | MAE | Logloss | poisson-nloglik	RMSE
Number of rounds	1 – 1000	100

Similarly with harsh brakings, a 5-fold RCV XGBoost with the Poisson cost function was trained on the urban arterial dataset for free flow conditions following Bischl et al. (2016). The resulting feature importance parameters are shown on **Table 7-36**:

Table 7-36: Feature importance of RCV XGBoost for harsh accelerations in free flow conditions

Independent variables – Features	Gain	Cover	Frequency
Mobile use seconds	0.2865	0.1928	0.0738
Average driver speed	0.1457	0.1598	0.0940
Pass count	0.1097	0.1019	0.1275
Segment length	0.0831	0.0706	0.1745
Average std. current traffic volume	0.0795	0.0340	0.0872
Speeding seconds	0.0711	0.0837	0.0470
Average traffic speed	0.0585	0.1017	0.0940
Gradient	0.0573	0.1070	0.1342
Average occupancy	0.0500	0.0610	0.0872
Curvature	0.0337	0.0488	0.0470
Lane number	0.0148	0.0351	0.0134
Traffic lights	0.0100	0.0026	0.0134
Bearing	0.0001	0.0010	0.0067

Algorithm evaluation metrics are shown on **Table 7-37** for the study area dataset.

Table 7-37: Evaluation metrics for the training of RCV XGBoost for harsh accelerations in free flow conditions

Metric	Value
RMSE	0.3974
MAE	0.1316
RMSLE	0.0776
CA	98.68%

7.4.2.2 Synchronized flow conditions

For harsh accelerations in synchronized flow conditions, results from hyperparameter optimization appear on **Table 7-38**:

Table 7-38: Hyperparameter optimization results for RCV XGBoost for harsh accelerations in synchronized flow conditions

Hyperparameter	Examined range	Optimized Value
Learning rate	0.000 – 1.000	0.850
Gamma	0 – 10	0
Maximum tree depth	1 – 50	6
Evaluation metric	RMSE | RMSLE | MAE | Logloss | poisson-nloglik	RMSE
Number of rounds	1 – 1000	100

Similarly with harsh brakings, a 5-fold RCV XGBoost with the Poisson cost function was trained on the urban arterial dataset for free flow conditions following Bischl et al. (2016). The resulting feature importance parameters are shown on **Table 7-39**:

Table 7-39: Feature importance of RCV XGBoost for harsh accelerations in synchronized flow conditions

Independent variables – Features	Gain	Cover	Frequency
Mobile use seconds	0.5723	0.2768	0.1429
Pass count	0.2169	0.2765	0.1837
Average driver speed	0.0604	0.0999	0.1224
Average occupancy	0.0457	0.0225	0.1429
Speed difference	0.0332	0.0239	0.0612
Gradient	0.0243	0.0755	0.0816
Segment length	0.0240	0.0825	0.1224
Curvature	0.0183	0.0863	0.0816
Average std. current traffic volume	0.0032	0.0390	0.0204
Speeding seconds	0.0017	0.0171	0.0408

Algorithm evaluation metrics are shown on **Table 7-40** for the study area dataset.

Table 7-40: Evaluation metrics for the training of RCV XGBoost for harsh accelerations in synchronized flow conditions

Metric	Value
RMSE	0.5000
MAE	0.1711
RMSLE	0.1638
CA	97.37%

7.4.2.3 Discussion of results

The application of XGBoost algorithms allows for a certain amount of information to be gleaned based on the ranking and magnitude of variable importance. Once again, it appears that the XGBoost tree ensemble draws information from the data in a different manner than frequentist or Bayesian modelling. Therefore, there are similarities but also differences in the statistical significance from the previous models.

For harsh accelerations in both traffic states, driver distraction as expressed by average seconds of mobile phone use per segment is the most important variable, a retained finding from harsh braking frequencies. This variable ranks first in gain scores in both free and synchronized flow states, though the gain magnitude in free flow conditions is lower. The ranking of subsequent variables is different across traffic states. Subsequently, driver speed and pass count seem to be the second and third most influential variables in free flow. Their places interestingly reverse when moving to synchronized flow conditions.

In free flow, segment length is an important factor, followed by traffic and behavioral variables such as average standardized current traffic volume, average speeding seconds of drivers and average traffic speed. Gradient, average occupancy and curvature were also found to exert modest amounts of influence on harsh acceleration frequency. Lastly, the ensemble draws minor amounts of information from lane number, the presence of traffic lights and bearing (direction) of road segments.

In synchronized flow conditions, the average occupancy and speed difference of traffic and drivers were found to offer a medium information gain. Geometric characteristics follow afterwards (gradient, segment length and curvature), while the ensemble closes with information gains from average standardized current traffic volume and speeding seconds of drivers.

The results are overall very consistent with the outputs of the CAR and GWPR statistical methods. However, some variables appearing as having a degree of feature importance in the XGBoost ensemble did not appear as significant in the previous models. Indicative examples are the number of lanes, for free flow conditions, and average standardized current traffic volume, for synchronized conditions.

Regarding the three error metrics and custom accuracy (CA), the effectiveness of XGBoost is proven again. The algorithm features an excellent performance on the urban arterial area, ranking better than both GWPR and Bayesian CAR models for harsh accelerations – although CAR models for harsh accelerations in synchronized flow performed comparably close. The XGBoost algorithms have effectively fitted the data for free flow (which was harder for GPWR and CAR) and synchronized flow, adapting to the urban arterial dataset.

7.4.3 Harsh braking SPCV XGBoost implementation

7.4.3.1 Free flow conditions

XGBoost algorithms are calibrated with spatial cross-validation (SPCV) following Lovelace et al. (2019) as in urban road networks. For free flow conditions, results from hyperparameter optimization appear on **Table 7-41**:

Table 7-41: Hyperparameter optimization results for SPCV XGBoost for harsh brakings in free flow conditions

Hyperparameter	Examined range	Optimized Value
Learning rate	0.000 – 1.000	0.310
Gamma	0 – 10	0.3
Maximum tree depth	1 – 50	6
Evaluation metric	RMSE | RMSLE | MAE | Logloss | poisson-nloglik	RMSE
Number of rounds	1 – 1000	100

The algorithms are then trained following Bischl et al. (2016). The resulting feature importance parameters are shown on **Table 7-42**:

Table 7-42: Feature importance of SPCV XGBoost for harsh brakings in free flow conditions

Independent variables – Features	Gain	Cover	Frequency
Mobile use seconds	0.5496	0.1924	0.1005
Average driver speed	0.0687	0.0779	0.0854
Segment length	0.0572	0.1012	0.1206
Speed difference	0.0548	0.0966	0.0905
Average traffic speed	0.0493	0.0377	0.0704
Average std. current traffic volume	0.0469	0.0636	0.1005
Gradient	0.0408	0.1420	0.1055
Pass count	0.0364	0.0990	0.1055
Speeding seconds	0.0330	0.0311	0.0503
Average occupancy	0.0310	0.0849	0.0804
Curvature	0.0183	0.0334	0.0452
Bearing	0.0067	0.0007	0.0101
Lane number	0.0027	0.0012	0.0050
Pedestrian crossing	0.0025	0.0080	0.0151
Traffic lights	0.0021	0.0304	0.0151

Algorithm evaluation metrics are shown on **Table 7-43** for the study area dataset.

Table 7-43: Evaluation metrics for the training of SPCV XGBoost for harsh brakings in free flow conditions

Metric	Value
RMSE	0.4730
MAE	0.1316
RMSLE	0.2105
CA	99.34%

7.4.3.2 Synchronized flow conditions

For harsh brakings in synchronized flow conditions, results from hyperparameter optimization appear on **Table 7-44**:

Table 7-44: Hyperparameter optimization results for SPCV XGBoost for harsh brakings in synchronized flow conditions

Hyperparameter	Examined range	Optimized Value
Learning rate	0.000 – 1.000	0.300
Gamma	0 – 10	0.3
Maximum tree depth	1 – 50	6
Evaluation metric	RMSE | RMSLE | MAE | Logloss | poisson-nloglik	RMSE
Number of rounds	1 – 1000	100

The algorithms are then trained following Bischl et al. (2016). The resulting feature importance parameters are shown on **Table 7-45**:

Table 7-45: Feature importance of SPCV XGBoost for harsh brakings in synchronized flow conditions

Independent variables – Features	Gain	Cover	Frequency
Pass count	0.4146	0.1920	0.1339
Mobile use seconds	0.2250	0.1663	0.0971
Speeding seconds	0.0781	0.1463	0.1076
Gradient	0.0638	0.0844	0.1312
Average driver speed	0.0610	0.0955	0.1102
Segment length	0.0533	0.0950	0.1417
Curvature	0.0269	0.0398	0.0577
Average std. current traffic volume	0.0266	0.0847	0.0919
Speed difference	0.0231	0.0351	0.0551
Average occupancy	0.0172	0.0539	0.0577
Bearing	0.0095	0.0020	0.0105
Pedestrian crossing	0.0006	0.0002	0.0026
Traffic lights	0.0002	0.0048	0.0026

Algorithm evaluation metrics are shown on **Table 7-46** for the study area dataset.

Table 7-46: Evaluation metrics for the training of SPCV XGBoost for harsh brakings in synchronized flow conditions

Metric	Value
RMSE	0.3441
MAE	0.0472
RMSLE	0.0921
CA	98.68%

7.4.3.3 Discussion of results

Results indicate that for urban arterials, SPCV XGBoost yielded very comparable outcomes to RCV XGBoost. In free flow conditions, the variable of average mobile phone use seconds continues to rank first regarding information gain, however in synchronized flow it is overtaken by the exposure variable of pass count. The rest of the variables are similarly rearranged in importance.

One noticeable small differentiation is that in SPCV XGBoost algorithms, slightly more features emerge as important. For the first time in urban arterials, pedestrian crossings also contribute a small amount of information gain for harsh braking frequency modelling both in free flow and in synchronized flow.

The performance of SPCV XGBoost is excellent, adapting to the data very adeptly. Error metrics are very low and CA has very high values. Nonetheless, it is apparent that SPCV did not surpass RCV XGBoost in urban arterials for harsh brakings. This is attributed to the difference in the study area: the largely homogeneous environment of urban arterials apparently makes different types of cross-validation produce less deviant results. In addition, there are fewer road segments – serving as data-points – in order to calibrate this potent Machine Learning method.

7.4.4 Harsh acceleration SPCV XGBoost implementation

7.4.4.1 Free flow conditions

XGBoost algorithms are calibrated with spatial cross-validation (SPCV) following Lovelace et al. (2019) as in urban road networks. For free flow conditions, results from hyperparameter optimization appear on **Table 7-47**:

Table 7-47: Hyperparameter optimization results for SPCV XGBoost for harsh accelerations in free flow conditions

Hyperparameter	Examined range	Optimized Value
Learning rate	0.000 – 1.000	0.305
Gamma	0 – 10	0.29
Maximum tree depth	1 – 50	6
Evaluation metric	RMSE | RMSLE | MAE | Logloss | poisson-nloglik	RMSE
Number of rounds	1 – 1000	100

The algorithms are then trained following Bischl et al. (2016). The resulting feature importance parameters are shown on **Table 7-48**:

Table 7-48: Feature importance of SPCV XGBoost for harsh accelerations in free flow conditions

Independent variables – Features	Gain	Cover	Frequency
Mobile use seconds	0.2212	0.1751	0.0994
Average std. current traffic volume	0.1281	0.0736	0.1014
Pass count	0.1183	0.1203	0.1136
Average driver speed	0.1136	0.1244	0.0994
Segment length	0.0759	0.0837	0.1156
Gradient	0.0756	0.0562	0.1055
Speed difference	0.0749	0.1045	0.0892
Average occupancy	0.0656	0.0700	0.0913
Speeding seconds	0.0631	0.1195	0.0933
Curvature	0.0413	0.0514	0.0609
Lane number	0.0092	0.0142	0.0122
Traffic lights	0.0058	0.0009	0.0081
Pedestrian crossing	0.0056	0.0051	0.0061
Bearing	0.0021	0.0010	0.0041

Algorithm evaluation metrics are shown on **Table 7-49** for the study area dataset.

Table 7-49: Evaluation metrics for the training of SPCV XGBoost for harsh accelerations in free flow conditions

Metric	Value
RMSE	0.3536
MAE	0.1118
RMSLE	0.0507
CA	99.34%

7.4.4.2 Synchronized flow conditions

For harsh accelerations in synchronized flow conditions, results from hyperparameter optimization appear on **Table 7-50**:

Table 7-50: Hyperparameter optimization results for SPCV XGBoost for harsh accelerations in synchronized flow conditions

Hyperparameter	Examined range	Optimized Value
Learning rate	0.000 – 1.000	0.650
Gamma	0 – 10	0.7
Maximum tree depth	1 – 50	6
Evaluation metric	RMSE | RMSLE | MAE | Logloss | poisson-nloglik	RMSE
Number of rounds	1 – 1000	100

The algorithms are then trained following Bischl et al. (2016). The resulting feature importance parameters are shown on **Table 7-51**:

Table 7-51: Feature importance of SPCV XGBoost for harsh accelerations in synchronized flow conditions

Independent variables – Features	Gain	Cover	Frequency
Pass count	0.5617	0.3197	0.2319
Mobile use seconds	0.2209	0.1771	0.1594
Average driver speed	0.0568	0.1395	0.1159
Average occupancy	0.0407	0.0581	0.1014
Segment length	0.0259	0.0168	0.0290
Gradient	0.0258	0.0626	0.0870
Average std. current traffic volume	0.0220	0.1067	0.0870
Curvature	0.0180	0.0403	0.0725
Speed difference	0.0144	0.0134	0.0580
Speeding seconds	0.0071	0.0647	0.0435
Bearing	0.0067	0.0012	0.0145

Algorithm evaluation metrics are shown on **Table 7-52** for the study area dataset.

Table 7-52: Evaluation metrics for the training of SPCV XGBoost for harsh accelerations in synchronized flow conditions

Metric	Value
RMSE	0.2810
MAE	0.0658
RMSLE	0.0842
CA	98.68%

7.4.4.3 Discussion of results

Similarly with harsh brakings, SPCV XGBoost yielded very comparable outcomes to RCV XGBoost. The first-ranking variable was found to be the same as well: In free flow conditions, the variable of average mobile phone use seconds continues to rank first regarding information gain, however in synchronized flow it is overtaken by the exposure variable of pass count. The rest of the variables are similarly rearranged in importance.

One noticeable small differentiation is that in SPCV XGBoost algorithms, slightly more features emerge as important. This time pedestrian crossings emerged as a minor contributor only in free flow conditions, and not in synchronized flow.

The performance of SPCV XGBoost is excellent, adapting to the data very adeptly. Error metrics are very low and CA has very high values. SPCV presented a slight improvement over RCV XGBoost for harsh accelerations, which might be conditional and dependent upon the dataset.

7.5 Overall urban arterial results

The present section conducted a series of analyses involving GWPR and Bayesian CAR models, and RCV XGBoost and SPCV XGBoost algorithms for harsh braking and harsh acceleration frequencies for free flow and synchronized flow conditions in urban arterial segments. The previous datasets, analyses and methodologies were augmented with variables describing traffic and driver behavior. The main findings can be summarized in the following points:

1. The methodology previously used in urban road networks can be meaningfully expanded and augmented with variables related to traffic and driver behavior. Nonetheless, the combination of individual driver, traffic and fixed infrastructure variables and the merging of the respective datasets for integration and utilization in road safety models remains a challenging task. As this is a quite specialized topic however, no particular approach has emerged as more appropriate in comparison to others.

2. It is clear that the road safety standpoint differs from the traffic flow optimization standpoint. As a middle-ground approach, this dissertation selected the separate examination of road segment datasets separately per traffic state (free flow, synchronized flow, congested flow). Of the three, harsh events occurred predominantly in free flow and synchronized flow states, and as such, modelling was conducted in these two states. The determination of different traffic, driver behavior, geometric and road network characteristic variables as significant in each of the two remaining traffic states is considered a finding which validates this approach.

3. From the initial spatial analyses, it was determined that there is large spatial autocorrelation in harsh braking and harsh acceleration frequencies of certain urban arterial 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 as well.

4. 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.

5. Once again, 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.

6. 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.

7. 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.

8. 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.

9. 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.

10. 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.

11. 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: It was not possible to create direction-based variograms, and GWPR models suffered reductions in their capabilities to adapt to the data more accurately.

12. 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.

8 Conclusions

8.1 Dissertation overview

8.1.1 Background, data collection & processing

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 replicated on Figure 8-1.

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.

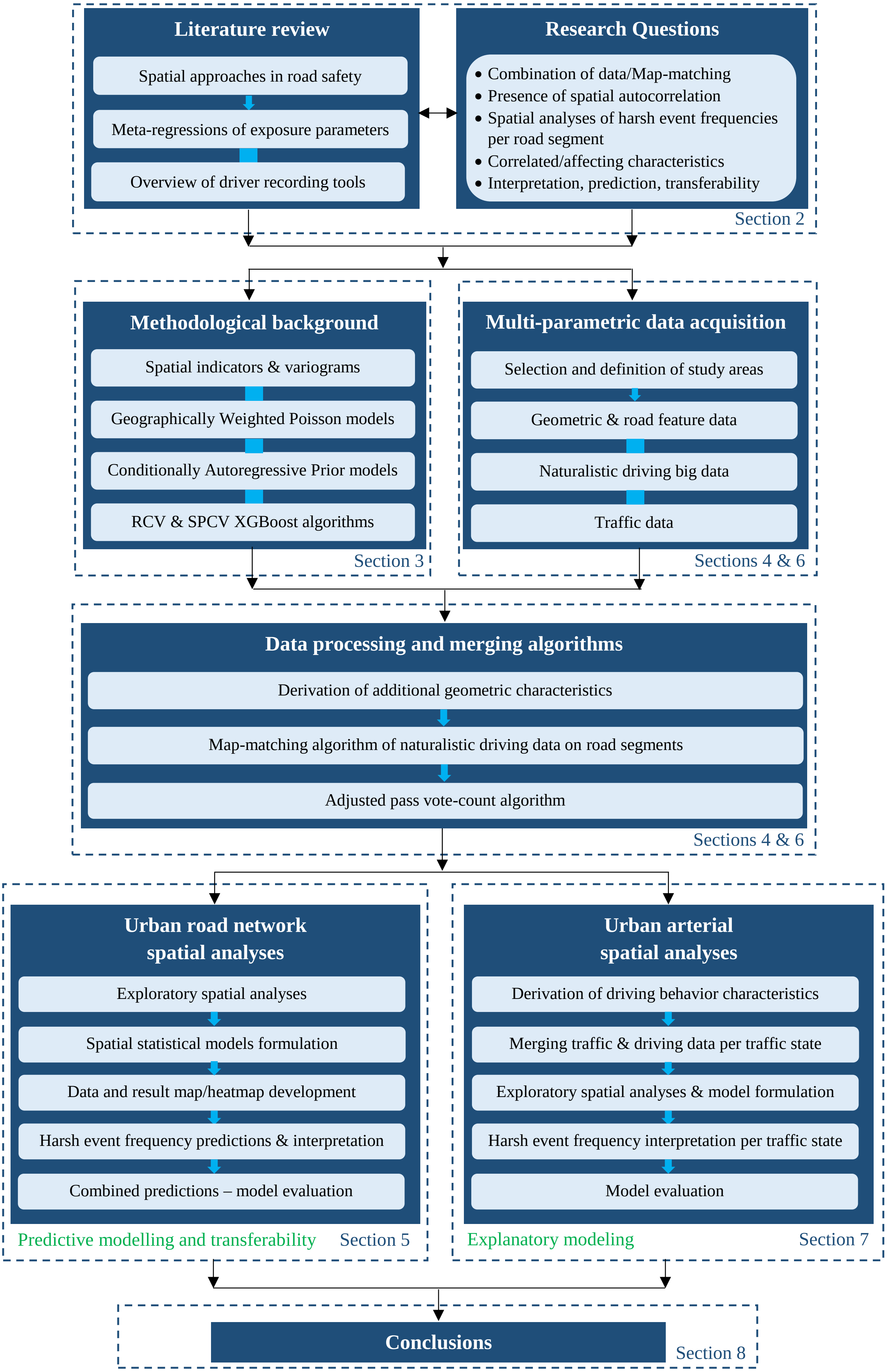

Literature review
Spatial approaches in road safety
Meta-regressions of exposure parameters
Overview of driver recording tools
Research Questions
• Combination of data/Map-matching
• Presence of spatial autocorrelation
• Spatial analyses of harsh event frequencies per road segment
• Correlated/affecting characteristics
• Interpretation, prediction, transferability
Section 2
Methodological background
Spatial indicators & variograms
Geographically Weighted Poisson models
Conditionally Autoregressive Prior models
RCV & SPCV XGBoost algorithms
Section 3
Multi-parametric data acquisition
Selection and definition of study areas
Geometric & road feature data
Naturalistic driving big data
Traffic data
Sections 4 & 6
Data processing and merging algorithms
Derivation of additional geometric characteristics
Map-matching algorithm of naturalistic driving data on road segments
Adjusted pass vote-count algorithm
Sections 4 & 6
Urban road network spatial analyses
Exploratory spatial analyses
Spatial statistical models formulation
Data and result map/heatmap development
Harsh event frequency predictions & interpretation
Combined predictions – model evaluation
Predictive modelling and transferability
Section 5
Urban arterial spatial analyses
Derivation of driving behavior characteristics
Merging traffic & driving data per traffic state
Exploratory spatial analyses & model formulation
Harsh event frequency interpretation per traffic state
Model evaluation
Explanatory modeling
Section 7
Conclusions
Section 8

Figure 8-1: Overall methodological framework of the doctoral dissertation

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.

The naturalistic trip data in this dissertation were provided by OSeven Telematics through innovative smartphone applications that seamlessly and non-intrusively record driving trips when users drive their vehicles normally. The applications enable the acquisition of a large number of naturalistic driving behavior metrics 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.

8.1.2 Urban road network results

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 can only 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 (CA > 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 4-18 (depicting harsh braking events in Omonoia area) and Figure 5-30 (depicting the respective combined prediction heatmap of harsh braking frequencies in Omonoia area) are mentioned.

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%).

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.

Non-count based modelling methods, including linear spatial methods such as Geographically Weighted Regression, and regression machine learning methods such as Support Vector Machines and Random Forests proved inappropriate to analyze harsh event frequencies either as count variables or as harsh event rates. The harsh event phenomena are highly non-linear, leading to poor model fits, poor CA and large error metrics. Additionally, road segment datasets contain zeros which do not allow for log-linear methods. Furthermore, harsh event rates lead to loss of information by forcing exposure variables to have coefficients bound to 1.

8.1.3 Urban arterial segment results

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 as well.

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, 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.

8.2 Innovative contributions

The present doctoral dissertation offers significant innovative contributions in the field of road safety and traffic behavior analysis, as shown in Figure 8-2:

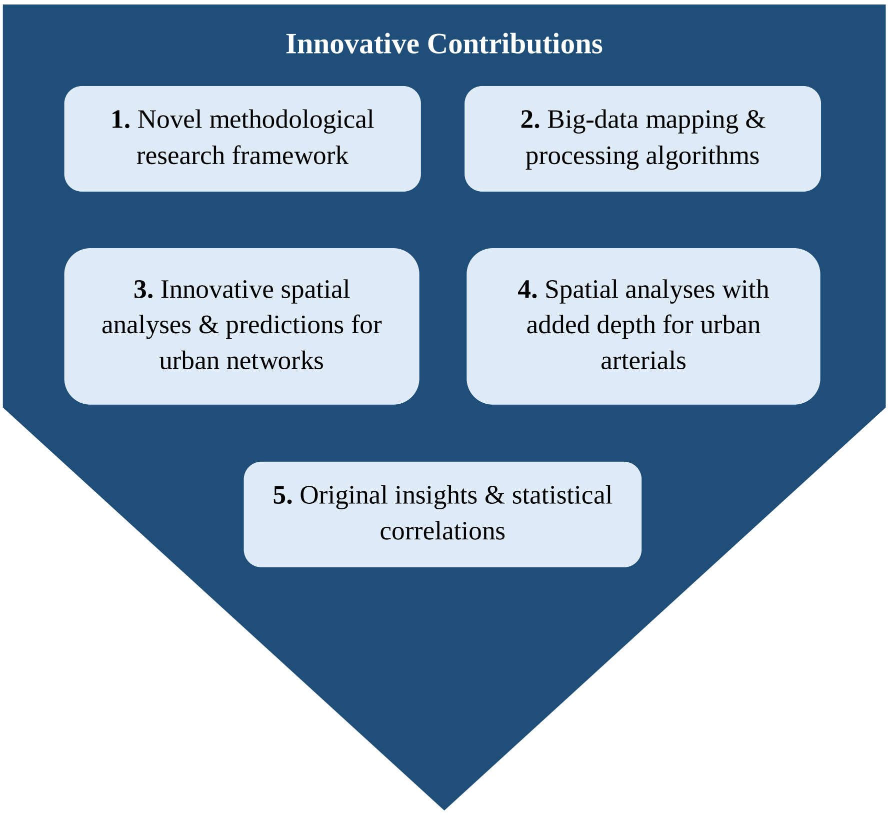

Innovative Contributions
1. Novel methodological research framework
2. Big-data mapping & processing algorithms
3. Innovative spatial analyses & predictions for urban networks
4. Spatial analyses with added depth for urban arterials
5. Original insights & statistical correlations

Figure 8-2: Innovative contributions of the dissertation

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 dissertation, original insights and statistical correlations were discovered between the frequencies of harsh braking and harsh acceleration events per segment and geometrical, road network, traffic and driver behavior variables.

8.2.1 Novel methodological research framework

Recent technological advancements have eased data collection and acquisition from several distinct sources, revealing new research opportunities that were previously inaccessible. This dissertation exploited high resolution naturalistic driving big datasets that were recorded via smartphone sensors. In total, 2,895,312 total trip-seconds were analyzed from more than 314 individual drivers, containing 3,927 harsh brakings and 2,892 harsh accelerations across three study areas. These data were projected on highly detailed digital map data describing 2,258 road segments in total, through which additional geometrical characteristics were calculated and network characteristics were obtained. High resolution traffic data of 90s intervals were collected during a three-month period from 54 locations, corresponding to 4,676,691 traffic measurements in total, were also acquired for specific analyses in urban arterial segments. The aforementioned volumes classify the utilized datasets as big data.

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.

Additional value is provided by this research via the creation of several informative maps and heatmaps based on (i) collected data and (ii) produced results. The level of detail and comprehensiveness of maps of harsh event frequencies allows for precision in hotspot detection by road management authorities, road safety stakeholders and even individual road users if the maps are released in the appropriate public domains. This entails a more informed selection of routes for individual road users, and a scientifically supported allocation of funding for targeted and effective road safety interventions by road safety stakeholders. Furthermore, a critical merit of harsh event analysis that is enhanced by the present work is that they are pro-active road safety indicators. In other words, by utilizing the methodology and results of this research, stakeholders can select and prioritize problematic locations before road crashes occur.

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, and 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.

8.2.2 Big data mapping and processing algorithms

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 and map-matching. As such, they provided the vehicle for realizing the envisioned innovative framework and prepared the spatial data-frames comprising of road segments that were analyzed afterwards.

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.

Subsequently, 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.

Overall, the algorithms utilized in the present doctoral dissertation enable the seamless transferability of the entire methodological and data processing framework followed in the present doctoral dissertation. With minimum adjustments, spatial data-frames can be obtained for different areas, which can then be analyzed utilizing the same or new variables, study periods and statistical methodologies.

8.2.3 Innovative spatial analyses & predictions for urban networks

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 constitute heavily understudied areas in road safety.

In direct response to the set research questions, several correlated variables were determined for harsh event frequency for urban road networks. In particular, apart from the exposure variable of pass count, geometrical characteristics were found to affect harsh braking frequencies per road segment: Segment length is positively correlated with harsh brakings, while gradient and neighborhood complexity are negatively correlated with harsh brakings. Curvature, road direction, traffic lights and pedestrian crossings were not determined as statistically significant.

Furthermore, apart from pass count, different geometrical characteristics were found to affect harsh acceleration frequencies per road segment: Segment length, curvature and the presence of traffic lights are positively correlated with harsh accelerations. Road direction was not statistically significant. For both harsh event types, lane number and road type have more unclear circumstantial effects, depending on the utilized models.

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 respective 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.

8.2.4 Spatial analyses with added depth for urban arterials

For urban arterials, the adopted approach included some common elements but also additional novelties compared to urban networks. An innovative methodological approach bridging the gaps between road safety and traffic flow theory was devised. This methodological approach entailed the determination of the traffic flow state for each trip-second, following limits determined by previous research, and the implementation of separate spatial data-frames and analyses per traffic state. The integration of traffic data as attributes of fixed locations for spatial analyses, such as road segments, remains an unanswered problem by the literature, to which the present doctoral dissertation provided its own answer.

In direct response to the set research questions, several correlated variables were determined for harsh event frequencies for urban arterial segments, which mostly originated from the newly introduced variable types of traffic and driver behavior. Furthermore, it was determined that different variables are significantly correlated with harsh event occurrence per traffic state.

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. In free flow conditions, speed difference between traffic and driver was found to exert a positive influence, while the influence of the averaged standardized current traffic volume was found to be negative. In some models, 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. In synchronized flow conditions, average occupancy assumes a statistically significant positive correlation for harsh braking frequencies, while the influence of traffic volume (standardized or hourly) was found to be circumstantially negative.

The exposure parameters of segment length and pass count were found to be positively correlated with harsh acceleration frequencies. In free flow conditions, the average mobile use seconds of drivers per road segment was found to have a significantly positive effect as well. Additionally, 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. In synchronized flow conditions, traffic volume (standardized or hourly) was found to be positively correlated with harsh accelerations as well. In both traffic states, geometric and road network characteristic variables were found to have very circumstantial effects.

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.

In addition to the previous, it was determined that the linearity of the study area of Kifisias Avenue affected the applied models differently. Specifically, GPWR models suffered a reduction in the degree to which they fit the data. However, the spatial effects in CAR models and the tree ensemble of XGBoost proved unaffected. This finding contributes to the assessment of spatial analysis methods that can be used in future planning of further spatial analyses.

8.2.5 Original insights and statistical correlations

The evaluation of the contributions of the present dissertation would not be complete without mentioning the implications of the outcomes of the spatial models, in the form of statistical correlations and results. 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.

A cornerstone of frequency analyses is the measurement of exposure. Two main exposure variables were used for the spatial analyses of the dissertation, namely road segment length and pass count. Within the present research, the influences of exposure variables on harsh event occurrence were found to be statistically significant and their impacts were quantified.

Additionally, the profound and complex non-linear manner in which traffic parameters impact harsh event frequencies was also highlighted by the present work. The considerably different model results determined in urban arterial segments under free flow and synchronized flow conditions justify the examination per separate traffic state.

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.

8.3 Further challenges

The present doctoral dissertation tackled several composite issues pertaining to data collection, merging and processing, spatial analysis, prediction, and modelling per traffic state. As such, it is natural that limitations arose throughout the entire research process, and open challenges remain, which ought to be mentioned.

A first point is the reliance on road segmentation as conducted by OpenStreetMap (OSM). The present dissertation relied on the existing OSM structure and protocols for segmentation of roads (ways). The presence of segments with fluctuating attributes was desirable to ensure the transferability of the methodology and results. However, the particular segmentation has not been proven to be optimal for road safety analysis (as certain zonal systems have been proven superior to others in the reviewed literature). Moreover, several different criteria can be set and examined (e.g. geometry based or traffic based). The creation and examination of alternative segmentation protocols is an arduous process, and requires its own algorithmic implementation. As such, it was outside the scope of the dissertation.

Harsh driving events are ultimately behavioral variables. This dissertation has treated them as point-data for investigation of their frequencies, and driver numbers were adequate by standards of the literature. Nonetheless, it is possible that this driver sample may have deviated from the mean driving behavior, and produced more or fewer harsh driving events than should be expected. A secondary research question also arises in tandem: The research periods of two and three months included voluminous data, however an investigation should be conducted on the amount of drivers, trips and harsh event numbers that are required for the creation of representative harsh event/road safety maps.

Lastly, the author is not a professional computer scientist or does not claim any official coding background or training. It is certain that the algorithms utilized for the needs of the present research could have been better optimized or written in a more straightforward manner. The computational times were also considerable at times, surpassing 3 days. Indicatively speaking, for the map-matching process a large number of 'for' loops and 'if' statements was implemented, which seemed essential for the examination of each trip-second separately. R operates much better when functions are vectorized (i.e. applied to data vectors simultaneously, as opposed to loops), similar to many other programming languages, and a way to achieve this might exist. However, this becomes a computer science problem, and as such was not tackled within the present research.

8.4 Future research directions

As the present dissertation is concluded, the author maintains the belief that the current research findings lead to a plethora of further research questions and additional issues which merit scientific investigation. The following indicative examples are provided.

The first direction that is apparent when considering road safety research is road crashes. It would be very interesting and highly fruitful to conduct analyses parallel with the ones of the present research, with the same research areas and methodologies, in order to obtain crash hotspot locations. The different types of hotspot locations – crash, harsh acceleration and harsh braking – would then be compared and examined for spatial overlap. Thus the informative capabilities of harsh events for spatial predictions of crashes would be assessed. This approach would require detailed data; considerable progress in crash recording protocols is thus required, in order to mitigate present-day uncertainties.

Another promising research venue is the examination through time. Essentially the present spatial analyses would be transformed to spatio-temporal analysis. The temporal evolution would (i) capture seasonal cyclical trends in harsh event hotspots, such as those caused by tourists, and (ii) permit the observation of hotspot mitigation, an issue frequently present in road safety, especially after the implementation of measures or other interventions.

In order to mitigate uncertainty from the driver sample, a cross-investigation of driver behavior could be undertaken. The process would entail categorizing drivers or trips into clusters (e.g. aggressive/average driving) and the creation of heatmaps from a homogeneous sample. Similarly, since the concept of spatial analysis of harsh events has proven successful and promising, harsh events could be further segregated per intensity for future research. The hotspots of events of different intensity could be compared to detect any overlap or lack thereof.

Additional spatial models could always be considered. Apart from XGBoost, other powerful machine learning methods, such as the many forms of Neural Networks, can be used with spatial cross-validation, for instance. Additional prior distributions can be examined in CAR models as well. The examination of different spatial weighting schemes for GWPR or Moran's *I* coefficients, even purpose-made ones, could also be a part of future research.

The combination of individual driver, traffic and fixed infrastructure variables for integration and utilization in road safety models remains a challenging task. It is clear that the road safety standpoint differs from the traffic flow optimization standpoint, however, no particular approach has emerged as more appropriate in comparison to the others. Dedicated road safety research should be conducted in that direction involving detailed modelling approaches, while taking the various heterogeneity issues into consideration as well.

The borders of the present work can also be expanded by replicating this research to cover additional areas, possibly in other countries as well. Different road types, such as rural roads, or area types, such as school/pedestrian zones with traffic calming measures in place remain to be investigated. In all new cases, the transferability of the methodology and the results should be measured and compared with the present findings.

It is obvious that not all aspects of the road environment have been covered in this dissertation. The integration of additional independent variables can provide unexplored insights, such as the presence and proximity of public transport to each road segment (e.g. sidewalk bus stops, metro stations), the presence of gas stations, roadworks etc. Likewise, maps and heatmaps can be produced for additional phenomena that would be analyzed as dependent variables. For instance, stakeholders could be interested in heatmaps for locations of speeding, mobile phone use, traffic light violations or similar indicators.

9 References

1. Aarts, L., & Van Schagen, I. (2006). Driving speed and the risk of road crashes: A review. Accident Analysis & Prevention, 38(2), 215-224.
2. Abdel-Aty, M. A., Lee, J., Eluru, N., Cai, Q., Al Amili, S., & Alarifi, S. (2016). Enhancing and generalizing the two-level screening approach incorporating the highway safety manual (HSM) methods, Phase 2. University of Central Florida, Department of Civil, Environmental and Construction Engineering.
3. Abdel-Aty, M., Lee, J., Siddiqui, C., & Choi, K. (2013). Geographical unit based analysis in the context of transportation safety planning. Transportation Research Part A: Policy and Practice, 49, 62-75.
4. Abdel-Aty, M., Siddiqui, C., Huang, H., & Wang, X. (2011). Integrating trip and roadway characteristics to manage safety in traffic analysis zones. Transportation Research Record, 2213(1), 20-28.
5. Abdel-Aty, M., & Wang, X. (2006). Crash estimation at signalized intersections along corridors: analyzing spatial effect and identifying significant factors. Transportation Research Record: Journal of the Transportation Research Board, (1953), 98-111.
6. Acharya, A. (2014). Are We Ready for Driver-less Vehicles? Security vs. Privacy-A Social Perspective. arXiv preprint arXiv:1412.5207.
7. Adamidis, F. K., Mantouka, E. G., & Vlahogianni, E. I. (2020). Effects of controlling aggressive driving behavior on network-wide traffic flow and emissions. International Journal of Transportation Science and Technology.
8. Aguero-Valverde, J., Wu, K.F., Donnell, E.T. (2016). A multivariate spatial crash frequency model for identifying sites with promise based on crash types. Accident Analysis and Prevention 87, 8–16.
9. Aguero-Valverde, J. (2014). Direct spatial correlation in crash frequency models: estimation of the effective range. Journal of Transportation Safety & Security, 6(1), 21-33
10. Aguero-Valverde, J. (2013). Multivariate spatial models of excess crash frequency at area level: Case of Costa Rica. Accident Analysis & Prevention, 59, 365-373.
11. Aguero-Valverde, J., & Jovanis, P. P. (2010). Spatial correlation in multilevel crash frequency models: Effects of different neighboring structures. Transportation Research Record, 2165(1), 21-32.
12. Aguero-Valverde, J., & Jovanis, P. P. (2008). Analysis of road crash frequency with spatial models. Transportation Research Record, 2061(1), 55-63.
13. Aguero-Valverde, J., & Jovanis, P. P. (2006). Spatial analysis of fatal and injury crashes in Pennsylvania. Accident Analysis & Prevention, 38(3), 618-625.
14. Aksan, N., Hacker, S. D., Sager, L., Dawson, J., Anderson, S., & Rizzo, M. (2016). Correspondence between simulator and on-road drive performance: implications for assessment of driving safety. Geriatrics, 1(1), 8.
15. Alarifi, S. A., Abdel-Aty, M. A., Lee, J., & Wang, X. (2018). Exploring the effect of different neighboring structures on spatial hierarchical joint crash frequency models. Transportation research record, 2672(38), 210-222.
16. Alarifi, S. A., Abdel-Aty, M. A., Lee, J., & Park, J. (2017). Crash modeling for intersections and segments along corridors: a Bayesian multilevel joint model with random parameters. Analytic methods in accident research, 16, 48-59.
17. Alessandrini, A., Cattivera, A., Filippi, F., & Ortenzi, F. (2012, August). Driving style influence on car CO2 emissions. In 2012 international emission inventory conference.

18. Amoh-Gyimah, R., Saberi, M., & Sarvi, M. (2017). The effect of variations in spatial units on unobserved heterogeneity in macroscopic crash models. Analytic methods in accident research, 13, 28-51.
19. Anderson, T. K. (2009). Kernel density estimation and K-means clustering to profile road accident hotspots. Accident Analysis & Prevention, 41(3), 359-364.
20. Anderson, T. (2007). Comparison of spatial methods for measuring road accident ‘hotspots’: a case study of London. Journal of Maps, 3(1), 55-63.
21. Anselin, L., Murray, A.T., Rey, S.J. (2014). The Oxford handbook of quantitative methods: Foundations (Vol. 2, Chapter 8). Oxford University Press, USA.
22. Anselin, L. (1995). Local indicators of spatial association - LISA. Geographical analysis, 27(2), 93-115.
23. Armstrong, K., Filtness, A. J., Watling, C. N., Barraclough, P., & Haworth, N. (2013). Efficacy of proxy definitions for identification of fatigue/sleep-related crashes: An Australian evaluation. Transportation Research Part F: Traffic Psychology and Behaviour, 21, 242-252.
24. Atubi, A. O. (2012). Determinants of road traffic accident occurrences in Lagos State: Some lessons for Nigeria. International Journal of Humanities and Social Science, 2(6), 252-259.
25. Backer-Grøndahl, A., & Sagberg, F. (2011). Driving and telephoning: Relative accident risk when using hand-held and hands-free mobile phones. Safety Science, 49(2), 324-330.
26. Backer-Grøndahl, A., Phillips, R., Sagberg, F., Touliou, K., Gatscha, M. (2009). Naturalistic driving observation: Topics and applications of previous and current naturalistic studies. PROLOGUE Deliverable D1.1. TØI Institute of Transport Economics, Oslo, Norway.
27. Balk, S. A., Moore, K., Steele, J. E., Spearman, W., & Duchowski, A. (2006). Mobile phone use in a driving simulation task: Differences in eye movements. J Vis, 6(6), 872.
28. 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.
29. Bao, J., Liu, P., & Ukkusuri, S. V. (2019). A spatiotemporal deep learning approach for citywide short-term crash risk prediction with multi-source data. Accident Analysis & Prevention, 122, 239-254.
30. Bao, J., Liu, P., Qin, X., & Zhou, H. (2018). Understanding the effects of trip patterns on spatially aggregated crashes with large-scale taxi GPS data. Accident Analysis & Prevention, 120, 281-294.
31. Bao, J., Liu, P., Yu, H., & Xu, C. (2017). Incorporating twitter-based human activity information in spatial analysis of crashes in urban areas. Accident Analysis & Prevention, 106, 358-369.
32. Barmpounakis, E. N., Vlahogianni, E. I., & Golias, J. C. (2016a). Vision-based multivariate statistical modeling for powered two-wheelers maneuverability during overtaking in urban arterials. Transportation Letters, 8(3), 167-176.
33. Barmpounakis, E. N., Vlahogianni, E. I., & Golias, J. C. (2016b). Extracting kinematic characteristics from unmanned aerial vehicles (No. 16-3429).
34. Barua, S., El-Basyouny, K., & Islam, M. T. (2016). Multivariate random parameters collision count data models with spatial heterogeneity. Analytic methods in accident research, 9, 1-15.
35. Barua, S., El-Basyouny, K., & Islam, M. T. (2014). A full Bayesian multivariate count data model of collision severity with spatial correlation. Analytic Methods in Accident Research, 3, 28-43.
36. Bellinger, D. B., Budde, B. M., Machida, M., Richardson, G. B., & Berg, W. P. (2009). The effect of cellular telephone conversation and music listening on response time in braking. Transportation Research Part F: Traffic Psychology and Behavior, 12(6), 441-451.
37. Benedetto, S., Pedrotti, M., Minin, L., Baccino, T., Re, A., & Montanari, R. (2011). Driver workload and eye blink duration. Transportation research part F: traffic psychology and behavior, 14(3), 199-208.

38. Bengio, Y., & Grandvalet, Y. (2004). No unbiased estimator of the variance of k-fold cross-validation. Journal of machine learning research, 5(Sep), 1089-1105.
39. Beratis, I. N., Pavlou, D., Papadimitriou, E., Andronas, N., Kontaxopoulou, D., Fragkiadaki, S., ... & Papageorgiou, S. G. (2017). Mild cognitive impairment and driving: does in-vehicle distraction affect driving performance?. Accident Analysis & Prevention, 103, 148-155.
40. Berge, T. J. (2015). Predicting recessions with leading indicators: Model averaging and selection over the business cycle. Journal of Forecasting, 34(6), 455-471.
41. Bertazzon, S., & Elikan, O. (2009). Alternative neighbourhood specifications of the spatial weight matrix; effects on spatial autocorrelation index and multivariate analysis of health data. In 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (ACM GIS 2009): 4-6 November 2009; Seattle.
42. Besag, J., York, J., & Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. Annals of the institute of statistical mathematics, 43(1), 1-20.
43. Besag, J., 1974. Spatial interaction and the statistical analysis of lattice systems. Journal of the Royal Statistical Society B 36(2), pp. 192-236.
44. Bíl, M., Andrášik, R., & Janoška, Z. (2013). Identification of hazardous road locations of traffic accidents by means of kernel density estimation and cluster significance evaluation. Accident Analysis & Prevention, 55, 265-273.
45. Birrell, S. A., Fowkes, M., & Jennings, P. A. (2014). Effect of using an in-vehicle smart driving aid on real-world driver performance. IEEE Transactions on Intelligent Transportation Systems, 15(4), 1801-1810.
46. Bischl, B., Lang, M., Kotthoff, L., Schiffner, J., Richter, J., Studerus, E., Casalicchio. G., & Jones, Z. M. (2016). mlr: Machine Learning in R. The Journal of Machine Learning Research, 17(1), 5938-5942.
47. Bivand, R., Altman, M., Anselin, L., Assunção, R., Berke, O., Bernat, G. A., & Müller, W. (2019). spdep: Spatial dependence: weighting schemes, statistics and models. Version 1.1-3, Updated 2019-09-18.
48. Bivand, R., Yu, D., Nakaya, T., Garcia-Lopez, M. A., & Bivand, M. R. (2017). Package 'spgwr'. R software package.
49. Bivand, R. (2017). Geographically Weighted Regression. Vignette of Bivand, R. S., Pebesma, E. and Gómez-Rubio V. (2008). Applied Spatial Data Analysis with R, Springer-Verlag, New York (pp. 305–308).
50. Bivand, R., Müller, W. G., & Reder, M. (2009). Power calculations for global and local Moran's I. Computational Statistics & Data Analysis, 53(8), 2859-2872.
51. Bivand, R. S., Pebesma, E. J., Gomez-Rubio, V., & Pebesma, E. J. (2008). Applied spatial data analysis with R – Second Edition (Vol. 747248717). New York: Springer.
52. Blana, E., & Golias, J. (2002). Differences between vehicle lateral displacement on the road and in a fixed-base simulator. Human Factors, 44(2), 303-313.
53. Bolstad, W.M. (2007). Introduction to Bayesian statistics (2nd ed.) New Jersey: Wiley.
54. Bonsall, P., Liu, R., & Young, W. (2005). Modelling safety-related driving behaviour —impact of parameter values. Transportation Research Part A: Policy and Practice, 39(5), 425-444.
55. Boquete, L., Rodríguez-Ascariz, J. M., Barea, R., Cantos, J., Miguel-Jiménez, J. M., & Ortega, S. (2010). Data acquisition, analysis and transmission platform for a pay-as-you-drive system. Sensors, 10(6), 5395-5408.
56. Bowers, A. R., Anastasio, R. J., Sheldon, S. S., O'Connor, M. G., Hollis, A. M., Howe, P. D., & Horowitz, T. S. (2013). Can we improve clinical prediction of at-risk older drivers?. Accident Analysis & Prevention, 59, 537-547.
57. Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

58. Brunsdon, C., & Comber, L. (2015). An introduction to R for spatial analysis and mapping. Sage.
59. Bu, L., Wang, F., & Gong, H. (2018). Spatial and factor analysis of vehicle crashes in Mississippi state. Natural Hazards, 1-22.
60. Cai, Q., Abdel-Aty, M., Sun, Y., Lee, J., & Yuan, J. (2019). Applying a deep learning approach for transportation safety planning by using high-resolution transportation and land use data. Transportation Research Part A: Policy and Practice, 127, 71-85.
61. Cai, Q., Abdel-Aty, M., Lee, J., Wang, L., & Wang, X. (2018). Developing a grouped random parameters multivariate spatial model to explore zonal effects for segment and intersection crash modeling. Analytic methods in accident research, 19, 1-15.
62. Cai, Q., Abdel-Aty, M., Lee, J., & Eluru, N. (2017a). Comparative analysis of zonal systems for macro-level crash modeling. Journal of safety research, 61, 157-166.
63. Cai, Q., Abdel-Aty, M., & Lee, J. (2017b). Macro-level vulnerable road users crash analysis: a Bayesian joint modeling approach of frequency and proportion. Accident Analysis & Prevention, 107, 11-19.
64. Cai, Q., Lee, J., Eluru, N., & Abdel-Aty, M. (2016). Macro-level pedestrian and bicycle crash analysis: Incorporating spatial spillover effects in dual state count models. Accident Analysis & Prevention, 93, 14-22.
65. Caird, J. K., Johnston, K. A., Willness, C. R., & Asbridge, M. (2014a). The use of meta-analysis or research synthesis to combine driving simulation or naturalistic study results on driver distraction. Journal of safety research, 49, 91-e1.
66. Caird, J. K., Johnston, K. A., Willness, C. R., Asbridge, M., & Steel, P. (2014b). A meta-analysis of the effects of texting on driving. Accident Analysis & Prevention, 71, 311-318.
67. Caird, J. K., Willness, C. R., Steel, P., & Scialfa, C. (2008). A meta-analysis of the effects of cell phones on driver performance. Accident Analysis & Prevention, 40(4), 1282-1293.
68. Carsten, O., Kircher, K., & Jamson, S. (2013). Vehicle-based studies of driving in the real world: The hard truth?. Accident Analysis & Prevention, 58, 162-174.
69. Chan, A. B., & Vasconcelos, N. (2009). Bayesian poisson regression for crowd counting. In 2009 IEEE 12th international conference on computer vision (pp. 545-551). IEEE.
70. Chang, L. Y., & Wang, H. W. (2006). Analysis of traffic injury severity: An application of non-parametric classification tree techniques. Accident Analysis & Prevention, 38(5), 1019-1027.
71. Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794). ACM.
72. Chen, T., He, T., Benesty, M., Khotilovich, V., & Tang, Y. (2015). Xgboost: extreme gradient boosting. R package version 0.4-2, 1-4.
73. Cheng, J., Karambelkar, B., Xie, Y., Wickham, H., Russell, K., et al. (2019). Package 'leaflet': Create Interactive Web Maps with the JavaScript 'Leaflet' Library. R package version 2.0.3
74. Chidester, A. C. D., Hinch, J., & Roston, T. A. (2001). Real world experience with event data recorders. In Proceedings of the Seventeenth International Technical Conference on the Enhanced Safety of Vehicles, Amsterdam, Netherlands (June 2001).
75. Chiou, Y. C., Fu, C., & Chih-Wei, H. (2014). Incorporating spatial dependence in simultaneously modeling crash frequency and severity. Analytic methods in accident research, 2, 1-11.
76. Chiou, Y. C., & Fu, C. (2013). Modeling crash frequency and severity using multinomial-generalized Poisson model with error components. Accident Analysis & Prevention, 50, 73-82.
77. Chung, W., Abdel-Aty, M., & Lee, J. (2018). Spatial analysis of the effective coverage of land-based weather stations for traffic crashes. Applied geography, 90, 17-27.

78. Consiglio, W., Driscoll, P., Witte, M., & Berg, W. P. (2003). Effect of cellular telephone conversations and other potential interference on reaction time in a braking response. Accident Analysis & Prevention, 35(4), 495-500.
79. Correia, J. T., Iliadis, K. A., McCarron, E. S., Smolej, M. A., Hastings, B., & Engineers, C. C. (2001). Utilizing data from automotive event data recorders. In Proceedings of the Canadian Multidisciplinary Road Safety Conference XII, London Ontario.
80. Cottrill, C. D., & Thakuriah, P. V. (2010). Evaluating pedestrian crashes in areas with high low-income or minority populations. Accident Analysis & Prevention, 42(6), 1718-1728.
81. Cressie, N. A. (1993). Statistics for Spatial Data. New York: John Willey & Sons. Inc.
82. Cressie, N. (1985). Fitting variogram models by weighted least squares. Journal of the International Association for Mathematical Geology, 17(5), 563-586.
83. Cui, G., Wang, X., & Kwon, D. W. (2015). A framework of boundary collision data aggregation into neighbourhoods. Accident Analysis & Prevention, 83, 1-17.
84. Daniels, S., Martensen, H., Schoeters, A., Van den Berghe, W., Papadimitriou, E., Ziakopoulos, A., ... & Weijermars, W. (2019). A systematic cost-benefit analysis of 29 road safety measures. Accident Analysis & Prevention, 133, 105292.
85. Dash, D., & Cooper, G. F. (2004). Model averaging for prediction with discrete Bayesian networks. Journal of Machine Learning Research, 5(Sep), 1177-1203.
86. De Oliveira, V. (2010). Bayesian analysis of conditional autoregressive models. Annals of the Institute of Statistical Mathematics, 64(1), 107-133.
87. De Romph, E. (2013). Using BIG data in transport modelling. Data & Modelling Magazine, vol.13., pp. 2013.
88. De Winter, J., Van Leuween, P., & Happee, P. (2012). Advantages and disadvantages of driving simulators: a discussion. In Proceedings of Measuring Behavior (pp. 47-50).
89. De Winter, J. C. F., & Dodou, D. (2010). The Driver Behaviour Questionnaire as a predictor of accidents: A meta-analysis. Journal of safety research, 41(6), 463-470.
90. Delmelle, E., & Thill, J. C. (2008). Urban bicyclists: spatial analysis of adult and youth traffic hazard intensity. Transportation Research Record: Journal of the Transportation Research Board, (2074), 31-39.
91. Desmond, P., Hancock, P., & Monette, J. (1998). Fatigue and automation-induced impairments in simulated driving performance. Transportation Research Record: Journal of the Transportation Research Board, (1628), 8-14.
92. Di Stefano, M., & Macdonald, W. (2003). Assessment of older drivers: relationships among on-road errors, medical conditions and test outcome. Journal of safety research, 34(4), 415-429.
93. Dingus, T. A., Guo, F., Lee, S., Antin, J. F., Perez, M., Buchanan-King, M., & Hankey, J. (2016). Driver crash risk factors and prevalence evaluation using naturalistic driving data. Proceedings of the National Academy of Sciences, 113(10), 2636-2641.
94. Dingus, T. A., Hankey, J. M., Antin, J. F., Lee, S. E., Eichelberger, L., Stulce, K. E., ... & Stowe, L. (2015). Naturalistic driving study: Technical coordination and quality control (No. SHRP 2 Report S2-S06-RW-1).
95. Dingus, T. A., Klauer, S. G., Neale, V. L., Petersen, A., Lee, S. E., Sudweeks, J. D., ... & Bucher, C. (2006). The 100-car naturalistic driving study, Phase II-results of the 100-car field experiment (No. HS-810 593).
96. Dong, N., Huang, H., Lee, J., Gao, M., & Abdel-Aty, M. (2016). Macroscopic hotspots identification: a Bayesian spatio-temporal interaction approach. Accident Analysis & Prevention, 92, 256-264.

97. Dong, N., Huang, H., & Zheng, L. (2015). Support vector machine in crash prediction at the level of traffic analysis zones: assessing the spatial proximity effects. Accident Analysis & Prevention, 82, 192-198.

98. Dong, N., Huang, H., Xu, P., Ding, Z., & Wang, D. (2014). Evaluating spatial-proximity structures in crash prediction models at the level of traffic analysis zones. Transportation Research Record: Journal of the Transportation Research Board, (2432), 46-52.

99. Dumitru, A. I., Girbacia, T., Boboc, R. G., Postelnicu, C. C., & Mogan, G. L. (2018). Effects of smartphone based advanced driver assistance system on distracted driving behavior: A simulator study. Computers in Human Behavior, 83, 1-7.

100. Eby, D. W., & Vivoda, J. M. (2003). Driver hand-held mobile phone use and safety belt use. Accident Analysis & Prevention, 35(6), 893-895.

101. Effati, M., Thill, J. C., & Shabani, S. (2015). Geospatial and machine learning techniques for wicked social science problems: analysis of crash severity on a regional highway corridor. Journal of Geographical Systems, 17(2), 107-135.

102. Efthymiou, A., Barmpounakis, E. N., Efthymiou, D., & Vlahogianni, E. I. (2019). Transportation Mode Detection from Low-Power Smartphone Sensors Using Tree-Based Ensembles. Journal of Big Data Analytics in Transportation, 1(1), 57-69.

103. El-Basyouny, K., & Sayed, T. (2011). A full Bayes multivariate intervention model with random parameters among matched pairs for before–after safety evaluation. Accident Analysis & Prevention, 43(1), 87-94.

104. El-Basyouny, K., & Sayed, T. (2009). Urban arterial accident prediction models with spatial effects. Transportation Research Record: Journal of the Transportation Research Board, (2102), 27-33.

105. Ellison, A. B., Bliemer, M. C., & Greaves, S. P. (2015). Evaluating changes in driver behavior: a risk profiling approach. Accident Analysis & Prevention, 75, 298-309.

106. Elvik, R., & Goel, R. (2019). Safety-in-numbers: An updated meta-analysis of estimates. Accident Analysis & Prevention, 129, 136-147.

107. Elvik, R., & Bjørnskau, T. (2017). Safety-in-numbers: a systematic review and meta-analysis of evidence. Safety science, 92, 274-282.

108. Elvik, R. (2015). Some implications of an event-based definition of exposure to the risk of road accident. Accident Analysis & Prevention, 76, 15-24.

109. Elvik, R. (2011). Effects of mobile phone use on accident risk: Problems of meta-analysis when studies are few and bad. Transportation research record, 2236(1), 20-26.

110. Elvik, R., Vaa, T., Hoye, A., & Sorensen, M. (Eds.). (2009). The handbook of road safety measures. Emerald Group Publishing.

111. Elvik, R. (2005). Introductory guide to systematic reviews and meta-analysis. Transportation Research Record: Journal of the Transportation Research Board, (1908), 230-235.

112. Elvik, R. (2001). Area-wide urban traffic calming schemes: a meta-analysis of safety effects. Accident Analysis & Prevention, 33(3), 327-336.

113. Enev, M., Takakuwa, A., Koscher, K., & Kohno, T. (2016). Automobile driver fingerprinting. Proceedings on Privacy Enhancing Technologies, 2016(1), 34-50.

114. Erdogan, S. (2009). Explorative spatial analysis of traffic accident statistics and road mortality among the provinces of Turkey. Journal of safety research, 40(5), 341-351.

115. Erdogan, S., Yilmaz, I., Baybura, T., & Gullu, M. (2008). Geographical information systems aided traffic accident analysis system case study: city of Afyonkarahisar. Accident Analysis & Prevention, 40(1), 174-181.

116. Filtness, A. J., Armstrong, K. A., Watson, A., & Smith, S. S. (2017). Sleep-related crash characteristics: Implications for applying a fatigue definition to crash reports. Accident Analysis & Prevention, 99, 440-444.
117. Flahaut, B. (2004). Impact of infrastructure and local environment on road unsafety: Logistic modeling with spatial autocorrelation. Accident Analysis & Prevention, 36(6), 1055-1066.
118. Flask, T., & Schneider IV, W. (2013). A Bayesian analysis of multi-level spatial correlation in single vehicle motorcycle crashes in Ohio. Safety science, 53, 1-10.
119. Fotheringham, A. S., & Oshan, T. M. (2016). Geographically weighted regression and multicollinearity: dispelling the myth. Journal of Geographical Systems, 18(4), 303-329.
120. Fotheringham, A. S., Brunsdon, C., & Charlton, M. (2002). Geographically weighted regression (pp. 159-183). West Atrium: John Wiley & Sons, Limited.
121. Fotheringham, S., Brunsdon, C., & Charlton, M. (2000). Quantitative Geography: Perspectives on Spatial Data Analysis. London: Sage.
122. Fotheringham, S., & Wegener, M. (1999). Spatial models and GIS: New and potential models (Vol. 7). CRC press.
123. Geary, R. C. (1954). The contiguity ratio and statistical mapping. The incorporated statistician, 5(3), 115-146.
124. Geroliminis, N., & Skabardonis, A. (2005). Prediction of arrival profiles and queue lengths along signalized arterials by using a Markov decision process. Transportation Research Record, 1934(1), 116-124.
125. Geurts, K. & Wets, G. (2003). Black spot analysis methods: Literature review. Flemish Research Center for Traffic Safety, Diepenbeek, Belgium.
126. Gomes, M. J. T. L., Cunto, F., & da Silva, A. R. (2017). Geographically weighted negative binomial regression applied to zonal level safety performance models. Accident Analysis & Prevention, 106, 254-261.
127. Goodchild, M. F. (2008). Commentary: whither VGI?. GeoJournal, 72(3-4), 239-244.
128. Greenshields, B. D., Bibbins, J. R., Channing, W. S., & Miller, H. H. (1935). A study of traffic capacity. In Highway research board proceedings (Vol. 1935). National Research Council (USA), Highway Research Board.
129. Gringarten, E., & Deutsch, C. V. (2001). Teacher's aide: variogram interpretation and modeling. Mathematical Geology, 33(4), 507-534.
130. Guadamuz-Flores, R., & Aguero-Valverde, J. (2017). Bayesian spatial models of crash frequency at highway–railway crossings. Transportation Research Record, 2608(1), 27-35.
131. 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.
132. Guo, Q., Xu, P., Pei, X., Wong, S. C., & Yao, D. (2017). The effect of road network patterns on pedestrian safety: A zone-based Bayesian spatial modeling approach. Accident Analysis & Prevention, 99, 114-124.
133. Guo, F., Wang, X., & Abdel-Aty, M. A. (2010). Modeling signalized intersection safety with corridor-level spatial correlations. Accident Analysis & Prevention, 42(1), 84-92.
134. Hadayeghi, A., Shalaby, A. S., & Persaud, B. N. (2010). Development of planning level transportation safety tools using Geographically Weighted Poisson Regression. Accident Analysis & Prevention, 42(2), 676-688.
135. Hadayeghi, A., Shalaby, A., & Persaud, B. (2003). Macrolevel accident prediction models for evaluating safety of urban transportation systems. Transportation Research Record: Journal of the Transportation Research Board, (1840), 87-95.

136. Hakkert, A. S., Braimaister, L., & Van Schagen, I. (2002). The uses of exposure and risk in road safety studies (Vol. 2002, No. 12). SWOV Institute for Road Safety.

137. Haklay, M., & Weber, P. (2008). Openstreetmap: User-generated street maps. IEEE Pervasive Computing, 7(4), 12-18.

138. Han, C., Huang, H., Lee, J., & Wang, J. (2018). Investigating varying effect of road-level factors on crash frequency across regions: a Bayesian hierarchical random parameter modeling approach. Analytic methods in accident research, 20, 81-91.

139. Handel, P., Skog, I., Wahlstrom, J., Bonawiede, F., Welch, R., Ohlsson, J., & Ohlsson, M. (2014). Insurance telematics: Opportunities and challenges with the smartphone solution. IEEE Intelligent Transportation Systems Magazine, 6(4), 57-70.

140. Hanowski, R. J., Perez, M. A., & Dingus, T. A. (2005). Driver distraction in long-haul truck drivers. Transportation Research Part F: Traffic Psychology and Behaviour, 8(6), 441-458.

141. Harris, R. (2013). An Introduction to Mapping and Spatial Modelling in R. School of Geographical Sciences, University of Bristol.

142. Hart, T., & Zandbergen, P. (2014). Kernel density estimation and hotspot mapping: Examining the influence of interpolation method, grid cell size, and bandwidth on crime forecasting. Policing: An International Journal of Police Strategies & Management, 37(2), 305-323.

143. Hauer, E. (1997). Observational before/after studies in road safety. Estimating the effect of highway and traffic engineering measures on road safety.

144. Hauer, E., Ng, J. C., & Lovell, J. (1988). Estimation of safety at signalized intersections (with discussion and closure) (No. 1185).

145. Hastie, T., Tibshirani, R., Friedman, J., & Franklin, J. (2009). The elements of statistical learning: data mining, inference and prediction – Second Edition. The Mathematical Intelligencer, 27(2), 83-85.

146. Hellenic Statistical Authority (ELSTAT). (2019). Road Crashes of year 2017. Press release. Available: https://www.statistics.gr/el/statistics/-/publication/SDT04/2017 [Accessed 22-11-2019] [In Greek].

147. Helman, S., & Reed, N. (2015). Validation of the driver behaviour questionnaire using behavioural data from an instrumented vehicle and high-fidelity driving simulator. Accident Analysis & Prevention, 75, 245-251.

148. Hendry, D. F., & Clements, M. P. (2004). Pooling of forecasts. The Econometrics Journal, 7(1), 1-31.

149. Hensher, D. A., & Stopher, P. R. (Eds.). (1979). Behavioural travel modelling. Taylor & Francis.

150. Hill, J., Aldah, M., Talbot, R., Giustiniani, G., Fagerlind, H., Jänsch, M. (2012). Final Report, Deliverable 2.5 of the EC FP7 project DaCoTA.

151. Horrey, W. J., & Wickens, C. D. (2006). Examining the impact of cell phone conversations on driving using meta-analytic techniques. Human factors, 48(1), 196-205.

152. Høye, A., & Elvik, R. (2010). Publication Bias in Road Safety Evaluation: How Can It Be Detected and How Common Is It?. Transportation Research Record: Journal of the Transportation Research Board, (2147), 1-8.

153. Hu, W., Yan, L., Liu, K., & Wang, H. (2016). A short-term traffic flow forecasting method based on the hybrid PSO-SVR. Neural Processing Letters, 43(1), 155-172.

154. Huang, H., Zhou, H., Wang, J., Chang, F., & Ma, M. (2017). A multivariate spatial model of crash frequency by transportation modes for urban intersections. Analytic methods in accident research, 14, 10-21.

155. Huang, H., Song, B., Xu, P., Zeng, Q., Lee, J., & Abdel-Aty, M. (2016). Macro and micro models for zonal crash prediction with application in hot zones identification. Journal of Transport Geography, 54, 248-256.

156. Huang, H., & Abdel-Aty, M. (2010). Multilevel data and Bayesian analysis in traffic safety. Accident Analysis & Prevention, 42(6), 1556-1565.
157. Huang, H., Abdel-Aty, M., & Darwiche, A. (2010). County-level crash risk analysis in Florida: Bayesian spatial modeling. Transportation Research Record: Journal of the Transportation Research Board, (2148), 27-37.
158. Huang, H., Chin, H., & Haque, M. (2009). Empirical evaluation of alternative approaches in identifying crash hot spots: naive ranking, empirical Bayes, and full Bayes methods. Transportation Research Record: Journal of the Transportation Research Board, (2103), 32-41.
159. Hughes, G. M., Rudin-Brown, C. M., & Young, K. L. (2013). A simulator study of the effects of singing on driving performance. Accident Analysis & Prevention, 50, 787-792.
160. Hijmans, R. J., Williams, E., Vennes, C., & Hijmans, M. R. J. (2017). Package ‘geosphere’. R package version, 3.
161. Imprialou, M. I. M., Quddus, M., Pitfield, D. E., & Lord, D. (2016). Re-visiting crash–speed relationships: A new perspective in crash modelling. Accident Analysis & Prevention, 86, 173-185.
162. Inagaki, Y., Shinkuma, R., Sato, T., & Oki, E. (2019). Prioritization of Mobile IoT Data Transmission Based on Data Importance Extracted From Machine Learning Model. IEEE Access, 7, 93611-93620.
163. International Traffic Safety Data Analysis Group (IRTAD), (2011). Reporting on Serious Road Traffic Casualties: Combining and Using Different Data Sources to Improve Understanding of Non-fatal Road Traffic Crashes. International Traffic Safety Data and Analysis Group.
164. Ivancic IV, K., & Hesketh, B. (2000). Learning from errors in a driving simulation: Effects on driving skill and self-confidence. Ergonomics, 43(12), 1966-1984.
165. Iqbal, M. U., & Lim, S. (2006). A privacy preserving GPS-based Pay-as-You-Drive insurance scheme. In Symposium on GPS/GNSS (IGNSS2006) (pp. 17-21).
166. Jackson, P., Hilditch, C., Holmes, A., Reed, N., Merat, N., & Smith, L. (2011). Fatigue and road safety: a critical analysis of recent evidence. UK Department for Transport, (21).
167. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, p. 18). New York: springer.
168. Jansen, R. J., & Wesseling, (2018). SWOV. Harsh Braking by Truck Drivers: A Comparison of Thresholds and Driving Contexts Using Naturalistic Driving Data. In Proceedings of the 6th Humanist Conference.
169. Jarvis, A., Reuter, H. I., Nelson, A., & Guevara, E. (2008). Hole-filled SRTM for the globe Version 4. Available from the CGIAR-CSI SRTM 90m Database (http://srtm.csi.cgiar.org), 15, 25-54.
170. Jensen, M., Wagner, J., & Alexander, K. (2011). Analysis of in-vehicle driver behavior data for improved safety. International journal of vehicle safety, 5(3), 197-212.
171. Jiang, X., Abdel-Aty, M., Hu, J., & Lee, J. (2016). Investigating macro-level hotzone identification and variable importance using big data: A random forest models approach. Neurocomputing, 181, 53-63.
172. Jurecki, R.S., Stańczyk, T.L., & Jaśkiewicz, M.J. (2014). Driver’s reaction time in a simulated, complex road incident. Transport, 32(1), 44-54.
173. Kaber, D. B., Liang, Y., Zhang, Y., Rogers, M. L., & Gangakhedkar, S. (2012). Driver performance effects of simultaneous visual and cognitive distraction and adaptation behavior. Transportation research part F: traffic psychology and behavior, 15(5), 491-501.
174. Kamarianakis, Y., Kanas, A., & Prastacos, P. (2005). Modeling traffic volatility dynamics in an urban network. Transportation Research Record, 1923(1), 18-27.
175. Kamla, J., Parry, T., & Dawson, A. (2019). Analysing truck harsh braking incidents to study roundabout accident risk. Accident Analysis & Prevention, 122, 365-377.

176. Kanarachos, S., Christopoulos, S. R. G., & Chroneos, A. (2018). Smartphones as an integrated platform for monitoring driver behavior: The role of sensor fusion and connectivity. Transportation Research Part C: Emerging Technologies.
177. Karlaftis, M. G., & Tarko, A. P. (1998). Heterogeneity considerations in accident modeling. Accident Analysis & Prevention, 30(4), 425-433.
178. Karney, C. F. (2013). Algorithms for geodesics. Journal of Geodesy, 87(1), 43-55.
179. Katrakazas, C., Quddus, M., & Chen, W. H. (2017). A simulation study of predicting real-time conflict-prone traffic conditions. IEEE Transactions on Intelligent Transportation Systems, 19(10), 3196-3207.
180. Kelley, K., Clark, B., Brown, V., & Sitzia, J. (2003). Good practice in the conduct and reporting of survey research. International Journal for Quality in health care, 15(3), 261-266.
181. Kerner, B. S. (2012). The physics of traffic: empirical freeway pattern features, engineering applications, and theory. Springer.
182. Kerner, B. (2004). The physics of traffic: Empirical freeway pattern features, engineering applications, and theory. Understanding complex systems series, J. A. Kelso, ed., Springer, New York.
183. Kim, K., Brunner, I. M., & Yamashita, E. Y. (2006). Influence of land use, population, employment, and economic activity on accidents. Transportation research record, 1953(1), 56-64.
184. Kitanidis, P. K. (1997). Introduction to geostatistics: applications in hydrogeology. Cambridge University Press.
185. Kyriakidis, M., Happee, R., & de Winter, J. C. (2015). Public opinion on automated driving: Results of an international questionnaire among 5000 respondents. Transportation research part F: traffic psychology and behaviour, 32, 127-140.
186. Ladron de Guevara, F., Washington, S., & Oh, J. (2004). Forecasting crashes at the planning level: simultaneous negative binomial crash model applied in Tucson, Arizona. Transportation Research Record: Journal of the Transportation Research Board, (1897), 191-199.
187. Lajunen, T., & Summala, H. (2003). Can we trust self-reports of driving? Effects of impression management on driver behavior questionnaire responses. Transportation Research Part F: Traffic Psychology and Behavior, 6(2), 97-107.
188. Laohasiriwong, W., Puttanapong, N., & Luenam, A. (2017). A comparison of spatial heterogeneity with local cluster detection methods for chronic respiratory diseases in Thailand. F1000Research, 6.
189. LaScala, E. A., Gruenewald, P. J., & Johnson, F. W. (2004). An ecological study of the locations of schools and child pedestrian injury collisions. Accident Analysis & Prevention, 36(4), 569-576.
190. LaScala, E. A., Johnson, F. W., & Gruenewald, P. J. (2001). Neighborhood characteristics of alcohol-related pedestrian injury collisions: a geostatistical analysis. Prevention Science, 2(2), 123-134.
191. LaScala, E. A., Gerber, D., & Gruenewald, P. J. (2000). Demographic and environmental correlates of pedestrian injury collisions: a spatial analysis. Accident Analysis & Prevention, 32(5), 651-658.
192. Lawson, A. B., Browne, W. J., & Rodeiro, C. L. V. (2003). Disease mapping with WinBUGS and MLwiN (Vol. 11). John Wiley & Sons.
193. Lee, J., & Abdel-Aty, M. (2018). Macro-level analysis of bicycle safety: Focusing on the characteristics of both crash location and residence. International journal of sustainable transportation, 12(8), 553-560.
194. Lee, J., Abdel-Aty, M., Huang, H., & Cai, Q. (2019a). Transportation Safety Planning Approach for Pedestrians: An Integrated Framework of Modeling Walking Duration and Pedestrian Fatalities. Transportation Research Record, 2673(4), 898-906.

195. Lee, J., Abdel-Aty, M., De Blasiis, M. R., Wang, X., & Mattei, I. (2019b). International transferability of macro-level safety performance functions: a case study of the United States and Italy. Transportation Safety and Environment.
196. Lee, J., Abdel-Aty, A., & Park, J. (2018a). Investigation of associations between marijuana law changes and marijuana-involved fatal traffic crashes: A state-level analysis. Journal of Transport & Health, 10, 194-202.
197. Lee, J., Yasmin, S., Eluru, N., Abdel-Aty, M., & Cai, Q. (2018b). Analysis of crash proportion by vehicle type at traffic analysis zone level: A mixed fractional split multinomial logit modeling approach with spatial effects. Accident Analysis & Prevention, 111, 12-22.
198. Lee, J., Abdel-Aty, M., Cai, Q., Wang, L., & Huang, H. (2018c). Integrated modeling approach for non-motorized mode trips and fatal crashes in the framework of transportation safety planning. Transportation research record, 2672(32), 49-60.
199. Lee, J., Abdel-Aty, M., & Cai, Q. (2017a). Intersection crash prediction modeling with macro-level data from various geographic units. Accident Analysis & Prevention, 102, 213-226.
200. Lee, J., Abdel-Aty, M., Wang, J. H., & Lee, C. (2017b). Long-term effect of universal helmet law changes on motorcyclist fatal crashes: comparison group and empirical Bayes approaches. Transportation Research Record, 2637(1), 27-37.
201. Lee, J., Abdel-Aty, M., & Jiang, X. (2015a). Multivariate crash modeling for motor vehicle and non-motorized modes at the macroscopic level. Accident Analysis & Prevention, 78, 146-154.
202. Lee, J., Abdel-Aty, M., Choi, K., & Huang, H. (2015b). Multi-level hot zone identification for pedestrian safety. Accident Analysis & Prevention, 76, 64-73.
203. Lee, J., Abdel-Aty, M., & Choi, K. (2014a). Analysis of residence characteristics of at-fault drivers in traffic crashes. Safety science, 68, 6-13.
204. Lee, J., Abdel-Aty, M., & Jiang, X. (2014b). Development of zone system for macro-level traffic safety analysis. Journal of transport geography, 38, 13-21.
205. Lee, I. J. (2014a). Big data processing framework of road traffic collision using distributed CEP. In Network Operations and Management Symposium (APNOMS), September 2014 16th Asia-Pacific (pp. 1-4). IEEE.
206. Lee, J. (2014b). Development of Traffic Safety Zones and Integrating Macroscopic and Microscopic Safety Data Analytics for Novel Hot Zone Identification. Doctoral dissertation, University of Central Florida. Electronic Theses and Dissertations. 4619.
207. Lee, D. (2013). CARBayes [version 5.1.2]: an R package for Bayesian spatial modeling with conditional autoregressive priors. Journal of Statistical Software, 55(13), 1-24.
208. Lenné, M. G., Triggs, T. J., & Redman, J. R. (1997). Time of day variations in driving performance. Accident Analysis & Prevention, 29(4), 431-437.
209. Leroux, B. G., Lei, X., & Breslow, N. (2000). Estimation of disease rates in small areas: a new mixed model for spatial dependence. In Statistical models in epidemiology, the environment, and clinical trials (pp. 179-191). Springer, New York, NY.
210. Levine, N., Kim, K. E., & Nitz, L. H. (1995). Spatial analysis of Honolulu motor vehicle crashes: II. Zonal generators. Accident Analysis & Prevention, 27(5), 675-685.
211. Li, P., Abdel-Aty, M., & Yuan, J. (2020). Real-time crash risk prediction on arterials based on LSTM-CNN. Accident Analysis & Prevention, 135, 105371.
212. Li, Z., Chen, X., Ci, Y., Chen, C., & Zhang, G. (2019). A hierarchical Bayesian spatiotemporal random parameters approach for alcohol/drug impaired-driving crash frequency analysis. Analytic Methods in Accident Research.
213. Li, Z., Wang, W., Liu, P., Bigham, J. M., & Ragland, D. R. (2013). Using geographically weighted Poisson regression for county-level crash modeling in California. Safety science, 58, 89-97.

214. 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.
215. Lighthill, M. J., & Whitham, G. B. (1955). On kinematic waves II. A theory of traffic flow on long crowded roads. Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences, 229(1178), 317-345.
216. Lin, L., Wang, Q., & Sadek, A. W. (2015). A novel variable selection method based on frequent pattern tree for real-time traffic accident risk prediction. Transportation Research Part C: Emerging Technologies, 55, 444-459.
217. Liu, C., & Sharma, A. (2018). Using the multivariate spatio-temporal Bayesian model to analyze traffic crashes by severity. Analytic methods in accident research, 17, 14-31.
218. Liu, J., Khattak, A. J., & Wali, B. (2017). Do safety performance functions used for predicting crash frequency vary across space? Applying geographically weighted regressions to account for spatial heterogeneity. Accident Analysis & Prevention, 109, 132-142.
219. Loo, B. P., & Anderson, T. K. (2015). Spatial Analysis Methods of Road Traffic Collisions. CRC Press.
220. Loo, B. P., Yao, S., & Wu, J. (2011). Spatial point analysis of road crashes in Shanghai: A GIS-based network kernel density method. In 2011 19th international conference on geoinformatics (pp. 1-6). IEEE.
221. 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.
222. Loukaitou-Sideris, A., Liggett, R., & Sung, H. G. (2007). Death on the crosswalk: A study of pedestrian-automobile collisions in Los Angeles. Journal of Planning Education and Research, 26(3), 338-351.
223. Lovegrove, G., Lim, C., & Sayed, T. (2009). Community-based, macrolevel collision prediction model use with a regional transportation plan. Journal of transportation engineering, 136(2), 120-128.
224. Lovegrove, G., & Sayed, T. (2007). Macrolevel collision prediction models to enhance traditional reactive road safety improvement programs. Transportation Research Record: Journal of the Transportation Research Board, (2019), 65-73.
225. Lovegrove, G. R., & Sayed, T. (2006). Macro-level collision prediction models for evaluating neighbourhood traffic safety. Canadian Journal of Civil Engineering, 33(5), 609-621.
226. Lovelace, R., Nowosad, J., & Muenchow, J. (2019). Geocomputation with R. CRC Press.
227. Lu, B., Charlton, M., Harris, P., & Fotheringham, A. S. (2014). Geographically weighted regression with a non-Euclidean distance metric: a case study using hedonic house price data. International Journal of Geographical Information Science, 28(4), 660-681.
228. Lu, B., Harris, P., Gollini, I., Charlton, M., & Brunsdon, C. (2013). GWmodel: an R package for exploring spatial heterogeneity. GISRUK 2013, 3-5.
229. Ma, T., Antoniou, C., & Toledo, T. (2020). Hybrid machine learning algorithm and statistical time series model for network-wide traffic forecast. Transportation Research Part C: Emerging Technologies, 111, 352-372.
230. Ma, Y. L., Zhu, X., Hu, X., & Chiu, Y. C. (2018). The use of context-sensitive insurance telematics data in auto insurance rate making. Transportation Research Part A: Policy and Practice, 113, 243-258.
231. Ma, X., Chen, S., & Chen, F. (2017). Multivariate space-time modeling of crash frequencies by injury severity levels. Analytic Methods in Accident Research, 15, 29-40.

232. Ma, X., Tao, Z., Wang, Y., Yu, H., & Wang, Y. (2015a). Long short-term memory neural network for traffic speed prediction using remote microwave sensor data. Transportation Research Part C: Emerging Technologies, 54, 187-197.
233. Ma, X., Yu, H., Wang, Y., & Wang, Y. (2015b). Large-scale transportation network congestion evolution prediction using deep learning theory. PloS one, 10(3), e0119044.
234. MacNab, Y. C. (2004). Bayesian spatial and ecological models for small-area accident and injury analysis. Accident Analysis & Prevention, 36(6), 1019-1028.
235. Manan, M. M. A., & Varhelyi, A. (2015). Motorcyclists' road safety related behavior at access points on primary roads in Malaysia–A case study. Safety Science, 77, 80-94.
236. Mannering, F. L., & Bhat, C. R. (2014). Analytic methods in accident research: Methodological frontier and future directions. Analytic methods in accident research, 1, 1-22.
237. Mantouka, E. G., Barmpounakis, E. N., & Vlahogianni, E. I. (2019). Identification of driving safety profiles from smartphone data using machine learning techniques. Safety Science.
238. Mantouka, E. G., Barmpounakis, E. N., & Vlahogianni, E. I. (2018). Mobile Sensing and Machine Learning for Identifying Driving Safety Profiles (No. 18-01416).
239. Martensen, H., Diependaele, K., Daniels, S., Van den Berghe, W., Papadimitriou, E., Yannis, G., Van Schagen, I., Weijermars, W., Wijnen, W., Filtness, A., Talbot, R. (2019). The European road safety decision support system on risks and measures. Accident Analysis & Prevention, 125, 344-351.
240. Martinussen, L. M., Hakamies-Blomqvist, L., Møller, M., Özkan, T., & Lajunen, T. (2013). Age, gender, mileage and the DBQ: The validity of the Driver Behavior Questionnaire in different driver groups. Accident Analysis & Prevention, 52, 228-236.
241. Matheron, G. (1963). Principles of geostatistics. Economic geology, 58(8), 1246-1266.
242. McCulloch, C. E. (2003). Generalized linear mixed models. In NSF-CBMS regional conference series in probability and statistics (pp. i-84). Institute of Mathematical Statistics and the American Statistical Association.
243. McFadden, D. (1977). Quantitative methods for analyzing travel behaviour of individuals: Some recent developments (Cowles Foundation Discussion Papers No. 474). Cowles Foundation for Research in Economics, Yale University.
244. Medina, S. A. O. (2018). Inferring weekly primary activity patterns using public transport smart card data and a household travel survey. Travel Behavior and Society, 12, 93-101.
245. Merat, N., Jamson, A. H., Lai, F. C., Daly, M., & Carsten, O. M. (2014). Transition to manual: Driver behavior when resuming control from a highly automated vehicle. Transportation research part F: traffic psychology and behavior, 27, 274-282.
246. Minis, I., & Tsamboulas, D. A. (2008). Contingency planning and war gaming for the transport operations of the Athens 2004 Olympic Games. Transport Reviews, 28(2), 259-280.
247. Mitra, S. (2009). Spatial autocorrelation and Bayesian spatial statistical method for analyzing intersections prone to injury crashes. Transportation research record, 2136(1), 92-100.
248. Miaou, S. P., & Lord, D. (2003). Modeling traffic crash-flow relationships for intersections: dispersion parameter, functional form, and Bayes versus empirical Bayes methods. Transportation Research Record: Journal of the Transportation Research Board, (1840), 31-40.
249. Miaou, S. P., & Song, J. J. (2005). Bayesian ranking of sites for engineering safety improvements: decision parameter, treatability concept, statistical criterion, and spatial dependence. Accident Analysis & Prevention, 37(4), 699-720.
250. Moeinaddini, M., Asadi-Shekari, Z., & Shah, M. Z. (2014). The relationship between urban street networks and the number of transport fatalities at the city level. Safety science, 62, 114-120.

251. Mohaymany, A. S., Shahri, M., & Mirbagheri, B. (2013). GIS-based method for detecting high-crash-risk road segments using network kernel density estimation. Geo-spatial Information Science, 16(2), 113-119.

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

253. Mountrakis, G., & Gunson, K. (2009). Multi-scale spatiotemporal analyses of moose–vehicle collisions: a case study in northern Vermont. International Journal of Geographical Information Science, 23(11), 1389-1412.

254. Nabatilan, L. B., Aghazadeh, F., Nimbarte, A. D., Harvey, C. C., & Chowdhury, S. K. (2012). Effect of driving experience on visual behavior and driving performance under different driving conditions. Cognition, Technology & Work, 14(4), 355-363.

255. Naderan, A., & Shahi, J. (2010). Aggregate crash prediction models: Introducing crash generation concept. Accident Analysis & Prevention, 42(1), 339-346.

256. Nakaya, T., Fotheringham, S., Charlton, M., & Brunsdon, C. (2009). Semiparametric geographically weighted generalised linear modelling in GWR 4.0.

257. Narayanamoorthy, S., Paleti, R., & Bhat, C. R. (2013). On accommodating spatial dependence in bicycle and pedestrian injury counts by severity level. Transportation research part B: methodological, 55, 245-264.

258. Nashad, T., Yasmin, S., Eluru, N., Lee, J., & Abdel-Aty, M. A. (2016). Joint modeling of pedestrian and bicycle crashes: copula-based approach. Transportation Research Record: Journal of the Transportation Research Board, (2601), 119-127.

259. Neale, V. L., Dingus, T. A., Klauer, S. G., Sudweeks, J., & Goodman, M. (2005). An overview of the 100-car naturalistic study and findings. National Highway Traffic Safety Administration, Paper, 5, 0400.

260. Ng, K. S., Hung, W. T., & Wong, W. G. (2002). An algorithm for assessing the risk of traffic accident. Journal of safety research, 33(3), 387-410.

261. Nielsen, D. (2016). Tree boosting with xgboost-why does xgboost win" every" machine learning competition? (Master's thesis, NTNU).

262. Nikias, V. A., Vlahogianni, E. I., Lee, T. C., & Golias, J. C. (2012). Determinants of powered two-wheelers virtual lane width in urban arterials. In Intelligent Transportation Systems (ITSC), 2012 15th International IEEE Conference on (pp. 1205-1210). IEEE

263. Noland, R. B., & Quddus, M. A. (2005). Congestion and safety: A spatial analysis of London. Transportation Research Part A: Policy and Practice, 39(7-9), 737-754.

264. Noland, R. B., & Oh, L. (2004). The effect of infrastructure and demographic change on traffic-related fatalities and crashes: a case study of Illinois county-level data. Accident Analysis & Prevention, 36(4), 525-532.

265. Noland, R. B., & Quddus, M. A. (2004). A spatially disaggregate analysis of road casualties in England. Accident Analysis & Prevention, 36(6), 973-984.

266. Oden, N. (1995). Adjusting Moran's I for population density. Statistics in Medicine, 14(1), 17-26.

267. Olson, R. L., Hanowski, R. J., Hickman, J. S., & Bocanegra, J. (2009). Driver distraction in commercial vehicle operations (No. FMCSA-RRT-09-042). United States. Federal Motor Carrier Safety Administration.

268. Orfanou, F. P., Vlahogianni, E. I., & Karlaftis, M. G. (2012). Associating driving behavior with hysteretic phenomena of freeway traffic flow. IFAC Proceedings Volumes, 45(24), 209-214.

269. Openshaw, S. (1984). The modifiable areal unit problem. Concepts and techniques in modern geography. Geo Books, Norwich.

270. OpenStreetMap. (2019). Official Wiki Website. Available: https://wiki.openstreetmap.org/wiki/About_OpenStreetMap [Accessed 12-11-2019]

271. Ord, J. K., & Getis, A. (1995). Local spatial autocorrelation statistics: distributional issues and an application. Geographical analysis, 27(4), 286-306.

272. OSeven Telematics (2019). Official Website. Available: https://oseven.io/ [Accessed 14-11-2019]

273. Ossenbruggen, P. J., Linder, E., & Nguyen, B. (2009). Detecting unsafe roadways with spatial statistics: point patterns and geostatistical models. Journal of Transportation Engineering, 136(5), 457-464.

274. Paefgen, J., Staake, T., & Fleisch, E. (2014). Multivariate exposure modeling of accident risk: Insights from Pay-as-you-drive insurance data. Transportation Research Part A: Policy and Practice, 61, 27-40.

275. Paefgen, J., Kehr, F., Zhai, Y., & Michahelles, F. (2012). Driving behavior analysis with smartphones: insights from a controlled field study. In Proceedings of the 11th International Conference on mobile and ubiquitous multimedia (pp. 1-8).

276. Page, S. J., & Meyer, D. (1996). Tourist accidents: an exploratory analysis. Annals of Tourism Research, 23(3), 666-690.

277. Palinko, O., Kun, A. L., Shyrokov, A., & Heeman, P. (2010). Estimating cognitive load using remote eye tracking in a driving simulator. In Proceedings of the 2010 symposium on eye-tracking research & applications (pp. 141-144). ACM.

278. Papadimitriou, E., Filtness, A., Theofilatos, A., Ziakopoulos, A., Quigley, C., & Yannis, G. (2019a). Review and ranking of crash risk factors related to the road infrastructure. Accident Analysis & Prevention, 125, 85-97.

279. Papadimitriou, E., Argyropoulou, A., Tselentis, D. I., & Yannis, G. (2019b). Analysis of driver behaviour through smartphone data: The case of mobile phone use while driving. Safety Science.

280. Papadimitriou, E., Tselentis, D. I., & Yannis, G. (2018). Analysis of Driving Behaviour Characteristics Based on Smartphone Data, Proceedings of 7th Transport Research Arena TRA 2018, April 16-19, 2018, Vienna, Austria.

281. Papadimitriou, E., Eksler, V., Yannis, G., & Lassarre, S. (2013). Modelling the spatial variation of road safety in Greece. In Proceedings of the Institution of Civil Engineers-Transport (Vol. 166, No. 1, pp. 49-58). Thomas Telford Ltd.

282. Papantoniou, P., Papadimitriou, E., & Yannis, G. (2015). Assessment of driving simulator studies on driver distraction. Advances in transportation studies, (35).

283. Patel, A., Katiyar, S. K., & Prasad, V. (2016). Performances evaluation of different open source DEM using Differential Global Positioning System (DGPS). The Egyptian Journal of Remote Sensing and Space Science, 19(1), 7-16.

284. Pebesma, E., & Graeler, B. (2013). gstat: Spatial and spatio-temporal geostatistical modelling, prediction and simulation. R package version, 1-0.

285. Pei, X., Wong, S. C., & Sze, N. N. (2012). The roles of exposure and speed in road safety analysis. Accident Analysis & Prevention, 48, 464-471.

286. Petraki, V., Ziakopoulos, A., Yannis, G. (2020 – in press). “Combined impact of road and traffic characteristics on driver behavior using smartphone sensor data.” Accident Analysis and Prevention

287. Petridou, E. T., Yannis, G., Terzidis, A., Dessypris, N., Germeni, E., Evgenikos, P., ... & Skalkidis, I. (2009). Linking emergency medical department and road traffic police casualty data: a tool in assessing the burden of injuries in less resourced countries. Traffic injury prevention, 10(1), 37-43.

288. Petzoldt, T., Bär, N., Ihle, C., & Krems, J. F. (2011). Learning effects in the lane change task (LCT)—Evidence from two experimental studies. Transportation research part F: traffic psychology and behavior, 14(1), 1-12.

289. Pirdavani, A., Bellemans, T., Brijs, T., & Wets, G. (2014a). Application of geographically weighted regression technique in spatial analysis of fatal and injury crashes. Journal of Transportation Engineering, 140(8), 04014032.

290. Pirdavani, A., Bellemans, T., Brijs, T., Kochan, B., & Wets, G. (2014b). Assessing the road safety impacts of a teleworking policy by means of geographically weighted regression method. Journal of transport geography, 39, 96-110.

291. Pirdavani, A., Brijs, T., Bellemans, T., Kochan, B., & Wets, G. (2013). Evaluating the road safety effects of a fuel cost increase measure by means of zonal crash prediction modeling. Accident Analysis & Prevention, 50, 186-195.

292. Polson, N. G., & Sokolov, V. O. (2017). Deep learning for short-term traffic flow prediction. Transportation Research Part C: Emerging Technologies, 79, 1-17.

293. 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.

294. Prato, C. G., Toledo, T., Lotan, T., & Taubman-Ben-Ari, O. (2010). Modeling the behavior of novice young drivers during the first year after licensure. Accident Analysis & Prevention, 42(2), 480-486.

295. Qin, X., Ivan, J. N., Ravishanker, N., & Liu, J. (2005). Hierarchical Bayesian estimation of safety performance functions for two-lane highways using Markov chain Monte Carlo modeling. Journal of Transportation Engineering, 131(5), 345-351.

296. Quddus, M. A. (2008). Modelling area-wide count outcomes with spatial correlation and heterogeneity: an analysis of London crash data. Accident Analysis & Prevention, 40(4), 1486-1497.

297. R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/.

298. Ratti, C. (2004). Space syntax: some inconsistencies. Environment and Planning B: Planning and Design, 31(4), 487-499.

299. Raykar, V. C., & Duraiswami, R. (2006). Fast optimal bandwidth selection for kernel density estimation. In Proceedings of the 2006 SIAM International Conference on Data Mining (pp. 524-528). Society for Industrial and Applied Mathematics.

300. Reason, J., Manstead, A., Stradling, S., Baxter, J., & Campbell, K. (1990). Errors and violations on the roads: a real distinction?. Ergonomics, 33(10-11), 1315-1332.

301. Reese, C. A., & Pash-Brimmer, A. (2009). North Central Texas pay-as-you-drive insurance pilot program. In Transportation, Land Use, Planning, and Air Quality: Selected Papers of the Transportation, Land Use, Planning, and Air Quality Conference 2009 (pp. 41-50).

302. Regan, M. A., Williamson, A., Grzebieta, R., & Tao, L. (2012). 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.

303. Region of Attica. (2012). Official Website. Available: http://www.patt.gov.gr/site/index.php?option=com_content&view=article&id=4874&Itemid=319&lang=el [Accessed 15-11-2019]

304. Regev, S., Rolison, J., Feeney, A., & Moutari, S. (2017). Driver distraction is an under-reported cause of road accidents: An examination of discrepancy between police officers' views and road accident reports. In DDI2017 e-Proceedings collection. The Fifth International Conference on Driver Distraction and Inattention.

305. Reimer, B., D'Ambrosio, L. A., Coughlin, J. F., Kafrissen, M. E., & Biederman, J. (2006). Using self-reported data to assess the validity of driving simulation data. Behavior research methods, 38(2), 314-324.

306. Reuter, H. I., Nelson, A., & Jarvis, A. (2007). An evaluation of void-filling interpolation methods for SRTM data. International Journal of Geographical Information Science, 21(9), 983-1008.

307. Rhee, K. A., Kim, J. K., Lee, Y. I., & Ulfarsson, G. F. (2016). Spatial regression analysis of traffic crashes in Seoul. Accident Analysis & Prevention, 91, 190-199.

308. Rosenbaum, P. R. (2010). Design of observational studies (Vol. 10). New York: Springer.

309. Rosenbloom, T., & Eldror, E. (2014). Effectiveness evaluation of simulative workshops for newly licensed drivers. Accident Analysis & Prevention, 63, 30-36.

310. Roshandel, S., Zheng, Z., & Washington, S. (2015). Impact of real-time traffic characteristics on freeway crash occurrence: Systematic review and meta-analysis. Accident Analysis & Prevention, 79, 198-211.

311. 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.

312. Sagberg, F., Backer-Grøndahl, A. (2010). A catalogue of applications and research topics for future naturalistic driving studies. PROLOGUE Deliverable D1.3. Oslo, Norway: TØI Institute of Transport Economics.

313. Sanocki, T., Islam, M., Doyon, J. K., & Lee, C. (2015). Rapid scene perception with tragic consequences: Observers miss perceiving vulnerable road users, especially in crowded traffic scenes. Attention, Perception, & Psychophysics, 77(4), 1252-1262.

314. Saunier, N., & Sayed, T. (2007). Automated analysis of road safety with video data. Transportation Research Record: Journal of the Transportation Research Board, (2019), 57-64.

315. Schleinitz, K., Petzoldt, T., Krems, J. F., & Gehlert, T. (2016). The influence of speed, cyclists' age, pedaling frequency, and observer age on observers' time to arrival judgments of approaching bicycles and e-bikes. Accident Analysis & Prevention, 92, 113-121.

316. Schratz, P., Muenchow, J., Iturritxa, E., Richter, J., & Brenning, A. (2018). Performance evaluation and hyperparameter tuning of statistical and machine-learning models using spatial data. arXiv preprint arXiv:1803.11266.

317. Shi, X., Wong, Y. D., Li, M. Z. F., Palanisamy, C., & Chai, C. (2019). A feature learning approach based on XGBoost for driving assessment and risk prediction. Accident Analysis & Prevention, 129, 170-179.

318. Shichrur, R., Sarid, A., & Ratzon, N. Z. (2014). Determining the sampling time frame for in-vehicle data recorder measurement in assessing drivers. Transportation research part C: emerging technologies, 42, 99-106.

319. Shinar, D., Tractinsky, N., & Compton, R. (2005). Effects of practice, age, and task demands, on interference from a phone task while driving. Accident Analysis & Prevention, 37(2), 315-326.

320. Siddiqui, C., Abdel-Aty, M., & Choi, K. (2012). Macroscopic spatial analysis of pedestrian and bicycle crashes. Accident Analysis & Prevention, 45, 382-391.

321. Siddiqui, C., & Abdel-Aty, M. (2012). Nature of modeling boundary pedestrian crashes at zones. Transportation Research Record, 2299(1), 31-40.

322. Simmons, S. M., Hicks, A., & Caird, J. K. (2016). Safety-critical event risk associated with cell phone tasks as measured in naturalistic driving studies: A systematic review and meta-analysis. Accident Analysis & Prevention, 87, 161-169.

323. Şimşekoğlu, Ö., & Lajunen, T. (2009). Relationship of seat belt use to health and driver behaviors. Transportation research part F: traffic psychology and behaviour, 12(3), 235-241.

324. Singh, S. (2018). Critical reasons for crashes investigated in the National Motor Vehicle Crash Causation Survey. (Traffic Safety Facts Crash Stats. Report No. DOT HS 812 506). Washington, DC: National Highway Traffic Safety Administration.

325. Singh, S. (2015). Critical reasons for crashes investigated in the National Motor Vehicle Crash Causation Survey. NHTSA Report (No. DOT HS 812 115).

326. Smith, T. E. (2009). Estimation bias in spatial models with strongly connected weight matrices. Geographical Analysis, 41(3), 307-332.

327. Soltani, A., & Askari, S. (2017). Exploring spatial autocorrelation of traffic crashes based on severity. Injury, 48(3), 637-647.

328. Song, J. J., Ghosh, M., Miaou, S., & Mallick, B. (2006). Bayesian multivariate spatial models for roadway traffic crash mapping. Journal of multivariate analysis, 97(1), 246-273.

329. Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. Journal of the royal statistical society: Series b (statistical methodology), 64(4), 583-639.

330. SRTM (Shuttle Radar Topography Mission) (2019). SRTM 90m DEM Digital Elevation Database. Official Website. Available: http://srtm.csi.cgiar.org/ [Accessed 13-11-2019]

331. St-Aubin, P., Saunier, N., & Miranda-Moreno, L. (2015). Large-scale automated proactive road safety analysis using video data. Transportation Research Part C: Emerging Technologies, 58, 363-379.

332. Stanojević, P., Lajunen, T., Jovanović, D., Sârbescu, P., & Kostadinov, S. (2018). The driver behaviour questionnaire in south-east europe countries: Bulgaria, romania and serbia. Transportation research part F: traffic psychology and behaviour, 53, 24-33.

333. Stavrakaki, A. M., Tselentis, D. I., Barmpounakis, E. N., Vlahogianni, E. I., & Yannis, G. (2019). How much driving data do we need to assess driver behavior? (No. 19-02956).

334. 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.

335. Stern, H., & Cressie, N. A. (1999). Inference for extremes in disease mapping.

336. Tango, T. (1995). A class of tests for detecting 'general' and 'focused' clustering of rare diseases. Statistics in Medicine, 14(21-22), 2323-2334.

337. Tasic, I., Elvik, R., & Brewer, S. (2017). Exploring the safety in numbers effect for vulnerable road users on a macroscopic scale. Accident Analysis & Prevention, 109, 36-46.

338. Taubman–Ben-Ari, O., Kaplan, S., Lotan, T., & Prato, C. G. (2016). The combined contribution of personality, family traits, and reckless driving intentions to young men's risky driving: What role does anger play?. Transportation research part F: traffic psychology and behavior, 42, 299-306.

339. Terrin, N., Schmid, C. H., & Lau, J. (2005). In an empirical evaluation of the funnel plot, researchers could not visually identify publication bias. Journal of clinical epidemiology, 58(9), 894-901.

340. Theofilatos, A., Ziakopoulos, A., Papadimitriou, E., & Yannis, G. (2018a). How many crashes are caused by driver interaction with passengers? A meta-analysis approach. Journal of safety research, 65, 11-20.

341. Theofilatos, A., Yannis, G., Antoniou, C., Chaziris, A., & Sermpis, D. (2018b). Time series and support vector machines to predict powered-two-wheeler accident risk and accident type propensity: A combined approach. Journal of Transportation Safety & Security, 10(5), 471-490.

342. Theofilatos, A., Ziakopoulos, A., Papadimitriou, E., Yannis, G., & Diamandouros, K. (2017a). Meta-analysis of the effect of road work zones on crash occurrence. Accident Analysis & Prevention, 108, 1-8.

343. Theofilatos, A., Yannis, G., Vlahogianni, E. I., & Golias, J. C. (2017b). Modeling the effect of traffic regimes on safety of urban arterials: The case study of Athens. Journal of traffic and transportation engineering (English edition), 4(3), 240-251.

344. Theofilatos, A. (2015). An advanced multi-faceted statistical analysis of accident probability and severity exploiting high resolution traffic and weather data (Doctoral dissertation, National Technical University of Athens).

345. Theofilatos, A., & Yannis, G. (2014). A review of the effect of traffic and weather characteristics on road safety. Accident Analysis & Prevention, 72, 244-256.

346. Thomas, I. (1996). Spatial data aggregation: exploratory analysis of road accidents. Accident Analysis & Prevention, 28(2), 251-264.

347. Tiefelsdorf, M., & Boots, B. (1997). A note on the extremities of local Moran's I_is and their impact on global Moran's I. Geographical Analysis, 29(3), 248-257.

348. Timmermann, A. (2006). Forecast combinations. Handbook of economic forecasting, 1, 135-196.

349. Ting, C. Y., Tan, N. Y. Z., Hashim, H. H., Ho, C. C., & Shabadin, A. (2020). Malaysian Road Accident Severity: Variables and Predictive Models. In Computational Science and Technology (pp. 699-708). Springer, Singapore.

350. Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. Economic geography, 46(sup1), 234-240.

351. Toledo, T., Musicant, O., & Lotan, T. (2008). In-vehicle data recorders for monitoring and feedback on drivers' behavior. Transportation Research Part C: Emerging Technologies, 16(3), 320-331.

352. Tselentis, D. I., Vlahogianni, E. I., & Yannis, G. (2019). Investigating the Temporal Evolution of Driving Safety Efficiency Using Data Collected from Smartphone Sensors (No. 19-03473).

353. Tselentis, D. I., Vlahogianni, E. I., & Yannis, G. (2018a). Comparative Evaluation of Driving Efficiency Using Smartphone Data. Presented at 97th Annual Meeting of the Transportation Research Board, Washington, D.C., 2018.

354. Tselentis D., Vlahogianni E., Yannis G., Koziris N. (2018b). "Quantifying the Need for Driving Data Collection in Driving Behavior Assessment Using Smartphone Data", Proceedings of the 7th Panhellenic Road Safety Conference, Larissa, Greece, 11-12 October 2018.

355. Tselentis D. (2018c). Doctoral dissertation, Benchmarking Driving Efficiency using Data Science Techniques applied on Large-Scale Smartphone Data, Department of Transportation Planning and Engineering, School of Civil Engineering, National Technical University of Athens, 2018.

356. 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.

357. Ukkusuri, S., Miranda-Moreno, L. F., Ramadurai, G., & Isa-Tavarez, J. (2012). The role of built environment on pedestrian crash frequency. Safety science, 50(4), 1141-1151.

358. Ukkusuri, S., Hasan, S., & Aziz, H. (2011). Random parameter model used to explain effects of built-environment characteristics on pedestrian crash frequency. Transportation Research Record: Journal of the Transportation Research Board, (2237), 98-106.

359. United Nations (2015). Resolution adopted by the General Assembly on 25 September 2015: 70/1. Transforming our world: the 2030 Agenda for Sustainable Development.

360. Van Houwelingen, H. C., Arends, L. R., & Stijnen, T. (2002). Advanced methods in meta-analysis: multivariate approach and meta-regression. Statistics in medicine, 21(4), 589-624.

361. Vardaki, S., & Karlaftis, M. G. (2011). An investigation of older driver road safety perceptions and driving performance on freeways. Advances in Transportation Studies, (Special Issue), 7-18.

362. Ver Hoef, J. M., Peterson, E. E., Hooten, M. B., Hanks, E. M., & Fortin, M. J. (2018). Spatial autoregressive models for statistical inference from ecological data. Ecological Monographs, 88(1), 36-59.

363. Victor, T., Dozza, M., Bärgman, J., Boda, C. N., Engström, J., Flannagan, C., ... & Markkula, G. (2015). Analysis of naturalistic driving study data: Safer glances, driver inattention, and crash risk (No. SHRP 2 Report S2-S08A-RW-1).

364. Victor, T. W., Harbluk, J. L., & Engström, J. A. (2005). Sensitivity of eye-movement measures to in-vehicle task difficulty. Transportation Research Part F: Traffic Psychology and Behaviour, 8(2), 167-190.

365. Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. Journal of statistical software, 36(3).

366. Vincenty, T. (1975). Direct and inverse solutions of geodesics on the ellipsoid with application of nested equations. Survey review, 23(176), 88-93.

367. Vlahogianni, E. I., & Barmpounakis, E. N. (2017). Driving analytics using smartphones: Algorithms, comparisons and challenges. Transportation Research Part C: Emerging Technologies, 79, 196-206.

368. Vlahogianni, E. I., Geroliminis, N., & Skabardonis, A. (2008). Empirical and analytical investigation of traffic flow regimes and transitions in signalized arterials. Journal of Transportation Engineering, 134(12), 512-522.

369. Vlachogiannis, D. M., Vlahogianni, E. I., & Golias, J. (2020). A Reinforcement Learning Model for Personalized Driving Policies Identification. International Journal of Transportation Science and Technology.

370. Wadley, V. G., Okonkwo, O., Crowe, M., Vance, D. E., Elgin, J. M., Ball, K. K., & Owsley, C. (2009). Mild cognitive impairment and everyday function: an investigation of driving performance. Journal of geriatric Psychiatry and Neurology, 22(2), 87-94.

371. Wakefield, J. C., Best, N. G., & Waller, L. (2000). Bayesian approaches to disease mapping. Spatial epidemiology: methods and applications, 104-107.

372. Wang, L., Abdel-Aty, M., Lee, J., & Shi, Q. (2019). Analysis of real-time crash risk for arterial ramps using traffic, geometric, trip generation, and socio-demographic predictors. Accident Analysis & Prevention, 122, 378-384.

373. Wang, Y., Veneziano, D., Russell, S., & Al-Kaisy, A. (2016a). Traffic Safety Along Tourist Routes in Rural Areas. Transportation Research Record: Journal of the Transportation Research Board, (2568), 55-63.

374. Wang, J., Kuffer, M., & Pfeffer, K. (2019). The role of spatial heterogeneity in detecting urban slums. Computers, environment and urban systems, 73, 95-107.

375. Wang, X., Yang, J., Lee, C., Ji, Z., & You, S. (2016b). Macro-level safety analysis of pedestrian crashes in Shanghai, China. Accident Analysis & Prevention, 96, 12-21.

376. Wang, J., & Huang, H. (2016). Road network safety evaluation using Bayesian hierarchical joint model. Accident Analysis & Prevention, 90, 152-158.

377. Wang, Y., & Kockelman, K. M. (2013). A Poisson-lognormal conditional-autoregressive model for multivariate spatial analysis of pedestrian crash counts across neighborhoods. Accident Analysis & Prevention, 60, 71-84.

378. Wang, C., Quddus, M. A., & Ison, S. G. (2011). Predicting accident frequency at their severity levels and its application in site ranking using a two-stage mixed multivariate model. Accident Analysis & Prevention, 43(6), 1979-1990.

379. Wang, C., Quddus, M. A., & Ison, S. G. (2009). Impact of traffic congestion on road accidents: A spatial analysis of the M25 motorway in England. Accident Analysis & Prevention, 41(4), 798-808.

380. Wang, X., & Abdel-Aty, M. (2006). Temporal and spatial analyses of rear-end crashes at signalized intersections. Accident Analysis & Prevention, 38(6), 1137-1150.

381. Washington, S. P., Karlaftis, M. G., & Mannering, F. (2010). Statistical and econometric methods for transportation data analysis. Chapman and Hall/CRC.
382. Watson, A., Watson, B., & Vallmuur, K. (2015). Estimating under-reporting of road crash injuries to police using multiple linked data collections. Accident Analysis & Prevention, 83, 18-25.
383. Wei, F., & Lovegrove, G. (2013). An empirical tool to evaluate the safety of cyclists: Community based, macro-level collision prediction models using negative binomial regression. Accident Analysis & Prevention, 61, 129-137.
384. Weiler, J. M., Bloomfield, J. R., Woodworth, G. G., Grant, A. R., Layton, T. A., Brown, T. L., ... & Watson, G. S. (2000). Effects of fexofenadine, diphenhydramine, and alcohol on driving performance: a randomized, placebo-controlled trial in the Iowa driving simulator. Annals of Internal Medicine, 132(5), 354-363.
385. Wen, H., Zhang, X., Zeng, Q., Lee, J., & Yuan, Q. (2019). Investigating spatial autocorrelation and spillover effects in freeway crash-frequency data. International journal of environmental research and public health, 16(2), 219.
386. Wier, M., Weintraub, J., Humphreys, E. H., Seto, E., & Bhatia, R. (2009). An area-level model of vehicle-pedestrian injury collisions with implications for land use and transportation planning. Accident Analysis & Prevention, 41(1), 137-145.
387. Winner, H., Hakuli, S., Lotz, F., & Singer, C. (Eds.). (2016). Handbook of driver assistance systems: Basic information, components and systems for active safety and comfort. Springer International Publishing.
388. World Health Organization – WHO. (2018). Global status report on road safety 2018. Available from:
https://www.who.int/violence_injury_prevention/road_safety_status/2018/en/
389. World Health Organization – WHO. (2015). Global status report on road safety 2015. Available from:
http://www.who.int/violence_injury_prevention/road_safety_status/2015/en/
390. XGBoost developer team (2019). XGBoost Documentation. Official website. Available from: https://xgboost.readthedocs.io/en/latest/index.html [Accessed 07/11/2019].
391. Xia, D., Wang, B., Li, H., Li, Y., & Zhang, Z. (2016). A distributed spatial–temporal weighted model on MapReduce for short-term traffic flow forecasting. Neurocomputing, 179, 246-263.
392. Xie, K., Ozbay, K., Kurkcu, A., & Yang, H. (2017). Analysis of traffic crashes involving pedestrians using big data: Investigation of contributing factors and identification of hotspots. Risk analysis, 37(8), 1459-1476.
393. Xie, K., Wang, X., Ozbay, K., & Yang, H. (2014). Crash frequency modeling for signalized intersections in a high-density urban road network. Analytic methods in accident research, 2, 39-51.
394. Xie, K., Wang, X., Huang, H., & Chen, X. (2013). Corridor-level signalized intersection safety analysis in Shanghai, China using Bayesian hierarchical models. Accident Analysis & Prevention, 50, 25-33.
395. Xie, Z., & Yan, J. (2008). Kernel density estimation of traffic accidents in a network space. Computers, environment and urban systems, 32(5), 396-406.
396. Xu, P., Huang, H., & Dong, N. (2018). The modifiable areal unit problem in traffic safety: basic issue, potential solutions and future research. Journal of traffic and transportation engineering (English edition), 5(1), 73-82.
397. Xu, P., Huang, H., Dong, N., & Wong, S. C. (2017a). Revisiting crash spatial heterogeneity: a Bayesian spatially varying coefficients approach. Accident Analysis & Prevention, 98, 330-337.
398. Xu, C., Li, H., Zhao, J., Chen, J., & Wang, W. (2017b). Investigating the relationship between jobs-housing balance and traffic safety. Accident Analysis & Prevention, 107, 126-136.

399. Xu, P., & Huang, H. (2015). Modeling crash spatial heterogeneity: random parameter versus geographically weighting. Accident Analysis & Prevention, 75, 16-25.

400. Yamakado, M., Takahashi, E. J., Saito, S., Abe, M., & Eng, D. (2009). G-vectoring new vehicle dynamics control technology for safe driving. Hitachi Review, 58(7), 347.

401. Yang, C. Y., & Morton, T. (2012). Trends of Transportation Simulation and Modeling Based on a Selection of Exploratory Advanced Research Projects: Workshop Summary Report (No. FHWA-HRT-12-040).

402. Yannis, G., Tselentis, D. I., Vlahogianni, E. I., & Argyropoulou, A. (2017a). Monitoring distraction through smartphone naturalistic driving experiment. In: 6th International Naturalistic Driving Research Symposium, The Hague, Netherlands, 7-9 June 2017.

403. Yannis, G., Theofilatos, A., & Pispiringos, G. (2017b). Investigation of road accident severity per vehicle type. Transport research procedia, 25, 2076-2083.

404. Yannis, G., Laiou, A., Papantoniou, P., & Gkartzonikas, C. (2016). Simulation of texting impact on young drivers' behavior and safety on motorways. Transportation research part F: traffic psychology and behavior, 41, 10-18.

405. Yannis, G., Laiou, A., Vardaki, S., Papadimitriou, E., Dragomanovits, A., & Kanellaidis, G. (2015). An analysis of mobile phone use by car drivers in Greece. In Proceedings of the Institution of Civil Engineers-Transport (Vol. 168, No. 2, pp. 161-171). Thomas Telford Ltd.

406. Yannis G., Theofilatos. A., Ziakopoulos. A., Chaziris. A. (2014). "Investigation of road accident severity and likelihood in urban areas with real-time traffic data." Traffic Engineering & Control, 55 (1), 31-35.

407. Yannis, G., Kondyli, A., & Mitzalis, N. (2013). Effect of lighting on frequency and severity of road accidents. In Proceedings of the Institution of Civil Engineers-Transport (Vol. 166, No. 5, pp. 271-281). Thomas Telford Ltd.

408. 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.

409. Yannis, G., Papadimitriou, E., & Antoniou, C. (2007). Multilevel modelling for the regional effect of enforcement on road accidents. Accident Analysis & Prevention, 39(4), 818-825.

410. Yasmin, S., & Eluru, N. (2016). Latent segmentation based count models: analysis of bicycle safety in Montreal and Toronto. Accident Analysis & Prevention, 95, 157-171.

411. Young, K., Regan, M., & Hammer, M. (2007). Driver distraction: A review of the literature. Distracted driving, 2007, 379-405.

412. Zaldivar, J., Calafate, C. T., Cano, J. C., & Manzoni, P. (2011). Providing accident detection in vehicular networks through OBD-II devices and Android-based smartphones. In Local Computer Networks (LCN), 2011 IEEE 36th Conference on (pp. 813-819). IEEE.

413. Zeeb, K., Buchner, A., & Schrauf, M. (2016). Is take-over time all that matters? The impact of visual-cognitive load on driver take-over quality after conditionally automated driving. Accident Analysis & Prevention, 92, 230-239.

414. Zeeb, K., Buchner, A., & Schrauf, M. (2015). What determines the take-over time? An integrated model approach of driver take-over after automated driving. Accident Analysis & Prevention, 78, 212-221.

415. Zeng, Q., & Huang, H. (2014). Bayesian spatial joint modeling of traffic crashes on an urban road network. Accident Analysis & Prevention, 67, 105-112.

416. Zhai, X., Huang, H., Xu, P., & Sze, N. N. (2019a). The influence of zonal configurations on macro-level crash modeling. Transportmetrica A: transport science, 15(2), 417-434.

417. Zhai, X., Huang, H., Sze, N. N., Song, Z., & Hon, K. K. (2019b). Diagnostic analysis of the effects of weather condition on pedestrian crash severity. Accident Analysis & Prevention, 122, 318-324.

418. Zhai, X., Huang, H., Gao, M., Dong, N., & Sze, N. N. (2018). Boundary crash data assignment in zonal safety analysis: an iterative approach based on data augmentation and Bayesian spatial model. Accident Analysis & Prevention, 121, 231-237.

419. Zhang, H., & Malczewski, J. (2019). Quality evaluation of volunteered geographic information: The case of OpenStreetMap. In Crowdsourcing: Concepts, Methodologies, Tools, and Applications (pp. 1173-1201). IGI Global.

420. Zhu, L., Guo, F., Krishnan, R., & Polak, J. W. (2018). The Use of Convolutional Neural Networks for Traffic Incident Detection at a Network Level (No. 18-00321).

421. Ziakopoulos A., Kontaxi A., Yannis G., Fortsakis P., Kontonasios K.N., Kostoulas G. (2020). Advanced driver monitoring using smartphone applications: The BeSmart project. Proceedings of the 8th Transport Research Arena TRA 2020, April 27-30, 2020, Helsinki, Finland.

422. Ziakopoulos, A., & Yannis, G. (2020). Meta-regressions of exposure parameters used in spatial road safety analyses. Advances in Transportation Studies.

423. Ziakopoulos, A., & Yannis, G. (2019). A review of spatial approaches in road safety. Accident Analysis & Prevention, 105323.

424. Ziakopoulos, A., Theofilatos, A., Papadimitriou, E., & Yannis, G. (2019). A meta-analysis of the impacts of operating in-vehicle information systems on road safety. IATSS Research.

425. Ziakopoulos, A., Theofilatos, A., Yannis, G., Margaritis, D., Thomas, P., Morris, A., Brown, L., Robibaro, M., Usami, D. S., Phan, V., Davidse, R., Buttler, I. (2018). “A preliminary analysis of in-depth accident data for powered two-wheelers and bicycles in Europe.” International Research Council on Biomechanics of Injury – IRCOBI, International Conference Proceedings, 12-14 September 2018, Athens, Greece.

426. Zohar, D., Huang, Y. H., Lee, J., & Robertson, M. (2014). A mediation model linking dispatcher leadership and work ownership with safety climate as predictors of truck driver safety performance. Accident Analysis & Prevention, 62, 17-25.