30th ITS World Congress, Dubai, UAE, 16-20 September 2024
Paper ID #287

Road accident prediction in south-eastern England: the advantages of using connected vehicle data

Thierry Castermans[1*], Esteban Hernandez Capel[1], Jean-François Meessen[1]
1. AISIN Technical Centre Europe, Belgium; Thierry.Castermans@aisin-europe.com

Abstract
For several years, connected vehicles produce massive amounts of data which represent a gold mine for road engineers who are in charge of improving road quality and safety. In this paper, we compare the road accident prediction performance level of two analyses: the first one is conducted based on historical road accidents and the other is based on harsh braking data only. We quantify the benefits in terms of accuracy and speed when it comes to predicting future accidents on a large-scale road network using real data from fleets of connected vehicles. We demonstrate that harsh braking clusters allow us to detect precise spots that are dangerous for the road users. Finally, we give some perspective for future work.

Keywords: Connected car data, crash prediction, road safety.

1. Introduction

1.1 Road safety: proactive vs reactive strategy

Despite notable progress over time, road safety remains an urgent global issue. As stated by World Health Organization, “Every year the lives of approximately 1.19 million people are cut short as a result of a road traffic crash. Between 20 and 50 million more people suffer non-fatal injuries, with many incurring a disability.” [17]. One of the challenges of casualty reduction is that we are usually working retrospectively. A cluster of KSI (Killed or Seriously Injured) crashes will attract the attention of safety engineers. However, to improve outcomes, it would be desirable to add a proactive strategy to the “traditional” methods. For more than a decade, this goal has motivated many initiatives and research projects.

The International Road Assessment Programme (iRAP) [14], for example, is dedicated to saving lives by eliminating high-risk roads throughout the world. The iRAP methodology provides a simple and objective measure of the relative risk associated with road infrastructure for the different road users. This risk is calculated by modelling the number of casualties mainly based on geometric characteristics of the road infrastructure (road attributes). This approach offers multiple advantages among which the capability to assess the effectiveness of a countermeasure brought to the road infrastructure.

A complementary approach consists in analysing driver behaviour. In the framework of the EUROFOT project [2], a reliable incident detection process has been presented as a surrogate measure of crashes and road

injuries. CAN-data from vehicles were used to generate safety indicators based on high lateral and longitudinal acceleration, the threshold values depending on speed and vehicle type. First experimental studies with drivers have shown that harsh events are clearly accumulating on certain hotspots of the road network [3], and this can be observed both using data collected by embedded car sensors or low-cost smartphone sensors [1]. Several authors have then clearly demonstrated a correlation between accumulation of harsh braking events and the occurrence of road crashes. In [10], evidence is given that "harsh braking records can be used to support accident modelling, they are a source of much more numerous data than accidents, and this may be important in considering changes or trends in accident risk over a much shorter time than for accident studies.". In [9], the analysis results "indicate a strong correlation between hard-braking events and rear-end crashes occurring more than 400 ft upstream of an intersection.". The authors suggest that agencies can use new hard-braking data sources to quickly address emerging issues, instead of waiting for 3–5 years of crash data. Similar correlation studies have been conducted to show the link between harsh braking activity and the safety level in and around road construction projects [5]. Finally, mathematical models have been developed to predict locations with high likelihood of accident based on harsh braking historical data [11, 7] and machine learning solutions have been proposed to monitor the accident risk in real time [12].

1.2 Objectives of this study

The objectives of this study are multiple. They are based on the following observations.

1. The literature clearly indicates the usefulness of driving behaviour analysis to support accident modelling. Most of the time, however, the correlation between high spatial density of harsh braking and accident-prone areas was established specifically on long road segments, on the order of 100 meters. In this paper, we want to show that harsh braking clusters, which are spreading over a (few) dozen meters instead of a (few) hundred meters are also valid to predict future accidents.
2. Although sophisticated accident prediction models have been published, traditional methods aiming at improving safety on road networks are still focusing on historical accident data (the reactive strategy prevails). In this paper, we want to quantify the benefit of using harsh braking clusters compared to a strategy based on historical accidents only. In particular, we want to show how much data is needed to predict future accidents. The proposed approach will be validated using a very large-scale road network, mixing different driving conditions like urban context, highways, national, and rural roads.

2. Data collection and method

2.1 Connected car data

This study leverages data from several fleets of commercial connected vehicles [15] equipped with dedicated telematics boxes. Two datasets were generated from those telematics boxes: (i) The **trace events**, which are recorded every 10 seconds and contain the position and the speed of the vehicle. (ii) The **harsh events**, which are recorded whenever the accelerometer signal crosses a pre-defined threshold. As the acceleration is measured on 3 axes, it is possible to detect and distinguish harsh acceleration, harsh braking, and harsh cornering events (see Figure 1A). For each harsh event, we determined the duration (i.e. the time interval

during which the acceleration has crossed the threshold level), the speed at the start and at the end of the event and the maximum value (in absolute value) of the acceleration during the event (see Figure 1B).

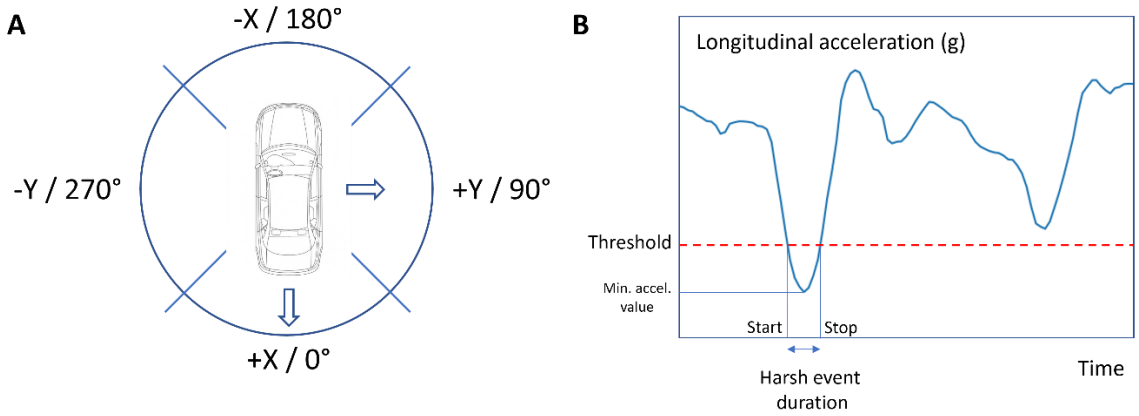

Figure 1 – A) Coordinate system adopted to define the harsh events depending on the direction of the accelerometer signal: harsh acceleration (0° ± 45°), harsh braking (180° ± 45°), left cornering (90° ± 45°), and right cornering (270° ± 45°). B) Example of a harsh braking event and illustration of the method used to determine its duration.

The data we analysed was recorded in a large region located in the southeast of England, comprising more than 60 counties and 100 000 km of roads (see Figure 2). In this area, more than 2 million harsh braking events were detected in 6 209 068 vehicle traces analysed between the 1st of January 2022 and the 31st March 2022. This dataset represents in total 134 million hours of driving. Passenger cars and light commercial vehicles (LCV) are comprised in the dataset, which is called "CCD" (Connected Car Dataset) in the following of this paper.

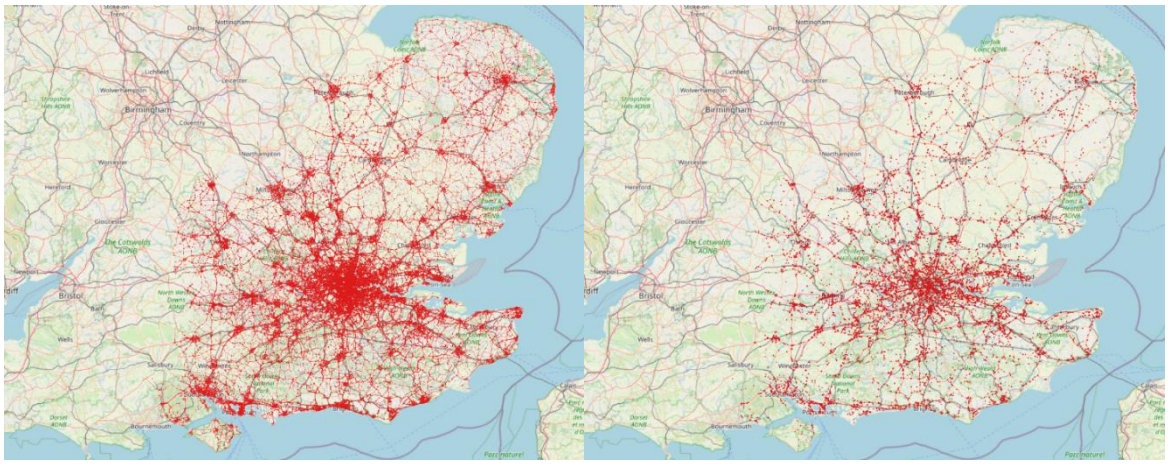

Figure 2 – The current study focused on an area of more than 60 counties in the Southeast of England. The road network length corresponds approximately to 100 000 km. On the left: the red dots depict the harsh braking events that were detected during the first 3 months of 2022. On the right: only the clusters of harsh braking that were extracted from our analysis are represented (see text for details).

2.2 Traffic accident datasets

Two different road accident datasets were used in this study. The first one contains the 263 455 accidents reported by the police between 2017 and 2021 (5 years) in our area of interest and is called the "road accident historical dataset". The second one contains the 50 840 accidents that were reported in the year 2022 only. All data are published by the British Department for Transport (DfT) [16]. In most cases, the circumstance of the accident is described by means of a series of data items e.g., the location and timing of the accident, the number of casualties/fatalities, the type of road users and vehicles involved, the speed limit and the weather conditions, to name but a few.

2.3 Connected Car Data processing pipeline

The full data processing pipeline is illustrated in Figure 3. In a first step, the raw data was map-matched to determine the most probable position of the car on the road network based on the raw GPS information. Our fast map-matching algorithm [8] was also used to derive the heading of the car, by using the position and speed information coming from several successive sampling points. The heading information enables us to distinguish on which side of the road the car was rolling.

After the map-matching step, all the points for which the longitudinal acceleration value was lower than a specific threshold were flagged as "harsh braking events". The threshold was fixed to 0.3 g, which is a value compatible with the ones used in other studies [1, 2]. During such braking events, all the objects in the car that are not well stowed will fly off the seat.

Figure 3 – Data processing pipeline. The map-matching procedure allows to convert raw GPS data into positions on the road network. The clustering algorithm itself is the core of our approach to predict future accident locations. The contextualisation step enables us to reject anomalies in the data.

As a next step in the processing pipeline, a severity factor was computed for each harsh braking event. This factor depends both on the acceleration value and the speed of the car at the starting point of the event. A map showing the location of the harsh braking events is shown in Figure 2 (on the left). At first sight, this map doesn't carry much useful information. However, applying a spatial clustering technique reveals locations characterized by a high density of harsh braking events. The fundamental interest of this step in the data processing is that the driving behaviour of different drivers is aggregated. This way, we are not focusing on the behaviour of a unique aggressive driver but rather on a collection of similar reflex actions. The spatial clustering was carried out using the DBSCAN algorithm [6]. This algorithm detects zones of high density based on two input parameters: first, a maximal distance to consider two points as being part of the same cluster and, second, a minimum number of points for the cluster to be considered as valid. The adequate choice of those parameters was guided both by domain knowledge and generally admitted procedures [13]. Each harsh braking cluster is to be considered as a dangerous spot on the road network. In practice, the clusters stretch on a zone of a few to several dozens of meters (see Figure 4).

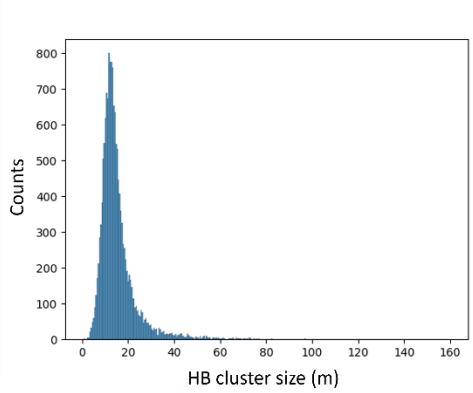

Figure 4 – Distribution of the harsh braking (HB) cluster size.

After the spatial clustering procedure, each harsh braking cluster was enriched with additional variables most of the time computed by combining the connected car data set with external databases. The aim of this procedure is to contextualise the conditions in which the different braking events occurred. For example, knowing the time and the location of the event allows us to determine the position of the sun and consequently the lighting conditions (daytime, nighttime, dawn, dusk). The data recorded by local weather stations give additional insight regarding the circumstances of the braking as well as the map itself, which indicates if the braking happened near a roundabout, a traffic light, or a school, for example. Additionally, the inclusion of contextualisation variables enables us to filter spurious events and select the most interesting harsh braking clusters. For instance, the distance between the raw and the map-matched GPS coordinates is a metric that helps to reject the points being either off-road or on a private domain.

2.4 Road accident prediction and validation

In this study, the harsh braking events recorded during the first 3 months of the year 2022 were used to predict future accidents occurring during the next 9 months of the same year in southeast of England. With this aim, the harsh braking (HB) events were clustered (as described in previous section) and the resulting clusters were considered as future accident locations (see Figure 5).

To characterize the performance of this approach in the anticipation of crashes, we defined two different metrics:

1) The **prediction rate** r is the number of accidents that were correctly predicted N_{ap} divided by the total number of accidents reported by the police in the area of interest N_{atot}: $r = \frac{N_{ap}}{N_{atot}}$.

In short, the prediction rate indicates to what extent the model can predict all the future accidents.

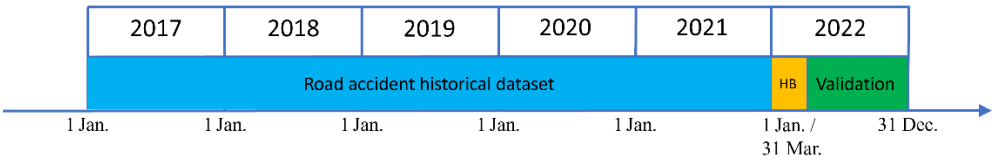

Figure 5 – Time periods used for data analysis and road accident prediction and validation: 5 years of road accident historical data, 3 months of harsh braking (HB) data only and 9 months of accident data for validation.

2) The **conversion rate** η represents the number of accidents that were correctly predicted N_{ap} divided by the total number of clusters extracted from the HB dataset N_{ctot}: $\eta = \frac{N_{ap}}{N_{ctot}}$.

In practice, a cluster of HB was considered as successfully predicting the location of a future accident if this accident was located at maximum 75 meters of the HB cluster centroid. This distance was adopted to consider the uncertainty coming both from the GPS data and the location of the accident as reported by the police (vehicles involved in a crash may stop far away from the impact point of a collision).

The conversion rate indicates to what extent the model can predict accidents given a certain number of candidate locations.

To verify that the prediction performance levels of our approach were significant, we developed four different baseline models:

1) **Model B0**: a random model where the prediction consists in picking N_{ctot} random locations on the road network of interest. The question to answer is: does our approach give better results than a pure random model?
2) **Model B0 weighted**: similar to the model B0 but giving a more important weight to the roads with more traffic. In other words, the locations that are randomly picked will have a higher probability of being on roads that are busier.
3) **Model B1**: a random model where the prediction consists in picking N_{ctot} random locations in the road accident historical dataset. Here the idea is to discover if a given number of HB clusters contains more information (or not) than the same number of locations taken from the past accidents.
4) **Model B2**: a model using all the locations in the road accident historical dataset. This will complete our analysis.

The results of the different approaches are presented and discussed in the following section.

3. Results and discussion

The main figures to consider here are the following:

- During the validation period (April to December 2022), **38 738 road accidents** were reported by the police. These are the target for our prediction task. Among those accidents, 30 719 (79%) caused slight injuries, 7633 (20%) serious injuries, and 386 (1%) were fatal.
- A total of **15 303 clusters of HB** were extracted from the CCD dataset. These locations represent candidates for future accidents according to our approach.
- Finally, **263 455 road accidents** were reported during the 5 years preceding 2022. These locations represent candidates for future accidents according to the model B2.

3.1 Performance of the model using the harsh braking clusters

It turns out that the HB clusters correctly predicted 3161 accidents (out of the 38 738), 2638 (83%) where people were slightly injured, 507 (16%) where people were severely injured and 16 (0.5%) which were fatal.

This leads to a prediction rate of 8.3%. As 15 303 clusters were used to make the prediction, the conversion rate is 21%.

3.2 Performance of the models B0 and B0 weighted

The performance of the model B0 can be determined using the binomial law. In practice, we divided the road network of our area of interest in N_{bins} small road segments of 20 meters. We then attributed a "status" to each road segment: 0 if no accident occurred during the validation period, and 1 if an accident did occur. At this stage, the problem is to compute the probability that attributing randomly the status 1 to N_{ctot} road segments will allow to predict the same number of accidents as our proposed method based on the HB clusters.

In practice, we used the binomial law to compute this probability. Indeed, this law allows to determine the probability $P(X = k)$ to get k "success" after realizing a number n of independent experiments having only 2 possible outcomes ("success" or "failure"), and for which the probability p of "success" is known and constant: $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$, with $\binom{n}{k} = \frac{n!}{k!(n-k)!}$. In our problem, we can identify that the probability p of "success" is the number of accidents that occurred during the validation period divided by the number of road segments N_{bins}; n corresponds to the number of clusters N_{ctot}; k corresponds to the number of accidents that were correctly predicted based on the HB clusters.

We computed $P(X \geq k) = 1 - \sum_{i=0}^{k} P(X = i)$, which is the probability to predict **at least** the same number of accidents as our proposed method and obtained 0.007, a value close to zero, meaning that our proposed method does not lead to results by chance.

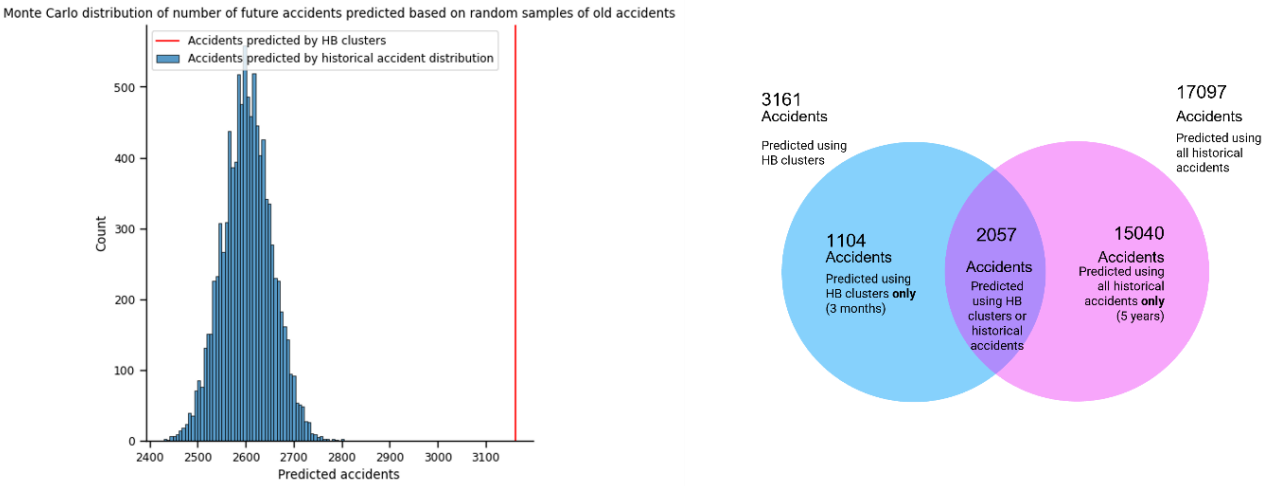

Figure 6 – On the left: To determine the model B1 prediction performance, we ran a Monte Carlo experiment 10 000 times. For each experiment, a sample of N_{ctot} locations were picked randomly in the historical road accident dataset. We then determined the number of accidents (belonging to the validation period) that were correctly detected and hence computed the prediction and conversion rates. The average performance of the model is listed in Table 1. On the right: Number of accidents detected with the HB and/or B2 model.

The model B0 considers implicitly that the traffic is the same on each road, which is obviously not true. We thus designed a model B0 weighted by the traffic values on each road segment. Those traffic values were determined by counting how many connected vehicles drove on each 20-meter road segment. The length of each of these road segments was then multiplied by the corresponding traffic values so that the weights of busier road segments were higher accordingly. Here also, the binomial probability was negligible (0.014), meaning that weighting each road segment by the traffic value did not change the conclusion obtained with the model B0.

3.3 Performance of the models B1 and B2

The performance levels of the models B1 (see Figure 6 on the left) and B2 are summarised in Table 1, as well as the performance reached by our method based on HB. We see that the best prediction rate is obtained by the model B2. However, the comparison seems not so fair, given that the model B2 leverages 263 455 points collected for 5 years, while our approach was based on 15 303 points acquired for 3 months only. This is the reason why we designed the model B1, using the same number of points as our approach. We can see that the average performance of this model B1 is the lowest in terms of prediction rate. Using the same number of points as B1, our approach leads to a prediction rate improved by 22%.

Additionally, our approach gives the best conversion rate: 21% of the HB clusters turn out to become real accidents in the 9 following months. This represents an improvement by a factor 3 compared to the conversion rate reached by the model B2. This means that utilising the HB clusters allows to predict much more quickly the future accidents than relying on the historical road accidents. Also, the quantity of useful information in HB clusters is bigger than in historical road accidents.

Table 1 – Comparison of the performance levels of the different models. 38 738 road accidents were reported during the validation period.

Model type	**Predicted**	**Using**	**Prediction rate r**	**Conversion rate η**
HB Clusters	3161 accidents	15 303 points	8.2%	21%
B1 (avg)	2602 accidents	15 303 points	6.7%	17%
B2	17097 accidents	263 455 points	44%	6.5%

3.4 Limitations and future work

This study relies on a very large-scale dataset covering a substantial region of England, but it also faces some challenges and limitations. Future research could include the investigation of differences across several regions, countries or driving contexts (on highways, national roads, secondary roads and in the city centre). Also, other vehicle types could be considered, like trucks or heavy good vehicles. The study could also be extended using other harsh events, like harsh acceleration. Indeed, harsh acceleration can reflect safety issues, as it can show erratic driving behaviour or situations where a driver is trying to quickly escape a possible crash site. Harsh cornering might also be considered, as an indicator of dangerous driving in curves or junctions.

In this study, we equally considered all the HB clusters as future accident locations to demonstrate in a simple way the advantages of this approach compared to the traditional reactive approach. In a future work, it will be useful to further enhance our prediction results by estimating the probability of crash of every cluster based on HB contextualisation variables like speed, speed difference, light and weather conditions, road geometry (presence of a junction or a sharp curve), the proximity of vulnerable road users on a cycle lane or near a school or a bus stop, for example. Finally, the temporal aspect of the harsh event clustering should be investigated as well to give further insight at each dangerous location.

4. Conclusion

One of the main goals of this study was to compare the accident prediction efficiency of two simple methods: (i) the traditional approach, which considers the locations of historic accidents as the most dangerous hotspots of the road network, and (ii) an approach which considers the clusters of harsh braking events as future accident location candidates. To conduct our analysis, we have used a massive set of data from connected vehicles driving over a large-scale road network in southeast of England, including all sorts of driving conditions and road types. In addition to this, we used in total a dataset of 6 years of validated accident reports. Our results indicate that harsh braking events bring comparatively more information than past accidents. Clusters of harsh braking bring an improvement of 22% in prediction rate compared to an equivalent number of historical accident locations. Additionally, we have shown that using harsh braking clusters allows a much quicker detection of future accidents. Moreover, the HB clusters enabled us to detect accidents that could not be predicted even using 5 years of historical road accident data. Finally, we have demonstrated that harsh braking clusters, which are spreading over a (few) dozen meters only are sufficient to predict future accidents. Future work should aim at combining driver behaviour analysis with road attributes analysis (following the iRAP methodology) to further enhance our accident risk prediction capabilities. Road managers will then have at hand better tools to improve road infrastructure and make better informed, risk-based decisions.

References

1. Aichinger C., Ph. Nitsche, R. Stütz, M. Harnisch (2016). Using low-cost smartphone sensor data for locating crash risk spots in a road network, *Transportation Research Procedia*, Vol. 14, pp. 2015-2024
2. Benmimoun, M., F. Fahrenkrog, A. Zlocki, L. Eckstein (2011). Incident detection based on vehicle can-data within the large-scale field operational test “euroFOT”. In Proceedings *22nd International Technical Conference on the Enhanced Safety of Vehicles (ESV)*.
3. Cao, G et al. (2015). Cluster-based correlation of severe braking events with time and location. In Proceedings of the *10th System of Systems Engineering Conference (SoSE)*, San Antonio, TX, USA
4. Castermans, T. (2019). Hybrid Driver Coaching (HDC): an eco-driving coaching system for hybrid car owners. In Proceedings *26th ITS World Congress*, Singapore.
5. Desai, J. et al. (2021). Correlating Hard-Braking Activity with Crash Occurrences on Interstate Construction Projects in Indiana, *Journal of Big Data Analytics in Transportation*, vol. 3, pp.27-41

6. Ester, M., H.-P. Kriegel, J. Sander, K. Xu (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of the *Second International Conference on Knowledge Discovery and Data Mining*, Portland, Oregon (KDD'96).
7. Feng S. et al. (2024). Exploring the correlation between hard-braking events and traffic crashes in regional transportation networks: A geospatial perspective, *Multimodal Transportation*, vol. 3(2), pp.100-128
8. Gonsette, J.-S. (2018). A fast and versatile map matching engine. In Proceedings of *25th ITS World Congress*, Copenhagen.
9. Hunter M. et al. (2021). A Proactive Approach to Evaluating Intersection Safety Using Hard Braking Data, *Journal of Big Data Analytics in Transportation*, vol. 3, pp.81-94
10. Kamla, J. et al. (2019). Analysing truck harsh braking incidents to study roundabout accident risk, *Accident Analysis & Prevention*, vol. 122, pp.365-377
11. Khanal, M. and Edelmann, N. (2023). Application of Connected Vehicle Data to Assess Safety on Roadways, *Eng*, vol.4(1), pp.259-275
12. Mussah A. et al. (2022). Machine Learning Framework for Real-Time Assessment of Traffic Safety Utilizing Connected Vehicle Data, *Sustainability*, vol. 14(22), 15348
13. Rahmah, N., I. S. Sitanggang (2016). Determination of Optimal Epsilon (Eps) Value on DBSCAN Algorithm to Clustering Data on Peatland Hotspots in Sumatra, *IOP Conference Series: Earth and Environmental Science*, vol.31
14. http://www.irap.org
15. https://datasolutions.bridgestone-emia.com/
16. https://www.gov.uk/government/organisations/department-for-transport
17. https://rb.gy/0ixqiw