# Machine Learning Analysis of Informal Minibus Taxi Driving

Nomfundo P. Cele, Alain Kibangou, Walter Musakwa

▶ **To cite this version:**

Nomfundo P. Cele, Alain Kibangou, Walter Musakwa. Machine Learning Analysis of Informal Minibus Taxi Driving. MAIH 2024 - International Conference on Mobility, Artificial Intelligence and Health, Nov 2024, Marrakech, Morocco. pp.1-6. ⟨hal-04818563⟩

**HAL Id: hal-04818563**

**https://hal.science/hal-04818563v1**

Submitted on 4 Dec 2024

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons CC BY 4.0 - Attribution - International License

# Machine Learning Analysis of Informal Minibus Taxi Driving

*Nomfundo* Cele[1,*], *Alain* Kibangou[2,3,**], and *Walter* Musakwa[3,***]

[1]University of Johannesburg, Department of Urban and Regional Planning, Doornfontein Campus,0184 Johannesburg, South Africa.
[2]Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP, Gipsa-Lab, F-38402 Saint Martin d'Hères, France.
[3]University of Johannesburg, Faculty of Science, GEMES, Auckland park Campus, 2006 Johannesburg, South Africa

**Abstract.** This paper presents a machine learning analysis of driving behaviors in informal minibus taxis, focusing on both controlled and uncontrolled environments. Informal minibus taxis play a crucial role in urban transportation, particularly in developing countries, yet their driving patterns and safety implications remain under-explored. We utilize exploratory factor analysis to analyze data collected from smartphone GPS carried by a passenger of a minibus taxi, identifying key driving behaviors and patterns. Our study highlights significant differences in driving styles between controlled and uncontrolled environments, offering insights into safety and efficiency. The findings provide valuable information for policymakers, transportation planners, and technology developers aiming to enhance urban mobility and safety in the informal transport sector.

## 1 Introduction

While there is growing interest in the health benefits of mobility (reduction of sedentary lifestyles, active mobility), we must not forget that mobility is also a source of pollution and accidents. According to the World Health Organization (WHO), each year road accidents kill over 1.35 million people. An increased burden of road accident deaths and injuries happens in low- and middle-income countries, and the problem is getting worse as a result of rapid motorisation and urbanisation [1]. In most of the developing world, transportation services are provided by independent stakeholders – the private sector – and constitute a greater share of road accident fatalities and injuries, which in turn account for an increasing and large amount of disease burden and death [2]. In South Africa, the road traffic accident mortality rate is 27 fatalities for a 100,000 population, which is over double the international average and well over the African region average. In South Africa, 75% of households depend on public transport, particularly with minibus taxis [3]. However, it is more regularly involved in road traffic accidents compared to other modes. Indeed, minibus taxis account for 70, 000 road traffic accidents each year, the double of the number of other public transport modes [4]. Pedestrian injury-related deaths account for 38% of all traffic deaths, followed by driver and passenger deaths at 26% and 33% [5].

Road safety and road accidents have become a serious public health issue, leading The United Nations to incorporate road safety into the 17 Sustainable development goals (SDGs), in particular, SDG 3 on good health and well-being and SDG 11 on sustainable cities and communities [6]. The persistence and extent of this public health issue demands new methods for road safety procedures [7]. Because of the second Decade of Action for Road Safety, announced by the World Health Organization as the 2021-2030 Decade of Action for Road Safety, road safety has become a major area of interest for researchers, policymakers, and transport engineers [8]. Traditional methods to improve road safety are centred on executing seminars and workshops on safe driving, awareness campaigns on safe driving, and implementing strict road rules to change the behavior and attitude of road users [9].

Research conducted by public institutions, private entities, and scholars on traffic safety discovered that human error was a contributing factor in more than 90% of road traffic accidents while being the sole cause of 57%. In contrast, only 2.4% of collisions were caused by mechanical problems and 4.7% were caused by environmental issues such as rain, storm, fog and hail [10]. According to [11], contributing factors to road accidents include unsafe driving behaviors (speeding, reckless driving, and drug or alcohol abuse-influenced driving), road curves, undulating terrain, and driver distraction. Unfortunately, the consequence of dangerous driving behavior is evident in sub-Saharan Africa, which has experienced an undesirable increase in road traffic accidents and fatalities. The majority of the road traffic accidents are largely linked to human elements such as gender, age, and education level, while other factors could be linked to drivers' attitudes and behaviors on the roads [12]. The authors of [13] define driving habits, or driving style, as how drivers decide to drive – the way they choose to handle themselves when they are driving. Furthermore, the characteristics of driving style include driving speed (choice of speed), braking or decel-

*e-mail: nomfundocele@gmail.com
**e-mail: alain.kibangou@univ-grenoble-alpes.fr
***e-mail: wmusakwa@uj.ac.za

eration, acceleration, choice to commit road traffic violations, and headway.

Literature review in [13] confirmed that driver behavior and driving style fall into two classifications, namely, vehicle motion research and survey studies. The latter uses self-reported questionnaires on driving habits to investigate relations between driver personality, driver behavior, and driving style factors [14–17]. Vehicle motion research collects and uses data on the movement of motor vehicles to categorize the driving style of vehicle drivers. These types of studies usually distinguish driving style into three groups: (1) aggressive, (2) normal, and (3) calm [18–20]. Harsh deceleration and acceleration, high speeds, harsh vertical and lateral movements, and sudden changes in motor steering angles are usually related to aggressive driving behavior [21]. Various algorithms and methods can be used to classify driving styles; the two main categories being rule-based approaches and machine learning-based approaches. Taking into consideration the challenges in acquiring driver style classifications in real application settings, current studies mainly employ unsupervised learning algorithms to detect and categorize driving styles. Random forest decision methods, support vector machines, and artificial neural networks are often the most utilised methods for the categorization of driving style [13, 19, 22, 23]. In contrast to the literature, the study presented in this article concerns smartphone data collected on the passenger side. By focusing on minibus taxis, our aim is to analyze the factors characterizing the driving of this means of transport, widely used in developing countries, and to compare different indicators in a controlled or uncontrolled context. A controlled context is one in which the presence of law enforcement agencies is evident. An unsupervised machine learning approach is adopted to solve the problem.

The remaining of the paper is organized as follows: the problem under study is formulated in Section 2 and the case study described in Section 3. The analysis and relevant findings are explained in Section 4 before concluding the paper.

## 2 Problem formulation

Consider two sets $\mathcal{S}_c$ and $\mathcal{S}_u$ of minibus taxis operating in two disjoints areas $C$ and $U$, respectively. Equipped with a smartphone, a passenger can collect speed, location and inertial measurements during trips. The collected data gives rise to a matrix $\mathbf{S} \in \mathfrak{R}^{N\times K}$, where $N$ stands for the number of trips using minibus taxis in $\mathcal{S}_c \cup \mathcal{S}_u$ and $K$ for the number of attributes extracted from the measurements. Since the study is related to driving behavior or style, we restrict the study to speed and acceleration as measurements. Therefore we assume that $\mathbf{S}$ can reveal $R$ hidden factors related to the driving behavior of minibus taxi drivers. Typically the mathematical objective is to carry out the decomposition

$$\mathbf{S} = \mathbf{A} + \mathbf{L}\mathbf{F}^T + \mathbf{E} \tag{1}$$

where $\mathbf{A} \in \mathfrak{R}^{N\times K}$ contains the observation means of each attribute, $\mathbf{L} \in \mathfrak{R}^{N\times R}$ stands for the loadings matrix, $\mathbf{F} \in \mathfrak{R}^{K\times R}$ the hidden factors matrix, and $\mathbf{E} \in \mathfrak{R}^{N\times K}$ the error term matrix. Interpretation of the obtained factors can be deduced by analyzing the attributes contributing the most to the considered factor. Then we will analyze how these factors are statistically distributed in the two considered sets $\mathcal{S}_c$ and $\mathcal{S}_u$ in order to highlight differences between the two considered environments if any.

Equation (1) stands for an exploratory factor analysis (EFA) problem. It is well known that the number $R$ of hidden factors can be estimated using a scree plot and/or the explained variance of data while $\mathbf{L}$ and $\mathbf{F}$ are in general obtained by solving a maximum likelihood problem.

## 3 Case study

The study concerns the city of Durban (eThekwini, in Zulu language), South Africa. It is the third-most populous city in South Africa, after Johannesburg and Cape Town. It is situated on the east coast of South Africa, on the Natal Bay of the Indian Ocean.

Minibus taxis are the standard form of transport for the majority of the population. With the high demand for transport by the working class of South Africa, minibus taxis are often filled over their legal passenger allowance, making for high casualty rates when they are involved in accidents. Minibuses are generally owned and operated in fleets. Although minibus taxis account for approximately 9% of the automobiles on Durban roads, they are involved in 18% of serious and fatal injury crashes in Durban [24].

Two routes were selected for this study. One in a controlled area $C$ and the second one in an uncontrolled area $U$. The controlled route stretched from Marine Parade to Suncoast Casino, and vice versa (see Fig. 1). The controlled route has high visibility of metro police, metro police station, speed humps, traffic robots, roundabouts, and traffic circles as measures to control driver behavior on the road. The uncontrolled area was a route stretched from


**Figure 1.** Controlled environment site.

KwaDabeka to Pinetown (see Fig. 2). It included a dis-

advantaged area (township). It had traffic lights and speed humps as measures to control driver behavior on the road.

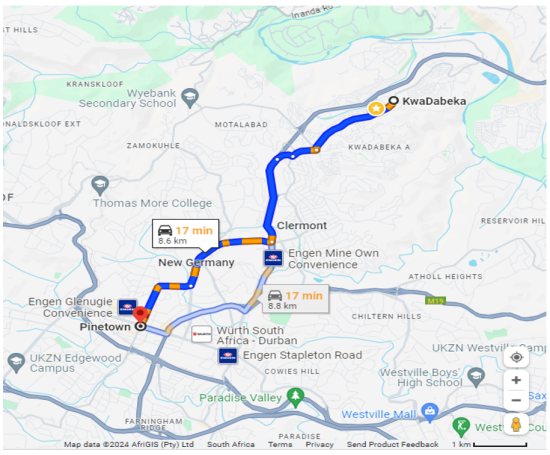


**Figure 2.** Uncontrolled environment.

**Table 1.** Difference between controlled and uncontrolled environments.

| Environment | Features |
|---|---|
| Controlled | Metro police, Police station<br>Speed humps, Traffic robots<br>Roundabouts, Traffic circles |
| Uncontrolled | Speed humps, Traffic robots |

Data was collected using smartphone GPS, specifically location and speed. The minibus taxis were randomly chosen, and drivers were unaware of the data collection campaign. The study did not require driver participation nor responses. The minibus taxis were randomly selected, and drivers were not informed of the data collection campaign to avoid cognitive bias. There was no interaction with the driver, and the smartphone was carried by the same passenger for all the trips. The researcher collected the smartphone data and was not at harm or risk. Measurements of smartphone sensors were collected by means of the App Senslogs[1]. Data was collected over a period of two months. Some of the collected data on trips had a high rate of missing data and were discarded from the research. Finally, 77 trips were analyzed.

## 4 Data analysis and results

Location and speed provided by the smartphone's GPS were collected with a frequency of 1 Hz. The location is used to confirm whether the route used by the minibus taxi is an urban road where the speed is limited to 60 km/h or an expressway where the speed is limited to 100 km/h for this category of vehicles. Fig. 3 depicts two examples of recorded speed profiles. Note that the measured speed is normalized by the speed limit as follows:

$$\bar{v}_t = \frac{v(x_t)}{V(x_t)} \tag{2}$$

[1]https://www.senslog.org/about/

where $x_t$, $v(x_t)$, $V(x_t)$, and $\bar{v}_t$ stand for the location at time $t$, the measured speed, the speed limit at location $x_t$, and the normalized speed. This normalization step allows to detect over-speeding whatever the speed limitation.

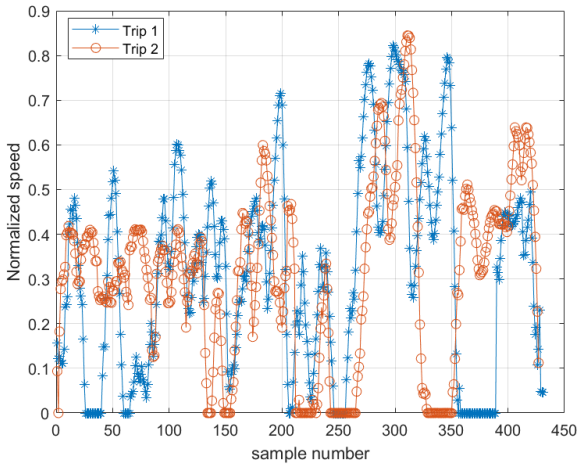


**Figure 3.** Normalized speed profile in a controlled environment.

Since data is collected at a 1 Hz frequency, acceleration is given by:

$$a_t = v(x_t) - v(x_{t-1}) \tag{3}$$

If $a_t > 0$, the vehicle is accelerating while $a_t < 0$ accounts for deceleration. According to the literature [25, 26], acceleration or deceleration is considered as risky if $|a_t| > 1.2m/s^2$. Fig. 4 depicts the magnitude of the acceleration and exhibits some risky cases (magnitude above the limits). In the sequel we will count the ratio of such events as a marker of risky driving. For risky acceleration:

$$R_a = \frac{\sum_{t=1}^{T} H\left(a_t - 1.2\right)}{\sum_{t=1}^{T} H\left(a_t\right)} \tag{4}$$

where $H(.)$ stands for the Heaviside step function while for risky deceleration we get:

$$R_d = \frac{\sum_{t=1}^{T} H\left(-a_t - 1.2\right)}{\sum_{t=1}^{T} H\left(-a_t\right)} \tag{5}$$

With speed and acceleration, one can also compute the stopping distance, a crucial quantity for safety concerns. It is the sum of the reaction distance with the deceleration distance

$$SD_t = v(x_t)T_r + \frac{v^2(x_t)}{|a_t|} \tag{6}$$

in deceleration mode ($a_t < 0$) where $T_r = 0.68s$ stands for the reaction time. We consider 70m as a safe stopping distance. Therefore, in the sequel $SD_t$ is normalized by 70m.

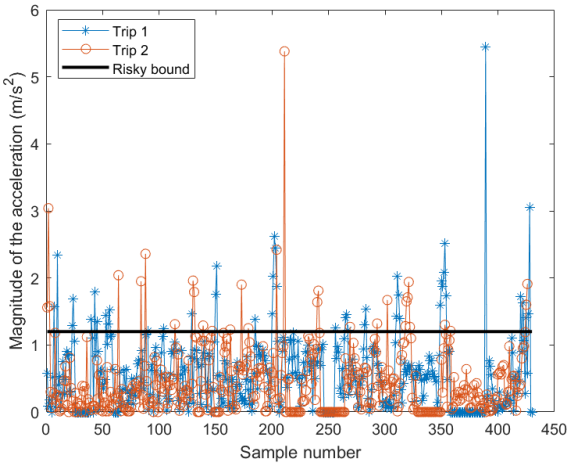


**Figure 4.** $|a_t|$ for two trips. Points above the bound in black solid line represent risky accelerations or decelerations.

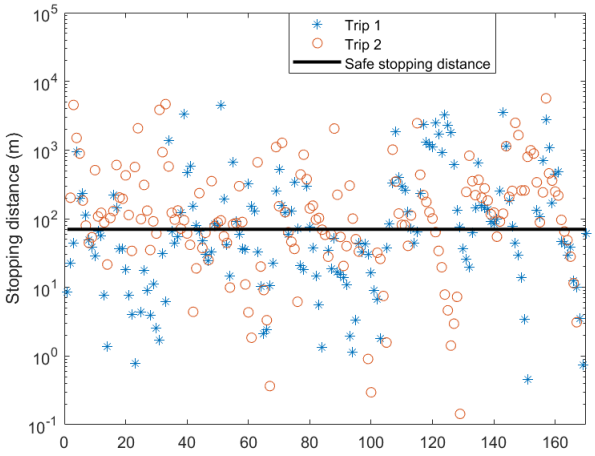


**Figure 5.** Stopping distance for two trips. The points below the black solid line are related to unsafe configurations.

For the analysis, as attributes, we consider the risky acceleration ratio, the risky deceleration ratio, the maximum normalized speed, and some statistics of the normalized speed, the acceleration, the deceleration, and the stopping distance; precisely, the 10th, 50th (median), and 80th percentiles. It appears that the only attributes allowing higher data suitability for factor analysis according to the KMO (Kaiser–Meyer–Olkin) test were:

1. Maximum normalized speed
2. Median of normalized speed
3. 80th percentile of normalized speed
4. Risky acceleration ratio
5. 10th percentile of acceleration
6. Median acceleration
7. 80th percentile of acceleration
8. Risky deceleration ratio
9. 10th percentile of deceleration
10. Median deceleration
11. 80th percentile of deceleration
12. Median stopping distance
13. 80th percentile of stopping distance.

With these attributes we get KMO=0.8062, meaning that data is meritorious for factor analysis.

Matlab was used to perform scree plot analysis and computation of factors of the EFA. Using the scree plot criterion, 2 factors can explain the collected data. To interpret these factors, we set a threshold of 0.5 to the obtained coefficients. After thresholding we get the results depicted in Fig. 6. It can be noticed that the first factor is related

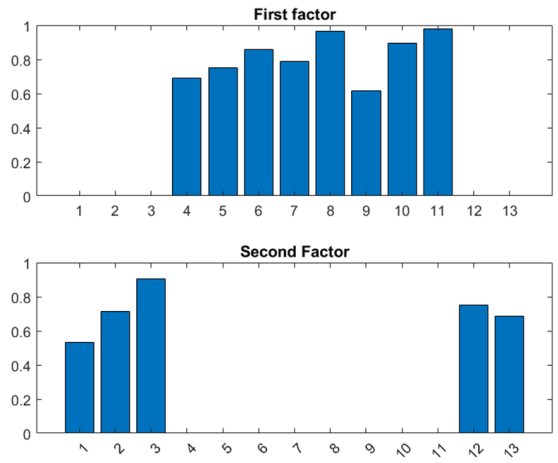


**Figure 6.** Threshold loading of quantitative data mode.

to acceleration and deceleration. Our interpretation is that it accounts for *aggressive driving*. The second factor is constituted with attributes related to speed and stopping distance. We interpret it as *speeding*.

**Table 2.** Definition of factors

| Factor | Definition |
|---|---|
| Aggressive driving | Harsh braking – to use more force than required to decelerate |
| Speeding | Risk taking – driving carelessly, above speeding limit |

Based on this, we can compute the score for each factor for each driver (trip). Let $\sigma_{if}$ be the loading of attribute $f$ in factor $i$ as provided by EFA. We define the threshold value as

$$\bar{\sigma}_{if} = \frac{\sigma_{if}}{2}\left(1 + sign\left(\sigma_{if} - \frac{1}{2}\right)\right),$$

where $sign(.)$ stands for the sign function. Hence, given $D_{jf}$ the score of the $j$th driver on the $f$th attribute and $F = 13$ the number of attributes, the scores of aggressive driving and speeding for the $j$th driver is given by:

$$A_j = \sum_{f=1}^{F} \frac{\bar{\sigma}_{1f}}{\sum_{\varphi=1}^{F} \bar{\sigma}_{1\varphi}} D_{jf} \tag{7}$$

$$S_j = \sum_{f=1}^{F} \frac{\bar{\sigma}_{2f}}{\sum_{\varphi=1}^{F} \bar{\sigma}_{2\varphi}} D_{jf} \tag{8}$$

Figure 7 depicts the statistical distribution of the aggressive driving score. The median value on the aggressive driving score in the controlled environment is below 0.5 while it is below 0.4 for the uncontrolled environment. It can be noticed that the driving style of minibus taxi drivers is more aggressive in a controlled environment where there is a high number of speed bumps, roundabouts, traffic circles, and traffic robots, than when there is less control.

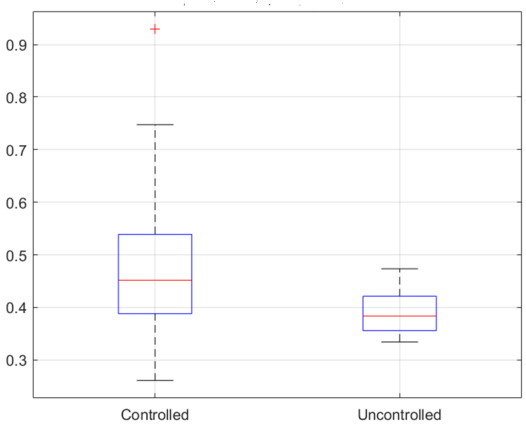


**Figure 7.** Statistical distribution of the aggressive driving score.

Figure 8 depicts the statistical distribution of the speeding score. We can observe that the median value on the speeding score in the controlled environment is 1 while it is above 1.5 in the uncontrolled environment. The results on the speeding score show that high visibility of metro police, speed bumps, roundabouts, traffic circles, and traffic robots reduces speeding unlike in uncontrolled environments which have traffic robots and a few speed bumps to reduce at-risk driving. The results on the uncontrolled environment speeding score provides evidence to support the findings of the Road Traffic Management Cooperation on their study where the Quantum (minibus taxi) was found to be one of the top three vehicles more involved in road traffic crashes in South Africa. It also confirms eThekwini Municipality concerns on harsh driving, and often overloaded minibus taxis [24]. Abrupt braking behaviour and accelerating are typical at-risk driving behaviours (see [25] and [26]) and can explain the problem of serious and fatal injury crashes in Durban.

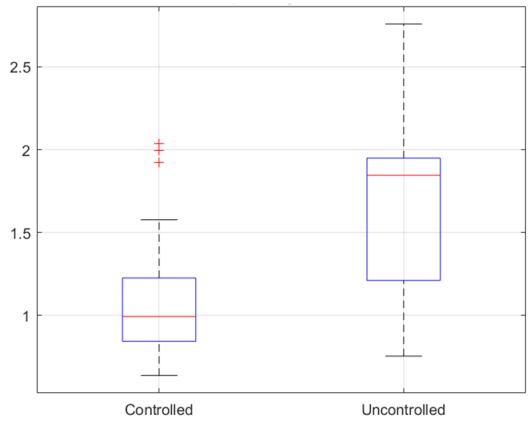


**Figure 8.** Statistical distribution of the speeding score.

## 5 Conclusion

Traffic accidents pose a significant public health challenge, especially in developing countries where many people rely on informal transport, such as minibus taxis. This paper aims to enhance our understanding of the driving behavior of minibus taxi drivers by employing machine learning techniques to compare driving behaviors in controlled versus uncontrolled environments. The primary finding is that a controlled environment fosters aggressive driving, while a less controlled setting encourages speeding. These findings support the eThekwini Municipality's stance in its road safety plan, which identifies driver aggressiveness as a major factor in road traffic accidents in Durban. However, one key factor was not considered in this study: lane-changing behavior. To properly analyze this behavior, attitude data or angular speed measurements are required. This aspect is part of ongoing research, as our database includes accelerometer and gyroscope measurements.

*This work is partly supported by the National Research Foundation South Africa, Grant 499 Number 129925, South Africa/France (Protea) Joint Research Programme-PercepTrans.

## References

[1] C. Staton, J. Vissoci, E. Gong, N. Toomey, R. Wafula, J. Abdelgadir, Y. Zhou, C. Liu, F. Pei, B. Zick et al., Road traffic injury prevention initiatives: a systematic review and metasummary of effectiveness in low and middle income countries, National Library of Medicine **11** (2016).

[2] J. Habyarimana, W. Jack, Heckle and chide: Results of a randomized road safety intervention in Kenya, Journal of Public Economics **95**, 1438 (2011).

[3] Statistics South Africa, Tech. rep., Statistics South Africa (2020)

[4] M. Sinclair, E. Imanirans, Aggressive driving behaviour: The case of minibus taxi drivers in Cape Town, South Africa, in *Proc. of the 34th Southern African Transport Conference* (2015)

[5] A. Sukhaia, P. Jones, A, Understanding geographical variations in road traffic fatalities in South Africa, South African Geographical Journal **95**, 187 (2013).

[6] United Nations Organization, Tech. rep., United Nations (2015)

[7] R. Naumann, L. Sandt, W. Kumfer, S. LaJeunesse, S. Heiny, L. K, Thinking in the context of road safety: Can systems tools help us realize a true "safe systems" approach?, Curr Epidemiol Rep **7**, 343 (2020).

[8] A. Sohail, M. Cheema, M. Ali, A. Toosi, H. Rakha, Data-driven approaches for road safety: A comprehensive systematic literature review, Safety Science, **158**, 105949 (2023).

[9] H. Safarpour, D. Khorasani-Zavareh, , R. Mohammadi, The common road safety approaches: A scoping review and thematic analysis, Chinese journal of traumatology **23**, 113 (2020).

[10] H. Gajjar, S. Sanyal, M. Shah, A comprehensive study on lane detecting autonomous car using computer vision, Expert Systems With Applications **233**, 120929 (2023).

[11] C. Bullard, S. Jones, K. Adanu, E, J. Liu, Crash severity analysis of single-vehicle rollover crashes in Namibia: A mixed logit approach, IATSS Research **47**, 318 (2023).

[12] I. Konkor, M. Kansanga, Y. Sano, K. Atuoye, I. Luginaah, Risk-taking behaviours and timing to first motorbike collision in the upper west region of Ghana, Journal of Transport and Health **12**, 105 (2019).

[13] C. Zhang, Y. Ma, J. Khattak, A, S. Chen, G. Xing, , J. Zhang, Driving style identification and its association with risky driving behaviors among truck drivers based on GPS, load condition, and in-vehicle monitoring data, Journal of Transportation Safety and Security pp. 1–35 (2023).

[14] L. Eboli, G. Guido, G. Mazzulla, G. Pungillo, R. Pungillo, Investigating car users' driving behaviour through speed analysis, PROMET – Traffic and Transportation **29**, 193 (2017).

[15] H. Hooft van Huysduynen, J. Terken, J. Martens, J. Eggen, Measuring driving styles: A validation of the multidimensional driving style inventory, in *Automotive UI '15 Proceedings of the 7th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (2015), pp. 257–264

[16] A. Useche, S, B. Cendales, F. Alonso, C. Pastor, J, L. Montoro, Validation of the multidimensional driving style inventory (MDSI) in professional drivers: How does it work in transportation workers?, Transportation Research Part F: Traffic Psychology and Behaviour **67**, 155 (2019).

[17] C. Wu, C. Sun, D. Chu, Z. Huang, J. Ma, H. Li, Clustering of several typical behavioral characteristics of commercial vehicle drivers based on GPS data mining: Case study of highways in China, Transportation Research Record: Journal of the Transportation Research Board **2581**, 154 (2016).

[18] Y. Feng, S. Pickering, E. Chappell, P. Iravani, C. Brace, Driving style analysis by classifying real-world data with support vector clustering, in *Proc. of the 3rd IEEE Int. Conf. on Intelligent Transportation Engineering (ICITE)* (Singapore, 2018), pp. 264–268

[19] B. Higgs, M. Abbas, A two-step segmentation algorithm for behavioral clustering of naturalistic driving styles, in *16th Int. IEEE Conference on Intelligent Transportation Systems (ITSC)* (The Hague, Netherlands, 2013), pp. 857–862

[20] Y. Ma, K. Tang, S. Chen, J. Khattak, A, Y. Pan, On-line aggressive driving identification based on in-vehicle kinematic parameters under naturalistic driving conditions, Transportation Research Part C: Emerging Technologies **114**, 554 (2020).

[21] W. Wang, J. Xi, A. Chong, L. Lin, Driving style classification using a semisupervised support vector machine, IEEE Transactions on Human-Machine Systems **47**, 650 (2017).

[22] K. Choudhary, A, K. Ingole, P, Smartphone based approach to monitor driving behavior and sharing of statistic, in *Fourth International Conference on Communication Systems and Network Technologies* (Bhopal, India, 2014), pp. 279–282

[23] A. Mohammadnazar, R. Arvin, A. Khattak, Classifying travelers' driving style using basic safety messages generated by connected vehicles: Application of unsupervised machine learning, Transportation Research Part C: Emerging Technologies **122**, 102917 (2021).

[24] eThekwiniMunicipality, Road safety plan, eThekwini Municipality pp. 1–16 (2017).

[25] Y. Li, L. Zhao, L. Rilett, Driving performances assessment based on speed variation using dedicated route truck GPS data, IEEE Access **7**, 51002 (2019).

[26] R. Fu, T. Liu, Y. Guo, S. Zhang, W. Cheng, A case study in China to determine whether GPS data and derivative indicator can be used to identify risky drivers, Journal of Advanced Transportation **16**, 9072531 (2019).