Machine Learning-based Road Crash Risk Assessment Fusing Infrastructure, Traffic and Driver Behaviour Data

Dimitrios Nikolaou

Civil Engineer NTUA
PhD Candidate
www.nrso.ntua.gr/dnikolaou
dnikolaou@mail.ntua.gr

Supervisor: George Yannis, Professor NTUA

March 2024

Introduction

- Road crashes are a **critical public health issue** with significant social and economic consequences.
  - 12th cause of death, 1st for young people aged 5-29.
  - 1.19 million road fatalities globally in 2021.
  - 20,400 in the European Union in 2023.
  - 621 in Greece in 2023 (provisional data).
- Road crashes are influenced by various parameters that can be divided into three distinct categories: (i) **road users**, (ii) **vehicles**, and (iii) **road infrastructure and environment**.
- Notably, a substantial percentage of road crashes, **up to 94%**, can be attributed to **human factors and errors**, either exclusively or partially (Singh, 2015).

Objective of the Dissertation

- Considering the **multifaceted nature** of road crashes, the main objective of this PhD Thesis is:

  to assess road crash risk by fusing infrastructure, traffic, and driving behaviour data.

- Furthermore, a critical aspect of this research entails thoroughly exploring the reliability of harsh driving behaviour events as **Surrogate Safety Measures (SSMs)** and their utilization for assessing the safety levels of road segments across various road environments where detailed road crash data are unavailable.

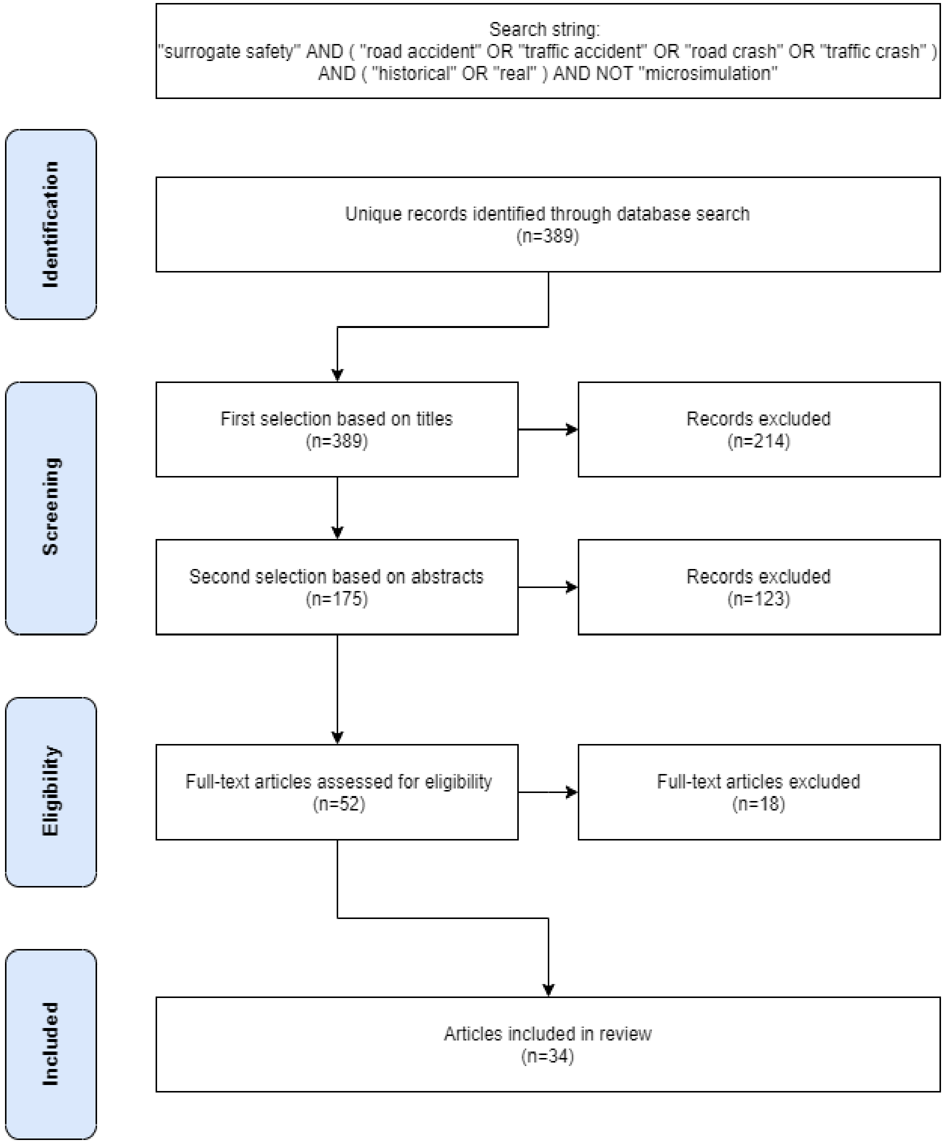
Literature Review: Road Crashes - SSMs

- Road crash data:
  - **long time period** for sufficient sample – rare events (Theofilatos et al., 2019).
  - **responsive approach** requiring good quality data (e.g. location) - not always available (Imprialou & Quddus, 2019).
  - **under-reporting** (Yannis et al., 2014; Janstrup et al., 2016).

- Surrogate Safety Measures (SSMs):
  - Metrics that are **not directly derived** from or rely on crash data (Tarko, 2018).
  - SSMs can either be an **alternative** to road safety analyses or even **complement** analyses that are based on historical crash records (Johnsson et al., 2018).
  - **Widely used SSMs**: TTC, PET, TA, DRAC, harsh brakings etc. (Bonela & Kadali, 2022).

Literature Review: Methodology

- **34 international studies** combining SSMs and historical road crash data were reviewed.
- **Driving under real road conditions** (simulation software and driving simulators are excluded).
- Results were extracted using the **PRISMA** flowchart (Moher et al., 2009).
- Search in **databases**: Scopus, TRID, Web of Science.
- Research studies written in **English** and **without restriction** on the date of publication.



Literature Review: Collection of SSMs (1/2)

- Recently, the use of **smartphone data** has begun to gain significant ground in studies featuring SSMs (e.g. Strauss et al., 2017; Paleti et al., 2017; Stipancic et al., 2019; Guo et al., 2021).
- The majority of the SSMs collected via smartphones are related to **harsh driving behaviour events**, especially harsh brakings.
- The **levels of deceleration** that define harsh braking events respectively may vary across different studies and transport modes (Kamla et al. 2019; Park et al. 2021).
  - Ranging **from 1.96m/s^2** for trucks (Blanco et al., 2011) **to as high as 8.43m/s^2** for passenger cars under dry surface conditions (Greibe, 2007).
  - Sometimes specific thresholds and calculation methods are not made public mainly due to **commercial reasons** (e.g. Guo et al., 2021; Kontaxi et al., 2021; Zhao et al., 2022).

Literature Review: Collection of SSMs (2/2)

- Naturalistic driving experiments using **instrumented vehicles** are another frequently selected option for collecting SSMs.
  - The majority of the SSMs collected through instrumented vehicles range in a similar concept to the data collected by smartphones and concern **harsh driving behaviour events** (e.g. Pande et al., 2017; Ambros et al., 2019; Kamla et al., 2019; Stipancic et al., 2021).

- The collection of traffic conflict-related SSMs under real road conditions in the majority of the examined studies is based on **video recordings** (e.g. Alhajyaseen, 2015; Zheng et al., 2019; Wang et al., 2019; Fu & Sayed, 2021).
  - In the studies reviewed, the most widely used SSMs are: TTC, PET, and DRAC.

- **Connected vehicles** are an additional emerging option for the collection of both harsh event and traffic conflict based SSMs (Xie et al., 2019; Hu et al., 2020; Yang et al., 2021).

Literature Review: Modelling Approaches

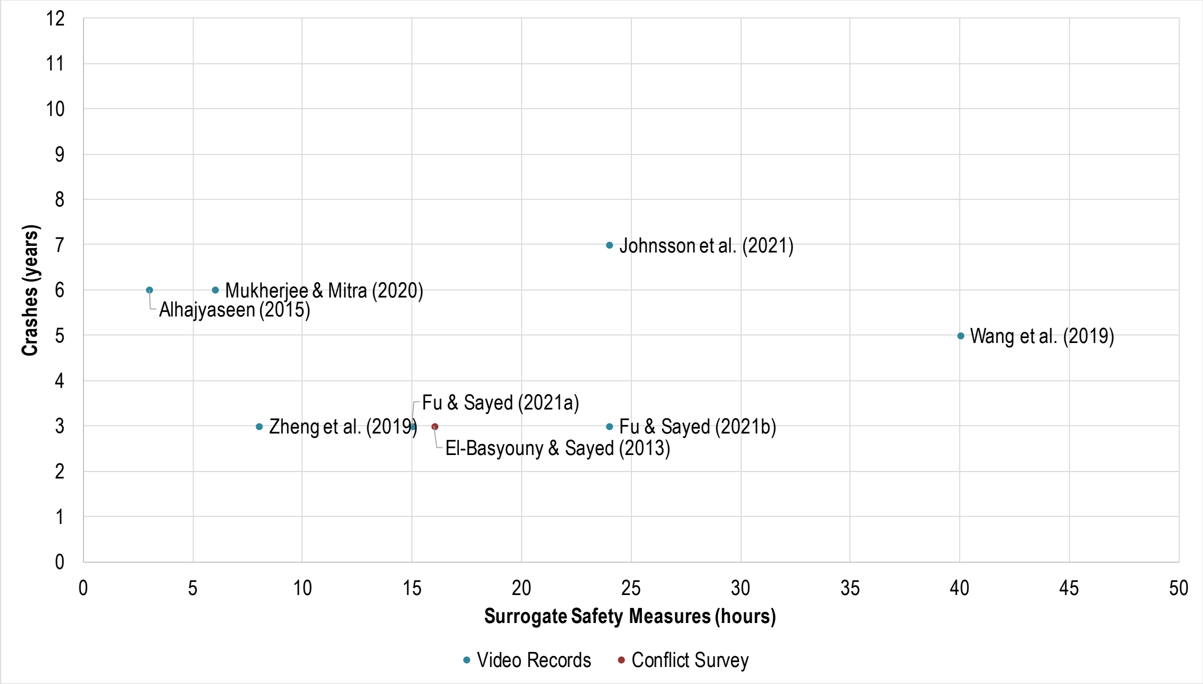
- **Correlation coefficients** of SSMs and road crashes - Pearson/Spearman (e.g. Kim et al., 2016; Strauss et al., 2017; Stipancic et al., 2018b; Xie et al., 2019)

- **Generalized Linear Models** (GLMs) – Poisson/Negative Binomial (e.g. Mukherjee & Mitra, 2020; Hunter et al., 2021; He et al., 2018; Johnsson et al., 2021)

- **Other methods:** e.g. Extreme Value Theory, Structural Equation Model, Bayesian models, etc.

- The selection of an appropriate modelling framework depends highly on the research questions being asked, the **available data** (e.g. count, rates, spatial autocorrelation etc.) and the specific context of each study.

Literature Review: Temporal Dimension



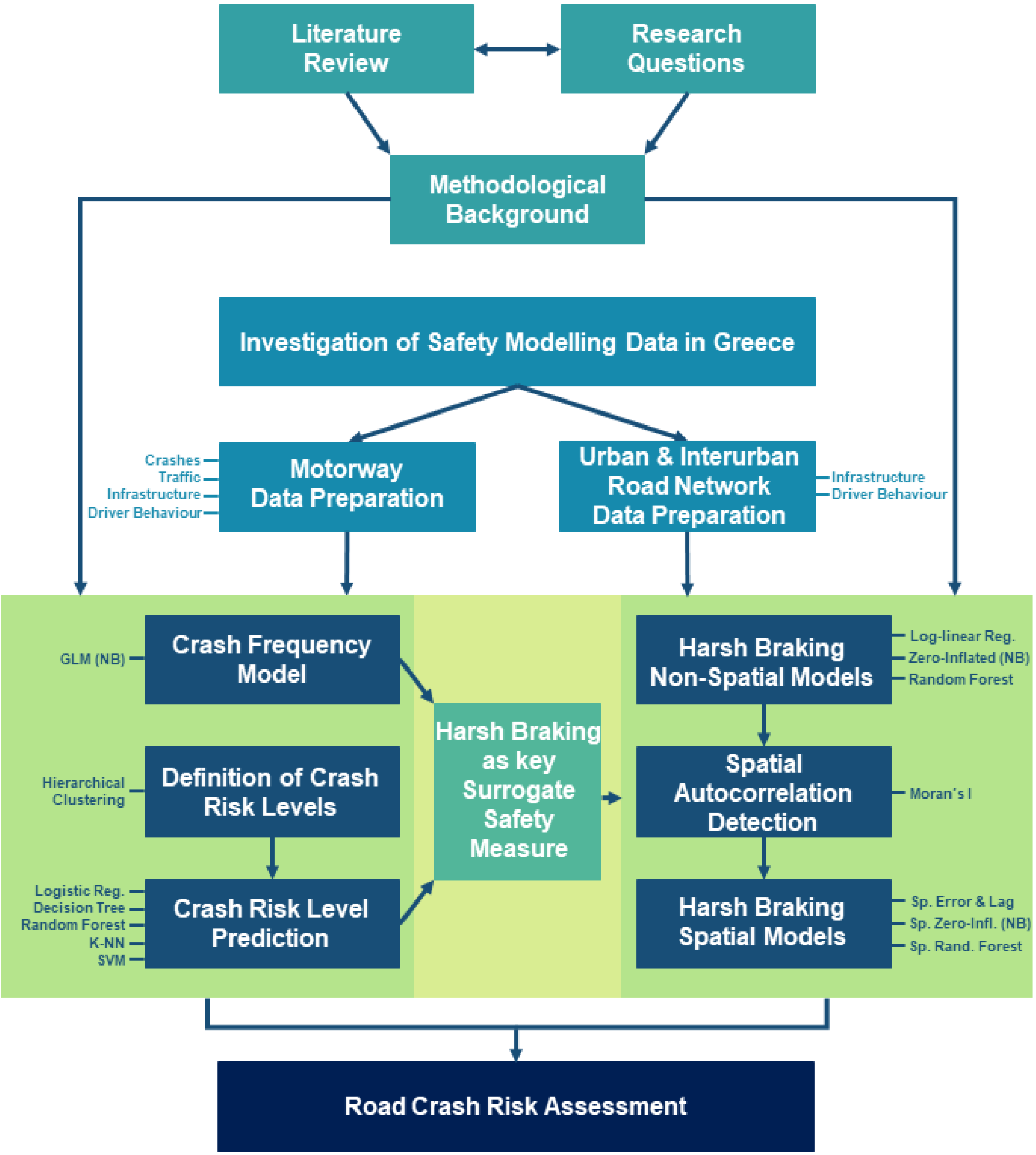
- Among all the examined studies the time period of crash data is **always greater than or equal** to the time period of collection of SSMs, highlighting the increased usability that SSMs provide.
- In the majority of the studies reviewed, the road crash data correspond on average to time periods that are **50 times longer** than the periods of collection of the SSMs.

Research Questions

1. How can **infrastructure, traffic and driver behaviour data** be fused and analyzed to derive meaningful conclusions for road crash risk assessment?
2. Can harsh driving behaviour events be meaningfully considered **reliable SSMs**?
3. Is it possible to **predict the crash risk level** of road segments by exploiting road geometry characteristics and driver-behaviour based SSMs?
4. Are **harsh braking events** more pertinent than harsh accelerations in predicting the crash risk level of road segments?
5. In the absence of highly detailed historical road crash data, **how can harsh braking events be analyzed** across various road environments?
6. Which road **infrastructure and driver behaviour** parameters exhibit a statistically significant impact on the number of harsh braking events per road segment?

Methodological Approach

- Literature Review
- Research Questions
- Methodological Background
- Investigation of Road Safety Modelling Data in Greece
- Motorway Analyses
- Urban and Interurban Road Network Analyses
- Road Crash Risk Assessment



Investigation of Road Safety Modelling Data in Greece

- Objective: to investigate the a**vailability and accuracy** of data that can be used in road crash prediction models.

- The **interurban road network** (excluding motorways) was examined for data on:
  - Road Crashes
  - Traffic
  - Geometric Design

Road Crash Data

- Official **national database**: ELSTAT
  (road crashes with at least one slight injury)
- Road crashes in the Regional Unit of **Viotia (2011-2015)**.
- In **51%** of road crashes the **road is unknown**.
- In a further **9%** (42/451) of total crashes although the road was available, the **exact station was unknown**.
- **14 rural roads** were isolated and the geo-located crashes were analyzed in order to identify whether the **infrastructure characteristics** as recorded in the crash database are identical to the actual characteristics of the site (intersection, curve – yes/no)
  - For almost half of these crashes (**46%**, 23/50) there are **obvious discrepancies.**
- Overall only **~20% of the available crash data** on interurban non-motorway roads is usable for microscopic analyses.
- **Motorway** concessionaires in Greece maintain their own databases (+crashes with material damage only).

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

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

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

Traffic Data

- In Greece, there is **no official national database** for traffic data, either traffic volumes or traffic synthesis.
- Regularly updated datasets exist only for **urban areas** (e.g., in Athens greater area) and on **toll-operated motorways**. (not openly and readily available to researchers)
- Traffic data on lower class rural roads (national and/ or regional) are usually collected on a **per-case** basis by regional road authorities, using spot traffic counts. (Viotia: 4 locations with available data for 2014)
- The **lack** of traffic data is a major **obstacle** to road safety research.

Geometric Design Data

- Examination of the road axis of **Patra-Pyrgos** National Road in the area "Vrachneika”.

- Comparison of road geometry data retrieved from **OPEN GIS sources** to the actual data as derived from a detailed **topographic survey** at scale 1: 500.

- **Small differences** (commonly less than 1m) were found in the comparison of the **horizontal alignment** → can potentially be used for road safety analyses.

- Street surface **elevations** obtained from Open GIS applications have **very large deviations** when compared to actual surveyed elevations (1m over 10m) → non accurate.

Data Collection - Motorway

- **Road Crashes**
  (injury and PDO)
  (Olympia Odos Operation SA)
- **Traffic**
  (Olympia Odos Operation SA)
- **Road geometry** characteristics
  (Open GIS, CAD, Google Earth)
- **Driver Behaviour** - SSMs
  (OSeven)

668 segments (200-600m length) of the Olympia Odos motorway.

- Average AADT (2018-2020): 10.786 vehicles/day
- Average trips per segment (6/2019-12/2020): 769
- Road Crashes (2018-2020): 80 injury & 1,270 PDO

Variable	Abbreviation	Descriptive Statistics
Number of Segment	no.	Count: 668
Direction	Direction	Frequencies: E: 337 T: 331
Segment Start (Chainage)	Seg_Start	-
Segment End (Chainage)	Seg_End	-
Number of through lanes	lanes	Frequencies: 2: 435, 3: 233
Length of motorway segment (km)	len_seg	Min.: 0.2000, Max.: 0.6000, Mean: 0.5284, Median: 0.6000
Average Annual Average Daily Traffic Volume of motorway segment (veh/day) 2018-2020	avg_AADT_18_20	Min.: 6,511, Max.: 22,079, Mean: 10,786, Median: 7,423
Posted speed limit (km/h)	speed_limit	Min.: 90.0, Max.: 130.0, Mean: 121.7, Median: 130.0
Number of Total Road Crashes (Injury & Property Damage Only) 2018-2020	TotCr18_20	Min.: 0.00, Max.: 13.00, Mean: 2.02, Median: 2.00
Number of Total Road Crashes (Injury & Property Damage Only) by segment length 2018-2020	TotCr18_20_len_seg	Min.: 0.00, Max.: 30.00, Mean: 3.88, Median: 3.33
Curve 1 - Radius R (m)	Curve1	Min.: 0, Max.: 50,000, Mean: 2,129, Median: 950
Curve 1 - Length of curve in segment (m)	Lcurve1_in_seg	Min.: 0.00, Max.: 600.00, Mean: 218.21, Median: 196.31
Lane width (m)	lane_width	Min.: 3.55, Max.: 3.95, Mean: 3.92, Median: 3.95
Paved inside shoulder width (m)	pav_ins_sh_width	Min.: 0.50, Max.: 1.75, Mean: 0.69, Median: 0.75
Median width (measured from near edges of traveled way in both directions) (m)	median_width	Min.: 2.25, Max.: 23.50, Mean: 4.96, Median: 4.88
Distance from edge of inside shoulder to barrier face (m)	dist_edginssh_barf	Min.: 0.00, Max.: 0.75, Mean: 0.04, Median: 0.00
Paved outside shoulder width (m)	pav_out_sh_width	Min.: 0.25, Max.: 4.50, Mean: 2.77, Median: 3.00
Distance from edge of outside shoulder to barrier face (m)	dist_edgoutsh_barf	Min.: 0.00, Max.: 3.25, Mean: 0.82, Median: 0.50
Number of recorded trips	rec_trips	Min.: 173, Max.: 1,689, Mean: 769, Median: 529
Average speed (all trips) (km/h)	avg_speed	Min.: 77.0, Max.: 153.0, Mean: 115.9, Median: 118.0
Number of harsh accelerations per trips	ha_per_trips	Min.: 0.0000, Max.: 0.1614, Mean: 0.0046, Median: 0.0020
Number of harsh brakings per trips	hb_per_trips	Min.: 0.0000, Max.: 0.1172, Mean: 0.0052, Median: 0.0022
Number of speeding events per trips	speeding_per_trips	Min.: 0.03, Max.: 2.56, Mean: 0.68, Median: 0.71

Methodological Background - Motorway (1/2)

Negative Binomial Regression

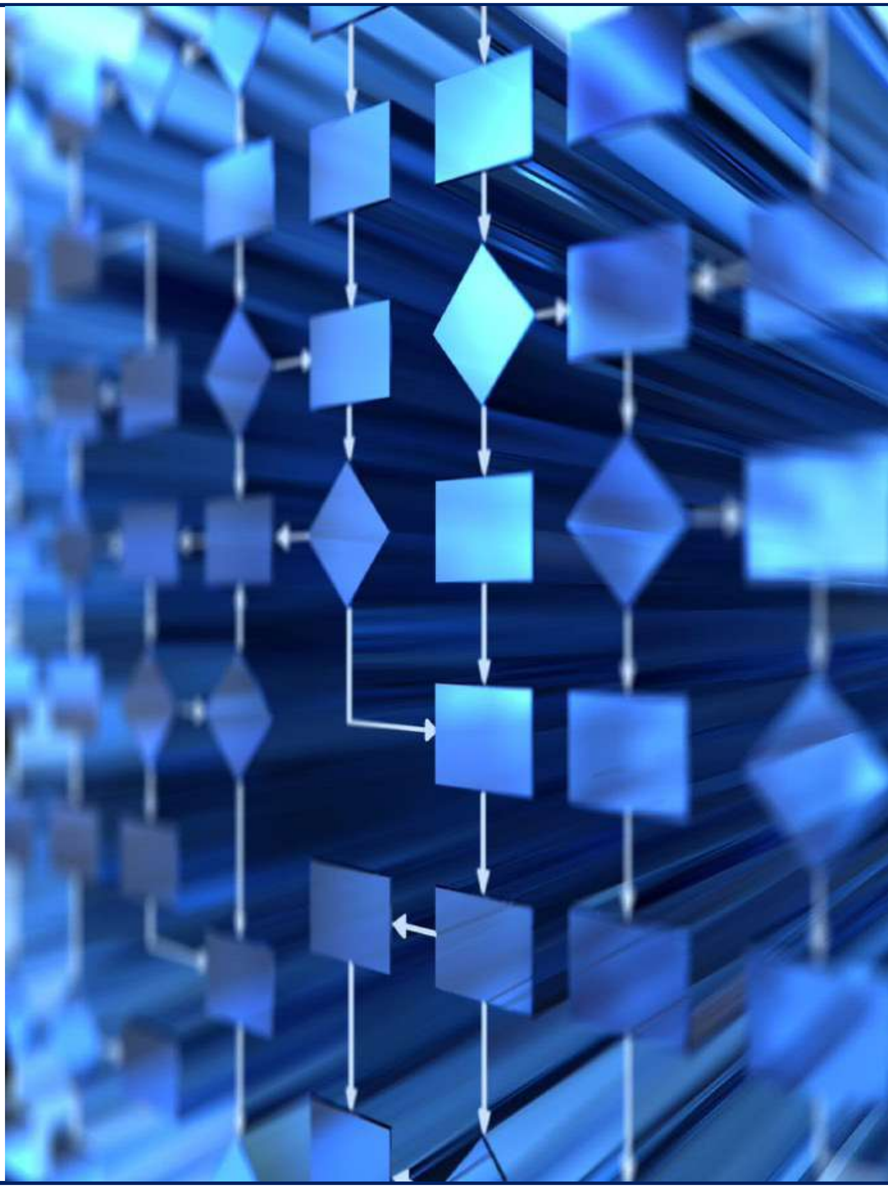
- Widely used for **count data** modelling.
- **Generalization** of Poisson regression.
- Preferred when **overdispersion** exists in crash count data.

Hierarchical Clustering

- Hierarchy of clusters based on the **agglomerative** approach.
- Each observation **starts in its own cluster** and pairs of clusters are merged as one moves up the hierarchy.
- Clusters are visually represented in a **dendrogram**.

Machine Learning Classification Algorithms

- **Logistic Regression**: linear classification model employing the logistic function.
- **Decision Tree**: non-parametric model with hierarchical structure (nodes - dataset features, branches - possible values, leaves - classification labels).
- **Random Forest**: ensemble learning technique with independent decision trees. DTs' outcomes are combined (majority vote or a vote of confidence).
- **Support Vector Machine**: finds the solution hyperplane for maximal separation of classes in high-dimensional feature space.
- **K-NN**: simple classifier based on the labels of K nearest neighbors.



Methodological Background - Motorway (2/2)

Classification Performance Metrics

- **Accuracy** (fraction of predictions that are correctly classified) ➔ (TP + TN)/P + N
- **Precision** (fraction of correct predictions for a certain class) ➔ TP/(TP + FP)
- **Recall** (fraction of instances of a class that were correctly predicted) ➔ TP/(TP + FN)
- **F1-Score** (harmonic mean of Precision and Recall) ➔ 2 * (Precision * Recall)/(Precision + Recall)
- **Macro-averaged**: Precision, Recall, F1-Score

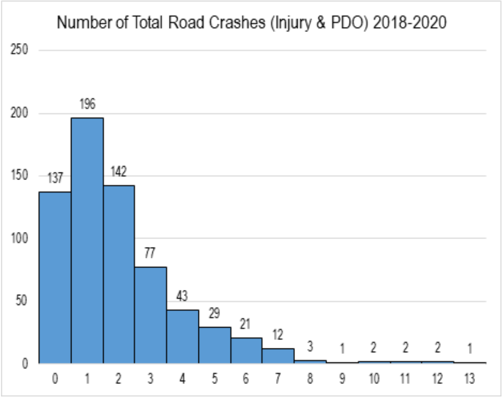
SHAP values

- **Model-agnostic method** drawing from coalitional game theory.
- Provide a **measure of contribution** of each feature to the prediction of a particular instance in a model.
- Defined as the **difference** between the expected model output and the output when that feature is excluded.

Crash Frequency Model - Motorway

- Negative Binomial Regression, dependent variable: "**Number of Total Road Crashes (Injury & Property Damage Only) 2018-2020**"

Independent Variables	Estimate	Std. Error	z value	Pr(|z|)	VIF
(Intercept)	-1.091	0.193	-5.667	**<0.001**	-
Average Annual Average Daily Traffic Volume of motorway segment (2018-2020)	**$6.67 * 10^{-5}$**	0.000	12.295	**<0.001**	1.014
Number of harsh accelerations per trips	**7.604**	2.174	3.499	**<0.001**	1.058
Number of harsh brakings per trips	**10.826**	2.541	4.261	**<0.001**	1.066
Length of motorway segment	**1.671**	0.325	5.144	**<0.001**	1.012
AICc	*2333.0*				

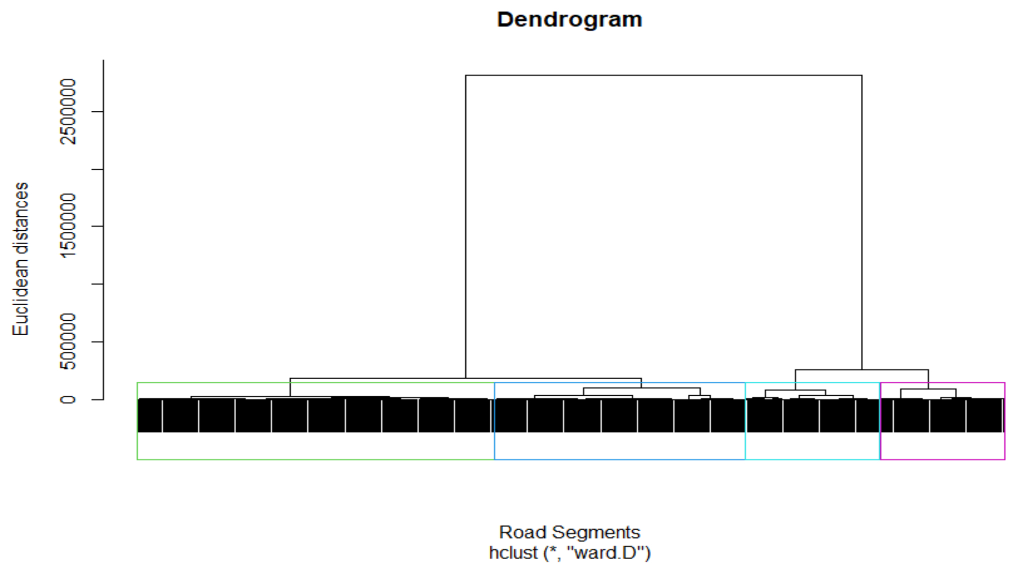


- Crash frequency is **positively correlated with the average AADT**, showing that as traffic volume increases, the number of road crashes increases as well.
- **Harsh accelerations and harsh brakings** have a **positive relationship** with the dependent variable, indicating that as the number of these two harsh driving behaviour events increases, crash frequency also increases → **harsh driving behaviour events: reliable SSMs**.
- Lastly, crash frequency is higher for motorway segments with **higher length**, as length serves as an exposure parameter.

Definition of Crash Risk Levels - Motorway

Agglomerative hierarchical clustering

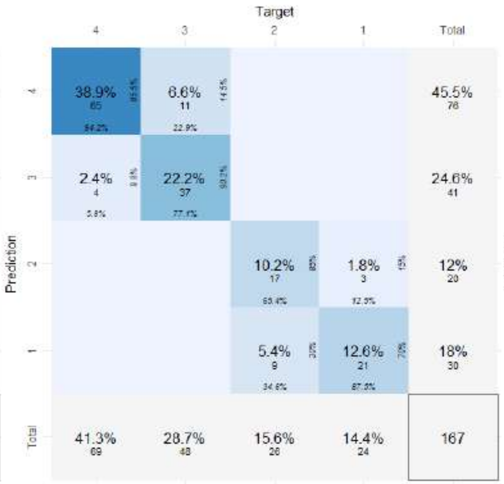
- The **Euclidean distance** between single observations of the dataset and **Ward's minimum variance** method as the linkage criterion were used.
- The variables considered for the formation of the risk level clusters of the motorway segments correspond to the number of **total road crashes by segment length** and the respective **AADT** of each segment.
- The selection of the number of clusters was based on the produced **dendrogram**.
- **Four distinct clusters** representing crash risk levels of the examined segments emerged from the hierarchical clustering procedure, ranging from more risk-prone, potentially unsafe locations to more safe locations.



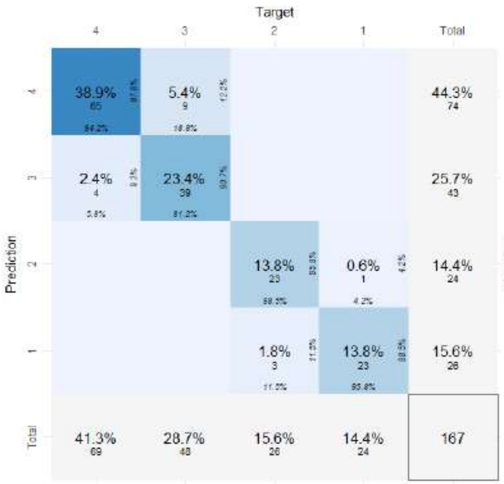
Crash Risk Level	Count of Segments	Average "TotCr18_20_len_seg"	Average "avg_AADT_18_20"
1	96	7.57	20,876
2	104	4.55	17,218
3	193	3.25	8,086
4	275	2.76	6,726
Total	**668**	**3.87**	**10,786**

Crash Risk Level Prediction - Motorway

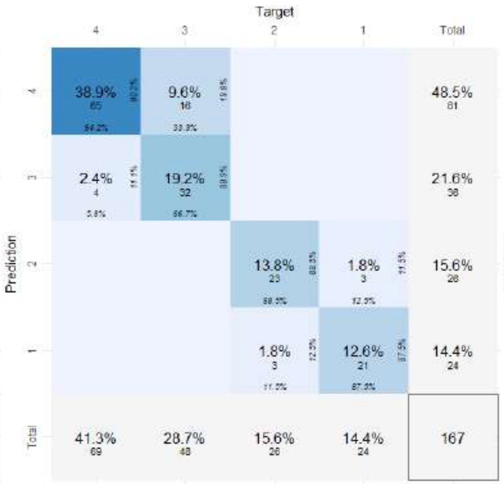
Logistic Regression (LR)



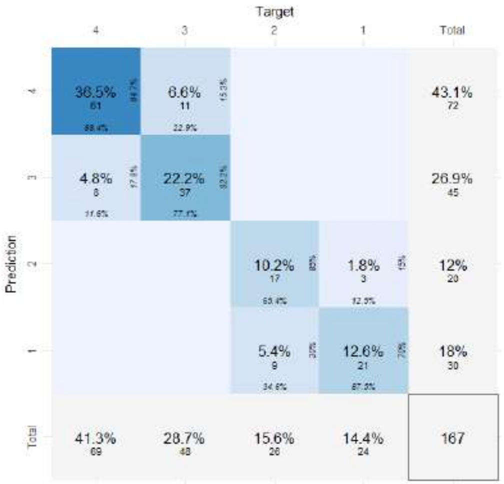
Decision Tree (DT)



Random Forest (RF)



Support Vector Machine (SVM)



K-Nearest Neighbour (K-NN)

- **Response variable**: Crash Risk Level
  Predictors: lanes, lane_width, Curve1, Lcurve1_in_seg, median_width, pav_ins_sh_width, pav_out_sh_width, dist_edginssh_barf, dist_edgoutsh_barf, speed_limit, avg_speed, speeding_per_trips, hb_per_trips, ha_per_trips
- The **training subset (75%)** was used to train the models, while the **test subset (25%)** was used to evaluate their performance.
- **Overall accuracies:** RF: 89.9%, LR: 85.1%, SVM: 84.5%, DT: 83.9%, K-NN: 81.5%.
- **RF classification model** was the best performing model, based on both the overall accuracy and the per-class metrics.

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

SHAP values - Motorway (1/2)

- SHAP values were provided for the RF model in order to deal with the difficult challenge of **interpreting** its results.
- To create a **representative instance of motorway segments**, the median values of the continuous predictors were used.
- Medians were preferred instead of the mean values, as it can be concluded that the **predictors are not normally distributed** based on the outcomes of Shapiro-Wilk normality tests, skewness and kurtosis values.

Variable	Shapiro-Wilk (p-value)	Skewness	Kurtosis	Median
Lane width (m)	<0.001	-2.42	10.48	3.95
Curve 1 - Radius R (m)	<0.001	5.74	42.56	950.00
Curve 1 - Length of curve in segment (m)	<0.001	0.49	2.27	197.65
Median width (measured from near edges of traveled way in both directions) (m)	<0.001	3.86	23.58	4.93
Paved inside shoulder width (m)	<0.001	1.64	11.43	0.75
Paved outside shoulder width (m)	<0.001	-0.85	3.68	3.00
Distance from edge of inside shoulder to barrier face (m)	<0.001	3.19	15.79	0.00
Distance from edge of outside shoulder to barrier face (m)	<0.001	0.96	3.13	0.50
Posted speed limit (km/h)	<0.001	-1.16	2.82	130.00
Average speed (all trips) (km/h)	<0.001	-1.27	6.31	118.00
Number of speeding events per trips	<0.001	0.24	2.68	0.71511
Number of harsh brakings per trips	<0.001	5.24	38.53	0.00215
Number of harsh accelerations per trips	<0.001	7.70	75.01	0.00197

SHAP values - Motorway (2/2)

- The SHAP values can be **positive** (green bars) **or negative** (red bars) for each crash risk level, depending on whether the feature has a positive or negative contribution to the prediction for that class.
- It can be observed that this representative motorway segment is **more likely to belong to the lowest crash risk level**, which corresponds to overall safer locations with lower traffic volumes and road crashes by segment length than the motorway segments between the first and the third crash risk level.
- The harsh acceleration related variable **does not make a significant contribution** to the prediction of the segment crash risk level.
- The results of this investigation suggest that **harsh brakings may be more pertinent** than harsh accelerations for predicting the crash risk level of motorway segments overall.

Data Collection - Urban & Interurban Road Network

The Region of **Eastern Macedonia and Thrace** was selected as a challenging location in terms of data availability.

Road Infrastructure (OpenStreetMap)
- Length, Curvature, Road Type
- **6103** road segments:
  (Mean Length: 288m, Total Length: 1763km)
- Road Types: (67.8% residential, 12.1% tertiary, 7.4% secondary, 3.8% motorway, 9% other)

Driver Behaviour – Telematics (OSeven)
- Harsh braking, Harsh acceleration, Speeding, Distraction
- Data from **5,129 trips** within the examined road network during 2021 were utilized.
  (mean duration: 634 sec, st.dev: 556 sec, 2889 harsh br.)

A spatial **map-matching** of the driver behaviour data and the examined road segments was carried out.

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

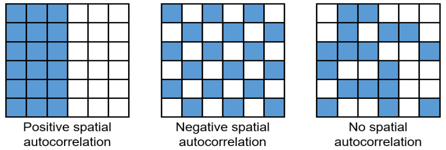
Methodological Background - Urban & Interurban Road Network

Zero-Inflated Negative Binomial Regression - ZINB

- Frequency modelling (positive integers or 0).
- **Overdispersion and excess zeros** in the dependent variable.
- Combination of Negative Binomial and Logistic Regression.

Detection of Spatial Autocorrelation

- **Moran's I** Index: calculated on a global scale [-1, 1].



Spatial Zero-Inflated Negative Binomial Regression - SZINB

- Addition of a **spatial lag** variable that essentially averages the neighbouring values of a location.
- It shows **how much a spatial feature is affected** by its neighbours.
- To address **spatial autocorrelation** in the dependent variable.

Spatial Random Forest - SRF

- **Spatial predictors** that take into account the spatial structure of the training data, minimizing the spatial autocorrelation of residuals and providing accurate variable significance scores.
- Adding the columns of the **distance matrix** of the considered road segments as explanatory variables (Hengl et al., 2018).

ZINB Model - Urban & Interurban Road Network

Dependent variable: **Number of harsh braking events →** Moran's I positive (0.0263) and statistically significant (p-value < 0.001)

1st part – Frequency

- **Segment length** and **number of trips** present a positive correlation with harsh brakings as they can be considered as exposure indicators.
- **Speeding** and **mobile phone use** are positively correlated with harsh braking events.
  Speeding → harsh brakings for collision avoidance or speed reduction
  Mobile phone use → distraction, reduction in reaction time, harsh braking
- Fewer harsh braking events on **motorways** compared to other road types.
  Motorway → smoother traffic flow, more lane options, better visibility
- **Spatial lag** term positive and statistically significant→ positive spatial autocorrelation

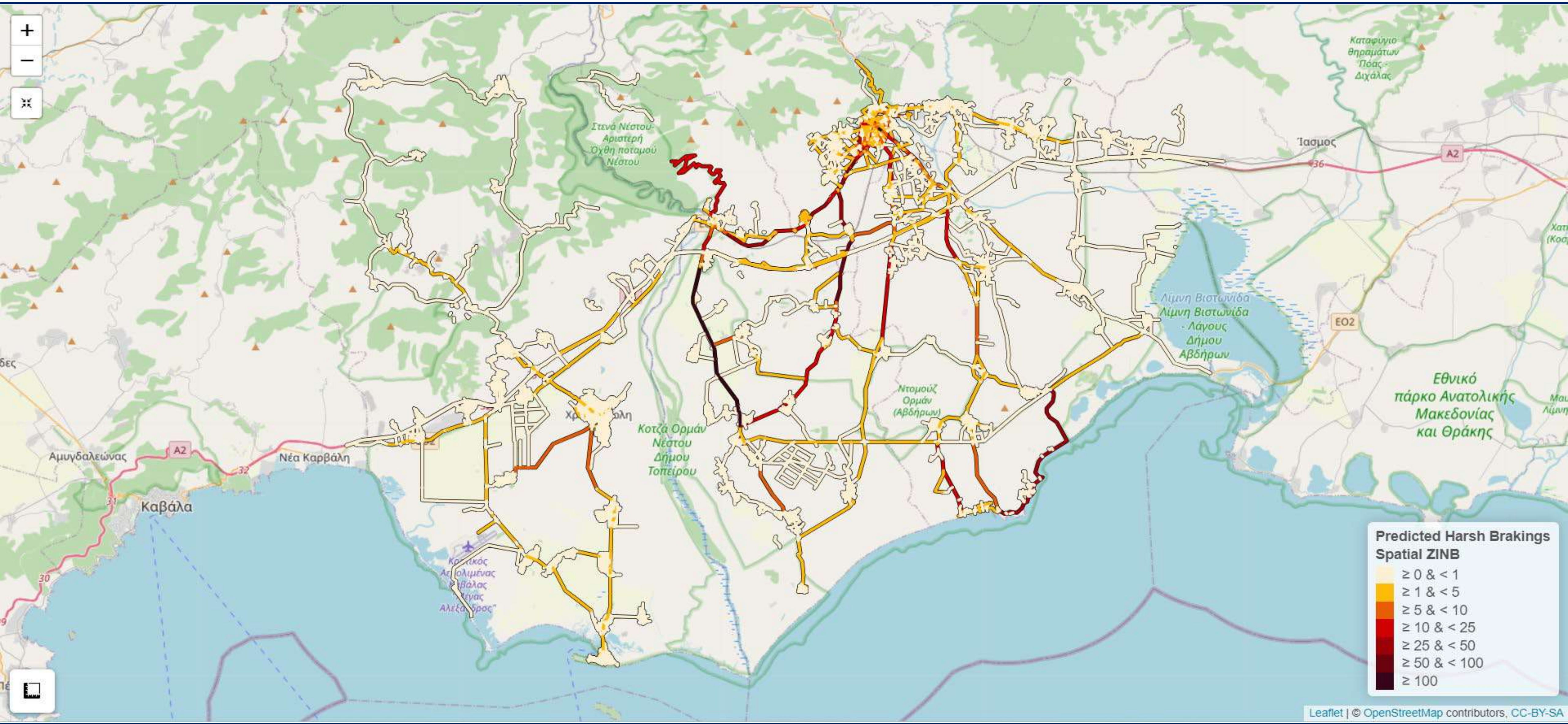
	Zero-Inflated Negative Binomial (ZINB)				Spatial Zero-Inflated Negative Binomial (SZINB)			
Count model coefficients (negbin with log link):								
Independent variables	Estimate	Std. Error	z value	Pr(>|z|)	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.527	0.112	-13.605	**<0.001**	-1.591	0.113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

2nd part - Possibility

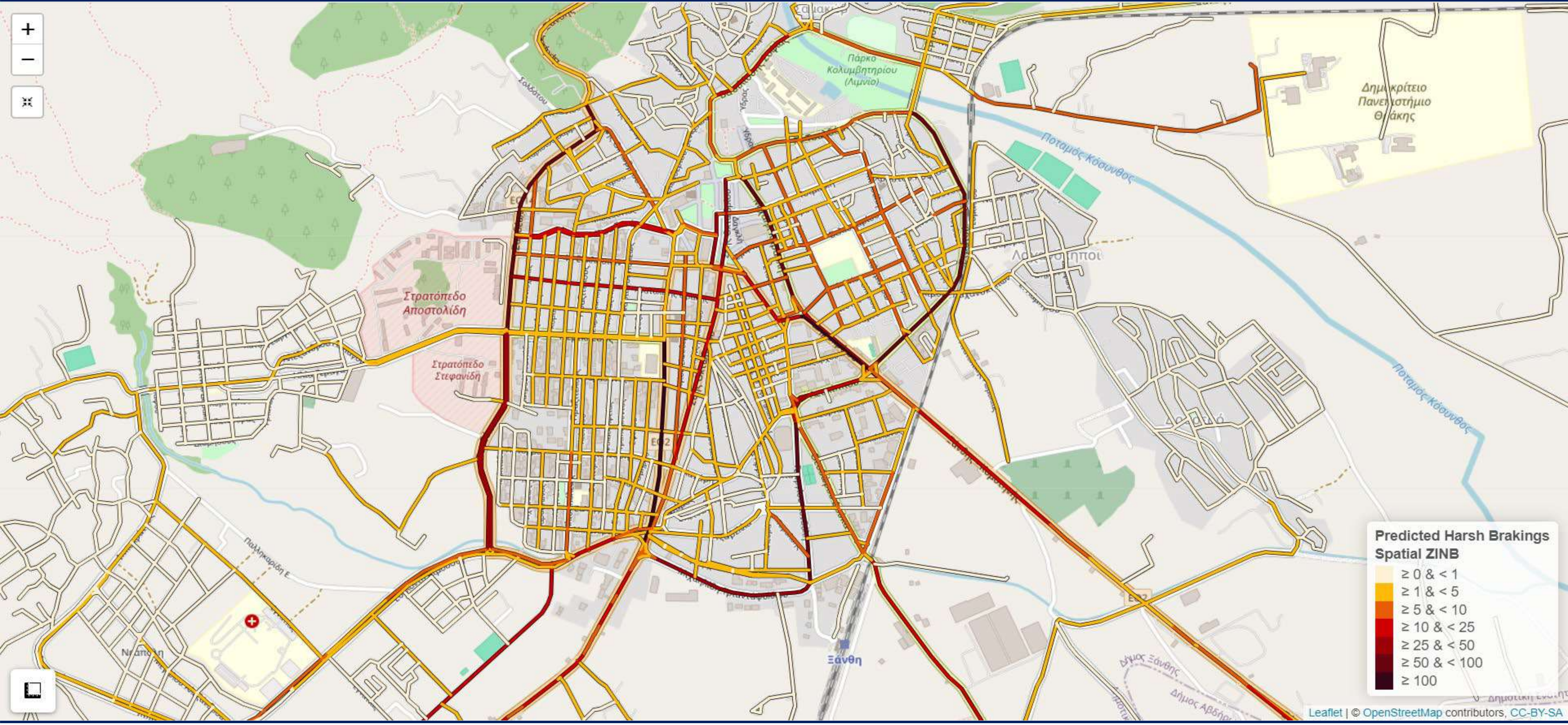
- The increase in the **number of trips** and **mobile phone use** lead to a reduced probability of zero harsh braking events on the examined road segments considered.

Comparison of Models: **Spatial model** demonstrated better fit than the non-spatial model, as shown by the lower AIC values.

Visualization of the SZINB Results



Visualization of the SZINB Results (zoomed-in view)



Random Forest - Urban & Interurban Road Network (1/3)

Response variable: **Harsh braking events**
"log (harsh_braking_count + 1)"

Predictors: number of trips, length, mobile phone use, speeding, linearity index "efficiency", motorway

Non-spatial Random Forest

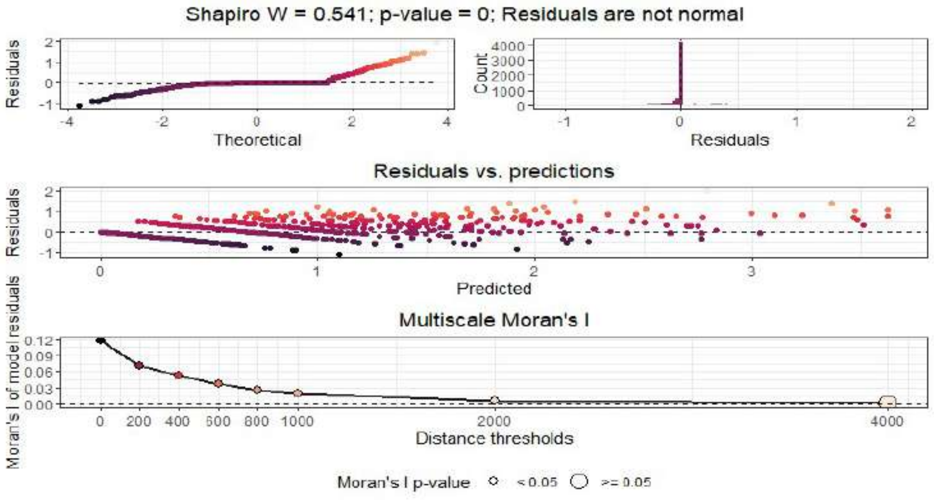
- **Positive and statistically significant** values of Moran's I index of residuals for distances 0-2000 m.

Spatial Random Forest

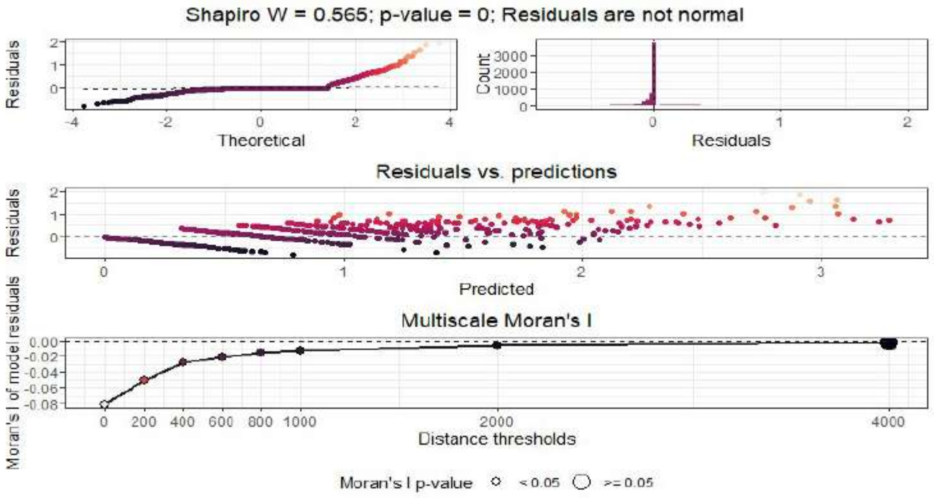
- Adding the columns of the **distance matrix** of the examined road segments as additional predictors in order to reduce the spatial autocorrelation of the residuals (Hengl et al., 2018).

- **Reduction** of the absolute values of Moran's I indices.

Non- spatial Random Forest



Spatial Random Forest



Random Forest - Urban & Interurban Road Network (2/3)

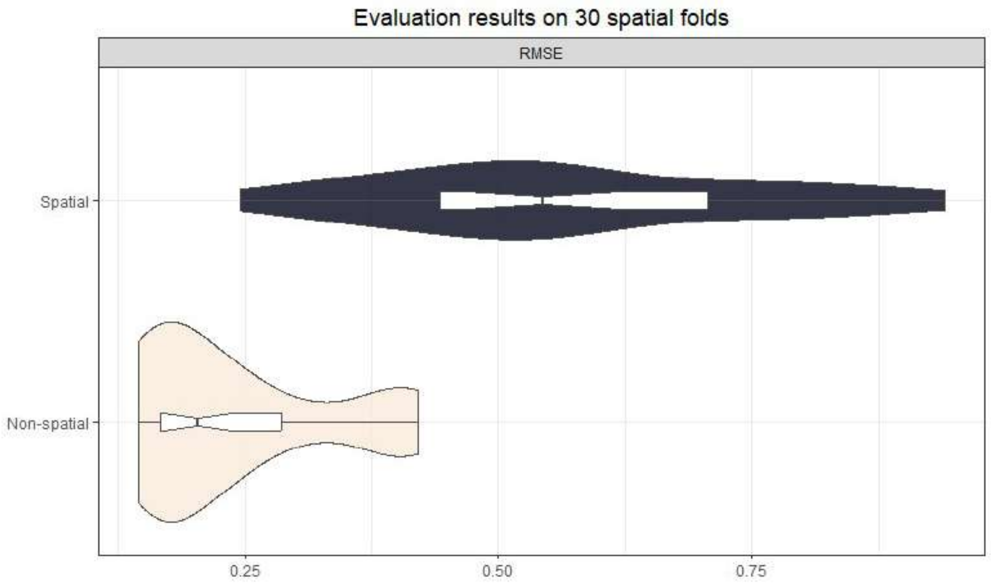
- The absolute values of the Moran's I index can provide some insight into the strength of spatial autocorrelation, but it is **not the sole criterion** for model evaluation.

- When examining **typical metrics** (not out-of-bag metrics), for instance, R2 and RMSE, it is observed that the SRF outperforms the non-spatial RF model.

- A spatial model can capture **spatial dependencies** among the considered data points leading to a better fit to the observed data compared to non-spatial model.

- However, based on the out-of-bag performance metrics, it is found that non-spatial RF model outperforms the SRF, declaring that the non-spatial model is likely performing better in terms of **generalization** on unseen data.

	Non-spatial RF	SRF
Number of trees	500	500
Sample size	6103	6103
Number of predictors	6	6109
Mtry	2	78
Minimum node size	5	5
R^2 (out-of-bag)	**0.526**	**0.440**
R^2 (cor (observed, predicted)^2)	**0.900**	**0.928**
RMSE (out-of-bag)	**0.309**	**0.336**
RMSE	**0.156**	**0.150**

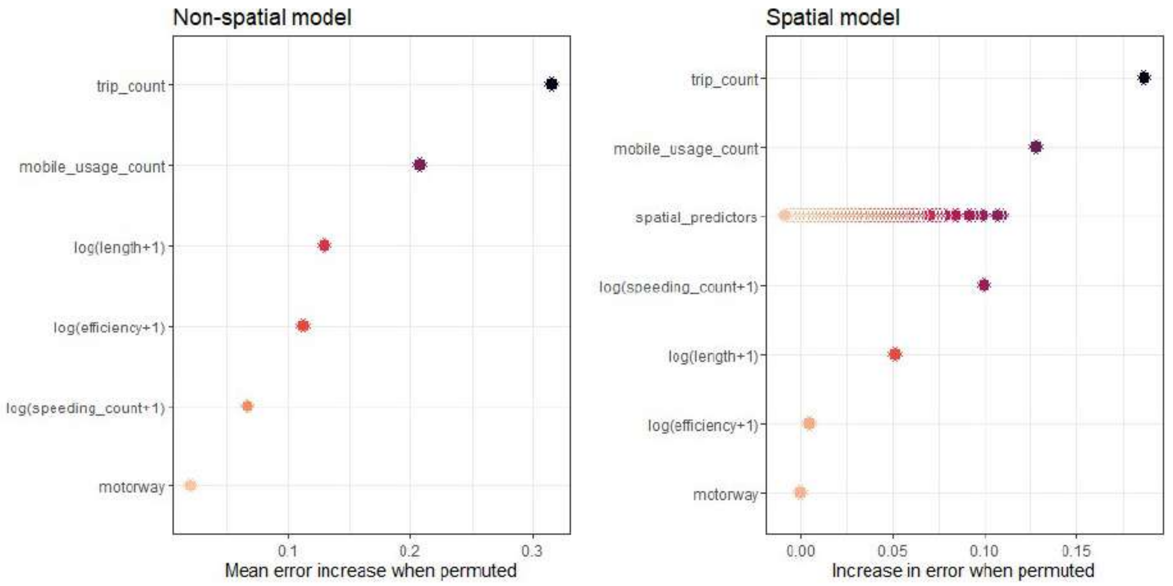


Random Forest - Urban & Interurban Road Network (3/3)

- In both RF models, the **number of trips** per examined road segment (which serves as a naturalistic driving exposure metric), was found to be the most influential predictor, highlighting its significant relevance in predicting the frequency of harsh braking events.

- On the other hand, the **motorway** variable exhibited the lowest importance in both RF model

- This finding may suggest that factors other than road type such as **driver distraction and speeding**, might play a more crucial role in influencing harsh braking events frequencies.



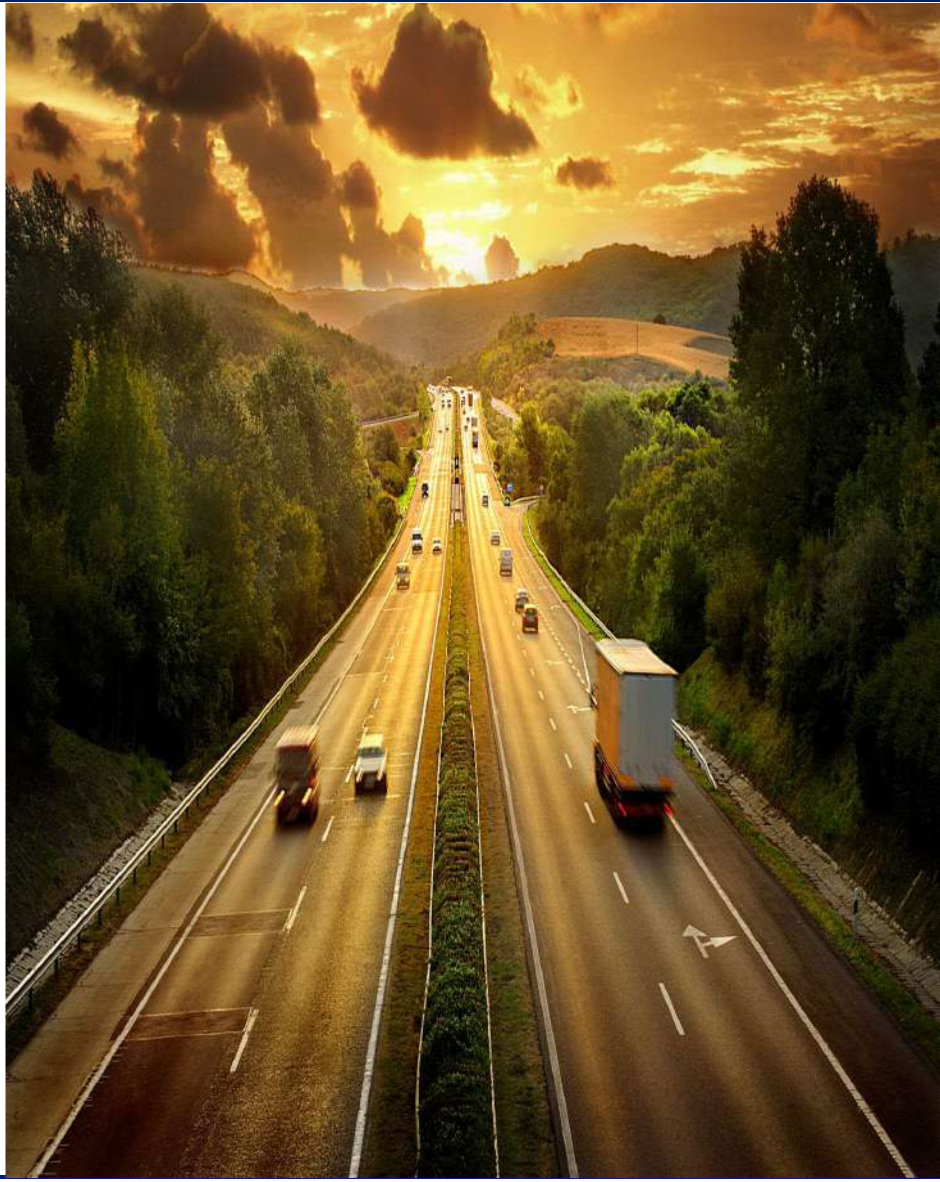
Conclusions of the Dissertation (1/2)

- The frequency of road crashes on motorway segments is **positively correlated** with the traffic volume, the length of the segment, the number of harsh accelerations and the number of harsh brakings per segment trips.

- The positive and statistically significant relationship between road crash frequency and events of harsh driving behaviour suggests that they can serve as a **valid subcategory** of naturalistic SSMs.

- The Random Forest classification model is a highly promising **proactive road safety tool**, capable of effectively identifying and prioritizing potentially hazardous motorway segments.

- **Harsh braking events** serve as a more suitable SSM than harsh accelerations in terms of crash risk level prediction.



Conclusions of the Dissertation (2/2)

- The number of harsh braking events is a SSM that can be analysed: i) either in various **proactive** road safety analyses before road crashes' occurrence ii) or in cases of unavailable detailed road crash data.
- Road segment **length** and **number of trips** were identified as proxy exposure indicators and are positively correlated with harsh brakings.
- Variables related to **speeding and mobile phone use** were also positively correlated with harsh brakings, while motorways had fewer harsh braking events compared to other road types.
- Statistically significant and positive **spatial autocorrelation** was identified in the frequencies of harsh braking events.
- Spatial models show a **better fit** to the data compared to non-spatial models; but they lack in generalization to unseen data.

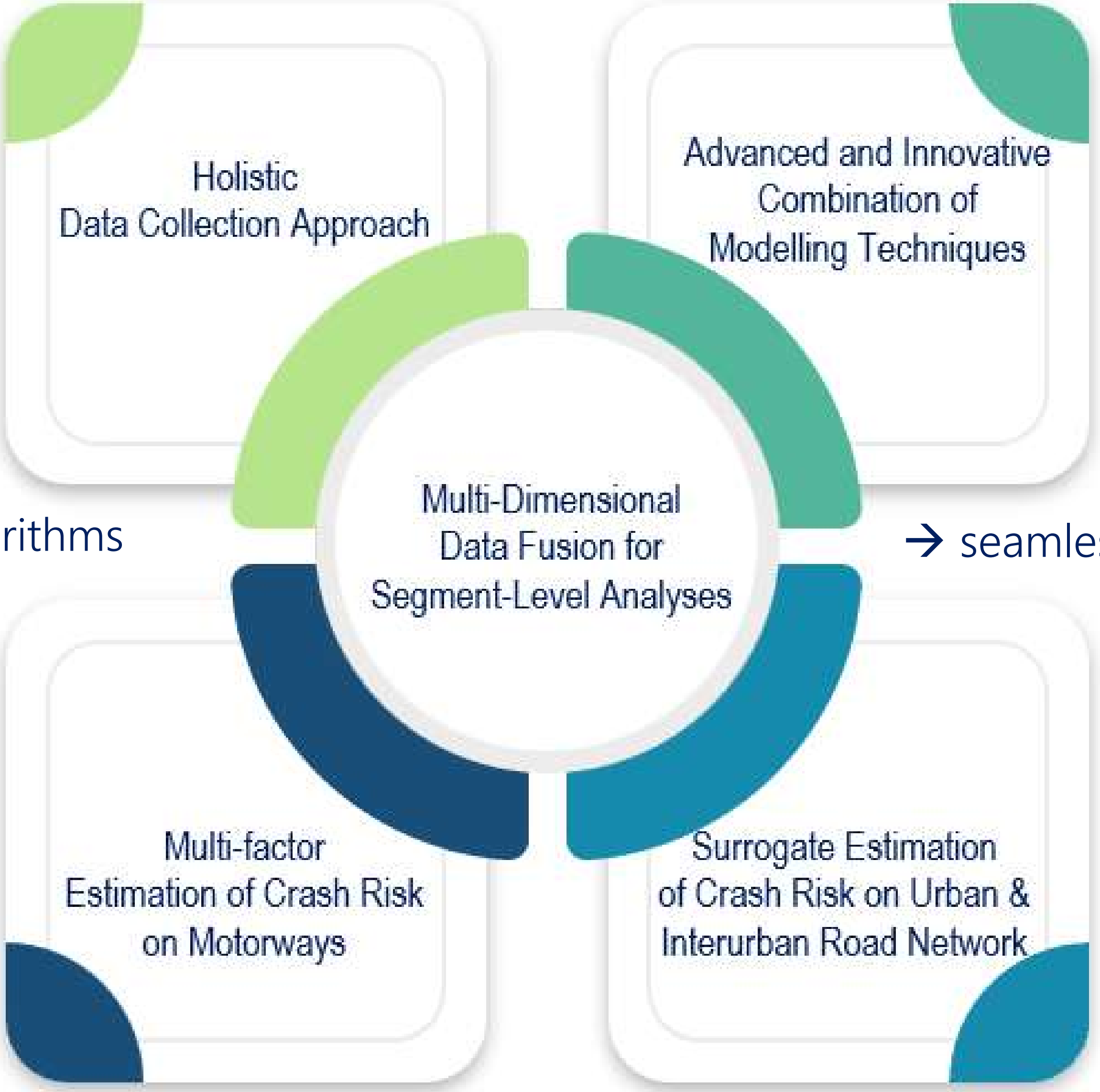
Innovative Contributions of the Dissertation

- Crashes, traffic, infrastructure, driver behaviour
- high-resolution big datasets, softwares

Series of processing algorithms

- Statistical, Spatial models, Machine Learning
- Methods being applied for the first time to harsh braking events

→ seamless transferability



- High-quality database
- Combination of crash, traffic, geometry and behaviour data

- SSMs modelling, considering spatial autocorrelation
- Development of crash risk maps

Limitations of the Dissertation

- The road **geometry characteristics** analysed are not an exact replication of the actual road design and minor differences could be expected if a comparison with the as-built drawings was made.

- The motorway segment analyses did not include toll sections and tunnels, resulting in some **discontinuities** in the research area.

- **Spatial autocorrelation** was not considered in the analyses of the motorway sections.

- **Lack of traffic data** (volumes, flow conditions) on the examined road network of the Eastern Macedonia and Thrace Region.

Further Research

- **Exploring temporal patterns** which would capture seasonal cyclical trends in both road crash and harsh braking hotspots.
- **Inclusion of additional parameters**: e.g. slopes, pavement conditions (wet/dry), presence of roadworks, weather conditions, land use etc.
- **Exploration of additional models**: e.g. Neural Networks, XGBoost etc.
- The scope of harsh braking analyses can be **expanded** by extending its application to include additional regions, potentially encompassing other countries.

Machine Learning-based Road Crash Risk Assessment Fusing Infrastructure, Traffic and Driver Behaviour Data

Dimitrios Nikolaou

Civil Engineer NTUA
PhD Candidate
www.nrso.ntua.gr/dnikolaou
dnikolaou@mail.ntua.gr

Supervisor: George Yannis, Professor NTUA

March 2024