Leveraging naturalistic Connected LCV data for spatial surrogate safety measure applications

Dr. Apostolos Ziakopoulos[1*], Nick Karahlis[2], Prof. George Yannis [1]

[1] National Technical University of Athens, Department of Transportation Planning and Engineering, 5 Iroon Polytechneiou Street, GR-15773, Athens, Greece

[2] Compass IoT Pty Limited, 608 Harris St, Ultimo 2007, NSW Australia

Email for correspondence: apziak@central.ntua.gr

1 INTRODUCTION

Road crashes remain a critical issue for all motorized societies in the developed and developing world, claiming 1.19 lives per year as well as constituting the most frequent cause of death for young age individuals (WHO, 2023). Progress to reduce crash occurrence and mitigate crash consequences when crashes do happen has been steady but slow, with reductions meeting resistance after some recent gains spurred by the recent COVID-19 pandemic and as such cannot be considered permanent (Ziakopoulos et al., 2025).

To support the continuous endeavor for fewer road crashes and road crash casualties, alternatives have emerged in the form of surrogate safety measures (SSM), which are measures that can be used to obtain road safety estimates in lieu of traditional crash and injury metrics (Tarko, 2018). SSMs may include kinematic measures, such as harsh brakings or harsh accelerations, and more position-oriented or relevant measures, such as Time-to-collision (TTC) or Post-Encroachment Time (PET). SSMs have numerous benefits, as they can be collected seamlessly and rapidly, usually in an automated manner. They are voluminous and can support road safety research without the need of excessive study periods (Nikolaou et al., 2023). In addition, SSMs can be exploited proactively, outlining the road safety profile of study areas of interest before crashes occur, and can also provide insights on problematic locations of the network (hotspots) or potential countermeasures. If crash data do exist in an overlapping space-time, SSMs can be linked with crash data in a binary fashion to model crash vs. non-crash occurrence (Wang et al., 2021). Progress on the field is also highly supported by new analytic technologies such as microsimulation advancing for road safety uses (Oikonomou et al., 2023) and driver telematics (Kontaxi et al., 2021). Nonetheless, it should also be noted that there are issues of compatibility for SSMs, and particularly discrepancies between countries and how they report or treat such data; indicatively, the reader can refer to (Laureshyn & Várhelyi, 2018).

Connected vehicles, with their namesake connectivity and a score of high-fidelity sensors, are well-placed to become a fruitful source for SSMs, complementing the aforementioned examples. Connected vehicles, and their accompanying data, have emerged as a transformative technology for road transport and infrastructure projects, offering unique opportunities for improved safety, efficiency, and resilience. By 2027, it is estimated that over 300 million connected vehicles will circulate across the globe. Such an expansive scale could present new sets of challenges and also disrupt the manner in which transport and infrastructure challenges are treated, and the speed at which they are solved.
Moreover, vehicle categories other than cars continue to be underrepresented in road safety research. Light Commercial Vehicles (LCVs) differ from cars in several noteworthy aspects, such as usage, wear effects and loading issues. On the human factors regard, LCV drivers, while professionals of various occupations, are normally not professionally trained drivers, with exceptions based on region and cargo type. Nonetheless, they also operate while encountering all the challenges of increased congestion and exposure as normal cars do, for longer hours, and with vehicles of the aforementioned different profile. Therefore, LCVs constitute a promising but unexplored research venue for SSM research, in order to open new investigation venues and to complement existing ones for cars.

Based on all the aforementioned, the aim of the present paper is to showcase both the contents and the findings of a large dataset of naturalistic connected vehicle data and to examine its potential for road safety analysis. Specifically, a large dataset of LCVs is presented, collected from the Greater London Area in the UK and their acceleration values are analysed, together with map visualizations which can support policymakers and stakeholders. The data is contrasted with LCV crashes, in an overlapping study area for the same period. Several diagnostic spatial analysis techniques are explored, such as cross-Nearest-Neighbour analysis for concentration comparison of two different datapoint patterns and Ripley's Cross K12 function, quantifying the spatial relation of the two different spatial datapoint sets.

2 THE COMPASS DATASET

The Connected Vehicle dataset utilized within the present research is provided by Compass IoT (https://www.compassiot.com.au/), which is a Road Intelligence company providing connected vehicle data for road,

traffic and transport applications through platforms (road intelligence) and custom data (Data Science outputs) for road and transport related projects in Australia and globally. The Compass connected vehicle data informs various related stakeholders and authorities, as well as to and support road safety research. Indicatively, for the twelve months up to November 2024 inclusive, Compass supported transport and road authorities, such as Transport for New South Wales (TfNSW) with more than 650 users on its platform, downloading more than 730 terabytes of road data through more than 21,500 dedicated queries. Similarly, road authorities at state, local and national level are using connected vehicle data technology to obtain insights from data originating directly from the vehicles and not from any external devices.

Within the Greater London Area in the UK, a 6-month period (June to December 2024) was isolated within the Compass dataset. The collected data included vehicle ID, timestamp, position (latitude, longitude), speed and acceleration of the connected vehicles in the axes of ground movement (namely x-axis and its perpendicular, y-axis), in other words, most simple trajectory features of the examined LCVs. Data was collected across a study period of 6 months, between June and December of 2024 in particular. It is worth noting that Compass data circumvents the typical black box characteristics of floating car data (FCD) which are agnostic of vehicle type (Ambros & Jurewicz, 2017).

The data containing harsh brakings were isolated for further analysis, with a cutoff of absolute values of 0.2 m/s2, as harsh brakings are a more direct surrogate road safety measure indicating an urgent action taken by drivers. An added control of acceleration values being below 10 m/s2 as absolute cutoff values was implemented as well, as an additional safeguard to ensure representation of circumstances found in real networks. The harsh braking dataset comprised 57,935 individual connected LCV IDs. An indicative visualization of the recorded values of harsh braking patterns through device sensors can be found in previous works, such as that of Ehsani et al. (2017) for instance.

After these controls, the harsh deceleration dataset was mapped and visualized in R-studio (R core team, 2024). An initial visualization of observations can be seen on Figure 1, created with the OSM/R-studio interface package and JavaScript library 'leaflet' (Cheng et al., 2019), which exploits Volunteered Geographical Information (VGI) from the openly available crowdsourced databases of OpenStreetMap (Goodchild, 2008; OSM, 2019).

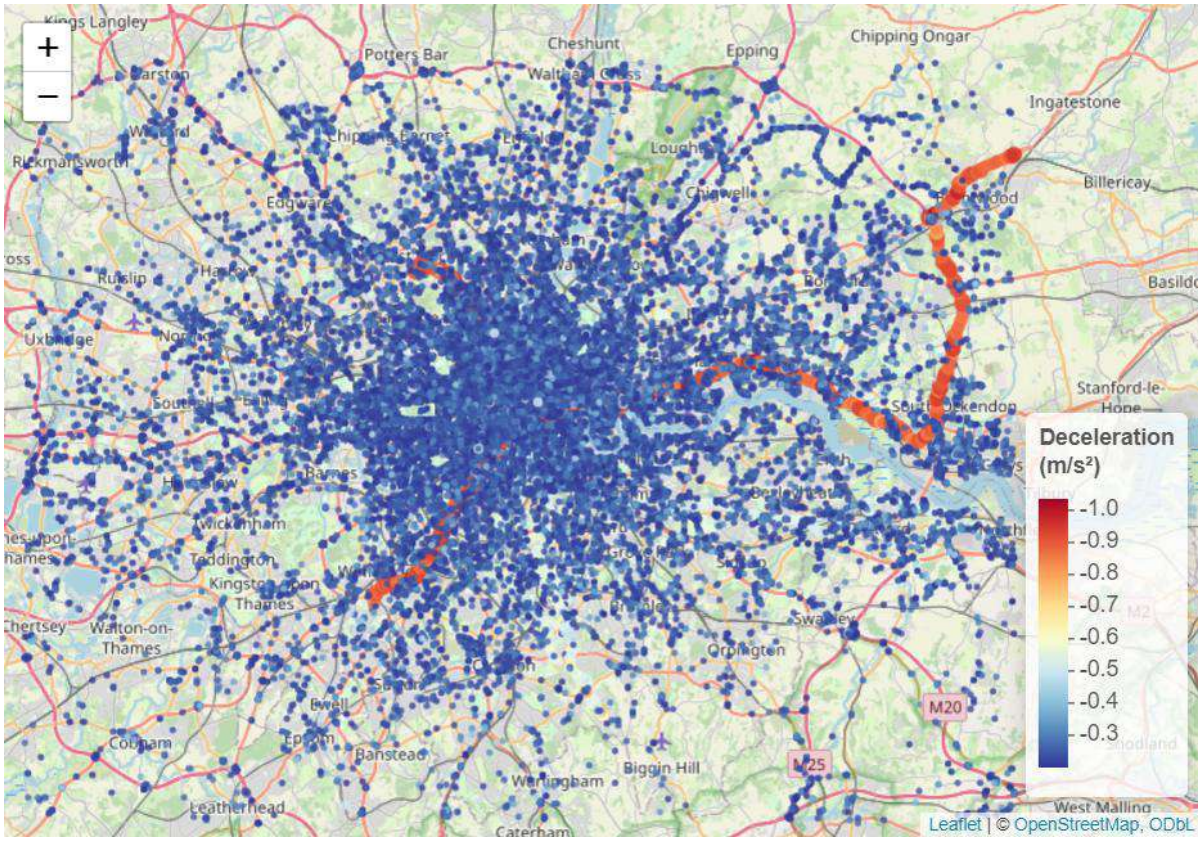

Figure 1. Initial visualization of harsh decelerations of connected LCVs

In addition to the Compass IoT data, a superimposition was made with UK crash data obtained from the STATS19 database. STATS19 offers a comprehensive database of road traffic crashes in the UK. Data in the system is collected by the traffic police and maintained by the UK Department for Transport (DfT). The datasets of STATS19 comprise complementary files regarding the crashes, as well as involved vehicles and casualties, and are used widely in road safety research (UK DfT, 2024). Initially, 25,541 crashes were identified for the study area overall. After filtering the involved vehicles through the vehicle type recorded in the corresponding vehicle database, 2,569 crashes involving LCVs remained in the study area. The spatial distribution of crashes is shown on Figure 2.

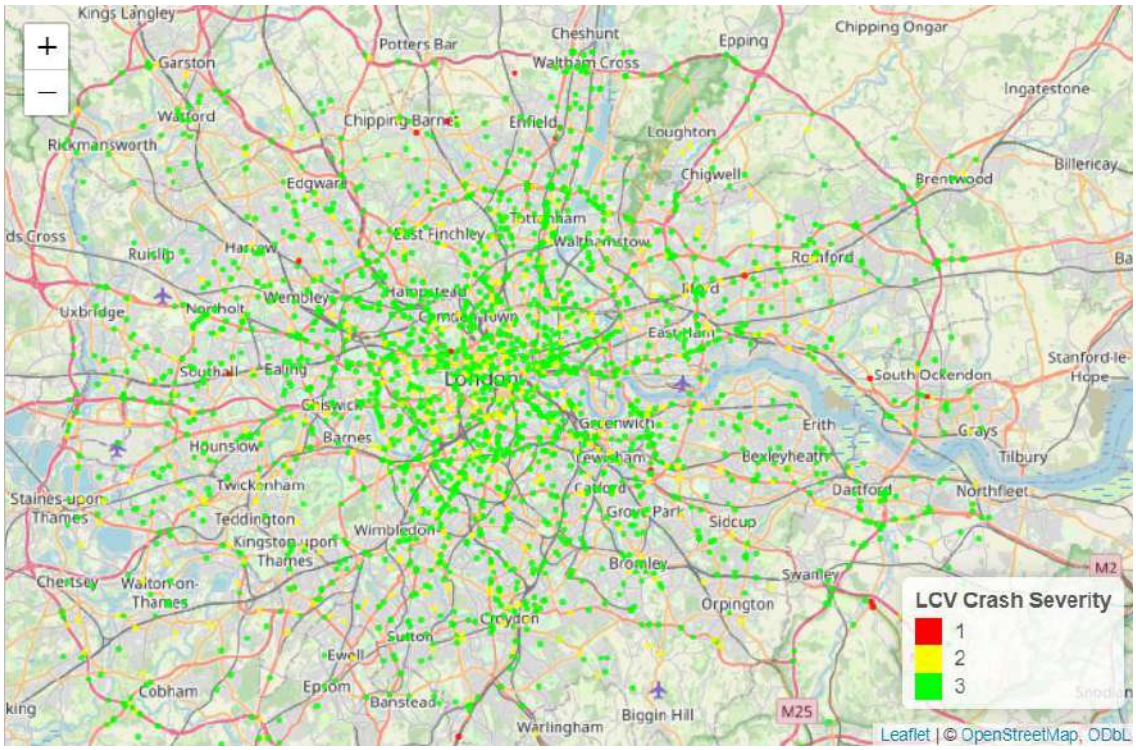

Figure 2. LCVs crash locations in Greater London Aread for 2023.

The crash datapoints are colored by severity. A severity of 1 denotes at least one fatality in the crash (17 crashes), while 2 denotes at least one fatality in the crash (418 crashes). A severity of 3 denotes only slight injuries to persons involved in the crash (2134 crashes). The traditional road safety pattern denoting higher injury severity in interurban and rural areas is retained in the present dataset as well, as almost all fatality crashes were recorded outside of the city centre. It should be noted that Compass has developed comparable analytics within industrial activities as part of its industrial research activities, featuring several dashboard options, such as the ones shown in Figure 3:

Figure 3. An earlier example of speed distributions across an OSM edge from the Compass IoT dashboard.

3 METHODOLOGY

The initial methodological approach involves the investigation of whether connected LCV harsh braking datapoints follow the same spatial distribution as the related crash datapoints. In other words, the superimposition of spatial concentrations of data is examined, via a cross nearest-neighbor (cNN) distance analysis. In essence, each spatial datapoint in the harsh braking dataset is compared with all the near datapoints in the crash dataset and the minimum Euclidean distance is obtained. The comparison required preparatory transformations to a common Coordinate Reference System (CRS), and specifically the British National Grid (EPSG:27700), ensuring consistency in spatial measurements.

Subsequently, Ripley's Cross-K Function (K12) is estimated. K12 quantifies how the points of one dataset (in particular, harsh brakings) are spatially related to the points of the other (in particular, crashes). The mathematical formulation of K12 is given by the Equation below:

$$K_{12}(r) = \frac{1}{N_1}\sum_{i-1}^{N_1}\frac{m_2(B_i(r))}{area(B_i(r))} \quad (1)$$

Where:

- $K_{12}(r)$ is the cross-K function value at a set distance r, denoting the spatial interaction between the two datapoint sets
- N_1 is the number of points of type 1 (i.e., deceleration points).
- $B_i(r)$ describes the circular region of radius radius r centered at the i-th point of type 1.
- $m_2(B_i(r))$ is the number of points of type 2 (i.e., crash points) within a distance r from the i-th point of type 1.
- $area(B_i(r))$ is the area size of the circular buffer of radius r centered at the i-th point of type 1.

4 RESULTS

As per the aforementioned, NN analysis of connected LCV and crash data constituted the first pillar of the present study. The calculated NN histogram appears on Figure 4.

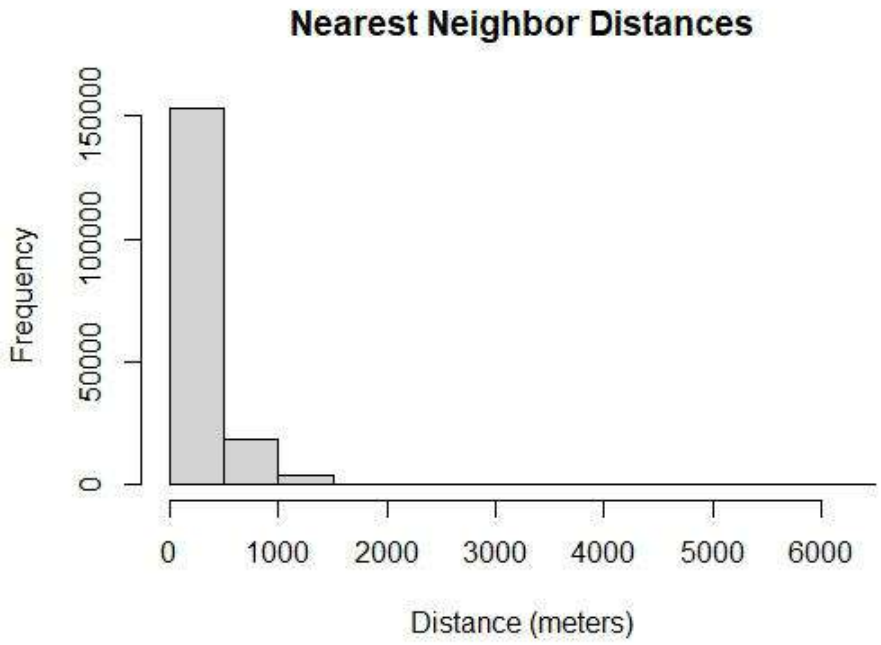

Figure 4: Nearest neighbor distances of connected LCV harsh brakings and crashes involving LCVs

Overall, the histogram of Figure 4 shows high concentrations in short distances, especially below 500m. This finding denotes that a significant number of datapoint pairs are located in close proximity to one another. There is a strong left-side skewness in the data, therefore only a very low number of total pairs are isolated spatially. Therefore, there is a strong spatial concentration observed in the examined datapoints.

Ripley's K12 function results are calculated for two separate distance thresholds. Initially, K12 is calculated for a distance of 1000 meters, with 50 meter increments. Results for the 1000-meter threshold appear on Figure 5.

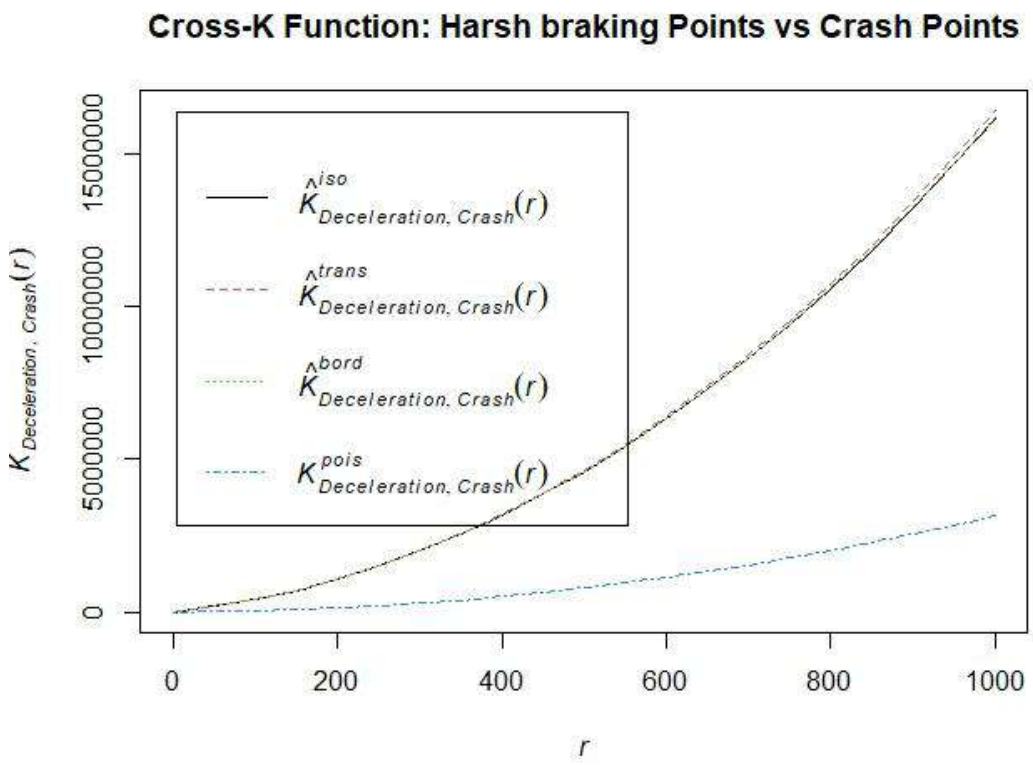

Figure 5. Cross-K Functions for connected LCV harsh brakings and crashes [Threshold = 1000 m]

Likewise, results for a tenth of the distance, namely a 100-meter threshold with 20 meter increments, appear on Figure 6.

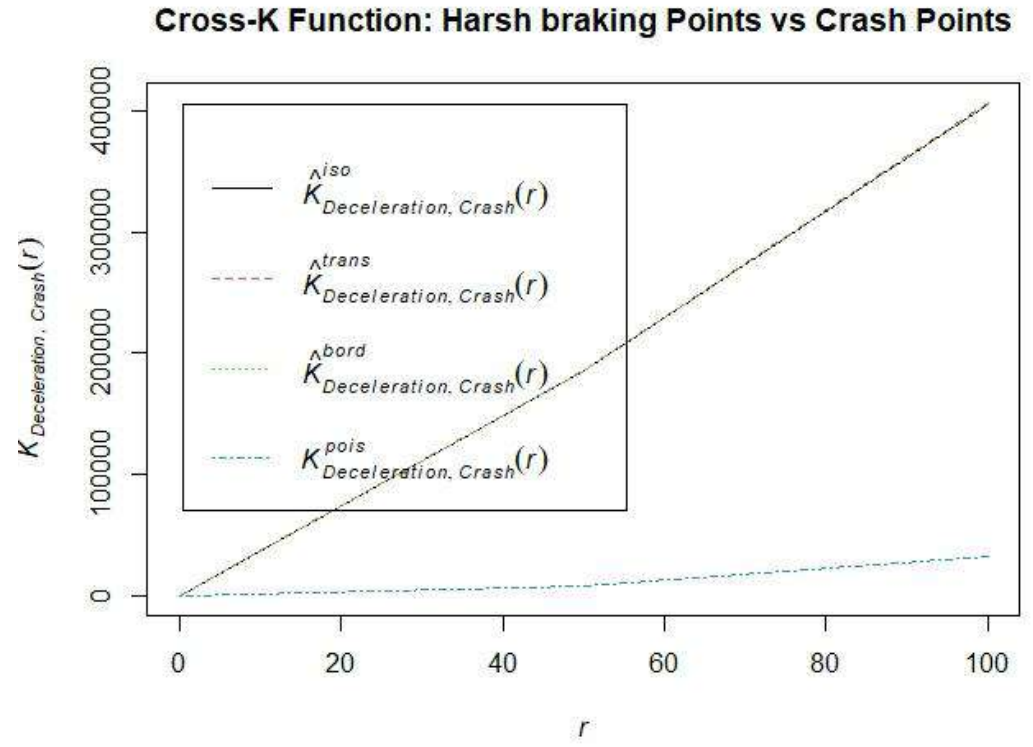

Figure 6. Cross-K Functions for connected LCV harsh brakings and crashes [Threshold = 100 m]

In the above graphs of Figures 5 and 6, the curve termed K12iso(r) represents the isotropic K12(r), which is the primary K12 function result, denoting the distribution of crash points as distances increase. The dashed curve of K12pois(r) represents the expectation under a completely random spatial distribution. The dashed curve of K12trans(r) takes edge corrections into consideration by using the translation method, in order to account for boundary effect bias (Ziakopoulos & Yannis, 2020). Likewise, the K12bord(r) dashed curve takes edge corrections into consideration by using the border method.

In both of the examined graphs, the number of crash points within r-distance of harsh braking points increases as distance increases. This is an expected outcome as larger distances create buffer points that enclose more crash locations. Moreover, and more interestingly, the K12iso(r) curve lies consistently above the expectation denoted by the K12pois(r) curve. This is an indicator of high spatial correlation and dependence. In other words, crash points are more likely to manifest near harsh braking points in the particular dataset compared to a random spatial distribution. Lastly, both of the explored edge-corrected versions indicate that the present analysis appears to be robust and not affected by boundary effects (Delmelle & Thill, 2008; Siddiqui and Abdel-Aty, 2012).

5 CONCLUSIONS

Rapid advances in the Connected Vehicle domain enable more data being collected and analyzed. Such data has immense potential including serving as surrogate safety measures. This study constituted an endeavor to investigate connected Lights Commercial Vehicle (LCV) data provided by Compass IoT and their correlation to the related LCV-involved crashes. LCV naturalistic driving data from the Greater London Area was collected for the year 2024 and contrasted with the LCV-involved crashes of the latest available year, namely 2023. Spatial diagnostic methods were applied in the form of cross-Nearest-Neighbor and Ripley's Cross K12 function analyses. The results reveal a significant spatial correlation between connected LCV harsh braking incidents and crashes involving LCVs. In other words, LCV crash locations are more likely to occur near harsh braking points in the dataset than would be expected from a random spatial distribution. Further research might involve detail spatial regression modelling for the data in order to discover its full potential for proactive road safety analyses and correlation with specific risk factors.

REFERENCES

Ambros, J., & Jurewicz, C. (2017). From big data to speed and safety: a review of surrogate safety measures based on speeds from floating car data. In 2017 Australasian road safety conference, Perth.

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.

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.

Ehsani, J. P., O'Brien, F., & Simmons-Morton, B. (2017). Comparing g-force measurement between a smartphone app and an in-vehicle accelerometer. In Driving Assessment Conference (Vol. 9, No. 2017). University of Iowa.

Goodchild, M. F. (2008). Commentary: whither VGI?. GeoJournal, 72(3-4), 239-244.

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

Laureshyn, A., & Várhelyi, A. (2018). The Swedish Traffic Conflict technique: observer's manual.

Nikolaou, D., Ziakopoulos, A. & Yannis, G. (2023). "A review of surrogate safety measures uses in historical crash investigations." Sustainability, 15(9), 7580.

Oikonomou, M., Ziakopoulos, A., Chaudhry, A., Thomas, P. & Yannis, G. (2023). "From conflicts to crashes: Simulating macroscopic connected and automated driving vehicle safety". Accident Analysis & Prevention, 187, 107087.

OpenStreetMap. (2019). Official Wiki Website. Available: https://wiki.openstreetmap.org/wiki/About_OpenStreetMap [Accessed 20-12-2024]

R Core Team (2024). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/

Siddiqui, C., & Abdel-Aty, M. (2012). Nature of modeling boundary pedestrian crashes at zones. Transportation Research Record, 2299(1), 31-40.

Tarko, A.P. (2018). Surrogate Measures of Safety, in Safe Mobility: Challenges, Methodology and Solutions; Emerald Publishing Limited: Bingley, UK, 2018.

UK Department for Transport (UK DfT). STATS19 Road Safety Data. Available: https://www.data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-accidents-safety-data

Wang, C., Xie, Y., Huang, H., & Liu, P. (2021). "A review of surrogate safety measures and their applications in connected and automated vehicles safety modeling." Accident Analysis & Prevention, 157, 106157.

WHO. (2023). Global status report on road safety. Retrieved from https://iris.who.int/bitstream/handle/10665/375016/9789240086517-eng.pdf on 06-12-2024.

Ziakopoulos, A., Sekadakis, M., Katrakazas, C., Kallidoni, M., Michelaraki, E. & Yannis, G. (2025). "Explainable macroscopic and microscopic influences of COVID-19 on naturalistic driver aggressiveness derived from telematics through SHAP values of SVM and XGBoost algorithms." Journal of Safety Research, 92, 393-407.

Ziakopoulos, A., & Yannis, G. (2020). A review of spatial approaches in road safety. Accident Analysis & Prevention, 135, 105323.