SENSOR DEVICE ON LATERAL PASSING DISTANCE: A CASE STUDY WITH UNSUPERVISED LEARNING MODEL TO ESTIMATE HAZARDOUS AREAS FOR BICYCLE MOBILITY

LUIZ MARCEL SILVA DE MELLO

FACULDADE DE TECNOLOGIA

UNIVERSIDADE DE BRASÍLIA

UNIVERSIDADE DE BRASÍLIA
FACULTY OF TECHNOLOGY
DEPARTMENT OF CIVIL AND ENVIRONMENTAL ENGINEERING

SENSOR DEVICE ON LATERAL PASSING DISTANCE: A CASE STUDY WITH UNSUPERVISED LEARNING MODEL TO ESTIMATE HAZARDOUS AREAS FOR BICYCLE MOBILITY

LUIZ MARCEL SILVA DE MELLO

ADVISOR: MICHELLE ANDRADE

MASTER'S THESIS IN TRANSPORTATION

BRASÍLIA/DF: October/2023

UNIVERSIDADE DE BRASÍLIA
FACULTY OF TECHNOLOGY
DEPARTMENT OF CIVIL AND ENVIRONMENTAL ENGINEERING

SENSOR DEVICE ON LATERAL PASSING DISTANCE: A CASE STUDY WITH UNSUPERVISED LEARNING MODEL TO ESTIMATE HAZARDOUS AREAS FOR BICYCLE MOBILITY

LUIZ MARCEL SILVA DE MELLO

MASTER'S THESIS SUBMITTED TO THE GRADUATE PROGRAM IN TRANSPORTATION OF THE DEPARTMENT OF CIVIL AND ENVIRONMENTAL ENGINEERING OF THE FACULTY OF TECHNOLOGY, AT THE UNIVERSITY OF BRASÍLIA, AS PART OF THE REQUIREMENTS TO OBTAIN THE MASTER'S DEGREE IN TRANSPORTATION.

APPROVED BY:

_____

MICHELLE ANDRADE, Ph.D. (UnB)
ADVISOR

_____

LI WEIGANG, Ph.D. (UnB)
INTERNAL EXAMINER

_____

JORGE TIAGO BASTOS, Ph.D. (UFPR)
EXTERNAL EXAMINER

BRASÍLIA/DF, October, 2023.

CATALOG FORM

MELLO, LUIZ MARCEL SILVA de

Sensor Device on Lateral Passing Distance: a case study with unsupervised learning model to estimate hazardous areas for bicycle mobility. Brasília, Distrito Federal, 2023. xii, 63p., 210 x 297 mm (ENC/FT/UnB, Master, Transportation, 2023).

Master's Thesis – University of Brasília. Faculty of Technology. Department of Civil and Environmental Engineering.

1. Bicycle Safety
2. Lateral Passing Distance
3. Machine Learning
4. Intelligent transportation
I. ENC/FT/UnB
II. Título (série)

REFERENCE

MELLO, L. M. S. DE (2023). Sensor Device on Lateral Passing Distance: a case study with unsupervised learning model to estimate hazardous areas for bicycle mobility. Publicação T.DM-009/2023. Departamento de Engenharia Civil e Ambiental, Universidade de Brasília, Brasília, DF, 74 p.

COPYRIGHT

Luiz Marcel Silva de Mello - marcelsmello@gmail.com

Anexo SG-12, 1° andar. Campus Universitário Darcy Ribeiro

Asa Norte: Brasília: Distrito Federal: 70910-900: Brasil; e-mail: ppgt@unb.br

“There is never any end. There are always new sounds to imagine and new feelings to get at. And always, there is the need to keep purifying these feelings and sounds so that we can see what we have discovered in its pure state. So that we can see more and more clearly what we are. In that way, we can give those who listen to the essence the best of what we are. But to do that at each stage, we have to keep on cleaning the mirror.”

- John Coltrane

ABSTRACT

Sensor Device on Lateral Passing Distance: a case study with an unsupervised learning model to estimate hazardous areas for bicycle mobility

Commuting by bicycle is widely increasing worldwide. As an active transportation mode, cycling can potentially reduce traffic congestion and air pollution. Also, promoting an active lifestyle can improve public health and make cities more human-friendly. Although, the quantity of occurrences and fatalities with cyclists is still worrisome.

The Surrogate Safety Measures (SSM) are promising indicators for assessing traffic safety with measures based on these traffic conflicts. The word "surrogate" is used because the measures are based not on crashes but on traffic conflicts. Also, network screening ensures an efficient identification of hazardous sites to reduce the number and severity of crashes. This methodology can be conducted using either a reactive or a proactive approach. Regarding the proactive approaches, bicycles instrumented with sensors became increasingly usable for research in the mobility field. By using a portable and multi-functional sensing device is possible to collect bicycle trajectory data and Lateral Passing Distance (LPD) using various sensors connected to a database system.

Therefore, the current research aims to estimate and define hazardous areas for active mobility by applying unsupervised machine learning algorithms (k-means and DBSCAN) based on a sensor device for data collection. The Lateral Passing Distance (LPD) results collected between bicycles and vehicles were related to the cyclist data. Beyond the clustering investigation, a correlation between the features has identified how the data interacted among them.

Some of this data includes velocity, curse elevation, altitude, accelerometer, and gyroscopic information from a field operational data collection on the street. The methodology was applied to a case study regarding the Brasília city center avenue with a shared pathway around the local City Park. Therefore, this study aims to propose a methodological and data-driven approach to bicycle safety using machine learning algorithms. Regarding the general data, 25% of the readings are less than 139.62cm for the LPD. When the clustering model was applied, 25% of the LPD readings were less than 100.13cm; for the second quartile, 50% were less than 193.69cm. It indicates critical LPD for one of the clusters with 75 readings, considering the threshold of 150cm for the minimal lateral clearance distance law adopted in Brazil.

Keywords: Bicycle Safety; Lateral Passing Distance; Machine Learning; Intelligent Transportation.

RESUMO

Dispositivo Sensor Na Distância Lateral De Ultrapassagem: Um Estudo De Caso Com Modelo De Aprendizado Não Supervisionado Para Estimar Locais De Risco À Mobilidade Por Bicicleta

A locomoção por bicicleta está aumentando significativamente em todo o mundo e como um modo de transporte ativo o ciclismo pode potencialmente reduzir o congestionamento do tráfego e a poluição do ar. Além disso, promover um estilo de vida ativo pode melhorar a saúde pública e tornar as cidades mais amigáveis para as pessoas. No entanto, a quantidade de ocorrências e fatalidades envolvendo ciclistas ainda é preocupante.

As Medidas de Segurança Substitutas (SSM) são indicadores promissores para avaliar a segurança no tráfego com base em conflitos, em vez de sinistros e acidentes. O termo "substitutas" é usado porque as medidas se baseiam em conflitos de tráfego e não em sinistros. Além disso, a triagem de rede garante uma identificação eficiente de locais perigosos para reduzir o número e a gravidade dos sinistros. Portanto, essa metodologia pode ser realizada tanto por uma abordagem reativa quanto por uma abordagem proativa. No que diz respeito às abordagens proativas, bicicletas instrumentadas com sensores tornaram-se cada vez mais úteis para pesquisas no campo da mobilidade. Por meio de um dispositivo de detecção portátil e multifuncional, é possível coletar dados de trajetória de bicicletas e a Distância Lateral de Ultrapassagem (LPD) usando variados sensores conectados a um sistema de banco de dados.

Dessa forma, o presente estudo tem como objetivo estimar áreas de risco para a mobilidade ativa, aplicando algoritmos de aprendizado de máquina não supervisionado (K-Means e DBSCAN) com base em dispositivos sensores para a coleta de dados. Os resultados da Distância Lateral de Ultrapassagem (LPD) coletados entre bicicletas e veículos foram então relacionados aos dados do ciclista. Além da pesquisa de clusterização, foi realizada a correlação entre as características para identificar como os dados interagiam entre si.

Alguns desses dados incluem velocidade, elevação do percurso, altitude, informações do acelerômetro e giroscópio a partir de uma coleta de dados naturalística na rua. A metodologia foi aplicada a um estudo de caso em uma avenida do centro da cidade de Brasília, em torno do Parque da Cidade. Por fim, este estudo objetiva uma aplicação metodológica e analítica de segurança cicloviária orientada a dados através da utilização de algoritmos de aprendizado de máquina. Em relação aos dados gerais, 25% das leituras relativas à distância lateral de passagem tiveram menos de 136,36 cm. Quando o modelo de agrupamento é aplicado, 25% dessas leituras tiveram menos de 100,13 cm; para o segundo quartil, 50% tiveram menos de 193,69 cm. Isso indica uma LPD crítica para um dos grupos com 75 leituras, quando considerado o limite de 150 cm para a distância mínima de afastamento lateral estabelecido pela legislação brasileira.

Keywords: Segurança da Bicicleta; Distância Lateral de Ultrapassagem; Aprendizado de Máquina; Transporte Inteligente.

CONTENTS

LIST OF TABLES:

LIST OF FIGURES:

1 INTRODUCTION

1.1 CONTEXT AND BACKGROUND

Commuting by bicycle is widely increasing worldwide. As an active transportation mode, cycling can potentially reduce traffic congestion and air pollution. Also, promoting an active lifestyle can improve public health and make cities more human-friendly (VANPARIJS *et al*., 2015; MUELLER *et al*., 2015; LI *et al*., 2015). Although, the quantity of occurrences and fatalities with cyclists is still worrisome (WORLD HEALTH ORGANIZATION, 2018). Due to the high number of crashes with active mobility, a practical approach is vital to provide greater confidence for decision-making regarding road safety involving the overtaking of bicycles by vehicles.

The types of crashes involving active mobility (e.g., pedestrians and bicyclists) yield small sample sizes that can result in inconclusive or unreliable crash-based safety evaluations. These small sample can be due to the lack of registration on accidents. Therefore, many safety professionals and engineers have adopted the Traffic Conflicts Technique (TCT) to measure the conflict's severity and recommend corrective actions for crash prevention. Thus, the Surrogate Safety Measures (SSM) are promising indicators for assessing traffic safety with measures based on these traffic conflicts (LORD *et al*., 2021). The word “surrogate” is used because the measures are based not on crashes but on conflicts.

In this regard, critical Lateral Passing Distance (LPD) events between motor vehicles and bicycles have the potential to be used as bicycle-oriented SSM indicators for safety evaluation. Some SSMs were already validated with crash frequency and the number of cyclists injured, demonstrating a positive and moderate correlation between critical LPD events and crashes (BERNADES *et al*., 2023).

The LPD between bicycles and motor vehicles is a crucial perspective for cyclist safety (DOZZA *et al*., 2016; LAMONDIA & DUTHIE, 2012). Feeling safe and comfortable is essential to the extensive use of the facilities. In this context, a recent systematic review focused on the factors influencing the LPD between bicycles and motorized traffic (Rubie *et al*., 2020). It indicates that on-road vehicle-cyclist passing distances have been investigated in several

previous studies (WALKER, 2007; LOVE *et al*., 2012; SAVOLAINEN *et al*., 2012; CHUANG *et al*., 2013; WALKER *et al*., 2014; MEHTA *et al*., 2015; HAWORTH *et al*., 2018; BECK *et al*., 2019; AMPE *et al*., 2020; FEIZI *et al*., 2021; MACKENZIE *et al*., 2021).

Moreover, concerning the overtaking measurement, various research objectives show that instrumented bicycles are a valuable, effective, and critical tool in cycling safety research. Also, it was possible to distinguish factors that have been indicated to affect passing behavior (GADSBY & WATKINS, 2020; FEIZI *et al*., 2021; MACKENZIE *et al*., 2021).

Furthermore, Machine Learning (ML) and Internet of Things (IoT) techniques in Intelligent Transport Systems (ITS) can conduct the studies for data-driven safety analysis. Thus, unsupervised Learning algorithms have been applied in hazardous site identification, like clustering models and Kernel Density Estimation (KDE), as some standard geospatial methods for analysis (NGUYEN *et al*., 2018; ZANTALIS *et al*., 2019; WANG *et al*., 2019; ZANTALIS, F. *et al*., 2019; LORD *et al*., 2021). The clustered areas are classified as hazardous spots for safety improvement.

The use of technology has made connectivity between various transportation system elements attainable. Likewise, with the proliferation of devices, sensors, and open-source information, it is possible to develop devices that address different data collection challenges and improve the ITS perspective (ANG & SENG, 2016; GUERRERO-IBÁÑEZ *et al*., 2018; LIM *et al*., 2018; OZBAY *et al*., 2018; BERNARDES *et al*., 2019; BERNARDES & OZBAY, 2023).

Lastly, network screening ensures an efficient identification of hazardous sites to reduce the number and severity of crashes. This methodology can be conducted using either a reactive or a proactive approach. The reactive approach relies on analyses of historical crash data. In contrast, the proactive approach relies on analyses and identification of geometric and operational characteristics highly associated with crash risk but not necessarily with crashes (AASHTO, 2010). The present research uses the proactive approach in the analysis to develop a methodology to identify high risk spots for cyclists, as could be checked in the following objectives session.

1.2 OBJECTIVE

Bicycles instrumented with sensors became increasingly usable for research in the mobility field. Using a portable and multi-functional sensing device is possible to collect bicycle trajectory data and Lateral Passing Distance (LPD) using various sensors connected to a database system or a memory card (GADSBY *et al*., 2020; BERNARDES *et al*., 2019).

Therefore, the current research aims to estimate and define hazardous areas for bicycle mobility by applying unsupervised machine-learning algorithms based on a sensor device for data collection. The Lateral Passing Distance results collected among bicycles and vehicles were related to the cyclist data. Some of this data includes velocity, curse elevation, accelerometer, and gyroscopic information through a field operational data collection on the street. The methodology was applied to a case study in Brasília city center avenue regarding a shared pathway around the local City Park. Specific objectives include:

- Collect field operational data using the BSafe360 Sensor Device for the specified research location;
- Procedure a data treatment and wrangling for exploratory data analysis (EDA);
- Apply unsupervised Machine Learning algorithms to cluster various features related to the dataset and estimating hazardous areas;
- Compare the performance for different machine learning models on clustering;
- Elaborate a Systematic Literature Review on the Lateral Passing Distance collection procedure with ultrasonic sensors and Internet of Things.
- Make all the algorithms and script manipulation of this study publicly available on the internet for future research.

1.3 METHODOLOGICAL RESEARCH AND SESSIONS

The last section presented the research objectives, introducing the context and methods employed. The following session presents the systematic literature review for the Lateral Passing Distance (LPD) collection procedure evolving the databases, criteria, and temporal approach used. Then the paper's key findings are presented as well as the data collection setups and data analysis procedures, and hardware and software chosen are addressed.

Also, a literature review regarding Machine Learning technics is addressed. The use of unsupervised learning models to cluster risk areas for mobility is presented. The following section shows the data collection methodology with the area characterization for the study and the description of the field procedure. Finally, a discussion about the results and approaches, as well as a comparison between clustering models is presented, besides the research conclusions and limitations, followed by topics for future research. The research schema is presented in the Figure 1.1, bellow.

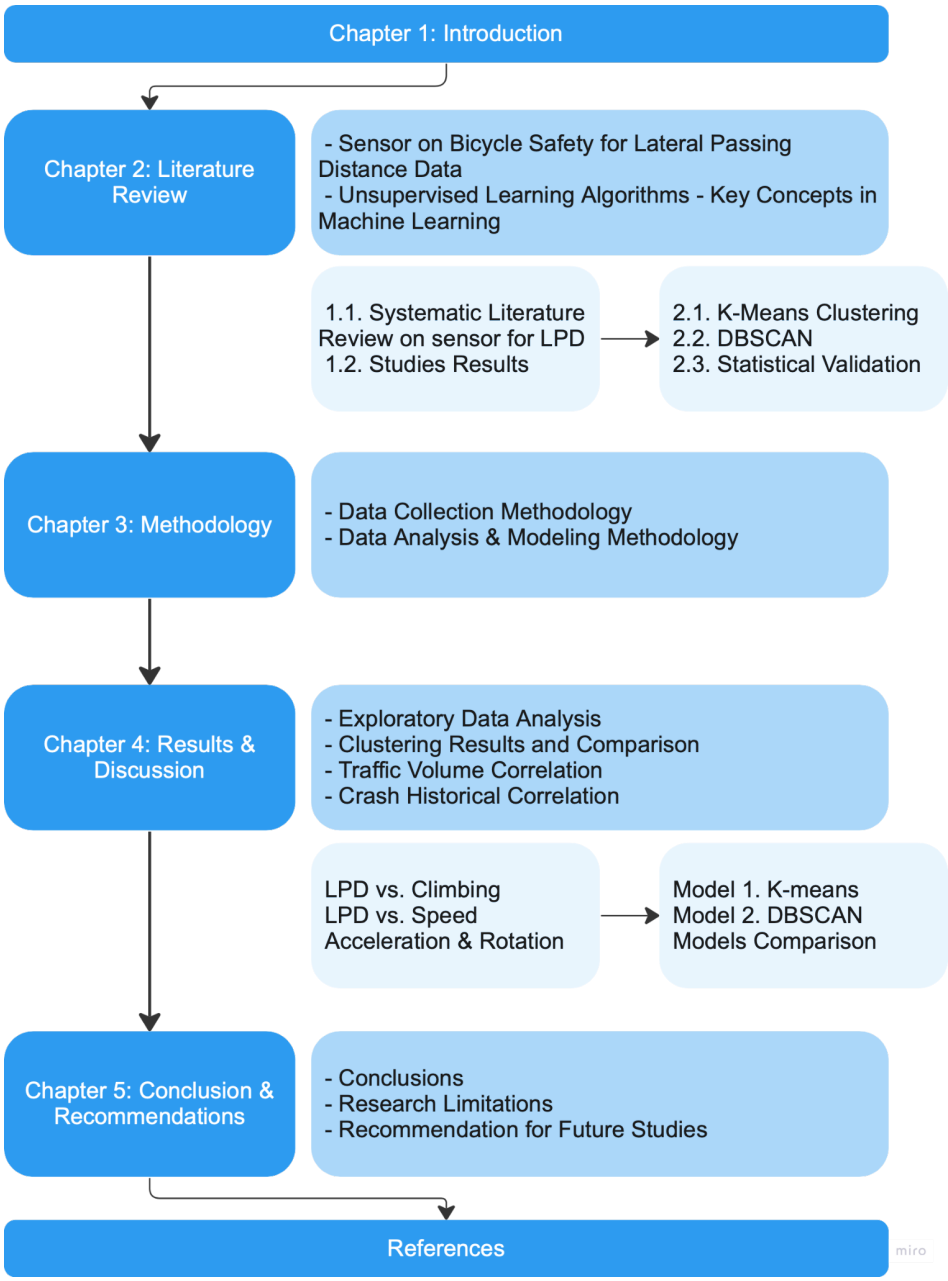


Figure 1.1 Research schema

2 LITERATURE REVIEW

2.1 REVIEW INTRODUCTION

Traffic conflict events were defined once as "*a driver takes evasive action, brakes or weaves to avoid a collision*" (PERKINS & HARRIS, 1967). The Traffic Conflicts Technique (TCT) procedure was conceived as a systemic approach for observing and measuring crash potential. Since Perkins & Harris, TCT has gained popularity as a diagnostic tool used to determine appropriate safety countermeasures at high crash locations and as an evaluative tool for safety treatments.

Thus, the Surrogate Safety Measures (SSM) are promising indicators for assessing traffic safety with measures based on these traffic conflicts (LORD *et al*., 2021; BERNARDES *et al*., 2023). The word "surrogate" is used because the measures are based not on crashes but on traffic conflicts. In this regard, critical Lateral Passing Distance (LPD) events between motor vehicles and bicycles have the potential to be used as bicycle-oriented SSM indicators for safety evaluation.

The following contents covered in this session are based on a systematic literature review regarding sensors in Lateral Passing Distance measurement, considering publications until 2021. This systematic review limited this data and aimed to include only studies with collecting procedures in field realized until the pandemic moment: march 2020. The isolation policies regarding the COVID-19 pandemic in the following years could influence and affecting practical researches on streets due to the increase of remote work and new routines, changing the commuting general behavior.

Furthermore, based specially on an extensive literature review on this theme (RUBIE *et al*., 2020), the book "Highway Safety Analysis and Modeling" (LORD *et al*., 2021); a paper review on Machine Learning and IoT in Smart Transportation (ZANTALIS *et al*., 2019); and in a validate proposal correlation between LPD and crashes historical data (BERNARDES *et al*., 2023).

2.2 SENSOR ON BICYCLE SAFETY FOR LATERAL PASSING DISTANCE DATA

2.2.1 Databases

For this study, three databases were used in the methodology: Google Scholar, Scopus, and Web of Science. The review followed the preferred reporting items for systematic reviews and meta-analysis protocols (PRISMA-P) 2015 statement (MOHER *et al*., 2015).

The first database consulted was Google Scholar. It was selected to broaden the search due to the more significant number of results. The term "*Bicycle Passing Distance*" was first tried and resulted in 165.000 documents. After that, the string combination "*bicycle* OR *cyclist*" AND "*passing distance* OR *overtaking*" was inserted and resulted in 408 documents. These documents are listed in order of relevance from keywords, and selected meaning 10% of the sample with 49 documents. This selection was exported in .CSV file. From these 49 documents, a new term, "*sensor*," was inserted to filter how many were explicitly measured regarding it. The new search resulted in 22 documents.

The second, the Scopus database, was searched into "TITLE-ABS-KEY" (titles, abstracts, and keywords) terms. The combination of strings and terms used was: (*bicycle* OR *cyclist*) AND (*passing* OR *overtaking* AND (*distance*)). It resulted in 164 documents. Moreover, the added term "*sensor*" generated a new result, delimiting the articles into twelve. The third one was the Web of Science database. The primary collection was searched through topics in the same combination: (*bicycle* OR *cyclist*) AND (*passing distance* OR *overtaking*), resulting in 133 articles. Furthermore, with the addition of the term "*sensor*," the final result was eight articles. The search was finalized in May 2021. This study limited this data and aimed to include only research with collecting procedures until the pandemic moment, march 2020, concerning the probable influence on that.

A keyword delimitation methodology is represented in Figure 2.1. It also tested the use of the keyword "*Passing behavior*," but it was perceived that it did not return a significant difference in the final result.

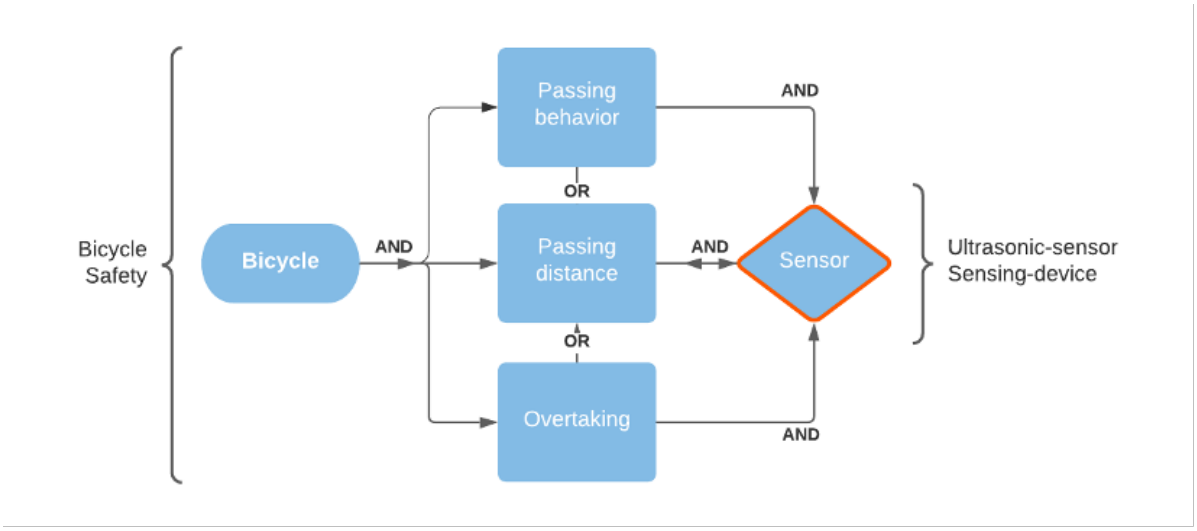


Figure 2.1 Keyword delimitation methodology

2.2.2 Inclusion Criteria

The results from the databases were merged, and 10 duplicate articles were excluded. The literature review included research published in English and in a peer-reviewed format. Additionally, the primary method for measuring passing distance was required to be the use of an ultrasonic sensor device. GPS sensors or video records were considered only as supplementary sources in this search, not as primary methods. GPS sensors and video recordings were considered supplementary in this search, rather than primary sources. Additionally, articles employing alternative approaches, such as simulators, or focusing on a driver's perspective or other perspectives unrelated to bicycle safety, were excluded. Reports, documentation, and Ph.D. these were likewise not taken into account.

Regarding the COVID-19 pandemic in the early 2020s, this research focused on studies that provided practical procedures in the field and limited the search for publications up until the beginning of 2021, focusing on procedures conducted prior to the pandemic. The isolation policies related to the COVID-19 pandemic in the subsequent years could have influenced and impacted some practical research on streets due to the rise of remote work and new routines, altering commuting behavior and limiting regular displacements. Therefore, a total of 15 papers were eligible for inclusion. The systematic review methodology through the databases is presented in Figure 2.2, and the totals follow the keywords and strings used.

The kind of studies concerning the perspective of the driver was not considered in the present review. Black *et al*. (2020) used vehicle-mounted ultrasonic sensors, and the findings suggest

that bicyclists should incorporate additional visibility aids to encourage safer passing distances of vehicles at nighttime. Rossi *et al*. (2021) found that an onboard real-time coaching program can improve the safety of maneuvers involving passing cyclists using vehicle-mounted ultrasonic sensors. These approaches can lead toward to Autonomous Vehicles (AV) theme regarding the interaction between cars and bicycles.

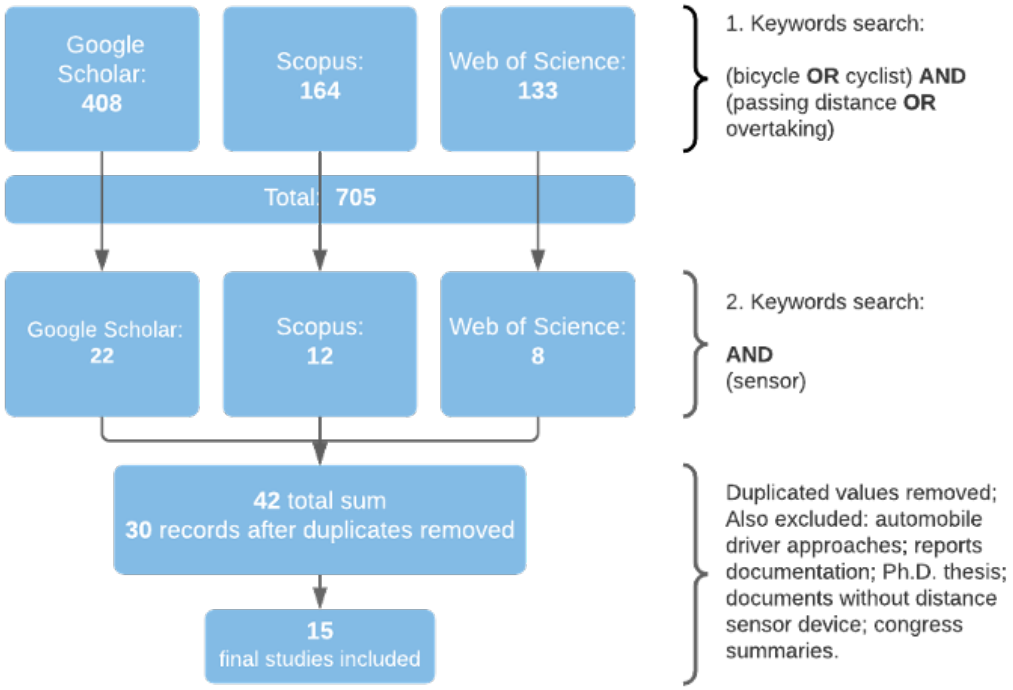


Figure 2.2 Systematic Review Methodology through bases, keywords, and strings

2.2.3 Keywords and temporal data analysis

This review covers a roughly decade-long timeframe. Since 2007, research involving instrumented bicycles has become more commonplace, but 2 articles were published yearly. However, a significant shift occurred in 2013 when the use of instrumented bicycles gained momentum, resulting in a minimum of nine articles per year (MACKENZIE *et al*., 2021; GADSBY, 2020).

Up until then, the majority of research involving instrumented bicycles primarily relied on video recordings and the use of Global Positioning System (GPS). The use of ultrasonic sensor devices to measuring passing distance is a relatively recent development. Figure 2.3 shows the correlation between the index and the author's keywords concerning included articles classified by time.

By using the VOS viewer software, a bibliographic data map was created. The totality of the articles has been joined into a unique base on the Scopus database, and a file was generated with the result of all searches from the combination of Google Scholar, Scopus, and Web of Science. It applied analysis of co-occurrence filtering the keywords (author's and Index's ones). It used the whole counting method, and the minimum number of keywords occurrences chosen was two. It results in a total of fifty-two keywords meeting the threshold. The lines between the points show the strength of the co-occurrences. The total author's keywords are 51, and the total Index keywords are 186.

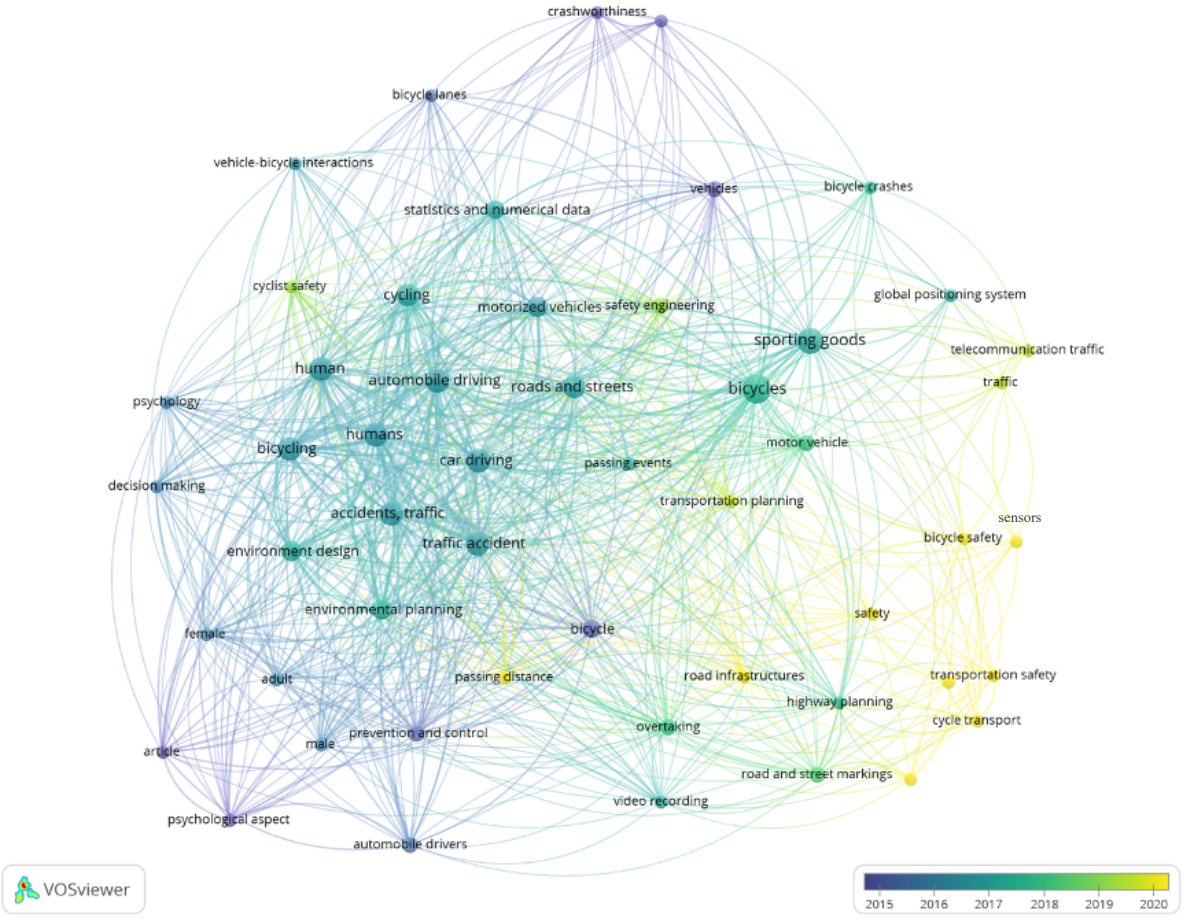


Figure 2.3 Temporal theme analysis regarding the author's and index keywords

By employing the freely available software VOS viewer (BLACK *et al*., 2020) was possible to trace the historical evolution of sensor usage in the context of the passing distance and bicycle safety. Notably, the yellow area on the right of the graph represents the most recent period, and the keyword "sensors" is closely associated with these recent papers, dating around 2019 and 2021. This trend illustrates the increasing research interest in recent years, likely influenced by the emergence of Smart Cities, the Internet of Things (IoT), Autonomous Vehicles, Machine

Learning Techniques, and the growing utilization of sensors for behavior studies, bicycle safety, and transportation planning. Figure 2.4 displays the most related terms linked to the keyword “sensors” during these recent periods.

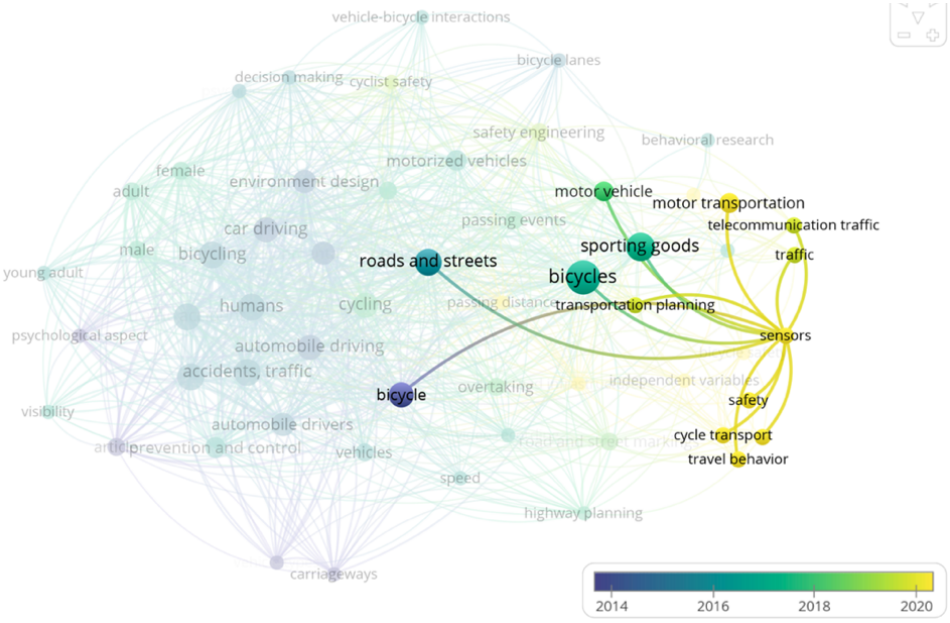


Figure 2.4 Keyword relationship regarding the timeline

Concerning the publication per country, Figure 2.5 illustrates the contributions of eight countries. The United States (USA) maintained a prominent position in research related to passing distance using ultrasonic sensor devices. Although, the United Kingdom (UK) has been widely cited in research throughout this decade, particularly for a naturalist experiment that employed an instrumented bicycle to collect proximity data from overtaking motorists (WALKER, 2007) with 158 citation indexes (SCOPUS, 2021).

Additionally, this author has numerous prior publications in the field of bicycle safety. In addition to the UK and USA, Australia, Canada, Belgium, Portugal, Taiwan, and Sweden have also made contributions to this area. Figure 2.5 further displays the publication by country in terms of the number of citations.

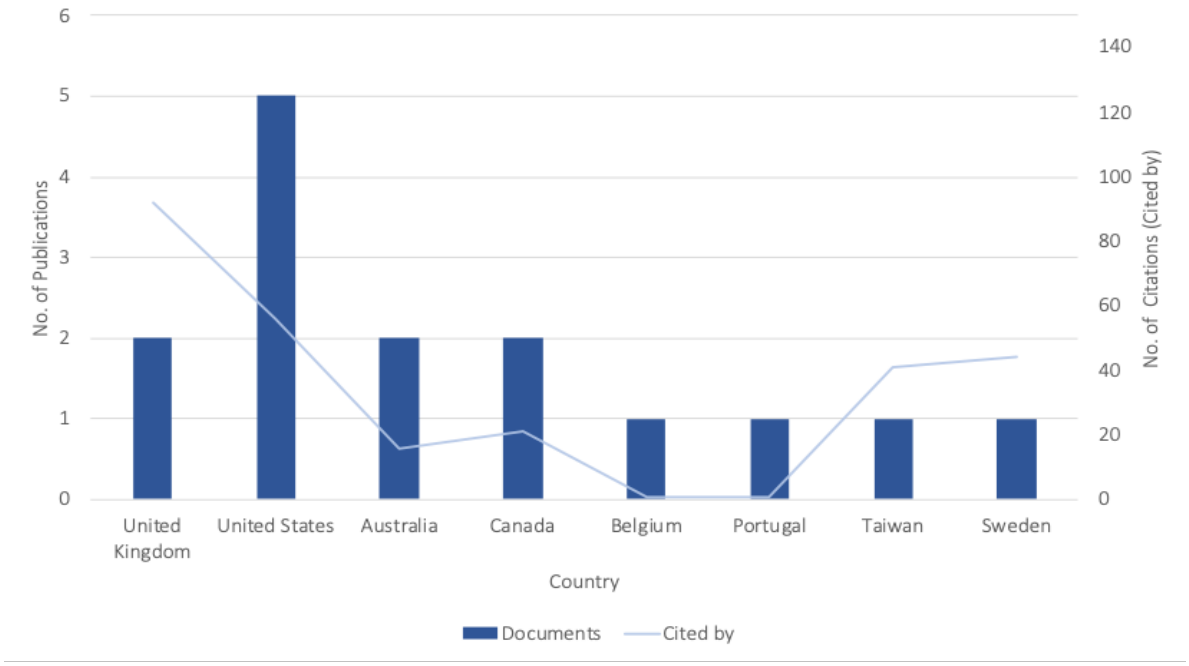


Figure 2.5 No. of publications and citations per country

The number of publications was organized in Figure 2.6 concerning the journal source. The SCImago Journal score was added to compare which of these publications has significant ratings around the last three years of ranking. The "Accident Analysis and Prevention*"* was the journal with the most significant number of publications in this review (7 papers); however, the "Transport Reviews" is the one with the highest SCImago score, and it keeps increasing through the years. Moreover, Figure 2.7 presents a geographic distribution of the publications.

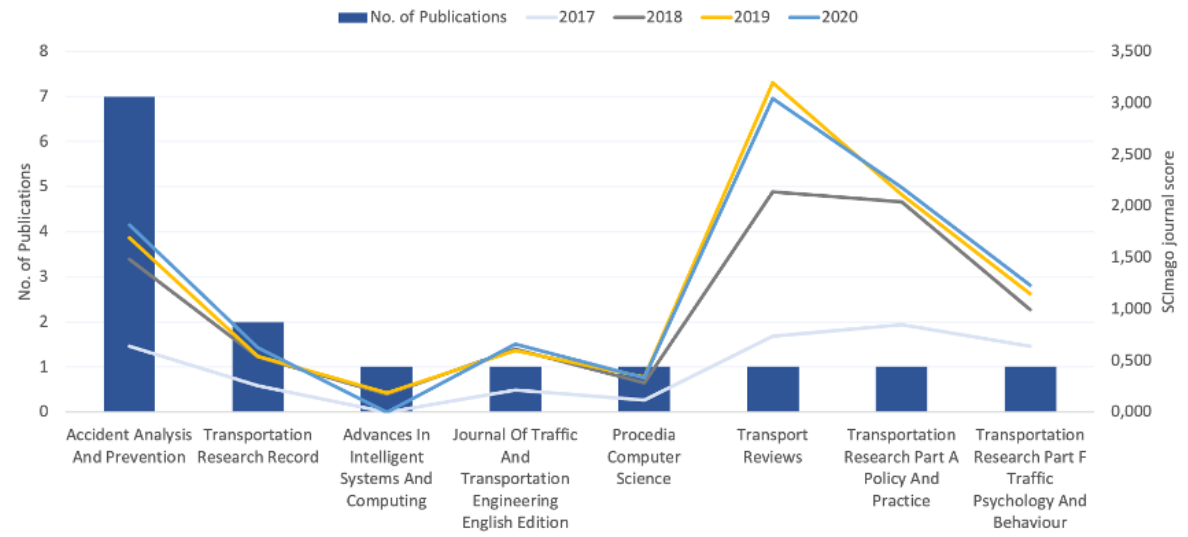


Figure 2.6 Number of Publications per Journal and year concerning the SCImago impact factor score

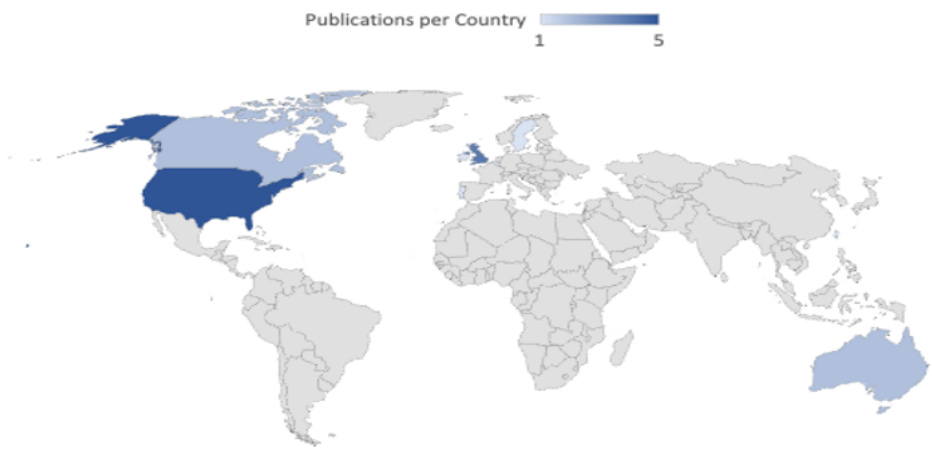


Figure 2.7 Geographic distribution of the countries during the last decade, per publication

2.2.4 Studies Results and Methodologies

The vehicle-cyclist passing distance has been investigated in several previous studies, and naturalistic driving data provide valuable information about how car drivers overtake cyclists (VAN ECK & WALTMAN, 2010). Rubie *et al*. (2020) identified some of these studies concerning the passing event. Gadsby & Watkins (2020), Feizi *et al*. (2021), and Mackenzie *et al*. (2021) focus on the factors influencing the passing distance. The majority of these papers are from the decade chosen, mainly from the last years until 2021. The ones which specifically approach the use of ultrasonic sensors to measure the lateral distance will be present in this section. Table 2.1 shows the main characteristic of the studies in this review.

The studies included in this review primarily focused on naturalistic experimental research that utilized instrumented bicycles before the COVID-19 pandemic to investigate various aspects of passing behavior between vehicles and cyclists in relation to Lateral Passing Distance (LPD). The rationale for selecting this specific period was explained earlier.

It is worth noting that several additional studies related to this approach were analyzed, which contributed to a deeper understanding of topics such as crashes between automobiles and bicycles, clearance distance laws, machine learning techniques for traffic crashes, driver behavior during overtaking, and bicycle safety. However, these studies were not included in the present review due to specific criteria and methodology employed for the systematic review. To present the studies in a chronological order spanning the decade from 2011 to 2021, this section provides a brief summary of each study.

Table 2.1 Characteristic methodologies of the studies

Study	No. of events	No. of trips	No. of riders	Distance (km)	Distance Sensor	GPS	Video
Mackenzie *et al*. (2021)	16,476	465	23	6,531	✓	✓	
Feizi *et al*. (2021)	2,838	Unknown	2	Unknown	✓		✓
Ampe *et al*. (2020)	1,423	19	2	Unknown	✓	✓	
Gadsby & Watkins (2020)	Unknown	75	Unknown	Unknown	✓	✓	✓
Beck *et al*. (2019)	18,527	422	60	5,302	✓	✓	✓
Mehta *et al*. (2015; 2019)	5,227	Unknown	Unknown	Unknown	✓	✓	✓
Bernardes *et al*. (2019)	74	1	1	1.45	✓	✓	✓
Dozza *et al*. (2016)	235	Unknown	2	84.6	✓	✓	✓
Walker *et al*. (2014)	5,69	67	1	Unknown	✓		✓
Chapman & Noyce (2012; 2014)	1,300	Unknown	Unknown	Unknown	✓	✓	✓
Shackel & Parkin (2014)	500	Unknown	Unknown	Unknown	✓		✓
Chuang *et al*. (2013)	1,380	34	34	Unknown	✓	✓	✓
Bahmankhah *et al*. (2020)	Unknown	Unknown	Unknown	Unknown	✓	✓	✓

Initially, Chapman & Noyce (2012) collected real-time interaction data between bicycles and motorized vehicles on rural roads. Vehicles types were also observed and counted. They found that drivers operated in a technically unsafe manner by frequently performing passing maneuvers outside designated areas. The study also found that bicycle lanes (paved shoulders) directly affected the likelihood of a driver committing a moving violation, with violation rates four to six times lower when a paved shoulder was available.

Then, Chuang *et al*. (2013) identified factors influencing motorists' passing decisions and cyclists' behaviors. Bicyclists exhibited less stability when passed by buses, and longer passing times led to more cautious but less stable riding behaviors.

In the next year, Walker *et al*. (2014) tested different outfits worn by bicyclists and found that altering appearance had limited influence on passing distances. Infrastructure improvements and education may be more effective in promoting safe passing. A motive for this study was to test Walker's (2007) hypothesis that the reduced passing proximities seen when a bicyclist wore a helmet might have been caused because drivers take helmeted riders to be more experienced or in control.

Right away, Chapman & Noyce (2014) developed a model showing how driver behavior varied with road characteristics during overtaking maneuvers involving bicycles. The model showed that driver behavior can be adjusted by including or excluding geometric elements. It is these geometric elements, such as road grade, shoulder presence and width, marked centerline, and road design speed that significantly affect how drivers use a rural roadway, especially when overtaking a bicycle.

Then, Shackel & Parkin (2014) also reported that the results provide evidence for the design and management of roads to better accommodate cyclists. The research presented is based on previous research and fills gaps considering the presence of bike lanes on 20 mph and 30 mph roads, different lane widths, different lane markings, type of vehicle, vehicle squad and traffic in the opposite direction. It concluded that lower speed limits and removing centerline markings could reduce overtaking speeds and increase comfort for cyclists.

Mehta *et al*. (2015) found that passing distances were significantly smaller on roads without dedicated bike lanes, leading to more unsafe passing maneuvers. The setup installed was capable to measure the lateral distance when overtaking takes place, as well as capturing the location, bicycle speed, and event time. It was found that the lateral separation between cyclists and motor vehicles is significantly smaller on facilities without exclusive bike lanes. For two-lane facilities without bike lanes, 12% of all passing maneuvers were unsafe, compared with only 0.2% unsafe passing maneuvers. These results suggest that introducing dedicated bike lanes not only improves safety for cyclists but also reduces the number of potential conflicts between motorized vehicles that arise from lane-changing or encroaching vehicles that are passing cyclists.

Afterwards, Dozza *et al*. (2016) used a LIDAR and two cameras to assess driver behavior during overtaking maneuvers. A LIDAR is a system consisting of a laser beam rotating at high speed to scan the environment. That sensor mounted on an instrumented bicycle provided continuous and high-resolution information about the overtaking maneuver, making it possible to identify and analyze critical phases of the overtaking maneuver along with their corresponding driver comfort zones. Oncoming vehicles had the greatest impact on passing maneuvers. Neither vehicle speed, lane width, shoulder width, nor posted speed limit significantly affected the driver comfort zone or the overtaking dynamics in this study.

Hereafter, Mehta *et al*. (2019) estimated the number of expected unsafe passing events on roads without bike lanes. Drivers provided smaller passing distances during restricted passing events, and the proposed method can also be used to evaluate the 'cycling safety level of service' on different road categories. It found that the probability of observing unsafe passing events on urban arterials without on street bicycle lanes is much higher when passing events are restricted. It was observed that when on-street bike lanes are not available: 1) drivers tend to provide smaller passing distances during restricted passing events; 2) a much higher proportion (29%) of restricted passing events were unsafe, compared to that of unrestricted passing events (11%); and 3) a much higher proportion of unrestricted passing events (73%) were encroachment or far lane passing compared to that of restricted passing events (38%).

Beck *et al*. (2019) found that a significant proportion of passing events were close passing events, and the introduction of dedicated bike lanes improved cyclist safety. Beck *et al.* (2019) have also identified that road infrastructure had a substantial influence on the distance that motor vehicles provide when passing cyclists. They used a hierarchical linear model to investigate the relationship between a motor vehicle and infrastructure characteristics and passing distance. It concluded that from a large sample of events in which a motor vehicle passed a cyclist, one in every 17 passing events was a close passing event (<100 cm), and in higher speed zones (over 60 km/h), one in every three was a close passing event (<150 cm).

Thereat, Bernardes *et al.* (2019) developed a portable sensing device to collect bicycle safety data and identified locations where drivers approached bicycles closely. The use of cheaper and smaller components resulted in a product that is portable and proper for mass data collection. Gadsby *et al.* (2020) also investigated the different tools of measurement to study behavior, safety, and maintenance using a range of sensors like GPS, cameras, ultrasonics, LIDAR, gyroscope, and go on. It mentioned that there are benefits and trade-offs for each choice. Lateral distance sensors cannot tell the researcher about the type of vehicle, but the data can be process faster.

At this time, Bahmankhah *et al.* (2020) proposed a methodology to estimate the human power required for cycling but lacked detailed information for further investigation. And, Ampe *et al.* (2020) used a mixed-effect regression representing a cyclist without a child, a cyclist with a child bike seat, and a cyclist with a child bike trailer and found that drivers adapt their passing

behavior when overtaking cyclists with children, but this effect varies depending on time of day and traffic density. At long last, Feizi *et al.* (2021) showed that passing distances were greater in locations with a five-foot passing law and on 3-lane roadways compared to 2-lane roadways.

Moreover, through regression analyses, Mackenzie *et al.* (2021) showed differences between passing distance and compliance with the minimum passing distance when associated with road classification, bike lane presence, and speed limit. 12.3 % of non-compliant passing events were identified on roads zoned more than 60 km/h (high-speed roads). On roads zoned 60 km/h or less (low-speed roads), there were 2.8 % non-compliant. Passing distances were generally greater on roads with a lower (hierarchy) classification. The presence of a bike lane was found to increase the average passing distance across all the road classifications.

Road classification is likely to be associated with a number of factors that have previously been identified as having an influence on passing distance such as number of lanes (MEHTA *et al.*, 2015), lane width (LOVE *et al.*, 2012), speed differential (CHUANG *et al.*, 2013), presence of parked vehicles (BECK *et al.*, 2019), and presence of oncoming vehicles (KAY *et al.*, 2014; MEHTA *et al*., 2015), and type of centreline (SAVOLAINEN *et al*., 2012; KAY *et al*., 2014).

Also, Feizi *et al*. (2021) explored the effects of bicycle facilities, the number of lanes, passing distance laws, and vehicle type. They demonstrated that overtaking distances in the locations with a five-foot passing law were significantly greater than those with a three-foot law or no specific law, in all types of roadway configuration. Besides, analysis using a two-sample t-test mean comparison indicated that the average passing distance in 2-lane roadways (M = 5.69 ft.) was significantly less than that in 3-lane roadways (M = 6.21 ft.). Whereas shared-use lanes or a higher share of heavy vehicles are associated with significantly closer passing distances. They also surveyed to study the driver's awareness of passing distance laws and drives the perception of a safe overtaking maneuver, which illustrated the drivers are poorly informed about the presence or passing laws.

Overall, the studies emphasize the importance of dedicated bike lanes. Therefore, road classification is likely to be associated with a number of factors that have previously been identified as having an influence on passing distance such as number of lanes (MEHTA *et al.*, 2015), lane width (LOVE *et al.*, 2012), speed differential (CHUANG *et al.*, 2013), presence of

parked vehicles (BECK *et al.*, 2019), and presence of oncoming vehicles (KAY *et al.*, 2014; MEHTA *et al.*, 2015), and type of centerline (SAVOLAINEN *et al.*, 2012; KAY *et al.*, 2014).

Moreover, factors that have been indicated to affect passing behavior include roadway and geometric design (SAVOLAINEN *et al*., 2012; SHACKEL & PARKIN, 2014; FOURNIER *et al*., 2020), whether the bicyclist was wearing a helmet (WALKER, 2007), type of vehicle (DE CEUNYNCK *et al*., 2017), traffic volume (LI *et al*., 2012), speed (LLORCA *et al*., 2017; CHUANG *et al*., 2013), the presence of a share the road sign (KAY *et al*., 2014; HØYE *et al*., 2016), and driver distraction (FENG *et al*., 2018).

In summary, these studies provide insights into passing behavior between vehicles and cyclists, emphasizing the influence of various factors such as road infrastructure, vehicle speed, lane width, presence of oncoming vehicles, and the presence of bike lanes. The findings highlight the importance of dedicated bike lanes, lower speed limits, and infrastructure improvements in promoting safer interactions between motorists and cyclists.

a Experimental Setup

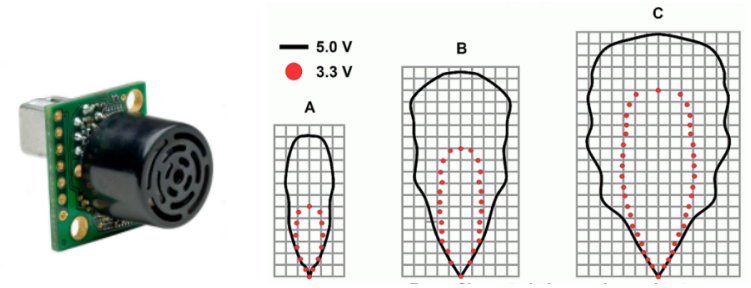
Different range approaches were utilized in the studies, employing various types of sensors. Mackenzie *et al*. (2021) used a dual ultrasonic distance sensor system that recorded distinct "footprints" during vehicle passing events, enabling automatic detection through a software algorithm. The sensors collected distance data at a frequency of 20 times per second (20 Hz) and were positioned on the bicycle axis.

Most studies employed a frequency of 10 Hz for ultrasonic surveys (WALKER *et al*., 2014; BECK *et al*., 2019; AMPE *et al*., 2020; FEIZI *et al*., 2019). Mackenzie *et al*. (2021) noted that the 20 Hz frequency might be insufficient to detect very fast passing vehicles, and the algorithm used for identifying passing events could be affected by a constant flow of varying data when multiple vehicles pass simultaneously. Dozza *et al*. (2016) also used a 20 Hz frequency but employed a LIDAR system (Hokuyo UXM-30LXH-EWA).

Regarding specific sensors, Feizi *et al*. (2021) used a Codaxus C3FT sensor positioned on the handlebar. Ampe *et al*. (2020) utilized a MaxBotix MB1200 XL-MaxSonar-EZ0 temperature-compensated ultrasonic distance sensor (±1cm accuracy), similar to the design in Walker *et al*.

(2014), placed on the axis of the luggage rack. Beck *et al*. (2019) employed a MaxBotix MB1230 XL-MaxSonar-EZ3 sensor positioned under the saddle.

Chapman & Noyce (2012) utilized a MaxBotix MaxSonar LV-EZ1 model, and Chuang *et al*. (2013) also used a MaxBotix sensor but did not specify the model. Figure 2.8 shows an example of the ultrasonic sensor used (a) and sample results of the measured beam pattern on a 30-cm grid (b), demonstrating the sensor's range capability when detecting dowels of varying diameters in front of it.



(a) Ultrasonic sensor MB1200 XL-MaxSonar EZ0

(b) Beam pattern

Figure 2.8 Ultrasonic sensor and sample result (a and b), MaxBotix Inc. (2021)

Global Positioning System (GPS) was utilized in several studies (DOZZA *et al*., 2015; CHUANG *et al*., 2013; MEHTA *et al*., 2015; BECK *et al*. 2019; AMPE *et al*., 2020; MACKENZIE *et al*., 2021; BERNARDES *et al*., 2019; RUBIE *et al*., 2020; SHACKEL & PARKIN 2014). The chosen frequency for the most of these papers was 1 Hz, except for one (FEIZI *et al*., 2019), which used 0.5Hz. Moreover, the GPS-related devices included a Garmim Forerunner 201 (MACKENZIE *et al*., 2021), and an Adafruit Ultimate GPS FeatherWing (BECK *et al*. 2019).

Additionally, videos recordings were also conducted. A single GoPro 9 camera was used on the handlebars (BECK *et al*. 2019; FEIZI *et al*., 2019). Two GoPro Hero cameras, operating at 30 frames per second (fps), were used - one facing forward and one backward (DOZZA *et al*., 2015). Moreover, video recordings were employed to calibrate the sensor array and investigate individual passing observations on the handlebar (MEHTA *et al*., 2015; SHACKEL & PARKIN, 2014). A Viosport POV 1.5 camera was positioned sideways, adjacent to the ultrasonic distance sensor, for vehicle type identification and passing speed calculation at 30 fps (CHAPMAN & NOYCE, 2014). Furthermore, five car DVR black boxes cameras,

operating at 30 fps were used (CHUANG *et al*., 2013). Lastly, two Oregon Scientific ATC2K helmet cameras (RUBIE *et al*., 2020).

The use of IoT technologies in instrumented bicycles enables connectivity between transportation system elements and addresses data collection challenges. With the availability of IoT devices, sensors, and open-source information, devices can be developed to handle these challenges (BERNARDES *et al*., 2019; Bahmankhah *et al*., 2020). These technologies generate significant amounts of data, requiring techniques for data acquisition, cleaning, aggregation, modeling, and interpretation in large-scale sensor-based systems. The application of big data models in networked systems has been demonstrated in studies focused on urban environments, facilitating the creation of smart and Intelligent Transportation Systems (ITS) (ANG & SENG, 2016; GUERRERO-IBÁÑEZ *et al*., 2018).

b Hardware systems

Portable logging devices like *Arduino* and *Raspberry Pi* have become cost-effective options for building instrumented bikes, enabling research in this field (Gadsby & Watkins, 2020). *Raspberry Pi*, a Linux-based single-board computer, has been widely used in research projects, including as a data acquisition system for riding dynamics in human-powered vehicles (DOZZA *et al*., 2016; OZBAY *et al*., 2018; AMBROZ, 2017). Bernardes *et al*. (2019) employed two ultrasonic sensors connected to a *Raspberry Pi* to collect bicycle trajectory data and lateral distances, utilizing a specially designed 3D-printed enclosure.

Arduino, another open-source hardware, is pre-programmed with a boot loader for easy program uploading. Amper *et al*. (2020) used an *Arduino Uno* prototyping computer with customized software, while Beck *et al*. (2019) utilized an *Arduino* microprocessor with Adafruit Feather M0 Adalogger. Walker *et al*. (2014) housed the sensor, *Arduino*, and batteries inconspicuously in a small plastic box mounted on the bicycle's luggage rack. The sensor data was recorded to an SD card using an *Arduino Uno* prototyping computer with dedicated software.

c Data Analysis and Software

Data from sensors and GPS were stored in a database and queried using SQL, specifically through SQLite3 (MACKENZIE *et al*., 2021; BERNARDES *et al*., 2019). This allowed for online data transfer when sensors had a network connection, enabling remote

analysis of historical and real-time data. A dashboard with a lateral distances data map was created using Tableau.

Data analysis was conducted using Stata, SAS, R Statistical Software, MATLAB, and SPSS (BECK *et al*., 2019; FEIZI *et al*., 2019; SAVOLAINEN *et al*., 2012). MATLAB was used to process CSV files from the SD card and apply a low-pass filter for bicycle accelerations (DOZZA *et al*., 2016). Python code was developed to read data from ultrasonic sensors and GPS receivers and combine them into a single file (BERNARDES *et al*., 2019). The ST-matching method in Python with OpenStreetMap was used for GPS data map matching (BECK *et al*., 2019). C programming was used to develop software for data collection (DOZZA *et al*., 2016).

Regression analyses, including generalized linear regression, logit regression, and Probit Model, were performed (MACKENZIE *et al*., 2021; FEIZI *et al*., 2021). Cohen's kappa statistics were used for assessment (BECK *et al*., 2019). Spectral and descriptive statistical analyses were conducted to characterize normal cycling dynamics (DOZZA *et al*., 2016). The operating system used was a Debian distribution of Linux (DOZZA *et al*., 2016). Moreover, VISSIM Traffic Microsimulation software was utilized for a simulation study on vehicle platooning and passing events (SHACKEL & PARKIN, 2014). Instruments were connected to a laptop running LabView software for data acquisition (CHUANG *et al*., 2013).

2.3 MACHINE LEARNING, UNSUPERVISED LEARNING ALGORITHMS AND STATISTICS FOR DATA ANALYSIS

Machine Learning is the development of algorithms and statistical models that enable computer systems to improve their performance on a specific task through learning from data, without being explicitly programmed. In essence, it's about computers learning patterns and making predictions or decisions based on data. In Machine Learning, Unsupervised Learning is an algorithm method in which the model is not given any labeled training samples. Instead, the model is only given a dataset and must learn to find patterns or relationships without guidance.

These algorithms discover patterns or relationships in a dataset that might not be immediately apparent. Some common applications of unsupervised learning include clustering, anomaly detection, and dimensionality reduction. One of the main distinctions between unsupervised and supervised learning is that the model does not have a predefined goal or objective in unsupervised learning. Instead, it must discover the structure of the data on its own and learn to identify patterns or relationships within the data. (KUBAT 2017; ZANTALIS *et al*., 2019)

Unsupervised learning algorithms are commonly used in cases where it is difficult or impossible to label the training data or where the goal is to discover patterns or relationships in the data that might not be directly evident.

Clustering models and Kernel Density Estimation (KDE) are typical geospatial methods applied in hazardous site identification (LORD *et al*., 2021). The present research uses K-Means and DBSCAN clustering algorithms to estimate and identify geospatial areas that might need special attention relating to the Lateral Passing Distance (LPD). Also, it uses KDE to analyze some cluster details.

2.3.1 K-Means Clustering Model

K-means clustering is a nonhierarchical clustering technique used to analyze patterns in the distribution of crashes and identify hazardous sites. It is an unsupervised method to classify elements into discrete groups based on their similarities or discovered conventions (JAIN *et al*., 1999; KIM & YAMASHITA, 2007; ANDERSON, 2009; MAURO *et al*., 2013; SELVI & CAGLAR, 2018; ZANTALIS, F. *et al*., 2019; LORD *et al*., 2021).

This algorithm, which Mac Queen introduced in 1967, is a cyclical algorithm in which clusters are constantly recalculated until the most suitable solution is acquired. The objective of the K-means algorithm is to divide a dataset composed of n data objects into k clusters determined depending on preliminary information or by using mathematical technics.

The mathematical technics help minimize a criterion known as inertia or within-cluster sum-of-squares. Each cluster is represented by the mean of samples in the clusters called "centroids." It aims to minimize the inertia or Within-Cluster Sum of Squares (WCSS) and maximize the Between-Cluster Sum of Squares (BCSS). In other words, for the intra-cluster similarity of events to be high and the inter-cluster similarity of events to be low. Inertia can be

acknowledged as a measure of how internally coherent clusters are. For that, the Euclidean distances between each centroid are calculated by equation 1.

$$deuc(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad (1)$$

While using K-Means, the research needs to define the number of K clusters before the procedure starts; and the metric used to calculate the distances (e.g., Euclidian). The present study uses the Euclidian distance due the spreading characteristic of the data, regarding being also a common choice straightforward to implement and measures the straight-line distance between two points in a multi-dimensional space. Given the "p" and "q" objects, the distance between the dimensions is calculated by equation 2.

$$d(p, q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_i - q_i)^2 + \cdots + (p_n - q_n)^2} \quad (2)$$

K-means has three iterative stages. The first stage chooses the initial centroids, with the primary method being to choose samples from the dataset. After initialization, K-means consists of looping between the two other steps. The first one assigns each sample to its nearest centroid. The second step creates new centroids by taking the mean value of all of the samples assigned to each previous centroid. The difference between the old and the new centroids is computed, and the algorithm repeats these last two steps until this value is less than a threshold. In other words, it repeats until the centroids do not move significantly (ARTHUR & VASSILVITSKII 2006).

a Discovering the Optimal Number of Clusters

The elbow method is a heuristic used to determine the optimal number of clusters for a K-means clustering analysis. It works by fitting the K-means model with different values of K and measuring the within-cluster sum of squared errors (WCSS) for each model. The WCSS measures the compactness of the clusters, with a lower WCSS indicating more compact clusters. The elbow method looks for an "elbow" in the plot of WCSS versus the number of clusters.

The idea is that the WCSS will decrease as the number of clusters increases, but at some point, the improvement in WCSS will begin to diminish, forming an "elbow" in the plot. The number

of clusters at the elbow is considered to be the optimal number of clusters for the K-means model. An example cab is seen below in Figure 2.9.

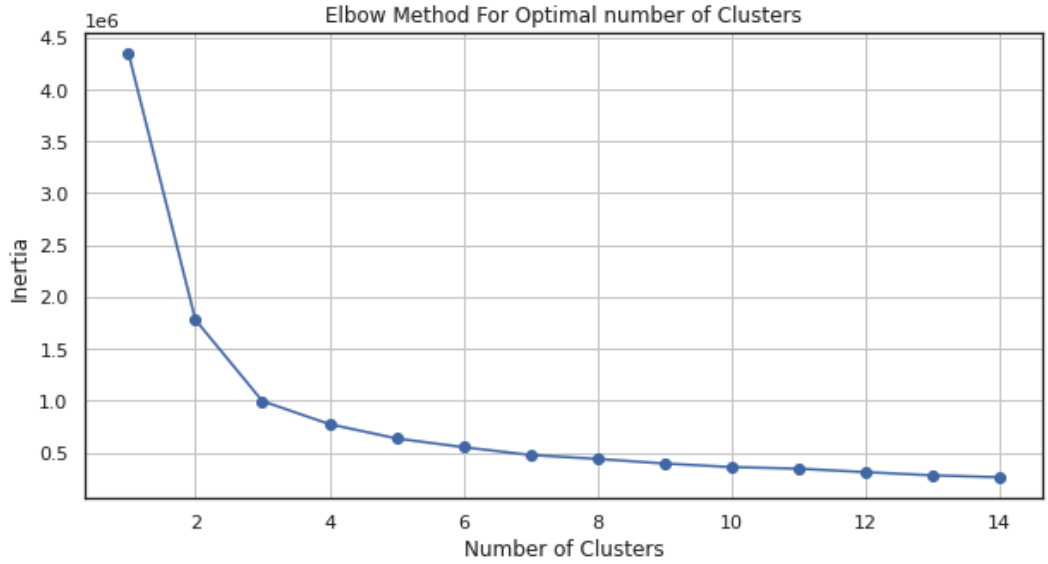


Figure 2.9 Inertias - WCSS vs. the number of cluster

To use the elbow method is needed to fit the K-means model with a range of values for K and plot the WCSS versus K. However, in some cases, the WCSS plot may not have a clear elbow, making it difficult to determine the optimal number of clusters. In these cases, it may be necessary to use additional techniques or domain knowledge to determine the appropriate number of clusters.

The Elbow Method to discover the optimal number of clusters is calculated using, where C_k is the k^{th} cluster and W(C_k) is the within-cluster variation (equation 3).

$$minimize\ (\sum_{k=1}^{k}\ W(C_k)) \qquad (3)$$

2.3.2 Density-Based Spatial Clustering of Applications with Noise - DBSCAN

The Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm views clusters as areas of high-density separated by low-density areas. Clusters found by DBSCAN can be any shape, as opposed to k-means which assumes that clusters are convex-shaped. It does not require you to specify the number of clusters in advance, as it can automatically determine the number of clusters based on the characteristics of the data (ZANTALIS *et al*., 2019).

The DBSCAN works regarding the core samples, which are examples that are in areas of high density. A cluster is therefore a set of core samples, each close to the other (measured by some distance measure) and a set of non-core samples that are close to a core sample (but are not themselves core samples).

There are two parameters to the algorithm: epsilon (Eps) and minimal samples (min_samples). Eps is the maximum distance between two points in the same cluster, and *min_samples* is the minimum number of points required to form a cluster. These parameters define what is dense mathematically. Higher *min*_samples or lower eps indicate the higher density necessary to form a cluster.

2.3.3 Statistical Validations

a T-test and ANOVA

One way to calculate the statistical significance of the results of a clustering model is to use a hypothesis test to determine whether the differences between the clusters are statistically significant. This can be done by low-density the means of the clusters and using a t-test or ANOVA (Analysis of Variance) test to determine whether the differences between the means are statistically significant.

To conduct a t-test firstly is needy to calculate the mean and standard deviation of the variables for each cluster. Then, these values are used to calculate the t-statistic and the p-value for each variable. If the p-value is less than a predetermined threshold (such as 0.05 for a 95% confidence), it can be concluded that the differences between the means are statistically significant. Nevertheless, to conduct an ANOVA test the same procedure is applied, but in this regard, it is possible to compare two or more groups; the F-statistic and the p-value is calculated in this case.

Moreover, these tests assume that the data is usually distributed and that the variances of the groups are equal. If these assumptions are not met, it may need to use a non-parametric test, such as the Mann-Whitney U test to determine the statistical significance of the clusters.

b Kolmogorov-Smirnov (K-S) test

The Kolmogorov-Smirnov (K-S) test is a nonparametric statistical test used to determine whether a sample comes from a population with a specific probability distribution. It is based on the maximum difference between the cumulative distribution function (CDF) of the sample and the CDF of the reference distribution.

The K-S test is a general test that can be used to compare any continuous distribution, including normal, uniform, exponential, and others. It is widely used in many fields, including physics, biology, economics, and engineering, to test hypotheses about the underlying distribution of a dataset.

The test compares the sample's empirical CDF with the theoretical CDF of the reference distribution. The empirical CDF is a proportion of the sample where it is less than or equal to each value in the sample. Theoretical CDF is the expected proportion of the sample that would be less than or equal to each value in the sample based on the reference distribution.

If the sample comes from the reference distribution, then the empirical CDF and the theoretical CDF should be very similar. If the sample does not come from the reference distribution, then the empirical CDF and the theoretical CDF will significantly differ. The K-S test measures the maximum difference between the two CDFs and calculates a p-value based on this difference. If the p-value is below a predetermined threshold (usually 0.05, for a 95% confidence), then the null hypothesis (that the sample comes from the reference distribution) is rejected.

2.3.4 Kernel Density Estimation

The Kernel Density Estimation (KDE) is a used approach for hazardous site selection because of its accuracy and consistency in prediction (ANDERSON, 2009; KUO *et al*., 2013; THAKALI *et al*., 2015). The KDE can define the extent of the risk of a threshold, like crashes in road safety. Using this method, the risk surrounding each target can be calculated, and the risk density is defined. When the distance is closer to zero, the risk density reaches the highest value and decreases with increased distance.

The KDE graph is a representation of the distribution of a continuous variable. The shape of the curve can give an idea of the distribution of the data, and the position of the peak(s) provides an idea of the center of the distribution. For example, a bell-shaped curve indicates a normal distribution, while a skewed curve indicates a skewed distribution.

The shape of a function can be estimated through its kernel density estimates as follows in equation 4.

$$\hat{f}_k(x) = \frac{1}{nh^d} \sum_{i=1}^{n} K(\frac{x-x_i}{h}) \quad (4)$$

Where x is the variable of interest; h is the bandwidth that controls the amount of smoothing; d is the dimension, and K(.) is the kernel function. Basically, the KDE smooths each data point xi into small density bumps and then adds all these small bumps together to obtain the final density estimate (LORD *et al*., 2021).

2.3.5 Features Correlation - Spearman Coefficient

The Spearman rank correlation coefficient (ρ) is a statistical measure of the strength and direction of the relationship between two variables. It is often used to determine the degree to which two variables are related and the direction of the relationship (positive or negative). This was used to see the relationship between different variables.

The coefficient is calculated using the ranks of the data rather than the raw data values. This makes it well-suited for use with ordinal data (data that is ranked or ordered but not necessarily evenly spaced) or data that is not normally distributed. It ranges from -1 to 1. A value of -1 indicates a perfect negative correlation, a value of 0 indicates no correlation, and a value of 1 indicates a perfect positive correlation. It is represented by the equation 5.

$$\rho = 1 - \frac{6 \sum_{i=1}^{n} (x_i - y_i)^2}{n(n^2 - 1)} \quad (5)$$

where *d* is the difference between the ranks of the two variables, *n* is the number of data points and ∑ is the sum of the squared differences.

3 METHODOLOGY

3.1 DATA COLLECTION METHODOLOGY

The present section consists of the collection and analysis methodology of data from trips made on a bicycle in Brasília city center. The area chosen for research was the avenue around the City Park at the city center. More details about this choice and the area characteristics are presented in the following sessions. The data were collected using the sensor system and platform BSafe360, developed by the researchers Suzana Duran Bernardes, Abdullah Kurkcu, and Kaan Ozbay from the New York University (NYU) Tandon School of Engineering (BERNARDES *et al*., 2019; BERNARDES & OZBAY, 2023). The data were manipulated and analyzed in Python and SQL programming and database languages.

3.1.1 Sensor device: BSafe360 Architecture

The BSafe360 is a lightweight, portable device with multiple sensors made to be installed on a bicycle, as shown in Figure 3.1 below. It is composed of: a microcomputer called Raspberry Pi, which performs the reading, storage, and transfer of data from the sensors; two ultrasonic sensors to measure the distance of the bike to objects to its left and right; a GPS receiver and antenna for tracking the bike's trajectory; a gyroscope and accelerometer sensor that reads the bike's position and acceleration vectors; and a portable charger that allows the BSafe360 to run for more than 2 hours.

(a) BSafe360 sensor installed on Bicycle (b) Bicycle used in the data collection

Figure 3.1 Equipment used for data collection

Concerning the data collection, the device was used along the entire route collecting information. Also, the data is collected and sent to the server when the device was connected to the internet.

Moreover, from the server, it is possible to see the device's location on a map, in real time. Also, the aggregated and non-aggregated readings in line graphs, and readings in tables if it is connected to internet during the trip.

It is possible to export and download the raw data from the platform in CSV format. However, if there is no internet connection when riding a bike is possible to obtain the data by connecting the Raspberry Pi device to a computer or turning it on in a pre-configured wi-fi connection. This was the option adopted by this research, regarding that all the data is written on a memory card.

The device includes an ultrasonic sensor Figure 3.2; an accelerometer, and a gyroscope (Figure 3.3). And a GPS, as shown in Figure 3.1, positioned outside the case.

Figure 3.2 Ultrasonic Sensor Unit

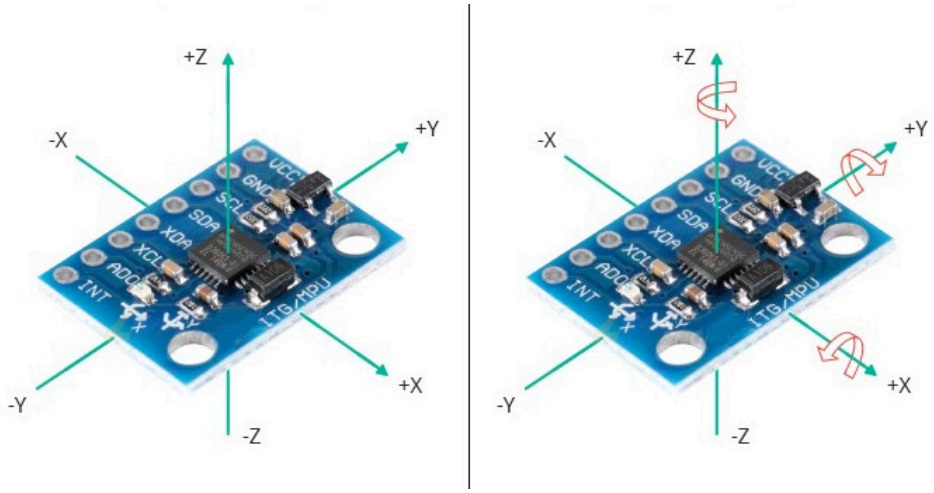


Figure 3.3 Accelerometer and Gyroscope MPU6050

3.1.2 Data Collection Procedures

a Environmental conditions

Aiming to get proper inputs to the research, the data collection was performed only on clean days, without rain, and with good visibility conditions. It was collected in the lunch hours, with a significant mobility flush for analysis between vehicles and cyclists: between 12:00 PM and 1:00 PM. This period of time was chosen regarding the high volume of commuting in this area from people getting off and arriving from their chores, configuring a local peak time.

Before using the sensor for collection, the researcher collected data through some exploratory rides to discover the area's characteristics and measure the average time, travel, road infrastructure, as well as the traffic conditions.

Moreover, this study collected 175 minutes of data, of which 100 minutes were used for the analysis, the rest of it was discarded due to not completely fill the research conditions or have been used only for exploratory analysis. Also, in the same way, approximately 65 km were traced in total, and 40 km were used in the analysis. The data collection procedures were always by completing an entire lap around the City Park. For each day of collection was procedure one lap collecting data.

Therefore, values not used in the analysis were removed because of the data adjustments, tests on the route by exploratory analysis, and system errors.

b Researcher Characteristics for Data Collection

Regarding maintaining the same behavior on cycling, this study proposed all the data collected by the same researcher: a male 34 years old, 1.91 meters tall. For every ride, it was used the same white bicycle, with a white T-shirt and helmet. This pattern's objective is to avoid any possible influence of clothes it could have in the research results (Walker, I., 2007). Regarding the ride, was tried to maintain the same distance from the edge throughout the route, around 60 cm.

c Survey Region Characterization: Brasília, Brazil

The Brasília City Park (Parque da Cidade Sarah Kubitschek de Brasília) is a public park at the city center of Brasília in the Asa Sul neighborhood, Federal District. It was founded in the seventies and has 420 hectares. It is wide used by cyclists for leisure and for performance sport cycling training. So, the park has an avenue surrounding the area, connecting the principal neighbors in the capital, like *Sudoeste*, *Asa Sul*, *Octogonal*, the Commercial Center Sector with shopping, hotels, schools, and big companies, and the central axis of the capital with governmental buildings and services centers. The avenue general characteristics are:

- The complete route is 9.80 kilometers long;
- The whole mean travel time is about 25 minutes;
- The shared cycle lane follows the right side of the road;
- There are cycle marks through the pathway indicating the shared lane;
- There are cycle signs through the road indicating the shared lane;
- The velocity limit is 60 kilometers per hour;
- Each lane has 3.50 meters, with two lanes (7.0m) in each direction;
- There are **six** intersections for the entrance and exit of the park area.

The road consists of a shared lane at the right corner of the road. The shared lane typology is one in which there is no physical separation between the cyclist's and the car's rolling path, having just an occasional marking on the pavement as it can see as follows. The road profile and the shared lane is presented in the Figure 3.4, Figure 3.5 and Figure 3.6.

Figure 3.4 Location of the survey region from the Federal District (Brasília great area) perspective

Figure 3.5 Location of the survey region from the Federal District (Brasília great area) perspective - Google Maps

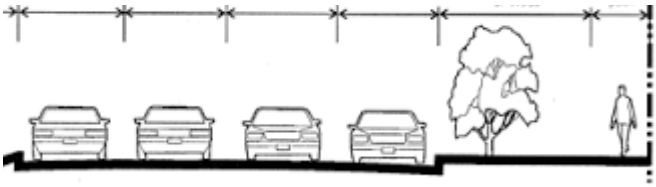


Figure 3.6 Road profile with four lanes of 3,5m

Regarding the land use, the City Park is surrounded by residential (*Asa Sul, Sudoeste, Octogonal*), and miscellaneous (residential, commercial, governmental) use. Also, there is a dense commercial and education (schools and universities) concentrated in some parts, provoking intense traffic rush three times a day: at the beginning of the morning, at the lunch hour, and in the evening.

This research procedures the collection at the lunch hour regarding it has an intense high-volume traffic concentrate at small period of time, involving different land-uses that impact considerably the avenue around. The survey location for the study is presented in Figure 3.7, the dots represent some data collected around the route. The land use, around the survey location is also presented in the Figure 3.8, as follows.

Figure 3.7 Location of the survey region from the Federal District (Brasília great area) perspective

Figure 3.8 Land use around the City Park

d Traffic Volume Correlation

For further analysis, the traffic volume was collected manually. Since, the Transportation Department of Brasília only have data from places with electronic inspection equipment and The Park area does not have an electronic traffic equipment, there was no volume information registered concerning this area.

The collections were made on the seventh of December 2022, the same hour as the collections made on bicycles using the BSafe360 equipment. The volunteers were divided into three points to collect at the same time for 30 minutes, separating the amount into fifteen minutes.

The Park has six entrance and exit points, indicated in orange circles on Figure 3.9. This study manually collected simultaneous data from three points, as shown in blue Figure 3.9. Regarding these points, it was possible to measure the volume between the orange points 1(orange) and 2(orange) for collection 1(blue). 3 (orange) and 4 (orange) for collection 2 (blue); And 6 (orange) to 1 (orange) for point 3 (blue). Therefore, it was not possible to make a full correlation analysis between the volume and the results from the Sensor Device considering the existence of other three points when considering the park area as a closed traffic system.

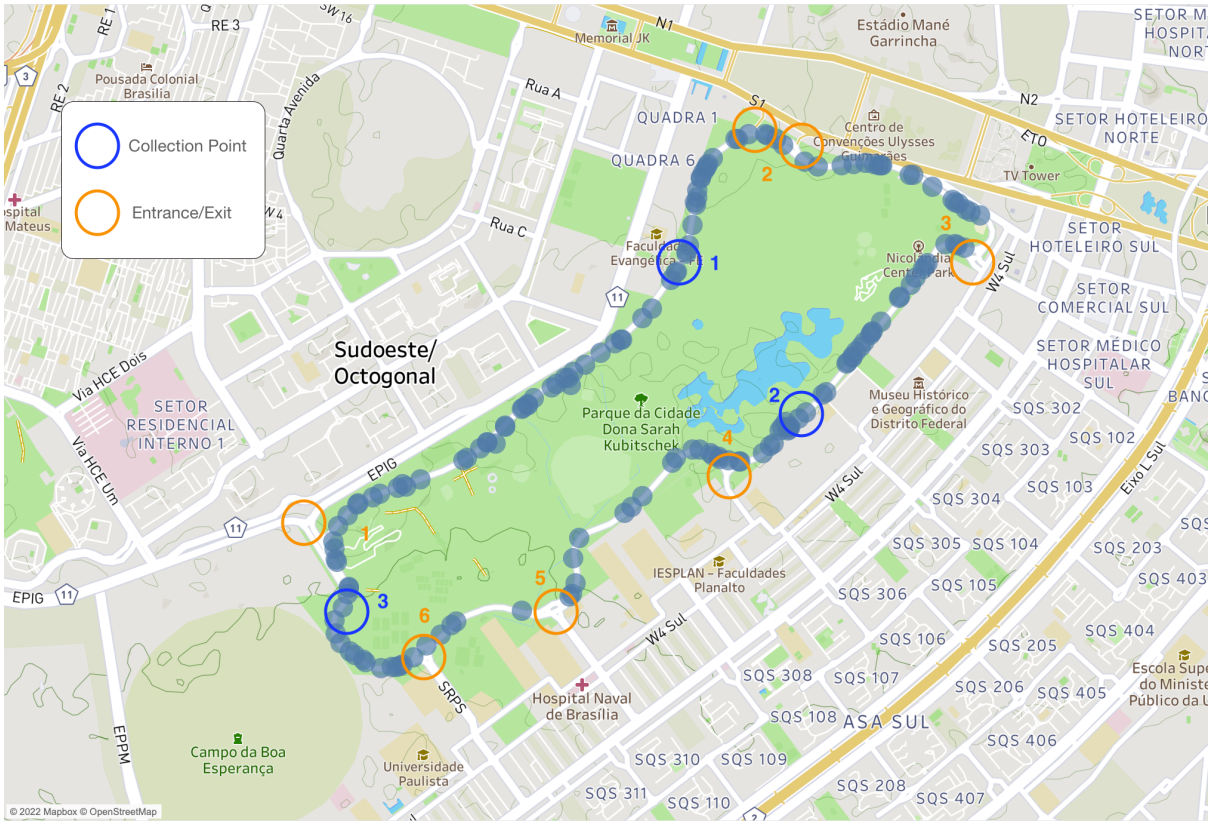


Figure 3.9 Collection points and Park entrances/exits

However, the traffic volume values are presented in Table 3.1. Also, a correlation between these points is presented below in Figure 3.10. These patches were isolated for correlation analysis.

Table 3.1 Traffic Volume in three points from the city park

Point 1	Point 2	Point 3
12:30 to 12:45	***12:30 to 12:45***	***12:30 to 12:45***
automobiles: 88	automobiles: 102	automobiles: 316
motorcycles: 5	motorcycles: 4	motorcycles: 28
others: 2	others: 2	others: 5
12:45 to 13:00	***12:45 to 13:00***	***12:45 to 13:00***
automobiles: 101	automobiles: 110	automobiles: 291
motorcycles: 8	motorcycles: 7	motorcycles: 20
others: 1	other: 1	other: 0
total: 189 automobiles y 13 motorcycles	total: 212 automobiles y 11 motorcycles	total: 607 automobiles y 48 motorcycles

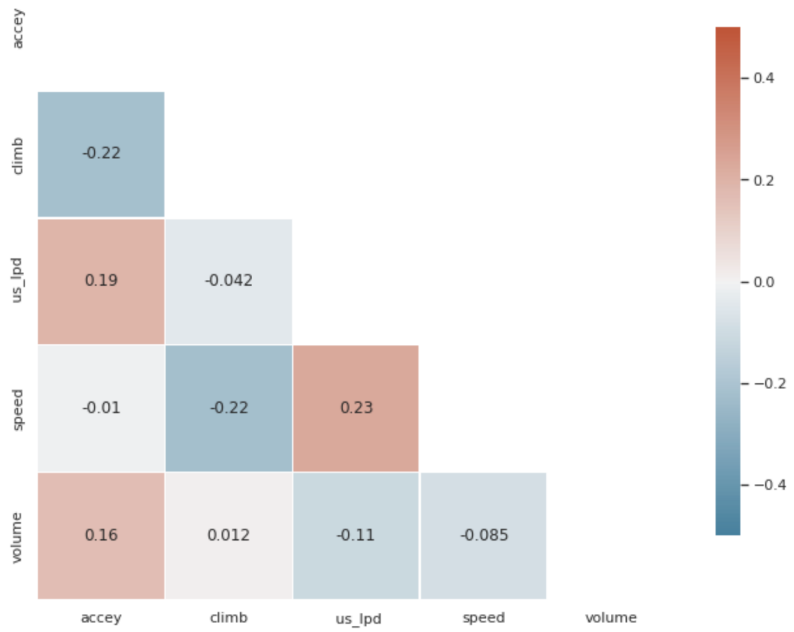


Figure 3.10 Spearman Correlation between Traffic Volume and other features

There is a weak negative correlation between the traffic volume and the lateral passing distance of -011. It means that, for this dataset, the more the traffic volume increases, the LPD decreases, indicating small distances between automobiles and bicycles. From these results, it can be inferred that high volume provokes more risk approximations.

However, more data need to be collected concerning all the points. Furthermore, some studies indicate a high correlation between traffic volume and hazardous lateral passing distances (MEHTA *et al*., 2015).

e Crash Historical Correlation

Additionally, was possible to obtain data for crashes between 2019 and 2020 from the Transportation Department of Brasília. The low quantity of historical crash data due to a possible lack of registration regarding the incidents, limited the analysis with the correlation feature. However, in the following Figure 3.11, and Figure 4.9**,** on the results topic**,** it is possible to see a similarity between the crashes that happened in these last two years against the location of the red cluster, which indicates the small lateral passing distance.

However, it cannot be correlated because it requires more crash data. Nonetheless, other studies seek this validation, demonstrating a correlation between the number of crashes and surrogate

safety measures, such as the lateral passing distance between vehicles and bicycles (BERNARDES *et al*., 2023).

Figure 3.11 Crash registration involving cyclists in 2019 and 2020

3.2 DATA ANALYSIS METHODOLOGY

This study aims for a data-driven safety analysis; all the methodology has the data as input to incorporate the principles and techniques of machine learning. The methodology starts with the field operational cycling data collection; this step was presented at the last session. After collecting the data, it can be sent directly to the database with a wi-fi connection or sent simultaneously to the server in a live connection. This research chooses the methodology of sending it asynchronously, after the collections.

An algorithm written in Python is responsible for group the data from the sensors and system and sending it to the database owned by New York University (NYU) as soon as it is connected to the internet. The data is queried from the database through SQL and manipulated in Tableau software to eliminate any segment outside the route and null or zero latitude and longitude records. After that, the dataset is downloaded and inserted into a Python notebook for manipulation. The following Table 3.2 presents the data feature description collected.

The data regarding the ultrasonic sensor for the lateral passing distance measure was manipulated to consider the accurate distance of the bicycle to the automobiles, regarding the

handlebar and the BSafe360 equipment mounted on the frame of the bicycle. Therefore, was subtracted 25 cm from it, 20cm from the distance of half the handlebar width, and more than 5cm of erroneous collections on cycling activities, for example, when sometimes a hand pass in front of the sensor registering some value. Therefore, LPD readings are also top-limited to 700cm, concerning the width of the one-direction way (two lanes of 3.5m) and the limited range of the ultrasonic sensor.

The data about traffic volume was collected manually at three points of the city park since the national traffic department did not have it for this location. Also, the national traffic department collected data about crashes involving cyclists. However, this data needed more quantity to apply to the models in this research. So, these data regarding crashes and traffic volume were not considered into the clustering model as follows.

Table 3.2 Unit and Data Description

FEATURE	UNIT	DESCRIPTION	SOURCE
latitude	degrees	Global Positioning System: Latitude	GPS
longitude	degrees	Global Positioning System: longitude	GPS
us_lpd	centimeters	Ultrasonic Sensor reading: Lateral Passing Distance (LPD)	Ultrasonic Sensor
speed	meters/second	Rate of movement - velocity	GPS
climb	meters	Elevation level	GPS
accex	meter per squared seconds	Accelerations of the bicycle in X axis as a function of gravity	Accelerometer MPU6050
accey	meter per squared seconds	Accelerations of the bicycle in Y axis as a function of gravity	Accelerometer MPU6050
accez	meter per squared seconds	Accelerations of the bicycle in Z axis as a function of gravity	Accelerometer MPU6050
gyrox	degrees per second	Rotation of the bicycle in the X axis	Gyroscope MPU6050
gyroy	degrees per second	Rotation of the bicycle in the X axis	Gyroscope MPU6050
gyroz	degrees per second	Rotation of the bicycle in the X axis	Gyroscope MPU6050

4 RESULTS AND DISCUSSION

This session presents the results concerning the clustering models and the data analysis. Beforehand, some data analysis will be presented to understand the approach and the data behavior from the bicycle safety perspective.

4.1 EXPLORATORY DATA ANALYSIS

Considering all the data utilized for this analysis, totalizing 160 readings with nine features, the mean of the principal target of the research, Lateral Passing Distance (LPD), was 259.52cm, with a standard deviation of 155.39cm. This data is the filtered result following the considerations presented in the last session. For further analysis, these results were clustered regarding the features relations for analysis using unsupervised learning algorithms in the next session, Table 4.1.

Table 4.1 Lateral Passing Distance (LPD) Statistics (160 readings)

LPD Statistics	LPD (cm)
Mean	261.13
Standard Deviation	154.53
Minimal	35.63
25%	139.62
50%	238.78
75%	351.08
Maximum	693.17

Concerning the variance, these values show a significant variation among the data. Nonetheless, the quartiles indicate that 25% of the readings are less than 139,62cm; for the second quartile, 238.78cm and 75% are less than 351.08cm. It demonstrates that values above it might be outliers and diverge from the distribution. It can be confirmed by looking at the histogram with a kernel density estimation below, Figure 4.1.

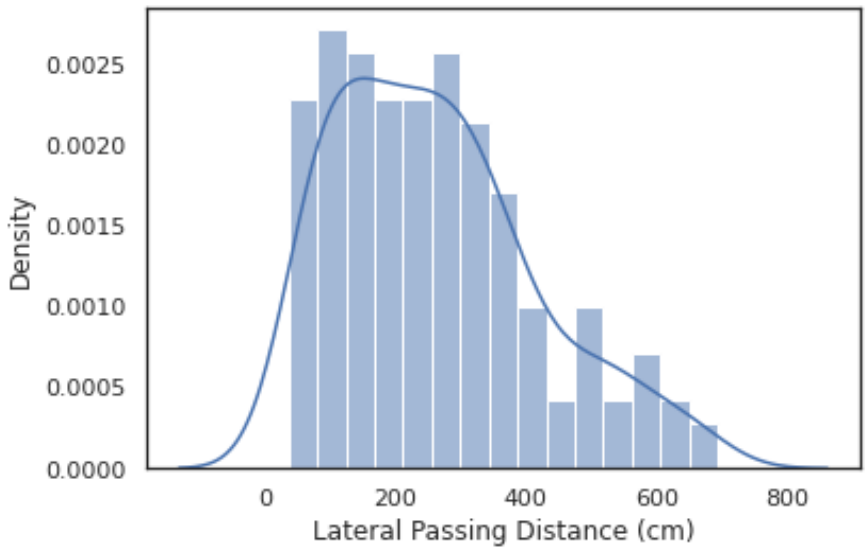


Figure 4.1 Histogram with a KDE for Lateral Passing Distance

Aiming to observe the behavior between variables, a scatterplot was used between them. When relating LPD to Climb and Speed, the following correlation was obtained by fitting a regression model to observe the resulting curve on a 95% confidence interval. The LPD in centimeters is on the y-axis, and the climb (meters per second) and speed (meters per second) data are on the x-axis, Figure 4.2 below. The curve shows the data behavior regarding some variables and it can be inferred that in ascendent movements (climb enhance from 0.0 m/s to +0.2 m/s), the LPD tends to decrease. Also, the LPD is greater on higher velocities ranges. It will be analyzed further forward.

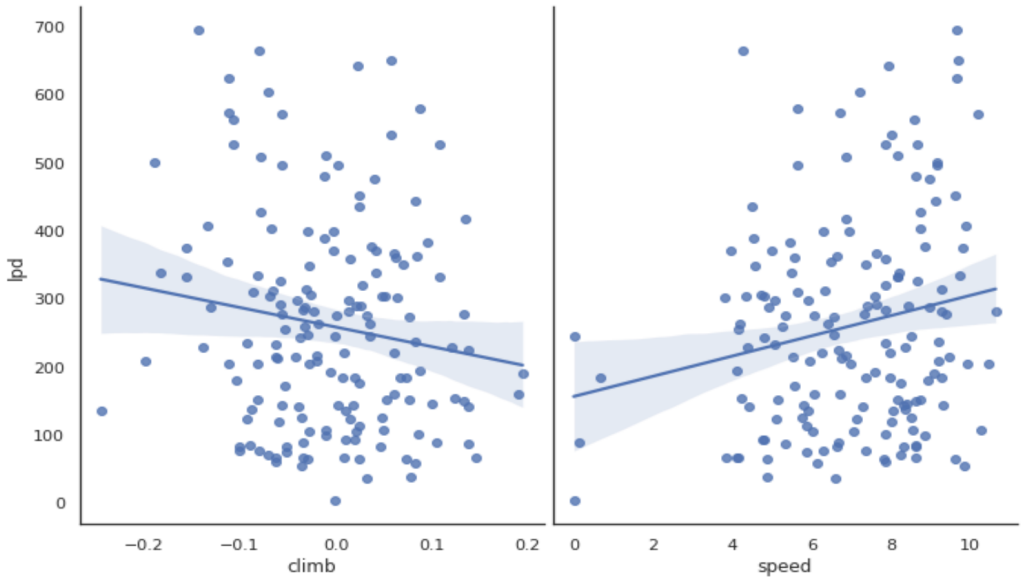


Figure 4.2 95% confidence interval regression between LPD and Climb/Speed variables

4.1.1 Correlation Analysis - Spearman Coefficient

Aims to understand the correlation between the features to access the best possibilities to apply the modeling and check if some feature has a particular influence, this study uses Spearman coefficient correlation.

The Spearman Coefficient ranges from -1 to 1. It indicating the strongest correlation at either direction (e.g., negative or positive). For the analysis, the coefficients close to zero indicate no correlation between variables. Coefficients between 0.10 and 0.20 has a weak correlation, between 0.20 and 0.40 has a moderate correlation, and above 0.40 has a strong correlation. The correlations are described as follows, in the Figure 4.3.

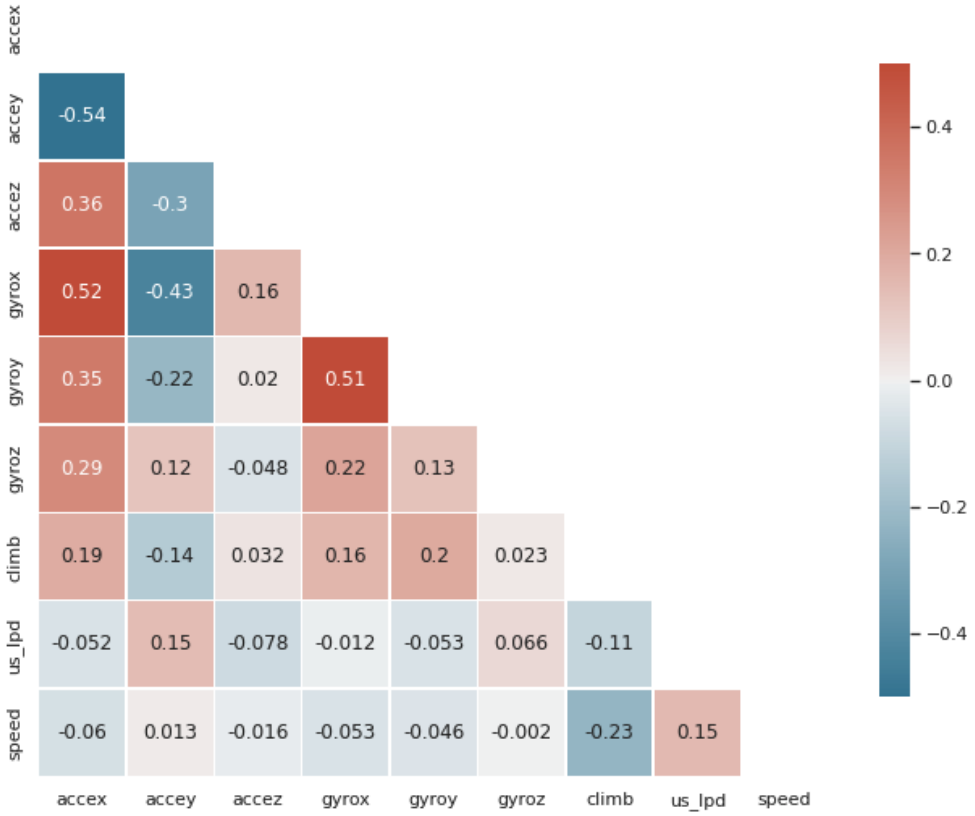


Figure 4.3 Spearman Correlation Matrix between variables

4.1.2 LPD vs. Climbing

Assuming the threshold of 150 cm, regarding the minimal lateral clearance distance law adopted in Brazil, the percentage of values in ascending movement is 4.38% bigger than in downhill routes. This relation is presented in Figure 4.3, above.

The "*us_lpd*" feature has a weak negative correlation (-0.11) compared to the "*climb*" feature. The more the bicycle goes on ascendent movement, the more likely the "*us_lpd*" decreases. It

represents fewer LPD measures on ascent movements, which means more possible risk approaches between bicycles and automobiles. Figure 4.4 shows more significant velocity values in a stronger red color. It shows light colors in ascendent routes. Figure 4.5 presents the altitude elevation, indicating the difference on the terrain impacting the speed.

Figure 4.4 Relationship between speed and LPD

Figure 4.5 Altitude elevation

The "*climb*" feature has a moderate negative correlation (-0.23) compared to the "*speed*" feature. The "*climb*" variable represents how much the terrain changes going up or down through the route. The more the bicycle goes on ascendent movement, the more likely the "*speed*" decreases, needing more effort by the cyclist. The terrain elevation varies from 1077.54 to 1143.42 meters altitude, a 65.88m variation (height above mean sea level).

4.1.3 LPD vs. Speed

The "*us_lpd*" feature has a weak positive correlation (0.15) compared to the "*speed*" feature. The more the bicycle goes fast, the more likely the "*us_lpd*" increases. It represents more significant LPD measures on faster movements, which means less possible risk

approaches between bicycles and automobiles. Figure 4.4. shows the relation between LPD and speed, with more significant velocity values in a stronger red. Smallest the speed, the lighter the red color. Also, the closer the motor vehicle passes to the cyclist, the biggest the red circle, alluding to high risk. The farthest the automobile passes, the smaller the red circle.

4.1.4 Acceleration and Rotation correlation

There is no significant correlation between the accelerometer and gyroscope output regarding the speed and the LPD. However, when the feature "*climb*" is analyzed, there is a weak positive correlation regarding the "*accex*" (+0.19) and "*gyrox*" (+0.16). A moderate positive correlation in "*gyroy*" (+0.20) and a weak negative correlation in "*accey*" (-0.14).

These results may indicate that from the cyclist's perspective, there is more bicycle movement in the axis mentioned above in ascend routes, influencing the clustering results proposed. It can be explained because of the movement of the cyclist applying force on the pedal, pendulum around bicycle. These relations can be better observable in the graphs in appendix A, presenting a 95% confidence interval regression.

Nevertheless, these values are used only concerning the clustering model to look for similarities between clusters; this study will not analyze the details of these parameters, indicating recommendations for future researches. Although these features can be used for plenty of analysis (Ghadge 2015).

4.2 CLUSTERING RESULTS AND COMPARISON

4.2.1 K-means

The K-mean clustering algorithm groups the features with more similar characteristics regarding the distance between these variables. Therefore, for the K-means model is necessary to choose the number of clusters beforehand applying the method. This was chosen using a mathematical formulation to find the minimal inertia possible. The Elbow Method for the optimal number of clusters was applied to the features.

Comparing Figure 4.6 and Figure 4.7 is possible to observe the difference in the inertias values and curve behavior when the data is normalized. In k-mean clustering, the normalization in the data is recommended because of the sensibility to variation. The Silhouette Score Method was

applied for the normalized Elbow graph because it is not showing a decisive number of clusters (a slope at the curve) (Table 4.2).

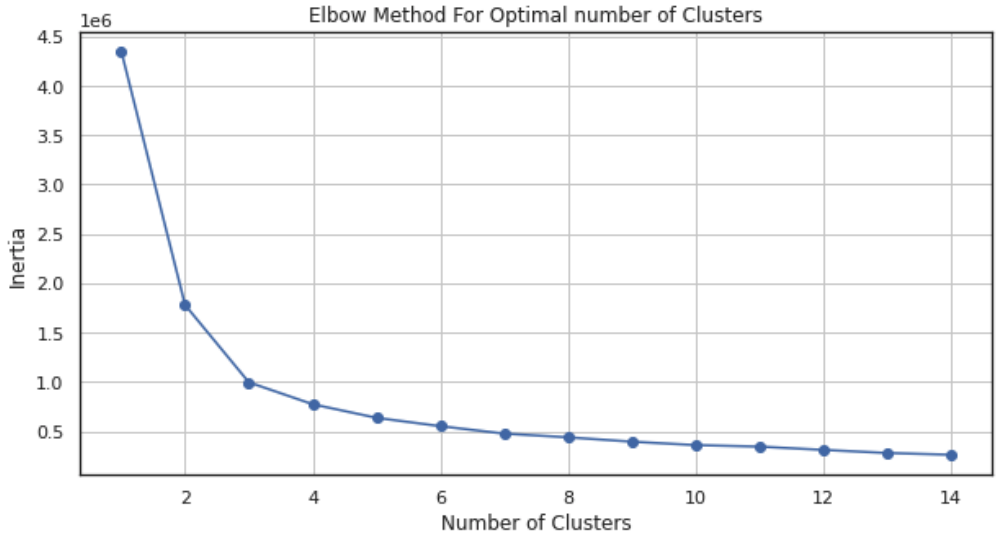


Figure 4.6 Elbow Method applied before the feature normalization

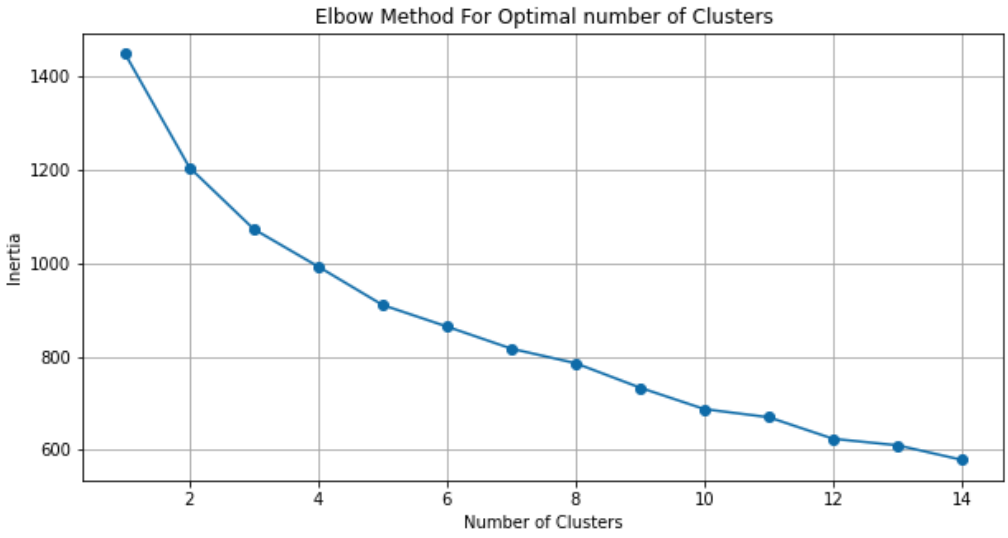


Figure 4.7 Elbow Method applied regarding the feature normalization

A sample regarding the five possibilities with greater silhouette score is showing bellow, Figure 4.8. It is possible to observe that the patterns in some locations are repeatedly coming up on clusters behaviors.

Table 4.2 Silhouette Score

Number of clusters	Average Silhouette Score
3	0.1373810613722346
4	0.1335622825469365
5	0.11906506152453945
6	0.1017479904810742
7	0.12336039022830236
8	0.11294442496204284
9	0.11252040848623253
10	0.13343487289993053
15	0.11902445668169791

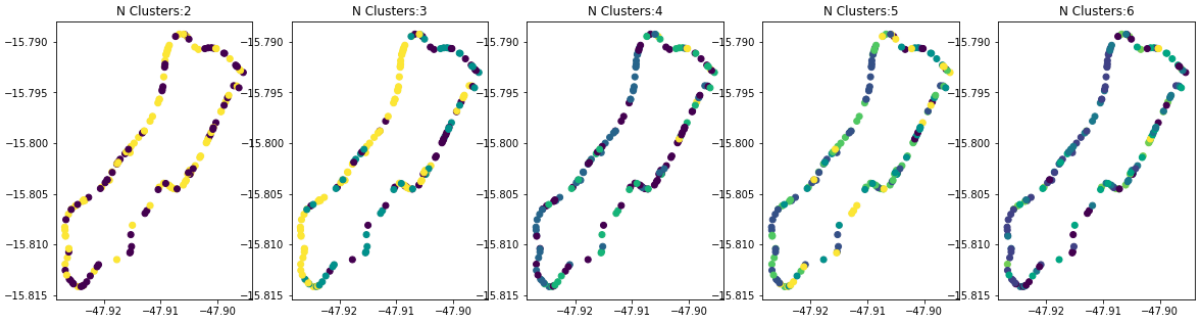


Figure 4.8 Comparison between clusters

Although the biggest silhouette score (0.1374) is the optimum number of clusters: three clusters. Therefore, the following results are obtained using three clusters (the three colors) in the algorithm model, relating all the features, Figure 4.9. The representation is in the same format as the route, indicating the data collection around the Brasília city park. The LPD indicates the distance through the motor vehicles and bicycles in centimeters, more the circumference is, more the lateral passing distance.

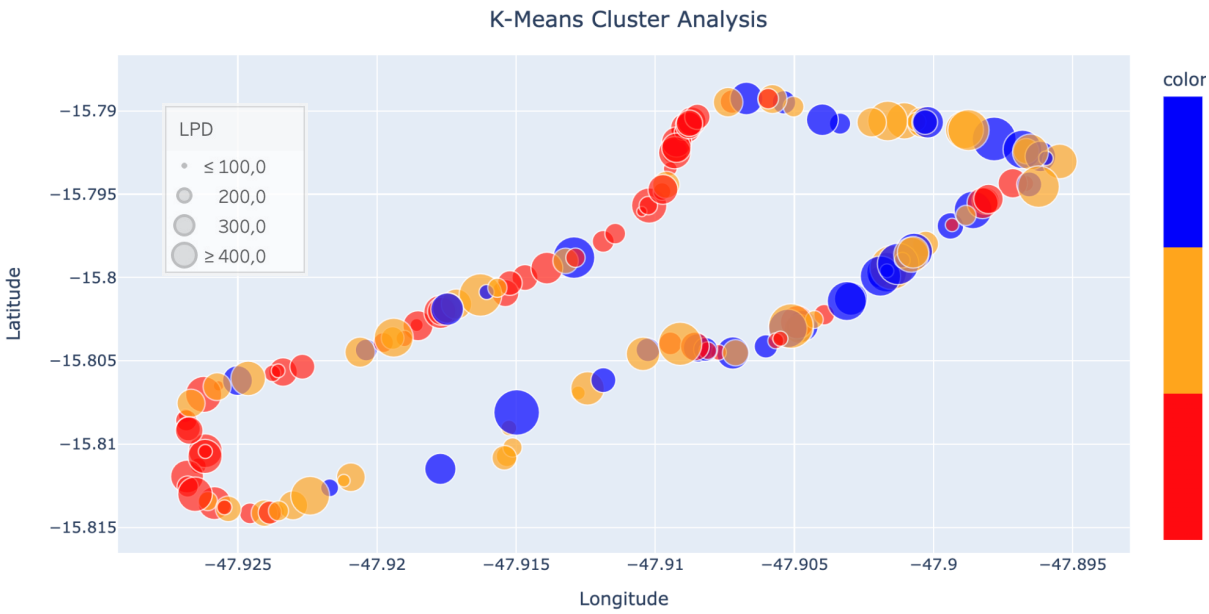


Figure 4.9 Clustering result for K-Means

Observing the clustering behaviors presented in Figure 4.9 is possible to infer some common patterns that will be discussed.

Moreover, the small circles in Clustering Results image indicate small LPDs; respectively, the more significant the circumference, the bigger the LPD. The colors represent the three clusters. The clusters are separately analyzed as follows, and Table 4.3 presents the amount of reading from each cluster. Also, Table 4.4 indicates the statistical information for each group.

Table 4.3 Readings of the clusters

Cluster	Readings
2. orange	46
1. red	75
3. blue	39

Table 4.4 LPD Statistics for the clusters

	samples	Mean(cm)	Std(cm)	Min(cm)	25%(cm)	50%(cm)	75%(cm)	Max(cm)
3. Blue Cluster	39.00	313.94	172.04	63.54	183.73	311.86	463.45	693.17
2. Orange Cluster	46.00	305.41	161.78	52.96	164.24	285.25	404.91	648.89
1. Red Cluster	75.00	206.51	120.35	35.63	100.13	193.69	287.83	663.03

The red cluster had 75 readings with 206.51cm of mean Lateral Passing Distances (LDP) and 120.35cm of standard deviation. Also, 25% of the readings were less than 100.13cm; for the second quartile, 50% were less than 193.69cm; and 75% of the reading was less than 287.83cm.

Aiming to distinguish the distribution between the cluster's results, a Kernel Density Estimation (KDE) was elaborated, Figure 4.10. A KDE graph is a graphical representation of the distribution for a continuous variable. It is a non-parametric way to estimate a random variable's Probability Density Function (PDF).

The orange cluster had 46 readings with 305.41cm of mean LPD and 161.78cm of standard deviations. With 25% of the readings less than 164.24cm; for the second quartile, 50%, 285.25cm; and 75% of the reading less than 404.91cm.

The blue cluster had 39 readings with 313.94cm of mean LPD and 172.04cm of standard deviations. With 25% of the readings less than 183.73cm; for the second quartile, 50%, 311.86cm; and 75% of the reading less than 463.45cm.

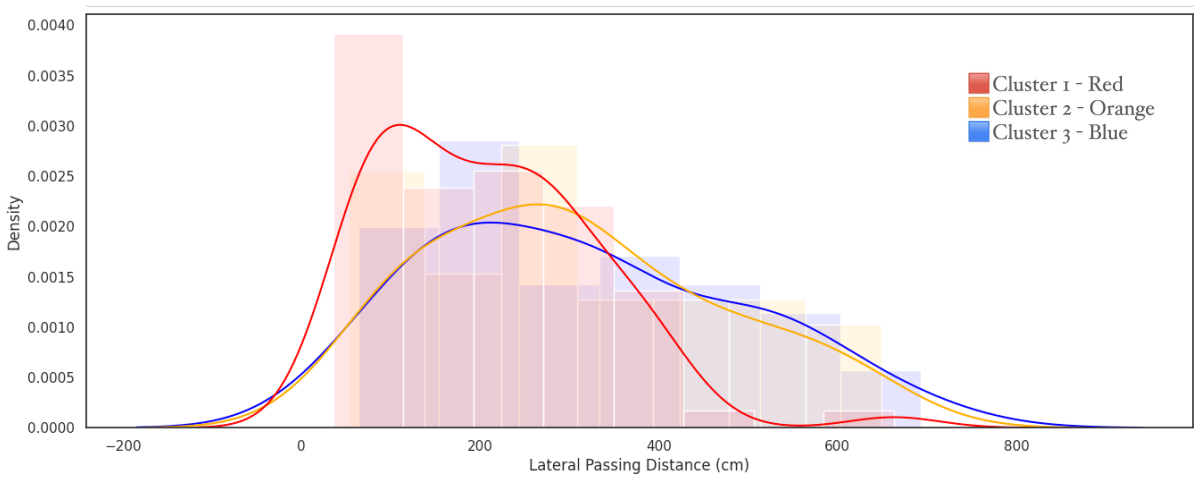


Figure 4.10 Kernel Density Estimation (KDE) for the Clusters

From the KDE graph, the shape of the curve and the position of the peak(s) can give an idea of the distribution of the data. For example, a bell-shaped curve indicates a normal distribution, while a skewed curve indicates a skewed distribution. The position of the peak(s) gives an idea of the center of the distribution. Looking at the Figure 4.10, the peak of the red cluster is smaller than 150cm, representing a value below the defined by law for LPD in Brazil.

Furthermore, the cluster orange and blue have similar distributions regarding the Lateral Passing Distance (LPD). Aimed to analyze the distribution of LPD in the clusters based on the Kernel Density Estimation (KDE) to assess the difference between them was process two-sample Kolmogorov-Smirnov test for goodness of fit between clusters, by an $\alpha = 0{,}05$, Table 4.5.

The results show a similarity between distributions on the orange and the blue cluster regarding the Lateral Passing Distance and the discrepancy between them to the red cluster. It indicates that the optimal number of clusters for this dataset might be two groups, as the mathematical silhouette method resulted before. The red cluster is presented in the Figure 4.11, as follows.

Table 4.5 Kolmogorov-Smirnov test between LPD of the Clusters

KS Test Between Clusters	Statistic	p-value
blue and orange	0.1276	0.8238
orange and red	0.3202	0.0042
blue and red	0.3394	0.0038

a) K-means Clustering Results - Red Cluster

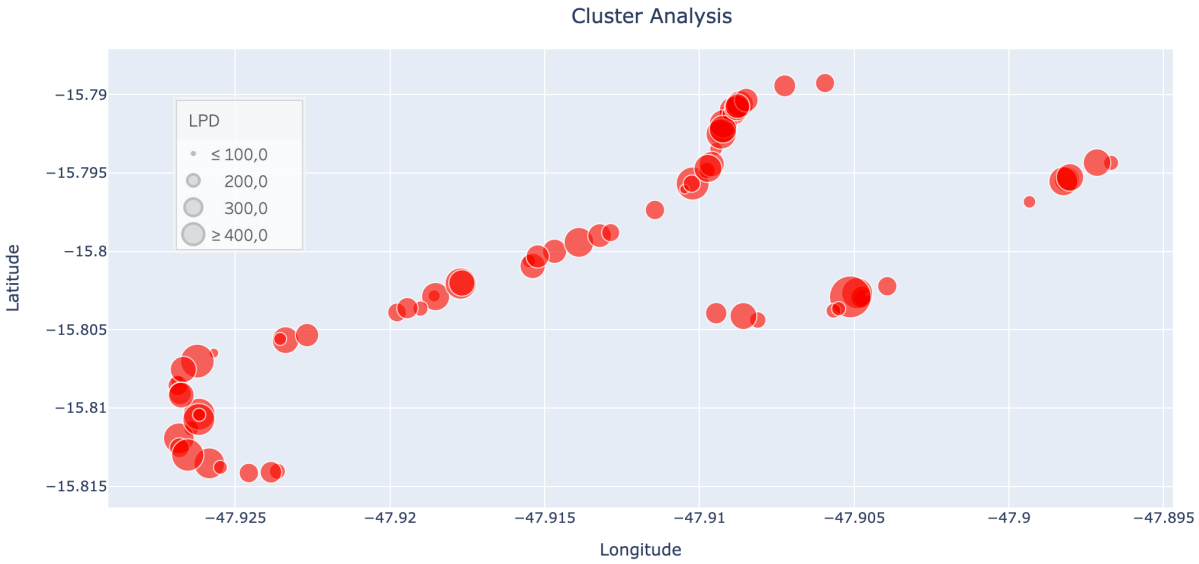


Figure 4.11 Red Cluster Visualization

The red cluster has the minor mean value for the Lateral Passing Distance (206.51cm) with a minimum value of 35.63cm and a std of 120.35cm, which might indicate more risk for the

cyclist compared to other clusters. This cluster is the focus of this research, focusing in areas with larger interactions between modals regarding the LPD.

Furthermore, it indicates the area with the most significant climb positive mean value; the elevation in this cluster varies from 1077 meters to 1143 meters, demonstrating that it is in an ascent path most of the way. Data description for the Red Cluster is in Table 4.6. As follows, the data description for the orange cluster (Table 4.7) and for the blue cluster (Table 4.8). Also, the orange cluster is presented in the Figure 4.12, and the blue cluster in the Figure 4.13.

Table 4.6 Data description for the Red Cluster

index	accex	accey	accez	climb	gyrox	gyroy	gyroz	us_lpd	speed
count	75.00	75.00	75.00	75.00	75.00	75.00	75.00	75.00	75.00
mean	2.71	8.72	-3.03	0.02	01.03	-1.05	4.67	206.51	5.73
std	2.94	2.99	3.44	0.06	19.56	8.93	11.01	120.35	1.80
min	-2.73	1.00	-11.69	-0.10	-75.81	-27.74	-20.53	35.63	0.00
25%	0.49	6.82	-5.69	-0.03	-8.87	-7.06	-2.36	100.13	4.67
50%	2.62	8.69	-3.27	0.02	-0.35	-1.98	4.49	193.69	5.75
75%	05.09	10.50	-0.90	0.06	14.14	4.15	11.96	287.83	6.84
max	8.45	16.99	6.96	0.15	51.44	22.73	34.89	663.03	8.91

b) K-means Clustering Results - Orange Cluster

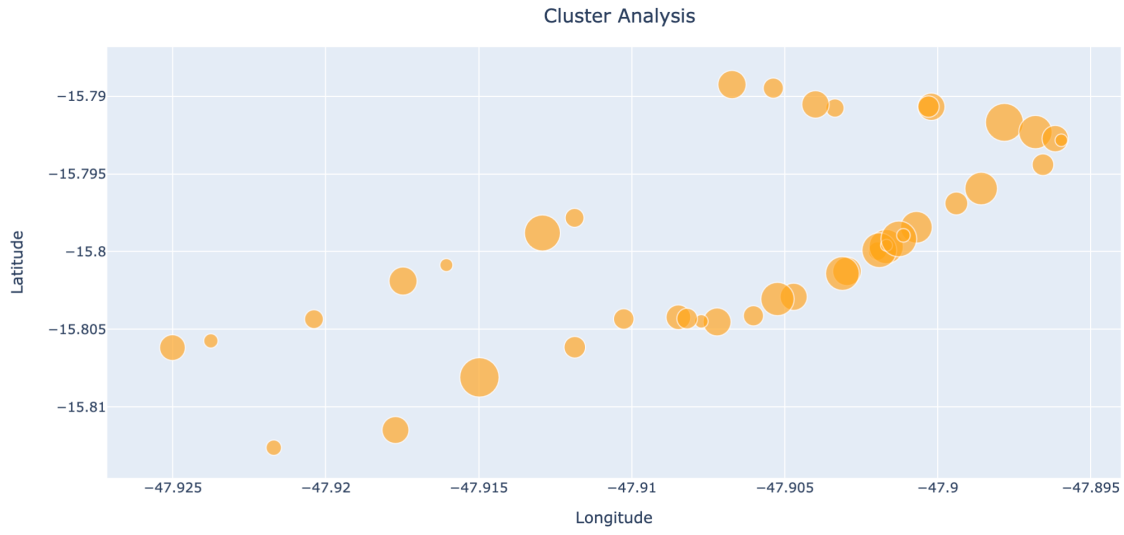


Figure 4.12 Orange Cluster Visualization

Table 4.7 Data description for the Orange Cluster

index	accex	accey	accez	climb	gyrox	gyroy	gyroz	us_lpd	speed
count	41.0	41.0	41.0	41.0	41.0	41.0	41.0	41.0	41.0
mean	5.8027	2.4324	-1.6007	0.0012	46.1031	11.6373	-3.7984	302.871	7.9931
std	3.6623	5.3804	5.8168	0.091	58.113	23.9871	26.6173	174.8356	1.4065
min	-3.5889	-13.1585	-1.4358	-0.199	-10.3969	-41.9618	-82.9771	63.54	4.761
25%	3.9265	0.6895	-4.4221	-0.065	16.5267	1.3664	-14.3664	160.03	6.757
50%	5.8778	3.5937	-2.1548	0.015	26.8931	7.3511	-2.3893	297.7	8.15
75%	7.9703	5.1882	0.6656	0.061	49.9542	18.4885	10.6947	451.22	9.167
max	14.224	10.0149	19.6127	0.195	250.1298	101.7176	56.771	693.17	10.257

c) K-means Clustering Results - Blue Cluster

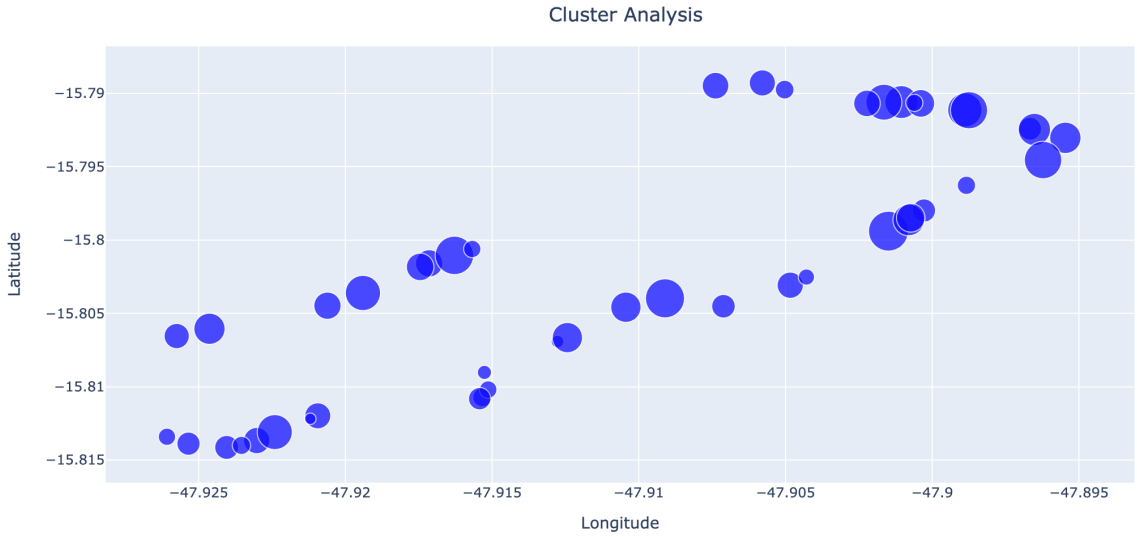


Figure 4.13 Blue Cluster Visualization

Table 4.8 Data description for the Blue Cluster

index	accex	accey	accez	climb	gyrox	gyroy	gyroz	us_lpd	speed
count	46.0	46.0	46.0	46.0	46.0	46.0	46.0	46.0	46.0
mean	-2.6107	11.4746	-6.0617	-0.0507	-27.6343	-12.3555	-12.9323	305.4096	8.2485
std	4.0497	3.7935	4.747	0.0755	38.597	14.4352	24.1817	161.7827	1.3665
min	-11.9471	1.7358	-17.03	-0.245	-190.6412	-69.2901	-88.2672	52.96	4.727
25%	-4.8794	9.1219	-9.396	-0.0928	-37.5973	-16.0172	-26.105	164.2375	7.6722
50%	-2.3571	12.0632	-5.527	-0.0585	-18.3588	-10.5382	-6.9962	285.25	8.516
75%	0.1808	14.1156	-2.5863	-0.0105	-6.7538	-3.1985	6.7118	404.915	9.239
max	7.9966	19.1177	5.3367	0.134	28.1756	7.2977	17.7405	648.89	10.626

4.2.2 Density-Based Spatial Clustering of Applications with Noise - DBSCAN

The DBSCAN model finds core samples of high density and expands clusters from them. Aimed to compare performance for a more appropriate cluster model to this study, the DBSCAN algorithm was also applied regarding the collected data. This model is good for data that contains clusters of similar density. For performing the DBSCAN is necessary to specify two parameters: Eps and MinPts. Eps is the maximum distance between two points in the same cluster, and MinPts is the minimum number of points required to form a cluster, the neighbor common distances.

It is worth noting that DBSCAN is sensitive to the choice of Eps and MinPts, and finding the optimal values for these parameters can be challenging. In practice, it may be necessary to experiment with different values to find the best results. This study uses an interactive mathematical algorithm to randomly combine six different Epsilons into fifteen Minimal Points given a ninety combination to get the optimum values for clustering, as shown below. The whole algorithm written in Python is presented in the annexes.

1. Input for the Epsilons:

```
epsilons = np.linspace(0.01, 1, num=15)
epsilons
```

Output:

```
array([0.01      , 0.08071429, 0.15142857, 0.22214286, 0.29285714,
       0.36357143, 0.43428571, 0.505     , 0.57571429, 0.64642857,
       0.71714286, 0.78785714, 0.85857143, 0.92928571, 1.        ])
```

2. Input for the Minimal Samples:

```
min_samples = np.arange(2, 20, step=3)
min_samples
```

Output:

```
array([ 2,  5,  8, 11, 14, 17])
```

3. Input for the number of combinations

```
import itertools
combinations = list(itertools.product(epsilons, min_samples))
N = len(combinations)
```

```
N
```

Output

```
90
```

Regarding the same dataset used in the K-means algorithm, the results to DBSCAN method is presented as follows (Table 4.9).

Table 4.9 DBSCAN results from clustering

Best Epsilon	0.4342
Best Minimal Samples	2
Best Score	-0.48985380

After running the DBSCAN algorithm, it is possible to examine each cluster label assigned to each data point. Which can be either a positive integer (indicating that the point belongs to a cluster) or a negative, -1 (indicating that the point is considered noise and does not belong to any cluster). To interpret the results, it is possible to count the number of points in each cluster and visualize the clusters using a scatter plot graph or another visualization method.

The results (Table 4.10) shows that the DBSCAN model needed better to fit this specific research question regarding the dataset. Most of the data fit into the noise values; furthermore, the silhouette score was -0.489853, indicating a bad clustering operation.

Table 4.10 DBSCAN number of points in each cluster

Noise values	-1	156
Borders values	0	3
Clustered	1	2

The visualization (Figure 4.14) shows that the clustering method does not work well in this case. It is further discussed in the next session, comparing the application of the two models.

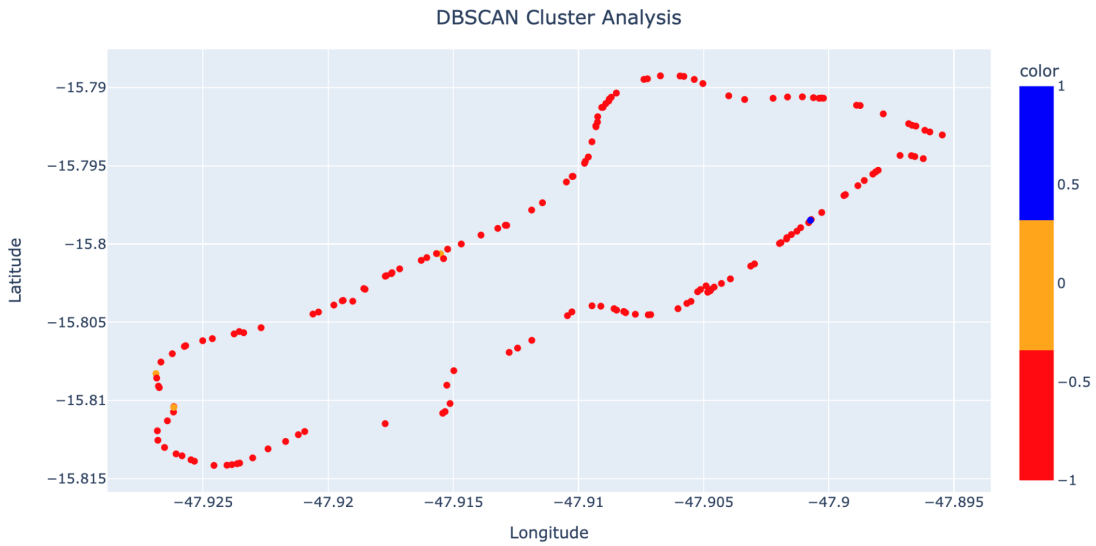


Figure 4.14 DBSCAN Clustering Visualization

4.2.3 K-Means VS. DBSCAN: Results summary

For this research and regarding the dataset provided from the sensor device collection, the K-Means clustering algorithm had a better performance than DBSCAN. The K-Means model clustered the dataset into 3 clusters with a silhouette score of 0.13826682 against a -0.48985380 score from DBSCAN (Table 4.11).

Table 4.11 Unsupervised Learning Models Comparison

Unsupervised Model	Model Score	No. Clusters
DBSCAN	-0.48985380	3
K-Means	0.13826682	3

Considering the calculus method mentioned earlier for DBSCAN, the limited amount of data can significantly impact the model's performance. The data were scattered along the route, making it impossible to create clusters using the nearby neighbor's clustering method, increasing the noise related.

Considering the perspective of data volume and the types of features involved, such as coordinates, velocity, and distances, as well as the use of georeferenced information, K-Means clustering proved to be a better choice for this context.

5 CONCLUSIONS AND RECOMMENDATIONS

5.1 CONCLUSIONS

This research aimed to estimate hazardous areas for bicycle mobility by utilizing an unsupervised machine-learning algorithm based on a sensor device for data collection. The Lateral Passing Distance (LPD) data collected among bicycles and vehicles were related to a variate cyclist data. Some of this data includes bicycle velocity, curse elevation, global positioning system (GPS) coordinates, acceleration, and gyroscopic information through a field operational data collection on the street.

The methodology was applied to a case study in Brasília's downtown avenue regarding a shared pathway around the principal local City Park, the *Parque da Cidade Sarah Kubitchek*. The road is 14 kilometers long, with two 3.50 meters lanes each way, one of it shared with cyclists. The total distance to complete all the lap was about 9.80 kilometers. Four entire laps on different days, always at the same time, were used in this research.

The data were collected using the BSafe360 Sensor Device from the C2SMART Center (NYU), employing a proactive research approach. It worth to note that field operational data collection practices can occasionally be hazardous, particularly when automobiles are passing too close to bicycles. However, implementing a sensor device proved valuable in gathering data for active mobility research. It enabled researchers to capture real-world data and gain insights into bicycle and vehicle interaction dynamics, enhancing the understanding of safety issues in cycling environments.

After data cleaning and exploratory analysis, 160 Lateral Passing Distances (LPD) readings were included. The descriptive statistics reveal that 25% of the LPD readings were less than 139.62cm, 50% were less than 238.78cm, and 75% of the readings were less than 351.08cm. These statistics provided a snapshot of the LPD values across the entire dataset and were obtained through an exploratory data analysis before applying the clustering models.

During the exploratory analysis, the LPD feature exhibited a weak negative correlation (-0.11) with the Climb feature during the general analysis. This suggests that as the bicycle encounters ascendent movements, the LPD tends to decrease. In other words, fewer LPD measurements

during ascents indicated a higher likelihood of risky proximity between bicycles and automobiles. Considering the threshold of 150cm, which aligns with the minimal lateral clearance distance law adopted in Brazil, the percentage of values in the ascending movement was found to be 4.35% greater than that in downhill routes. It can be seen more significantly in the clustering analysis. These findings highlight the importance of addressing safety concerns and implementing bigger measures to ensure sufficient lateral clearance during ascents, where the risk of close encounters between bicycles and vehicles appears to be higher.

These results indicate that from the cyclist's perspective, there is more movement in the lateral bicycle axis on ascending routes. It can be explained regarding the lateral movement of the bicycle and cyclist by applying force on the pedal and pendulum around it to move uphill.

Moreover, relating the LPD with speed, it had a positive correlation (+0.14). It assumes a small LPD with lower rates, as in climbing movements. The speed vs. climb relation had a moderate negative correlation (-0.23), with less rate in ascendant moves. However, these values were better correlated with the cluster models applied.

The application of an unsupervised learning model proved to be a valuable research tool for the analysis of lateral passing distance, enabling clustering information based on multiple variables.

Through the modeling process, distinct patterns were identified in the data, forming three clusters. The particular interest is the red cluster, which exhibited the lowest indices for LPD and was predominantly located in ascendent areas with high traffic volume.

Further analysis of the red cluster revealed essential statistics. It consisted 75 readings with a mean LPD by 206.51cm and a standard deviation of 120.35cm. Additionally, 25% of the LPD readings were less than 100.13cm, while 50% were below of 193.69cm in the second quartile. These statistics indicate critical LPD within this cluster, particularly when considering the minimal lateral clearance distance law adopted in Brazil, which has a threshold of 150cm. These were concentrated in strategy locals for the local mobility.

The volume of automobiles at a specific point is much higher than at other locations. This area is entirely located in the red cluster, indicating the lowest LPD and highest risk to bicycle-shared mobility in general.

Overall, these findings highlight the significance of the clustering model in identifying hazardous areas with insufficient LPD, especially in ascendant regions with high traffic volume. It emphasizes the need for safety improvements in these areas to ensure compliance with the minimal lateral clearance distance regulations, in a infrastructure perspective.

The methodology employed could be applied to other spots in the city to estimate hazardous areas for active mobility. The findings of this study can help transportation planners and policymakers applying the method aiming to make informed decisions to improve the safety of cycling infrastructure, and road safety in general.

Finally, all the script used in this research is available online. The access link is in the appendix session.

5.2 RESEARCH LIMITATIONS AND RECOMMENDATIONS FOR FUTURE RESEARCH

Some limitations were observed in this research. It will be presented in this session, aiming to recommend best practices for future research, leading to a more effective and evidence-based interventions to improve road safety regarding the cycling infrastructure.

First, using a single researcher for data collection aimed to minimize behavioral variation from a mobility perspective was a limitation. This approach resulted in the limited time available for more data collection, since obtaining a larger volume of data would enhance the machine learning model's performance and provide a more robust statistics analysis.

Moreover, low significance of crash registered data. The crash register obtained from the public organization had limited significance within the provided temporal window. Validating the results obtained by the clustering model through additional temporal analyses would be valuable (BERNARDES *et al*., 2023). This would enhance the confidence and reliability of the findings.

Furthermore, challenges in measuring traffic volume. Traffic volume is a relevant feature for this analysis. However, accurately measuring the volume at the six entrances and exits of the City Park area simultaneously proved to be challenging. Since the avenue around the City Park operates as a closed system, being possible the complete measurement of vehicles within it, a larger number of volunteers is required. This research collected data from three points simultaneously. Still, this study did not consider it in the analysis as a feature of the cluster model, limiting a comprehensive understanding of the entire system. Future research should explore methods to measure traffic volume effectively or consider the complete system, whether in this location, for a more conclusive analysis.

Since some studies suggest that introducing dedicated bicycle lanes improves cyclists' safety and reduces potential conflicts between motorized vehicles. These conflicts arise from lane-changing or encroaching vehicles that are passing cyclists (MEHTA *et al*., 2015).

Therefore, one recommendation for future research is to use the presented methodology to compare different bicycle lane infrastructures. Also, it is recommended to proceed on more locations with a significant historical crash to compare the results with the number of cyclists involved in collisions with vehicles. Regarding the unsupervised learning model, it is recommended to apply DBSCAN with more data collected. Thus, it is possible to analyze the behavior regarding the density incidence. It can result in better clustering results regarding the clustering score by the algorithm.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The author would like to express gratitude to Suzana Duran Bernardes, researcher at the New York University (NYU), and her advisor, Kaan Ozbay, from the Tandon School of Engineering, for providing the equipment through a research partnership between the universities. Additionally, appreciation is extended to the University of Brasília for enabling this study and numerous other researches endeavors.

REFERENCES

AASHTO (2010) Highway Safety Manual, first ed. American Association of State Highway and Transportation Officials, Washington, D.C.

ALDRED, R. & A. GOODMAN (2018) Predictors of the frequency and subjective experience of cycling near misses: Findings from the first two years of the UK Near Miss Project. *Accident Analysis and Prevention*, v. 110, p. 161-170.

AMBROŽ, M. (2017) Raspberry Pi as a low-cost data acquisition system for human powered vehicles. Measurement, v. 100, p. 7–18.

AMPE, T.; B. de GEUS; I. WALKER; B. SERRIEN; B. TRUYEN; H. DURLET & R. MEEUSEN (2020) The impact of a child bike seat and trailer on the objective overtaking behavior of motorized vehicles passing cyclists. *Transportation Research Part F*: *Traffic Psychology and Behaviour*, v. 75, p. 55-65.

ANDERSON, T.K., (2009) Kernel density estimation and k-means clustering to profile road accident hotspots. *Accident Analysis and Prevention*. 41 (3), 359e364.

ANG L-M., K.P. SENG (2016) Big sensor data applications in urban environments. *Big Data Research*, 4 , pp. 1-12.

ANN FORSYTH & KRIZEK, K. (2011) Urban Design: Is there a Distinctive View from the Bicycle?, *Journal of Urban Design*.

ARTHUR, D. & VASSILVITSKII, S. (2006) "k-means++: The advantages of careful seeding". Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms, Society for Industrial and Applied Mathematics.

BACCHIERI, G., J D BARROS, A, MOTTA, J & GIGANTE, D *et al*. (2009) Cycling to work in Brazil: Users profile, risk behaviors, and traffic accident occurrence. *Accident Analysis and Prevention*. TY - JOUR

BERNARDES, S. D.; A. KURKCU & K. OZBAY (2019) Design, Implementation and Testing of a New Mobile Multi- Function Sensing Device for Identifying High-Risk Areas for Bicyclists in Highly Congested Urban Streets. Mobi SPC, Halifax, Canada.

BERNARDES, S. D. & OZBAY, K. (2023) Derivation Of Surrogate Safety Measures From Lateral Passing Distance Between Vehicles And Bicycles. TRB Annual Meeting 2023, TRBAM-23-04188.

BERNARDES, S. D. & OZBAY, K. (2023) BSafe-360: An All-in-One Naturalistic Cycling Data Collection Tool. Sensors. 2023, 23, 6471.

BAHMANKHAH, B.; P. FERNANDES; J. FERREIRA; J. BANDEIRA; J. SANTOS & M. C. COELHO (2020) Assessing the overtaking lateral distance between motor vehicles and bicycles-influence on energy consumption and road safety. *Advances in Intelligent Systems and Computing*, 1083 AISC, p. 174-189.

BECK, B.; D. CHONG; J. OLIVIER; M. PERKINS; A. TSAY; A. Rushford & M. Johnson (2019) How much space to drivers provide when passing cyclists? Understanding the impact of motor vehicle and infrastructure characteristics on passing distance. *Accident Analysis and Prevention*, v. 128, p. 253–260.

BLACK, A. A.; R. DUFF; M. HUTCHINSON; I. NG; K. PHILLIPS; K. ROSE; A. USSHER & J. M. WOOD (2020) Effects of night-time bicycling visibility aids on vehicle passing distance. Accident Analysis and Prevention, v. 144, p. 105636.

CHAPMAN, J. & D. A. NOYCE (2012) Observations of driver behavior during overtaking of bicycles on rural roads. *Transportation Research Record: Journal of the Transportation Research Board*, v. 2321, p. 38–45.

CHAPMAN, J. R. & D. A. NOYCE (2014) Influence of roadway geometric elements on driver behavior when overtaking bicycles on rural roads. *Journal of Traffic and Transportation Engineering* (English Edition), v. 1, p. 28–38.

CHUANG, K. H.; C. C. HSU; C. H. LAI; J. L. DOONG & M. C. JENG (2013) The use of a quasi-naturalistic riding method to investigate bicyclists' behaviors when motorists pass. *Accident Analysis and Prevention*, v. 56, p. 32–41.

DE CEUNYNCK, T.; B. DORLEMAN; S. DANIELS; A. LAURESHYN; T. BRIJS; E. HERMANS & G. Wets (2017) Sharing is (s)caring? Interactions between buses and bicyclists on bus lanes shared with bicyclists. Transp. *Res. Part F: Traffic Psychology and Behaviour*, v. 46, 301–315.

DOZZA, M.; R. SCHINDLER; G. BIANCHI-PICCININI & J. Karlsson (2016) How do drivers overtake cyclists? *Accident Analysis and Prevention*, v. 88, p. 29-36.

FEIZI, A.; J.-S. Oh; V. KWIGIZILE and S. Joo (2019) Cycling environment analysis by bicyclists' skill levels using instrumented probe bicycle (IPB). *International Journal of Sustainable Transportation*, v. 14, n. 9, p. 722-732.

FENG, F.; S. BAO; R. C. HAMPSHIRE & M. DELP (2018) Drivers overtaking bicyclists—An examination using naturalistic driving data. *Accident Analysis and Prevention,* v. 115, p. 98–109.

FORSYTH, A. & K. J. KRIZEK (2011a) Promoting walking and bicycling: Assessing the evidence to assist planners. *Built Environment*. v. 37, n. 4, p. 429-446.

FORSYTH, A, and K. J. KRIZEK (2011b) Urban Design: Is there a Distinctive View from the Bicycle? *Journal of Urban Design*, v. 16, n. 4, p. 531-549.

FOURNIER, N., BAKHTIARI, S., VALLURU, K.D., CAMPBELL, N., CHRISTOFA, E., ROBERTS, S., KNODLER Jr, M. (2020). Accounting for drivers' bicycling frequency and familiarity with bicycle infrastructure treatments when evaluating safety. Accident Analysis and Prevention, 137, 105410.

GHADGE, M. D. PANDEY & D. KALBANDE (2015) "Machine learning approach for predicting bumps on road," International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), 2015, pp. 481-485.

GADSBY, A. & K. WATKINS (2020) Instrumented bikes and their use in studies on transportation behaviour, safety, and maintenance. *Transportation Reviews*, v. 40, p.774–795.

GUERRERO-IBÁÑEZ, J.; S. ZEADALLY & J. CONTRERAS-CASTILLO (2018) Sensor Technologies for Intelligent Transportation Systems. *Sensors*, v. 18, n. 4, p. 1212.

GUSTAFSSON, L., & ARCHER, J. (2013). A naturalistic study of commuter cyclists in the greater Stockholm area. Accident Analysis and Prevention, 58, 286–298.

HAWORTH, N.; K. C. HEESCH; A. SCHRAMM & A. K. DEBNATH (2018) Do Australian drivers give female cyclists more room when passing? *Journal of Transport and Health*, v. 9, p. 203–211.

HØYE, A.; A. FYHRI & T. BJØRNSKAU (2016) Shared Road is double happiness: evaluation of a "Share the road" sign. *Transportation Research Part F: Traffic Psychology and Behaviour*, v. 42, p. 500–508.

HUERTAS-LEYVA, P., DOZZA, M., & BALDANZINI, N. (2018) Investigating cycling kinematics and braking maneuvers in the real world: E-bikes make cyclists move faster, brake harder, and experience new conflicts. Transportation Research Part F: Traffic Psychology and Behaviour, 54, 211–222.

JAIN, A.K.; MURTY, M.N.; FLYNN, P.J. (1999). Data clustering: A review. ACM Comput. Surv. (CSUR) 1999, 31, 264–323.

KAY, J. J.; P. T. SAVOLAINEN; T. J. GATES & T. K. DATTA (2014) Driver behavior during bicycle passing maneuvers in response to a share the road sign treatment. *Accident Analysis and Prevention*, v. 70, p. 92–99.

KOVACEVA, JORDANKA., Gustav NERO, Jonas BÄRGMAN, Marco DOZZA (2019) Drivers overtaking cyclists in the real-world: Evidence from a naturalistic driving study, Safety Science, Volume 119, 2019, Pages 199-206.

KIM, K., YAMASHITA, E.Y., (2007) Using a k-means clustering algorithm to examine patterns of pedestrian involved crashes in honolulu, Hawaii. J. Adv. Transport. 41 (1), 69e89.

KUBAT, M. (2017) An Introduction to Machine Learning; Springer: Cham, Switzerland, 2017.

KURKCU, A. & K. OZBAY (2017) Estimating Pedestrian Densities, Wait Times, and Flows with Wi-Fi and Bluetooth Sensors. Transportation Research Record. 2644(1): p. 72-82.

LIM, C., Kwang-Jae KIM, Paul P. MAGLIO, (2018) Smart cities with big data: Reference models, challenges, and considerations, Cities, Volume 82, 2018, Pages 86-99.

LAMONDIA, J. & J. DUTHIE (2012) Analysis of factors influencing bicycle-vehicle interactions on urban roadways by ordered probit regression. *Transportation Research Record*, v. 2314, p. 81–88.

LI, C.; S. SUN & J. GUO (2015) Evaluation the impacts of bicycle-sharing systems on carbon emission reductions-empirical study in Beijing. Presented at *94th Annual Meeting of the Transportation Research Board*, Washington, D.C. 14 p.

LI, Z.; W. WANG; P. LIU; J. BIGHAM and D. R. RAGLAND (2012) Modeling bicycle passing maneuvers on multilane separated bicycle paths. *Journal of Transportation Engineering, v.* 139, n. 1, p. 57–64.

Lim, CHIEHYEON., Kwang-Jae KIM, Paul P. MAGLIO (2018), Smart cities with big data: Reference models, challenges, and considerations, Cities, Volume 82, 2018, Pages 86-99.

LLORCA, C.; A. ANGEL-DOMENECH; F. AGUSTIN-GOMEZ & A. GARCÍA (2017) Motor vehicles overtaking cyclists on two-lane rural roads: analysis on speed and lateral clearance. *Safety Science, v.* 92, p. 302–310.

LORD, D.; X. QIN & S. R. GEEDIPALLY (2021) *Highway Safety Analysis and Modeling*. Elsevier: Oxford, UK.

LOVE, D. C.; A. BREAUD; S. BURNS; J. MARGULIES; M. ROMANO & R. LAWRENCE (2012) Is the three-foot bicycle passing law working in Baltimore, Maryland? *Accident Analysis and Prevention*. V. 48, p. 451-456.

MACKENZIE, J. R. R.; J. K. DUTSCHKE and G. Ponte (2021) An investigation of cyclist passing distances in the Australian Capital Territory, *Accident Analysis and Prevention*, Volume 154, 2021.

MAURO, R., De LUCA, M., DELL'ACQUA, G., (2013) Using a k-means clustering algorithm to examine patterns of vehicle crashes in before-after analysis. Mod. Appl. Sci. 7 (10), 11.

MEHTA, K.; B. MEHRAN and B. HELLINGA (2015) Evaluation of the Passing Behavior of Motorized Vehicles When Overtaking Bicycles on Urban Arterial Roadways. *Transportation Research Record: Journal of the*

Transportation Research Board, No. 2520, *Transportation Research Board*, Washington, D.C., 2015, pp. 8–17.

MEHTA, K.; B. MEHRAN & B. HELLINGA (2019) A methodology to estimate the number of unsafe vehicle cyclist passing events on urban arterials. *Accident Analysis and Prevention*, v. 124, p. 92–103.

MOHER, D.; L. SHAMSEER; M. CLARKE; D. GHERSI; A. LIBERATI; M. PETTICREW & L. A. Stewart (2015) Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Systematic Reviews*, v. 4, v. 1.

MUELLER, N.; D. ROJAS-RUEDA; X. BASAGAÑA; M. CIRACH; T. COLE-HUNTER; P. DADVAND; D. DONAIRE-GONZALEZ; M. FORASTER; M. GASCON; D. MARTINEZ, & M. NIEUWENHUIJSEN (2015) Health impacts related to urban and transport planning: A burden of disease assessment. *Environment International*, v. 91, p. 138-146.

NGUYEN, H., KIEU, L.-M., WEN, T. & CAI, C. (2018) Deep learning methods in transportation domain: a review. IET Intell. Transp. Syst., 12: 998-1004.

OZBAY, K.; A. KURKCU & H. YANG (2018) Portable and Integrated Multi-Sensor System for Data-Driven Performance Evaluation of Urban Transportation Networks. *Transport Research International Documentation* - TRID.

ROSSI, R.; F. ORSINI; M. TAGLIABUE; L. L. Di STASI; G. De Cet and M. Gastaldi (2021) Evaluating the impact of real-time coaching programs on drivers overtaking cyclists. *Transportation Research Part F: Traffic Psychology and Behaviour*, v. 78, n. 1, p. 74-90.

RUBIE, E.; N. HAWORTH; D. TWISK & N. YAMAMOTO (2020) Influences on lateral passing distance when motor vehicles overtake bicycles: A systematic literature review. *Transport Reviews*, v. 40 n. 6, pp. 754 773.

SAVOLAINEN, P.; T. GATES; R. TODD; T. DATTA & J. MORENA (2012) Lateral placement of motor vehicles when passing bicyclists. *Transportation Research Record: Journal of the Transportation Research Board*, v. 2314, n. 1.

SELVI, H.Z., CAGLAR, B., (2018) Using cluster analysis methods for multivariate mapping of traffic accidents. Open Geosci. 10 (1), 772e781.

SCHRAMM, A., HAWORTH, N., HEESCH, K., WATSON, A., & DEBNATH, A. (2016) Evaluation of the Queensland minimum passing distance road rule. Department of Transport and Main Roads.

SHACKEL, S. C. & J. PARKIN (2014) Influence of road markings, lane widths and driver behaviour on proximity and speed of vehicles overtaking cyclists. *Accident Analysis and Prevention*, v. 73, p. 100–108.

VANPARIJS, J.; L. I. PANIS; R. MEEUSEN & B. de GEUS (2015) Exposure measurement in bicycle safety analysis: A review of the literature. *Accident Analysis and Prevention*, v.84, p.9-19.

VAN ECK, N. J. & L. WALTMAN (2010) VOSViewer: Visualizing Scientific Landscapes [Software]. Available from https://www.vosviewer.com

ZANTALIS, F. S.; KOULOURAS, G.; KARABETSOS, S.; KANDRIS, D. (2019). A Review of Machine Learning and IoT in Smart Transportation. Future Internet 2019, 11, 94.

WESTERHUIS, F., & de WAARD, D. (2016) Using Commercial GPS Action cameras for Gathering natura- listic cycling data. Journal of the Society of Instrument and Control Engineers, 55(5), 422–430.

WALKER, I. (2007) Drivers overtaking bicyclists: Objective data on the effects of riding position, helmet use, vehicle type and apparent gender. *Accident Analysis and Prevention*, v. 39, n.2, p.417-425.

WALKER, I.; I. GARRARD & F. JOWITT (2014) The influence of a bicycle commuter's appearance on drivers' overtaking proximities: An on-road test of bicyclist stereotypes, high-visibility clothing and safety aids in the United Kingdom. *Accident Analysis and Prevention*, v. 64, p. 69–77.

WANG, Y.; ZHANG, D.; LIU, Y.; BO DAI, LEE, L, H.; Enhancing transportation systems via deep learning: A survey, Transportation Research Part C: Emerging Technologies, Volume 99, 2019.

WORLD HEALTH ORGANIZATION (2018) *Global status report on road safety 2018*. Geneva, Switzerland.

APPENDIX A

All the Python script used for the exploratory data analysis, unsupervised learning model application, statistics, and visualizations are available in this GitHub repository, as well as the data:

https://github.com/marcellmello/msc-bicycle-safety

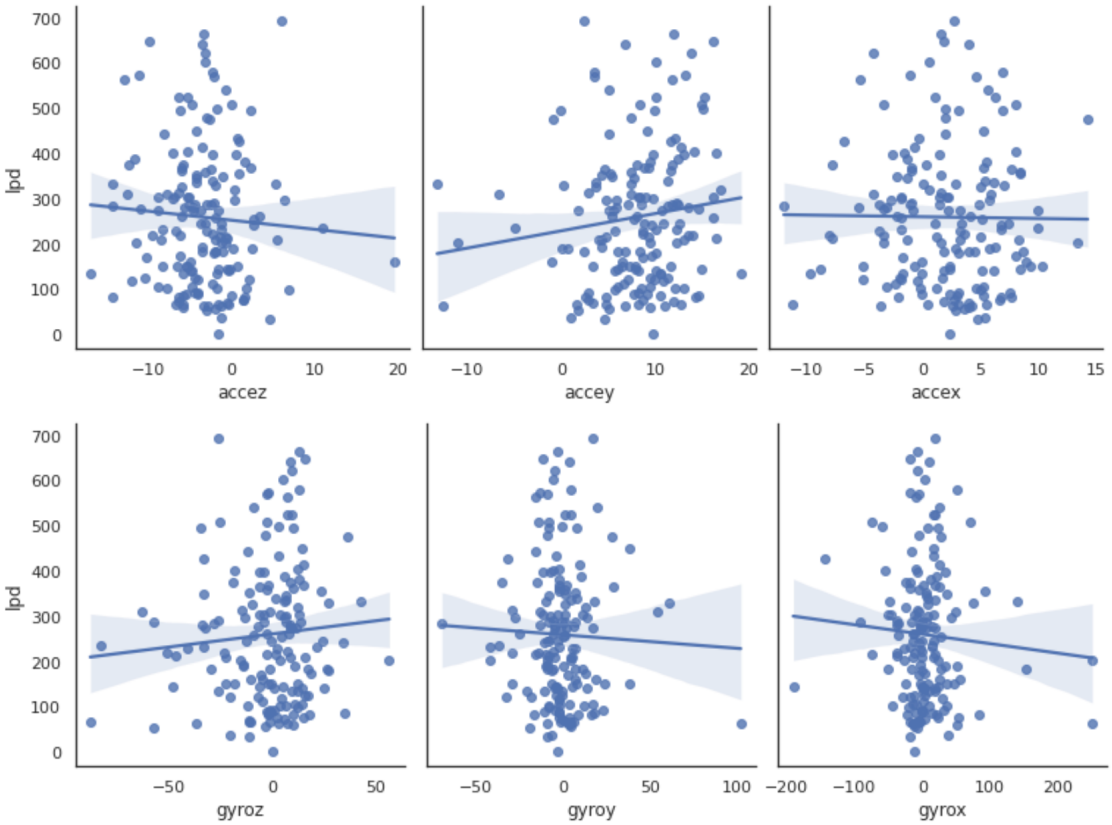


Figure A.1 Relation between the 3-axis accelerometer, gyroscope, and Lateral Passing Distance (LPD)

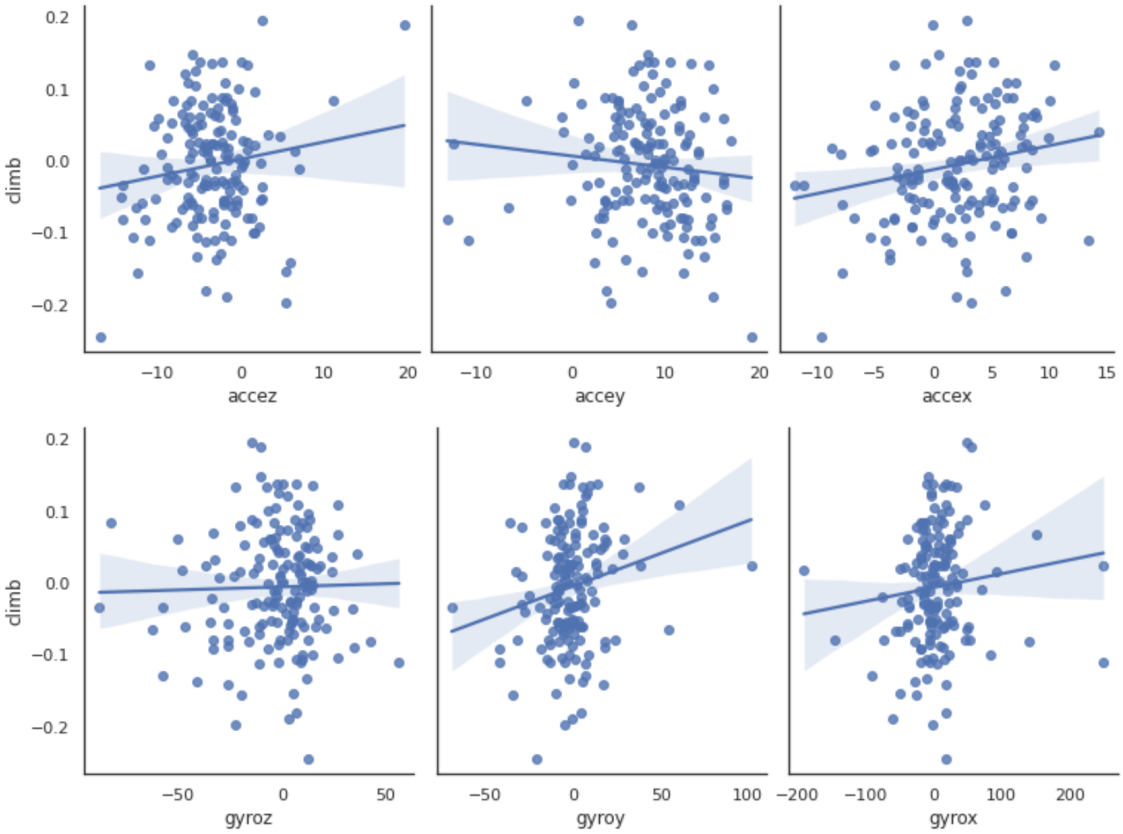


Figure A.2 Relation between the 3-axis accelerometer, gyroscope and Climb