Escuela Superior
de Ingeniería y Tecnología
Universidad de La Laguna

TRABAJO FIN DE GRADO

DEVELOPMENT OF REPRESENTATIVE DRIVING CYCLES OF THE TENERIFE METROPOLITAN AREA THROUGH CLUSTERING METHODS

GRADO EN INGENIERÍA MECÁNICA

Carlos Enrique Armas Palmero

Tutor: Oscar García Afonso

La Laguna, Julio 2021

CONTENTS

LIST OF FIGURES

ABSTRACT

In recent years, owing to the evolution of science and technology, more efficient methods of analysing high-dimensional data have been developed. Drawing on this progress, data science can be applied to the environmental sector, helping to determine more accurately the impact of vehicles emissions on the environment through representative driving cycles. This study aims to develop a methodology that helps to build representative driving cycles from a data set collected in the metropolitan area of Tenerife. The methodology proposed in this study consisted of the division of driving cycles into segments (Microtrips) and subsequently applying various clustering algorithms (k-means and Hierarchical clustering), following the application of a dimensionality reduction methodology (t-SNE and PCA). The results were split into groups with similar acceleration-related variables, representing the driving behaviours. . It was found that the highest quality clusters, assessed through silhouette coefficients, Calinski-harabasz and Davies Bouldin index, resulted from the utilization of a combination of t-SNE and k-means. The representative Microtrips were then merged to obtain the final cycle. This methodology seemed to be unable to satisfy the desired cycle duration without affecting the data's representativeness. However, when the final cycle was compared to the data set, the resulting discrepancies were deemed acceptable

CHAPTER 1

INTRODUCTION

Since decades before the beginning of the 21st century, joined by a progressively more intense climate change and more notorious contamination, the environment has played an increasingly important role in society and human behaviour. It has been reported that, in Spain (2013), about 23.940 people died prematurely due to air pollution. Additionally, 4.280 died due to the effects of NOx. Both contaminants are usually emitted by internal combustion engine (ICE) vehicles, which according to studies [1] produce about 13% of the air contamination in the European Union.

Many measures have been taken by different administrations around the world, some of which regulate the emission produced by ICE. In order to decrease such emissions, by law, car manufacturers have to test their new vehicle models. The most recent EU regulation is under the name EURO 6, which prevents light-duty vehicles from emitting more than 95 g CO_2/km (cars) and 147 g CO_2/km (vans) (2020-2024 objectives), supposing a reduction of 15% in CO_2 emissions.

Following the homologation procedures, vehicles are tested, in laboratories, under specific driving conditions while connected to an emission measurement system. Those vehicles with lower emissions than the limit can be driven without any specific restriction.

As can be inferred, test driving conditions play an important role when it comes to the emissions produced by a vehicle. For that reason, it is fundamental to develop the most realistic controlled driving environment of a certain country or region. This can be achieved by developing representative driving cycles. Theoretically, this driving cycle needs to contain the different driving conditions of the region considering its statistical proportion. This driving cycle needs to illustrate the average daily cycle of the vehicle through its life.

Currently, the most used driving cycle around the world is the WLTP (*New Worldwide Harmonized Light Vehicles Test Procedure*), which represents different driving conditions for different types of vehicles. It is important to highlight the fact that previously, the homologation procedure in the EU was carried out through the implementation of the New European Driving Cycle (NEDC), developed at the end of the 20th century. According to the European Commission, from 2021 onwards, the emissions targets for manufacturers will be based on the WLTP. One of the main reasons argued in favour of withdrawing the NEDC from the homologation procedures is its dissimilarity to real driving conditions as it will be shown later in this study. Since then, many methods have been proposed to recreate the most representative driving conditions through the driving cycles development.

Taking advantage of the increasingly advancing technology, it is possible to recreate representative driving conditions from statistical analysis applied to a large amount of data by using computational algorithms and data analytics tools. The driving cycles obtained from the statistics process must be the most representative of a large set of driving cycles. One of the most widely used methods to obtain driving cycles consists of the application of clustering algorithms to cycle subdivisions after obtaining the driving features. Finally, as can be inferred, driving cycles will vary depending on different factors, mainly influenced by road infrastructure and region.

AIM AND OBJECTIVES

This study aimed to develop a new methodology to obtain representative driving cycles from a data set of specific driving conditions in the Tenerife metropolitan area.

The following specific objectives are enumerated:

1. Conduct a statistical study of 490 cycles, evaluating representativeness and different parameters.

2. Determine the most representative driving features based on driving conditions representativeness and variables dispersion, obtaining relationships between them.

3. Establish the best cluster conditions considering the proposed dimensionality reduction methods for the dataset by applying different clustering standards.

4. Evaluate and compare the results obtained from the cluster analysis with current driving cycles.

BACKGROUND

Many have been the methods employed to analyse the performance of ICE vehicles considering real driving conditions. The most usual methodology consists of testing prototypes in test benches by the manufacturers.

As it was said, the main objective of testing vehicles through driving cycles is to compare different vehicles in specific driving conditions, so it is possible to calculate the automobile's fuel economy, which can be defined as the amount of fuel consumed to travel a distance [2].

According to the U.S. Department of Energy, the fuel economy of light-duty vehicles ranks between 10 to 140 Miles per gallon (23.52 – 1.68 l/100km) and also explains that the methodology followed to carry out those calculations is to test the vehicles in dynamometers following standardized driving cycles (Federal Test Procedures) that mainly consist of EPA city, EPA highway, US06, SC03 and cold temperature cycle, where each cycle determines different road and operating conditions.

The FTP city cycle represents urban driving where the vehicle is driven in stop-and-go rush hour traffic. The FTP highway cycle represents a mixture of rural and interstate highways in free-flowing traffic. Finally, the US06 represents a city and highway driving at higher speeds with more aggressive acceleration and braking.

The aforementioned tests are defined by mean speed, top speed, acceleration, time, Idling, lab temperature and the possibility of having the vehicle's air conditioning/heater running. It is important to highlight the fact that this methodology is currently taking place in nowadays testing procedures.

As explained in [1], there are two main ways to develop a driving cycle according to its shape: representing the driving features in a highly stylized driving cycle (such as the Japanese 10.15 and NEDC) and selecting real-world fragments of driving cycles (according to its representativeness) for different driving conditions. The latter is used more often (as seen in the WLTP, US06 and LA4).

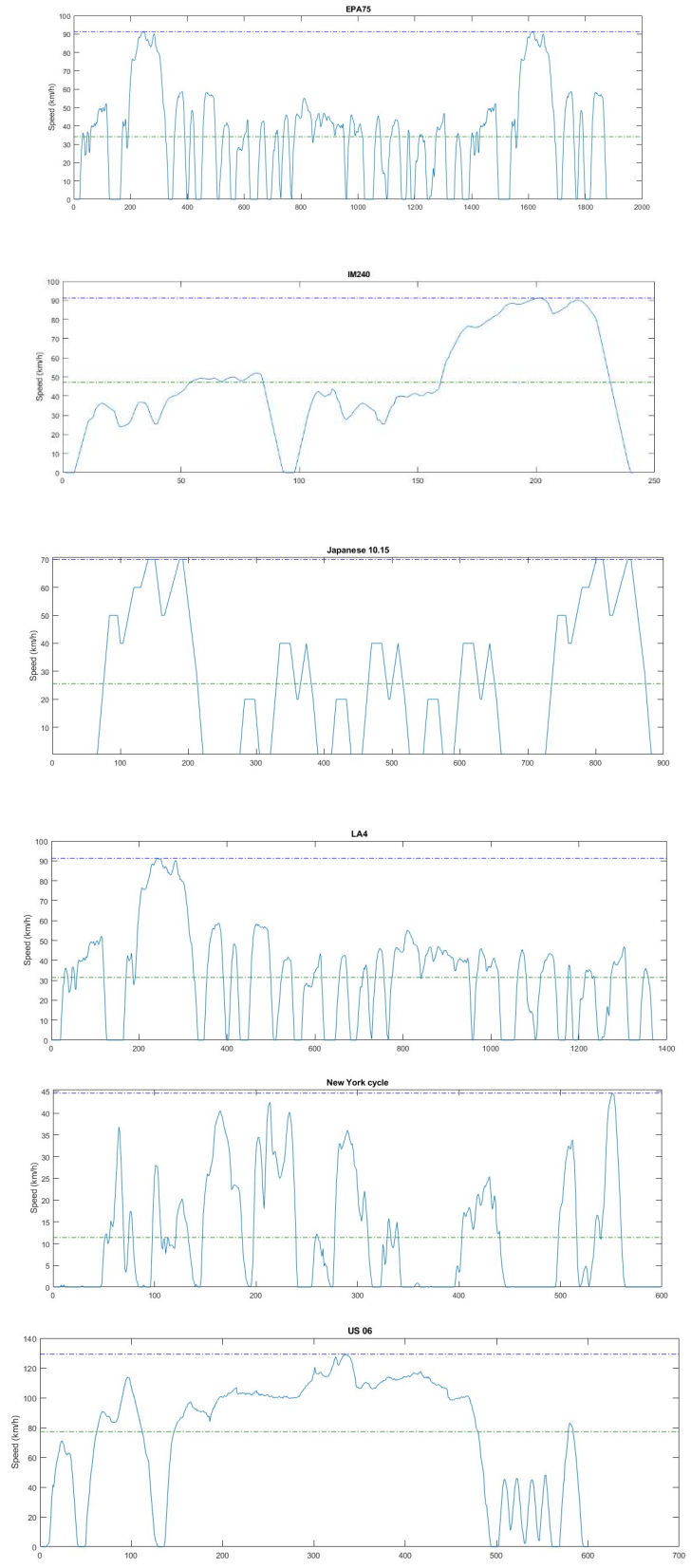

Figure 1.1: Standardized driving cycles. Source: U.S. Department of Energy.

On the other hand, the NEDC (New European Driving Cycle), last updated in 1997, was designed to represent light-duty vehicle usage in Europe which is highly criticized for not delivering real driving conditions; for this reason, the European Commission has introduced a new world harmonized cycle: WLTC (World Light Test Cycle) [5]. Another reason that boosts the NEDC withdrawal is its flexibility when it comes to interpreting different parameters of the cycle, such as some operating laboratory conditions that might affect the emissions produced. Additionally, it was proved [6] that the WLTP is more restrictive (better-defined boundaries) than NEDC in determinant factors such as the vehicle load, test mass, driving resistance forces and tire selection. As an example of the previously said, the electricity consumption of PHEV (Plug-in hybrid electric vehicles) under the WLTP is 26% higher than NEDC, which makes the electric range lower when the first procedure is followed.

Up to this point, the above-mentioned cycles were legislative, employed in type-approval tests whose only objective was to certify vehicles emissions. Another type of cycle is the non-legislative one, which is only used in research and will be the focus of the present thesis. As shown in figure 1.1, the non-legislative cycles are mainly focused on specific regions under specific driving conditions (i.e., New York cycle).

It is also highlightable the existence of pseudo-steady state cycles whose only aim is to determine the emissions produced at a regular speed without periods of significant acceleration. As mentioned in [4], these types of cycles are being excluded from current research due to several limitations, especially relevant in the case of catalyst-equipped modern vehicles, that can produce a large proportion of emissions in a short period during large accelerations.

RELATED WORK

Many methodologies have been developed to obtain regional representative driving cycles. As shown in [8] Fotouhi and Montazeri-Gh (2012) developed the Tehran representative driving cycle through clustering methods by selecting the most characteristic features of real-world data collected from Tehran roads. The methodology proposed the division of the driving cycles into small cycles decreasing the dispersion between clusters and increasing the number of observations, bringing more illustrative statistical models. The final driving cycle was obtained by linking the small trips into a single cycle following some representative parameters. Additionally, this article defines Microtrips (MTs) as subdivisions of the main driving cycle between idles.

Daniel Förster and Robert Inderka [9] in 2019 followed a similar methodology to the explained before, detailing the boundaries of the MTs and defining them as subdivisions of cycles between idling with a minimum duration of 60 seconds. It also dismisses MTs with unrealistic features (velocity, Idle and acceleration). It also establishes dependent driving features based on velocity and acceleration in order to decrease the number of variables. Finally, it determines a methodology to classify the driving behaviour, explaining its influence on specific features highly dependent on acceleration. This classification takes place assigning a final score to each MT.

J. Liu and X. Wang, [10] 2016 developed a model to determine the fuel efficiency of alternative fuel vehicles in order to ease the vehicle selection by users according to their region. They applied the clustering algorithm k-means to the data after employing a dimensionality reduction algorithm (PCA) to find the most influential parameters involved. The results showed that the clustering adjusts better to a dimensional-reduced dataset than to specifically selected features, displaying more defined boundaries.

In 2019 J. Huertas, L. Quirama, M. Giraldo and J. Díaz conducted a comparative study analyzing the different methodologies to assemble regional representative driving cycles. Those methods are Micro-Trips, Markov chains and MWD-CP, concluding that the most accurate methodology to develop the cycles is the last one. The MWD-CP (deterministic approach) is based on the selection of parameters related to the energy consumption in each cycle, and consequently, the cycle closest to the average energy consumption of the dataset is chosen as representative. It is mentioned that the drawback of the MWD-CP methodology is its lack of accuracy in some cases when the result's verification is needed through the employment of any other method.

Finally, A. Kabra at Blekinge Institute of Technology, Karlskrona, Sweden (2019), conducted an investigation related to clustering algorithms of driving data (from a pre-existing dataset), comparing the different algorithms and dimensionality reductions methods, concluding that the best performance was given by the k-means and the t-SNE method.

CHAPTER 2

DEFINITIONS

Driving features definition

In order to develop a statistical analysis applied to the driving cycles, it is necessary to determine the most representative driving features, defining the variables that can help the algorithm to identify the differences between driving conditions. Those variables can be determined from a bibliographic study.

Many studies have proposed mean speed, Idle and acceleration as representative driving features. In this study, as it was exposed in the methodology, the driving features will be collected from different scientific studies [6-9]. After the collection, features will be studied in order to select those with more variability and less relationship between them. However, the statistical analysis will be performed considering the most prominent number of possible features to reach a more detailed comparison between the raw data and the final representative cycle.

A table with the driving features studied in this thesis is presented as follows:

Feature	Abbreviation	Units
Mean speed	Vm	km/h
Mean driving speed	V	km/h
1st quartile speed		km/h
2nd quartile speed		km/h
3rd quartile speed		km/h
Maximum speed	Vmax	km/h
Specific kinetic energy	Ek	J/kg
Mean acceleration	Am	m/s2
Average positive acceleration	APA	m/s2
Average negative acceleration	ANA	m/s2
Relative positive acceleration	RPA	m/s2
Relative negative acceleration	RNA	m/s2
Maximum acceleration	Amax	m/s2
Minimum acceleration	Amin	m/s2
Time accelerating	Tacc	%
Time decelerating	Tdec	%
Time with high acceleration		%
Time with high deceleration		%
Idle time	Idle	%
Time with a positive grade		%
Time with a negative grade		%
Distance		km

Mean speed

Consists of the arithmetic mean calculated for every cycle as shown in equation 1, where *n* is the number of observations (in this case the duration of the driving cycle) and *v* the vehicle speed in the instant *i*:

$$\frac{1}{n}\sum_{i=1}^{n} vi \quad \text{(eq. 1)}$$

Mean driving speed

As it is described in the procedure followed for the WLTC construction [8], it is important to determine the real driving speed by isolating the instants where the vehicle is not moving (idle) as it lowers the mean speed giving unrealistic parameter values since the idle time is already being considered as a separate parameter. After the identification of the instants where the vehicle is not moving (v=0 km/h), the sum of the speed vector will be divided by the length of the vector minus the number of instances with idle, as expressed in equation 2:

$$\frac{1}{n-n_0} \sum_{i=0}^{n} vi \qquad \text{(eq.2)}$$

Speed quartiles

Theoretically, in order to prove the accuracy given by the speed-related features, speed quartiles are calculated. As can be inferred, the first quartile (or 25th percentile) separates the values from the lowest 25% of the mean speed, whereas the second quartile represents the median of the data set and, finally, the third quartile represents the 75% highest speed. As it is known, the second quartile corresponds to the median speed.

Maximum speed

This parameter is extracted from the speed vector. It is expected to obtain more than one frequency peak when it is evaluated in a histogram since different driving conditions are being considered. It is important to highlight its prominent influence on the driving profile since it is directly influenced by the road speed limit.

Kinetic energy

As it is described in the bibliography [8], it is relevant to determine a variable associated with kinetic energy. It is proposed to calculate the kinetic energy in every instance, given by equation 3.

$$\frac{1}{2} m \sum_{i=0}^{n} v_i^2 \quad \text{(eq. 3)}$$

As it was explained before, it is not known the vehicle technical description employed in each cycle, so it is not possible to consider the vehicle weight which is needed in the previous equation. It is proposed to calculate a variable dependent on only the squared speed.

$$\frac{1}{2} \sum_{i=0}^{n} v_i^2 \quad \text{(eq. 4)}$$

From this equation, it is noticeable that the proposed feature is strongly linked with the vehicle speed, so it may not give relevant information to the proposed study. From this variable, it will be possible to determine the variance through the employment of the next equation.

$$\sigma E = \frac{1}{n} \sum_{i=0}^{n} (v_i^2 - Ek)^2 \quad \text{(eq.5)}$$

Mean acceleration

As it was calculated for the driving speed, the acceleration average is determined for each cycle. It is expected to obtain near-zero values since acceleration and deceleration are considered for this calculation.

$$\frac{1}{n} \sum_{i=0}^{n} a_i \quad \text{(eq.6)}$$

Average positive/negative acceleration

The construction of two vectors takes place isolating positive and negative values of the main acceleration vector, consequently, each vector average is calculated. This parameter is commonly used in the identification of driving behaviours and on-road factors.

$$APA = \frac{1}{n}\sum_{i=1}^{n} a_i \ \ (a_i > 0) \quad \text{(eq. 7)}$$

$$APA = \frac{1}{n}\sum_{i=1}^{n} a_i \ \ (a_i < 0) \quad \text{(eq. 8)}$$

Relative negative/positive acceleration

It was found in different studies [3] that the RPA and RNA are parameters that describe the dynamics of the cycle. Additionally, it is widely employed for the construction of representative driving cycles. As defined in [3], the RPA and RNA are the integrals of the acceleration multiplied by the velocity and divided by the total cycle distance when the vehicle is accelerating and decelerating.

$$RPA = \frac{1}{Dist.}\int_0^t a_i . v_i \, dt \ (a_i > 0) \quad \text{(eq. 9)}$$

$$RNA = \frac{1}{Dist.}\int_0^t a_i . v_i \, dt \ (a_i < 0) \quad \text{(eq. 10)}$$

As it is inferred, the RPA and RNA associate the acceleration with the vehicle speed of an instant considering the distance travelled (of the analysed cycle).

Time accelerating/decelerating

This feature can be defined as the percentage of time spent increasing and decreasing the vehicle velocity. The length of the vectors employed to calculate APA and ANA are calculated and divided by the cycle's total duration.

Time with high/low acceleration

This parameter is calculated with the objective of determining the driving behaviour by calculating the percentage of time where the acceleration is higher and lower than the third quartile of the acceleration and deceleration vector respectively.

Idle time percentage

This feature represents the percentage of time where the vehicle's engine is idling (has a velocity of zero with the engine running). This usually occurs when the vehicle stops at a red light, waiting while parked or when the traffic conditions force it.

It is important to highlight that this is one of the most determinant parameters when constructing a driving cycle since the engine still produces emissions and the driven distance does not change. It is also strongly influenced by the road infrastructure, road type and traffic conditions.

Positive and negative grade

Another parameter provided by the GPS is the vehicle's grade. The grade is calculated by the device calculating the tangent of the difference between the elevation in two instances of time divided by the distance travelled in that period. It is highlightable that this parameter does not depend on the driving conditions nor driver's behaviour but on the road infrastructure which is strongly linked to the studied region/city. It can be noted that this parameter is not considered for the elaboration of the WLTC.

In this study, it is calculated the time when the grade is positive and negative and then divided by the total duration of the cycle. Additionally, it is calculated the time when the grade is higher and lower than 5%, receiving the name of Large negative/positive grade.

Statistical analysis

Before performing the clustering algorithm, it is important to build a statistical model in order to have an idea of how the different parameters behave and the distribution of said variables. The objective of this section is to demonstrate the representability of the data collected and the most influential variables. After this analysis, it will be possible to infer the ideal cluster results before the algorithm execution. In order to evaluate such representativeness, different parameters are plotted in histograms and bar charts.

Histograms

The invention of the histogram in the seventeenth century marked the beginning of modern statistics. Before the mentioned invention, statistical data came in form of large lists that made analysis and data interpretations less efficient and more arduous.

According to [13,14] a histogram is one of the main visual tools employed in descriptive statistics, that allows the recognition of outliers, gaps, shape (identifying peaks, symmetry and skewness) when quantitative data is analysed. A histogram can also provide an estimate of the underlying probability density function.

Finally, it is important to highlight that the skewness is the direction of the longer of the two tails of the distribution. It is also said that there is not a formal way of estimating the number of bins nor correct bin width, however, it is usually employed, in rough calculations, the Sturges' formula (eq. 11) and Cross-Validation.

$$Nº\ bins = 1 + \log_2 n \quad \text{(eq. 11)}$$

In the previous equation n is the number of observations in the dataset. In MATLAB [15] different aspects of the histogram can be changed, by assigning a specific number of bins, bar width, edges, and limits. When the number of bins is not specified by the user, MATLAB automatically calculates the ideal number of bins to be used for the processed data.

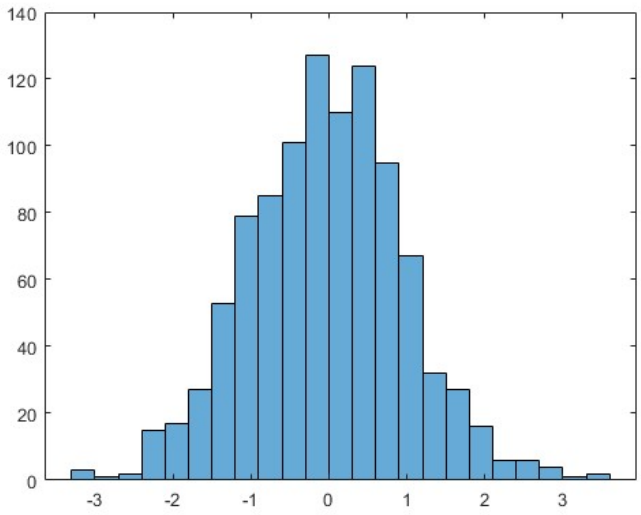

Figure 2.1: Sample histogram. Source: MathWorks

Cumulative distribution functions

The cumulative distribution determines, for each value, the fraction of the data less than or equal to the said value. It calculates the probability of selecting a value lower than a specific point. This type of distribution will be employed to determine the driving behaviour of the cycles. In order to find the correct type of distribution (when the data is not normally distributed) is necessary to plot a histogram and perform different distribution tests [16].

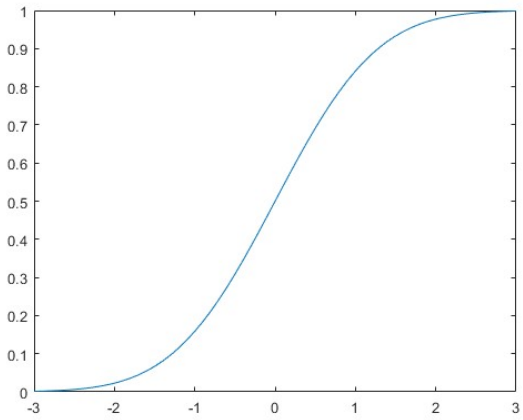

Figure 2.2: Sample CDF. Source: MathWorks

Scatterplots

Scatterplots can be two or three-dimensional data, composed of two or three variables that allow the visualization of the correlation between said variables. It also gives an insight into the dispersion of said variables. The shape of the scatterplot can describe linear correlations, curved relationships, and clusters (the most common shape for this study). The scatter plot will be essential to execute the clustering algorithm regarding cluster grouping and visualization.

Box plots

Box plots allow the identification of the different quartiles, medians and outliers of a dataset that can be grouped into different classes. Box Plots can be defined as a standardized way of displaying the data based on the five-number summary: The sample minimum, first quartile, median (second quartile), third quartile and the sample maximum. It is important to highlight that the sample maximum and minimum need not be outliers if they are not unusually far from other observations.

In a Box plot, outliers are defined as those points above and below the stated in equations 12 and 13 respectively, where IQR is the interquartile range described as the third quartile minus the first one [14].

$$Q3 + \frac{3}{2} . IQR \quad \text{(eq. 12)}$$

$$Q1 - \frac{3}{2} . IQR \quad \text{(eq. 13)}$$

Finally, outliers are displayed as circles or points at the top and bottom of the boxplot.

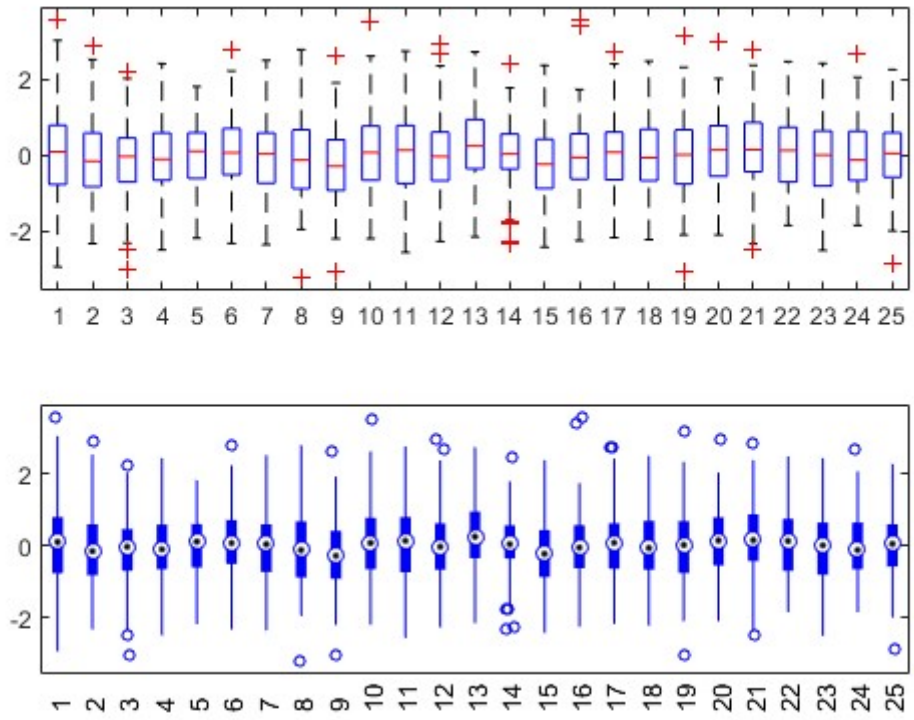

Figure 2.3: Sample box plot. Source: MathWorks

Outliers treatment

As it can be inferred, there are different ways of treating outliers found in raw data. The simplest and most usual one is the univariate method, which consists of analysing each variable separately through the employment of a box plot as it was stated in the previous paragraph, following the definition given by equations 12 and 13. Another way of treating outliers [17] is through the multivariate method, similar to the first one exposed but considering the relationship between at least two variables. It is useful when it comes to analysing the impact of different variables in the dataset when the data follows several linear or curved relationship.

Finally, the increasingly used Minkowski method is proposed, especially applied to machine learning processes, it assigns an error based on the standard mean squared and does not exclude the outlier, but it preponderates it with a lower weight in the dataset, which allows preserving more values in the said dataset.

In many cases [18], all the outliers are not supposed to be removed since the number of datapoints decreases and so does the representativeness, however, outliers identified as the result of spurious activity should be removed. Many approaches have been proposed in the last century, such as replacing missing data (or outliers) with the mean of the remaining data, nevertheless, this method will strongly reduce the data spread (dispersion) and increase the probability of committing a type-I error (false positives).

Other studies propose more sophisticated methods to deal with outliers, such as the one explained by M.R. Elliot [19] in 2007, which proposed to replace outliers with possible values.

As can be seen, there is not a single solution to the outliers detection and treatment. In this study, it is highlightable the assumption that isolated clusters will be composed of outliers.

Correlations

In this study, correlations are employed to determine the statistical relationship between two variables. The variables that show less dependence on each other will be selected to reduce the number of dimensions.

Microtrips Definition

As it was mentioned in the previous chapter, many authors have defined Microtrips as the division of a cycle between Idling [8-12]. As expressed, the cycles can be compounded by different driving situations belonging to different driving profiles, for this reason, its evaluation must take place separately. Those authors explain that the linkage between two driving profiles usually takes place in Idling, due to red lights, stop signs and motorway exits.

Despite the same definition being employed by most authors, in [9] a more restrictive definition is considered, where it is exposed to the requirement of a minimum MT duration of 60 seconds, followed by an outlier exclusion. This derives from a minor number of outliers, excluding the possibility of existing MT with non-representative driving times.

Alongside the previously mentioned restrictions, a minimum mean driving speed of 5 km/h was set. Additionally, a minimum period of 10 seconds of driving was required, and finally, a minimum distance of 200 m.

Ultimately, as it can be seen, the selection of the restrictions considered for the MT construction will influence the average duration of each cluster, especially the urban/low speed due to the existence of an inverse correlation between time and number of cycles. The procedure followed to identify an appropriate number of clusters will be exposed in chapter 3 (methodology).

Dimensionality Reduction

As exposed in [12], many methods have been proposed to deal with the crowding problem existing in multivariate datasets. Reducing dimensionality is important to visualise the results and acquire a better understanding of the machine learning algorithms applied to the dataset.

It is important to highlight the fact that machine learning algorithms can operate with large amounts of multivariate datasets without requiring the reduction of its dimensionality. However, as it was stated in the previous paragraph, the maximum recommended number of variables is three due to its ease to represent the algorithm results in a three-dimensional plot.

Crowding problem

When the dimensionality of a dataset is being reduced, the distance between the data points and any other specific point needs to be preserved, and for this reason, the distance between the cited data points will be lessened due to their gathering. Hence, the points will get compressed in the lower dimensional space producing crowding. The treatment of this problem will depend on the reduction method employed. One of the solutions proposed is the t-SNE algorithm designed to deal with the crowding problem.

Feature scaling

It is important to normalize the features before applying any dimensionality reduction method, especially before PCA, since it is quite sensitive to the variances of initial variables, and for this reason, it could wrongly preponderate the values with higher variability.

An example of the previously exposed are the APA and Driving speed, where a common value for the first one is around 0.1 m/s2 and the second one could easily fluctuate between 100 and 120 km/h. As it is possible to see, units of measurement play a crucial role in the variable variability and, for this reason, a scaling process is proposed.

There are two main ways of scaling features in machine learning: through standardization and normalization. The difference between both scaling methods can be seen in equations 14 and 15. The results obtained from the first one (standardization) will not be restricted to a particular range, where the mean becomes zero and the standard deviation 1. On the other hand, the results from the second equation (normalization) will be contained in a range between 0 and 1, where its mean and standard deviation does not have to follow any rule.

$$\chi' = \frac{\chi - \chi_{min}}{\chi_{max} - \chi_{min}} \quad \text{(eq.14)}$$

$$\chi'' = \frac{\chi - \mu}{\sigma} \quad \text{(eq.15)}$$

It is said that the data set should be standardized when its shape follows a Gaussian distribution and normalized when the distribution is unknown (methodology followed by most machine learning algorithms).

In this case, it will be assumed that the distribution is unknown and consequently, a variable normalization will be applied.

Principal Component Analysis

Usually known as PCA, is a mathematical methodology applied to multivariate datasets to reduce the number of dimensions. It is commonly employed before the implementation of a clustering algorithm. According to [20] the purpose of this analysis is to transform the multivariate data set into a new set of uncorrelated variables and ordered in a way that the first ones retain most of the variation of the original variables.

As it was mentioned, PCA does not ignore covariances and correlations, however, it prioritizes variances. In the next paragraphs, a brief explanation of the methodology followed by PCA is given.

Initially from a high-level point of view, the covariance matrix of the dataset is calculated. The objective of the covariance matrix is to demonstrate how much two variables are correlated and how this correlation is. When the correlation is positive, it is assumed the increment of a variable when the other one is decreasing and vice versa.

$$\sigma(x,y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - x)(y_i - y) \quad \text{(eq. 16)}$$

$$Corr = \begin{matrix} V_a & C_{ab} & C_{ac} \\ C_{ba} & V_b & C_{bc} \\ C_{ca} & C_{cb} & V_c \end{matrix} \quad \text{(eq. 17)}$$

Consequently, the eigenvectors and eigenvalues of the exposed covariance matrix are calculated. The eigenvalues will represent the magnitudes of the eigenvectors which will characterize the directions in the newly obtained feature space. As it can be inferred, the eigenvalues will quantify the variability of each vector.

Finally, the eigenvalues and eigenvectors will help to select the most important initial features based on their variability, those features will be given priority. The eigenvectors are sorted in descending order based on their respective eigenvalues.

According to [12], the quality and representativeness of the PCA components will be assessed through the variance represented by each principal component (eigenvector), for this, a Pareto graph is proposed to select several principal components that combined represents at least 80% of the total variability [21].

t-SNE

As it was said, the t-distributed stochastic neighbour embedding is a methodology used to reduce the dimensionality of a data set. This algorithm was introduced by Van der Maaten in 2009 [22] improving the known SNE (Stochastic neighbour embedding, Hinton and Roweis, 2002) by reducing the tendency to crowd points together, optimizing the visualization of the data. Mateen stated that the results obtained by the t-SNE algorithm were considerably better than those obtained with other dimensionality reduction algorithms for most distributions.

One of the main differences between t-SNE (2009) and its predecessor SNE (2002), is the employment of a student-t distribution rather than a gaussian distribution to compute the similarity of two points in the low dimensional space. A brief description of the process followed by this algorithm is explained in the following paragraphs.

According to [22], the t-SNE starts by converting the high-dimensional euclidean distances into conditional probabilities which are based on the probability that a point xi would choose a point xj as its closest neighbour (equation 18) if neighbours were selected under a Gaussian probability distribution, where is the variance of the data.

After performing the previous step, the data points should be spread randomly on a new low-dimensional space, where the goal of this algorithm is to find a similar probability distribution in said space. In this case, the student-t distribution is employed (equation 19) to reduce the crowding problem that exists if a Gaussian distribution is followed. The t-SNE aims to find a low-dimensional representation of the data that minimizes the discrepancy between pij and qij.

One of the ways to estimate this discrepancy is through the Kullback-Leibler divergence (also known as relative entropy), which is a measure of how one probability distribution is different from the other one (equation 20). As it can be inferred, KLD needs to be as small as possible (reducing the disparity between p and q).

$$P_{ij} = \frac{\exp\left(\frac{-\left\|x_i - x_j\right\|^2}{2\sigma^2}\right)}{\sum k \neq l \ \exp\left(-\|x_k - x_l\|^2 / 2\sigma^2\right)} \quad \text{(eq. 18)}$$

$$q_{ij} = \frac{\left(1+\left\|y_i - y_j\right\|^2\right)^{-1}}{\sum k \neq l \ \left(1+\|y_k - y_l\|^2\right)^{-1}} \quad \text{(eq. 19)}$$

$$C = KL(P\|Q) = \sum_i \sum_j P_{ij} \log \frac{P_{ij}}{q_{ij}} \quad \text{(eq. 20)}$$

Additionally, it is known that the algorithm requires more entry variables and supplementary features alongside the dataset. These features are called *hyperparameters* since they have a significant influence on the t-SNE result. One of the main hyperparameters is perplexity.

According to Van der Maaten, perplexity (equation 21) is a value specified by the user that depends on the Shannon entropy of the dataset and ideally must be located within the range 5-50, a value that will be discussed consequently.

$$Perp(Pi) = 2^{H(Pi)} \quad \text{(eq. 21)}$$

Perplexity can be defined as a measure of the number of effective neighbours of a data point; hence, a larger number of data points will require a higher perplexity. Y. Cao (2017) [23], proposed a practical approach to determine the optimal perplexity according to the data distribution and several data points. Said approach suggests calculating a score that needs to be as slow as possible and will depend on the number of data points and the KL resulting from the t-SNE iteration. Equation 22 describes the score introduced by Cao, where Perp is the perplexity used, and n is the number of data points evaluated. Finally, it is considered important to highlight the fact that the default perplexity employed by most data analysis functions is 30, which is not always correct.

In order to illustrate the above said, if the perplexity used for a dataset is insufficient (too low), the algorithm would not be able to separate the clusters found in the data but would instead create homogeneously distributed small groups, assigning more importance to local neighbourhoods than global groups. On the other hand, if the perplexity is excessive, the clusters would not be separated from each other, and clustering algorithms would not be able to differentiate them correctly.

$$S(perp) = 2KL(P||Q) + \log(n)\frac{perp}{n} \quad \text{(eq. 22)}$$

Another important hyperparameter that needs to be considered is the number of iterations [15] that takes place when running the algorithm, where the higher is the number of iterations, the better the results obtained. However, this will require more running time, producing a delay and consuming more resources.

Ultimately, the default method to measure the distance between two data points can be changed, however, the preferred one is the Euclidean distance. Also, at the beginning of the process, it is possible to exaggerate the plotting distance between points (in the first 5 iterations) to recreate more accurate clusters.

Clustering Algorithms

Machine learning is usually described as the study of computer algorithms that improve automatically through models built from large amounts of data, being part of what is known as artificial intelligence. This field of study can be divided into two main streams: supervised and unsupervised learning.

Supervised learning takes place through the study of data where the input and output are known (training data) and are employed to develop predictive models (i.e., regressions). On the other hand, unsupervised learning employs datasets where only the input is known and aims to find groups (clusters) in data, to develop structures.

Clustering can be defined as the machine learning (unsupervised learning) task of grouping data points according to specific variables. Different clustering techniques can be used depending on the type of data and data distribution. Most of them calculate the distance between data points to find the closest ones. Clustering is the main task for exploratory data analysis that brings a visual comprehension of how the data is being grouped. Two main methods are exposed: K-means and Hierarchical clustering since those are commonly used for this type of data.

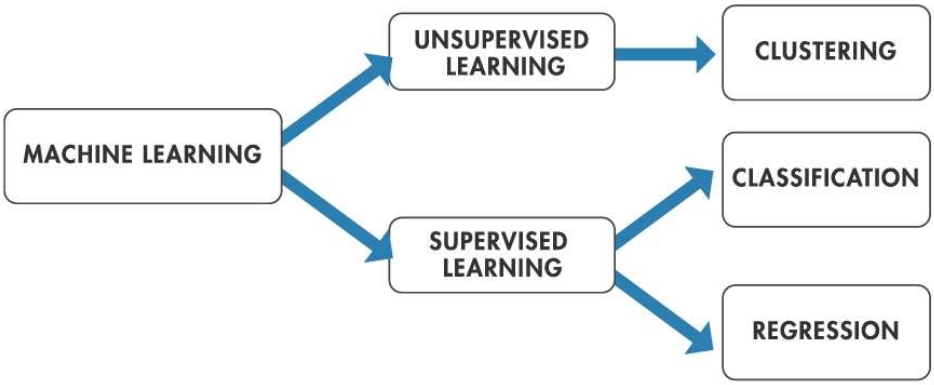

Figure 2.4: Machine learning methods. Source: MathWorks

k-Means

K-means is a clustering method that divides the data set into k clusters (specified by the user) based on the distance between each data point and a randomly selected data point (cluster centre).

Before the employment of this or any other clustering algorithm, the user must identify the most appropriate number of clusters, which will be explained later in this chapter (clustering coefficients). The ideal number of clusters will be an entry parameter of the clustering function.

After the k number of clusters is specified, the algorithm randomly selects k data points from the data set to start the clustering iteration (X1, X2,.. Xk), (figure 2.5 b).

Consequently, the euclidean distances between each point and the randomly selected points are calculated. The closest Xn to each data point will be selected and afterwards, said data point will be assigned to the Xn group (figure 2.5 c). After finishing the first assignment, it will be found in the space k groups. The next step is to calculate the centroid of the cluster (figure 2.5 d) and identify the closest data point to the said centroid. This data point would be the new Xn and the process would be repeated until there are no changes in the position of the centroid.

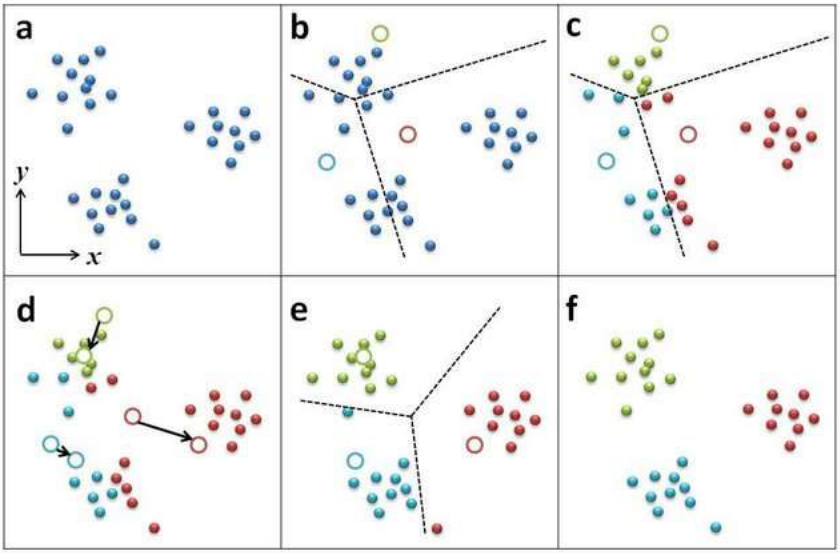

Figure 2.5: Schematic representation of K-means algorithm. Source: Y. Chen, Y Lai, "Universal structural estimator and dynamics approximation for complex networks", 2016.

Hierarchical clustering

Like the K-means, hierarchical clustering is a methodology followed to divide the dataset into different groups (clusters), but this time based on a hierarchy among data points. The Euclidean distance between each data point is calculated and, consequently, it is linked to the closest point. This process will be repeated until the desired number of clusters is reached.

There are two main ways of grouping data following hierarchical clustering methods: agglomerative hierarchical and divisive hierarchical [24]. The divisive method is selected for this study, where each observation starts with its cluster and pairs of clusters are merged when the hierarchy moves forward. The distance between two subsets is called linkage distance.

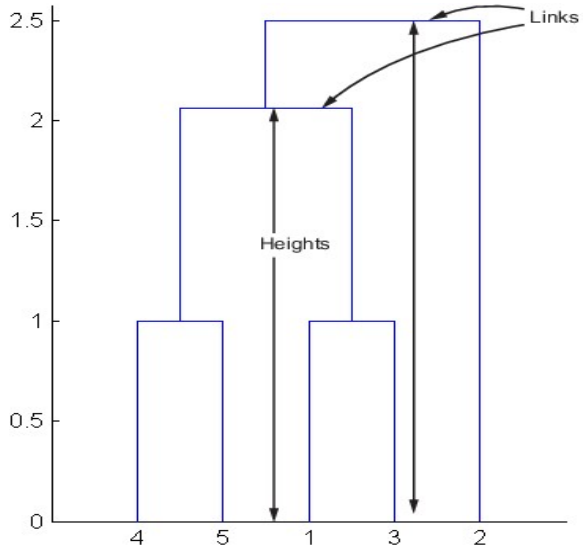

Figure 2.6: Graphic representation of hierarchical dendrogram. Source: MathWorks

Performance Metrics

In this study, cluster coefficients are employed to evaluate the quality of the cluster, find the ideal number of clusters and select the clustering algorithm that best fits the data set. The implemented coefficients will be the *Silhouette coefficient, Calinski-Harabasz criterion* and *Davies-Bouldin Index*. Those coefficients are usually employed to measure the cohesion of the data points.

Silhouette Coefficient

The silhouette coefficient represents graphically how well the data has been classified and how similar are the points of a cluster compared to other clusters. As it is explained in [25] the objective of calculating the clusters silhouette is to measure the dissimilarities between the components of a cluster and other clusters. Initially, a point i in cluster A is selected. As seen in figure 2.7, it is possible to calculate the Euclidean distances between the data points that belong to the cluster. Consequently, after calculating the mean distance between those data points, the average distance between the point i and points from the cluster (B) is computed (since it is the closest cluster).

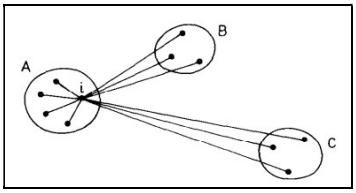

Figure 2.7: Illustration of the elements involved in different clusters. Source: [25]

The result of the above explained is stated in equation 23, where the distances between a(i) and b(i) are compared. It is possible to see that -1S(i)1.

$$S(i) = \frac{b(i)-a(i)}{max\{a(i),b(i)\}} \quad \text{(eq. 23)}$$

After performing the evaluation of the dataset, it is possible to organize the data points by silhouette coefficient and, consequently, plot silhouettes for each cluster. The aim is to find the most similar silhouettes among clusters with the highest coefficient, indicating a high similarity among the points of a cluster, measured through the Euclidean distance.

Calinski-Harabasz

Also called the variance ratio criterion (VRC), was introduced in 1974 by Calinski Harabasz with the aim of finding the most accurate number of clusters for a dataset.

The Calinski Harabasz Index is given by equation 24 where k is the number of clusters, N is the number of data points, *SSw* is the overall variance of a cluster (within-cluster variance) and SSD_B is the overall variance between clusters (between cluster variance) [26].

$$VRC = \frac{SS_B}{SS_W}\frac{N-K}{K-1} \quad \text{(eq. 24)}$$

SSB measures the variance of the centroids of the clusters compared to the dataset centroid, whereas SSW will measure the variance of the data points that belong to the same cluster.

When VRC is applied, the objective is to obtain the highest index, indicating the number of clusters is accurate. From equation 24, it is observed that a lower variance of the data points in a cluster, in comparison to the variance of the dataset, will result in a lower index.

$$SS_W = \sum_{j=1}^{k} \sum_{i=1}^{n_j} \left(O_j^i - \bar{O}_j\right)^T \left(O_j^i - \bar{O}_j\right) \text{ (eq. 25)}$$

$$SS_B = \sum_{j=1}^{k} n_j \left(O_j - \bar{O}\right)^T \left(O_j - \bar{O}_j\right) \text{ (eq. 26)}$$

In order to determine the variance of a dataset [26], SSw and SSB are calculated as shown in equation 25 and 26 respectively, where *Ōj* denotes the n-dimensional vector of means within the *jth* cluster (cluster centroid), and Ō denotes the n-dimensional vector of overall means (dataset centroid). K-1 is the degree of freedom of the cluster variations. According to said reference, a more separated dataset will tend to have a lower *SSw* and a large value of SSB. The ratio (n-k)/(k-1) prevents the VRC from increasing with the number of clusters.

Davies–Bouldin index

The Davies- Bouldin index (DBI) is a metric that seeks to evaluate the quality of the clusters; Was developed by David Davies in 1979.

A similarity measure Rij between clusters Ci and Cj is established based on the dispersion of the cluster Ci and the similarity between the two clusters named dij. [27]

$$DB = \frac{1}{nc}\sum_{i=1}^{nc} R_i \text{ (eq.27)}$$

$$R_{ij} = \frac{S_i + S_j}{M_{ij}} \text{ (eq. 28)}$$

The DBI is defined [28] in equations 27 and 28, where *Ri* is the maximum of *Rij* and *sj* and *si* are the dispersion of clusters *i* and j. *Mij* is defined as the distance between vectors that follow several are chosen as characteristic of clusters *i* and *j.*

As it can be seen from the previous expression, a higher distance *Mij* and a lower dispersion (*Si*+*Sj*) will result in a lower value of *R*, consequently, it is possible to state that according to the DBI definition, the most accurate number of clusters is given by the lowest DBI.

Driving Behaviour

In this study, the driver's aggressiveness will play an important role when it comes to developing representative driving cycles since it has an important influence on the changes of speed, acceleration and maximum speed reached during the cycle. Many studies have presented methodologies to determine the driving behaviour of a cycle based on data collected from the main cycle.

As it was demonstrated in [29] the driving behaviour and aggressiveness are strongly related to fuel consumption and, hence, to the emissions produced by an ICE. An aggressive driving style will always result in a higher CO2 emission, whereas a mild driving style will lead to higher energy efficiency. According to said reference, adopting an efficient driving style would have an impact of 15-20% on fuel consumption.

For the previous reason, despite it is not usually considered in the precursory studies, the driving behaviour will be considered in this thesis, alongside the fact that the purpose of this investigation framework is to determine the emissions produced in the metropolitan area of Tenerife (research purposes), introducing the possibility of studying the percentage of the drivers (in the said area) considered aggressive, average and mild improving the model results.

As can be inferred, measuring the behaviour of a driver based on the driving parameters previously mentioned is not easy. Due to this, in [30] it is demonstrated the existence of a high correlation between the driving style, the acceleration and velocity of a cycle. Liessner proposed the construction of a parameter dependent on the positive acceleration reached and the third quartile of the velocity during the driving cycle. As it can be inferred, more aggressive drivers will tend to reach higher accelerations and speeds.

Förster in [31] develops a system used to classify driving behaviours through the designation of variables that depend on acceleration and velocity. Said variables are arranged in cumulative distributions, defining three ranges for each distribution: 0 to 25th quantile, 25th to 75th quantile, and over 75th quantile. After said assignment, it is possible to rearrange the data points into a score, calculated from the average position in the previous cumulative distributions. As shown in figure 2.8, the cycles can be categorized as mild, average, and aggressive according to their position in the calculated score.

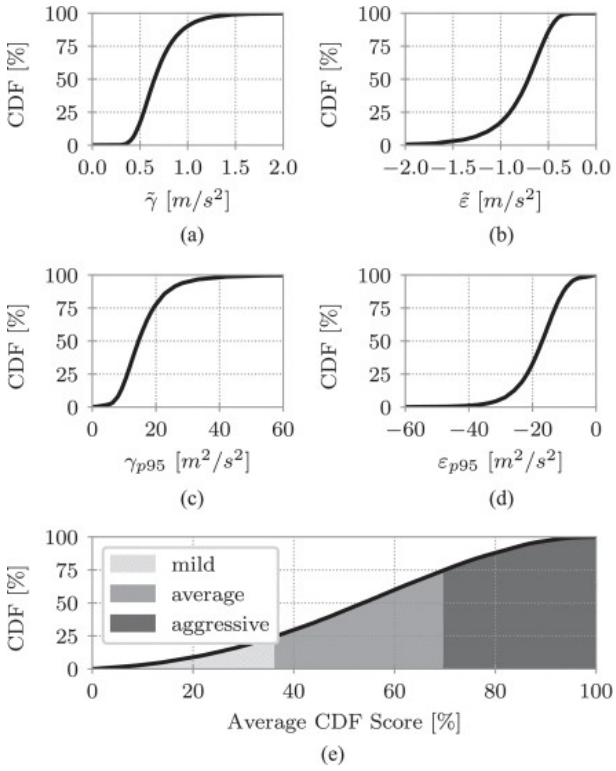

Figure 2.8: Driving behaviour according to variables in CDF. Source: [31]

Due to the previously stated, in this thesis, the driving behaviour will be studied through the analysis of behaviour-related features: APA, ANA, RPA and RNA. This analysis will begin with the calculation of the data point position in a CDF of the mentioned parameters and, consequently, arranged in an average score CDF according to their average position in the previous diagrams (behaviour-related features). Those cycles above the 75th percentile will be considered aggressive, whereas those under the 25th will be taken as mild.

CHAPTER 3

METHODOLOGY

This chapter aims to illustrate the methodology and process followed to develop the representative driving cycles, achieving the objectives proposed earlier in this thesis. Consequently, a brief explanation of the process is given.

As can be seen in the diagram exposed in figure 3.1 the totality of the data employed to develop the driving cycles was extracted from real driving conditions, geographically located in the metropolitan area of Tenerife: municipalities of La Laguna, Santa Cruz, El Rosario, and Tegueste.

The data extraction took place through the employment of a mobile application as GPS (GPS SpeedView, available in Google Play), being able to estimate the vehicle velocity, acceleration, distance, grade and geographical position each second (1 Hz). The data were taken by at least 10 different subjects after receiving instructions regarding its usage.

After collecting the data, it is necessary to apply a filter in order to reduce noise and delete outliers due to the application inaccuracy. The filter *Savitzky-Golay* was applied to perform this task.

It is important to highlight that up to this point, the first two steps of the first stage were performed and controlled by the Industrial Engineering department (ULL) in the investigation framework.

After receiving the clean driving cycles, two different paths are taken: statistical analysis of clean driving cycles and their division into micro trips (explained in detail later). As mentioned in the previous chapter, the driving features are defined to provide more statistically accurate results, facilitating the comparison of more

metrics between cycles.

The importance of the statistical analysis in this study can be described in two main points:

- Evaluating the representativeness of the data collected, making possible the designation of the scope and applicability of this study.
- Describing the data distribution, providing a brief insight of how the data points are issued and how the different parameters are correlated.

After this analysis, it will be possible to infer how the representative driving cycles should look like.

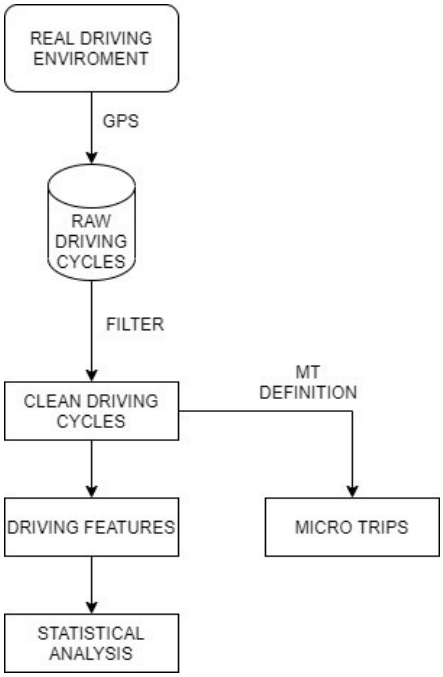

Figure 3.1: Diagram of the first stage of the study. Source: Own elaboration

The second stage of this thesis starts with the division of the driving cycles into micro trips (MT) as defined in the previous chapter. As can be inferred, the number of data points will increase, providing more accurate results [8]. As it was stated before, a driving cycle can be composed of different driving conditions (corresponding to different driving profiles), from here, it is crucial to separate said profiles.

After defining the MT and describing the driving features, the data needs to be standardized before performing any dimensionality reduction algorithm.

One of the questions that need to be addressed is why the dimensionality of the dataset needs to be reduced. As exposed in [20], it is possible to lose information if a group of representative variables is selected. Additionally, it needs to be considered the possibility of committing errors when choosing said variables. On the other hand, if a clustering algorithm is applied to all the variables, it will be necessary to analyse the behaviour of groups of variables separately and achieve a general understanding of how said variables behave in a multidimensional space.

As shown in figure 3.2, in order to compare the results, two dimensionality reduction methods will be used: PCA and t-SNE.

After performing the dimensionality reduction, two clustering algorithms are used: k-Means and Hierarchical clustering, attending the exposed in [12], where both methods were compared.

To perform said comparison, the quality evaluation of the clusters takes place through the employment of performance metrics. For this study, as argued in the previously referenced thesis, the Silhouette coefficient, Calinski-Harabasz and Davies–Bouldin index are used.

Finally, after selecting the most appropriate dimensionality reduction/clustering methodology combination, it is necessary to separate the clusters according to their driving behaviour (through the average score explained before). Consequently, the final number of groups will depend on two factors: the driving behaviour category and the driving profile. The expected final result is obtained from the merge of the different driving profiles within the same driving behaviour, resulting in three representative driving cycles.

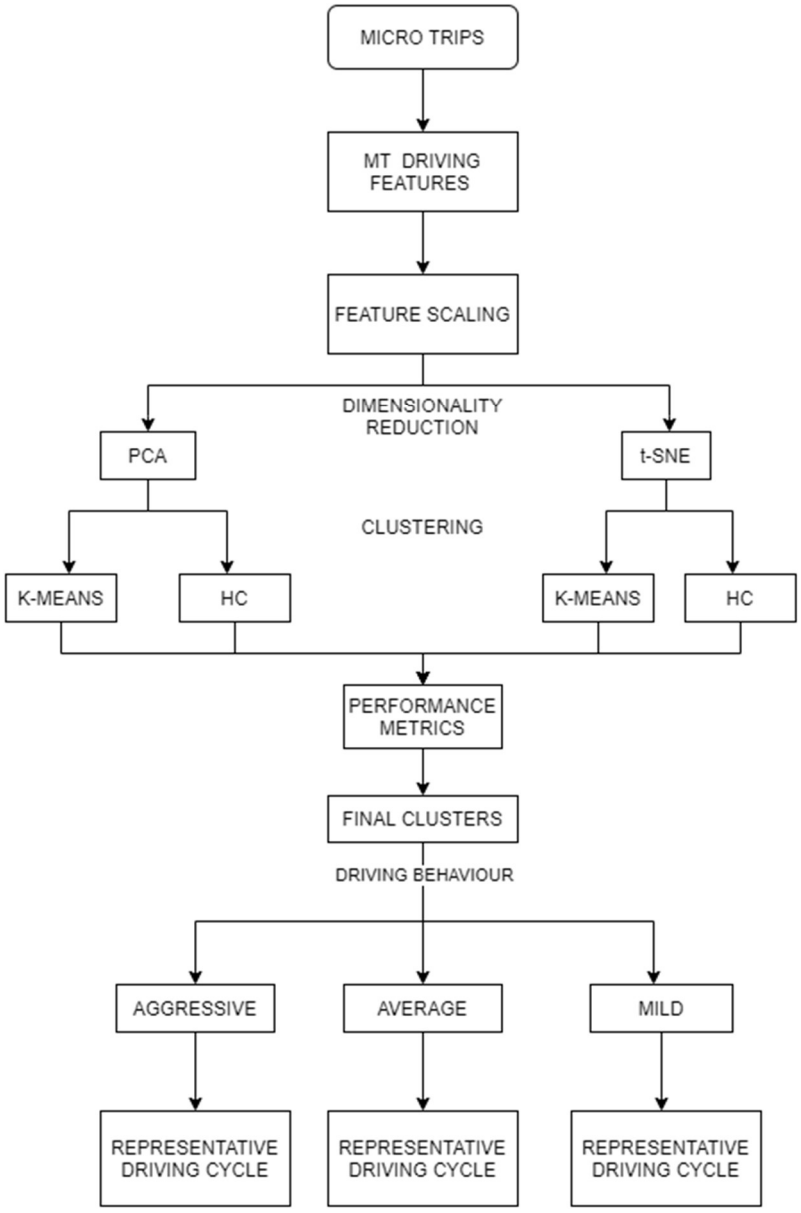

Figure 3.2: Diagram of the second stage of the study. Source: Own elaboration

CHAPTER 4

RESULTS

Statistical Analysis

The results of the statistical analysis are presented in this chapter. As explained before, in the methodology, it is possible to divide the statistical analysis into two main sections: evaluating the representativeness of the dataset intending to address the possible influences from external factors and, representing the distributions of the driving features previously exposed. Initially, to evaluate the representativeness of the data set it will be studied the location characteristics, collection dates-hours and weather conditions.

Geographical Characteristics

As it is possible to see in figure 4.1, the cycles are plotted on a geographical map, where it can be concluded that most of the cycles belong to the metropolitan area, previously defined as municipalities of La Laguna, Santa Cruz, El Rosario and Tegueste.

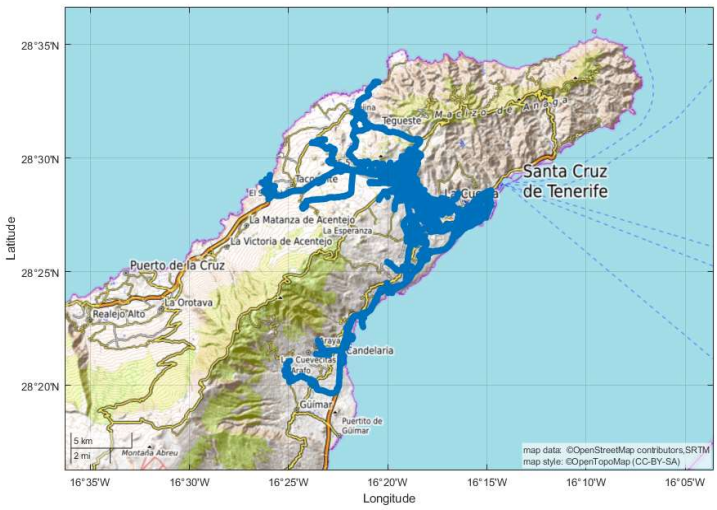

Figure 4.1: routes covered by the driving cycles

However, some cycles took place outside the metropolitan area, such as those beginning or finishing in the municipalities of Candelaria, Arafo and Tacoronte. It will be important to address those cycles carefully since they may modify the final results by increasing the driving distances and mean driving speed.

In order to illustrate how the data is acquired, figure 4.2 shows an area limited by specific coordinates that belong to a highway and a roundabout. Each data point is represented by a single point in a scatter plot, where the geographical position is also taken into consideration. Each point collected by the application (1 Hz) contains information about velocity, grade, acceleration, time, distance, and latitude/longitude, hence, the importance of treating the data through machine learning algorithms instead of basic spreadsheets.

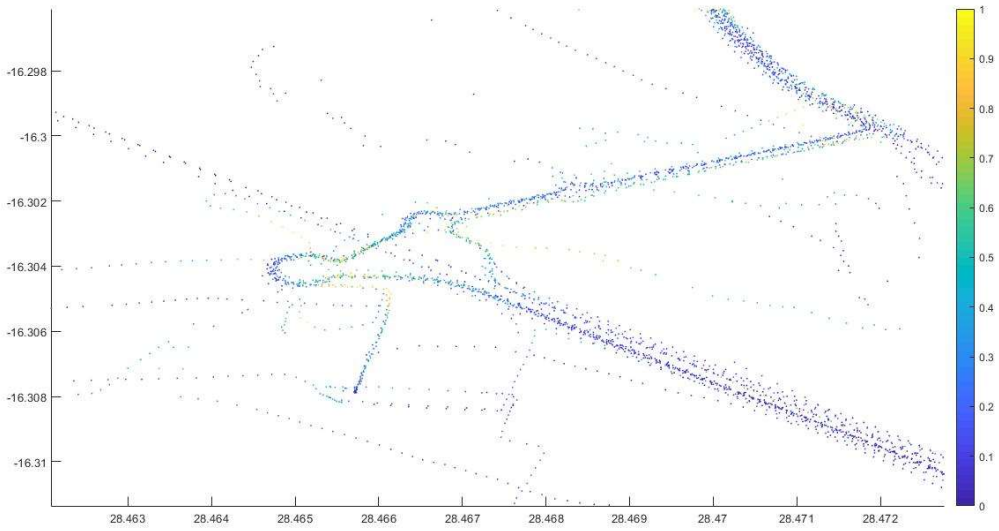

Figure 4.2: Collected data points in motorway

In order to provide a general overview about the driving time and geographical location, the heatmap of figure 4.3 is shown, where the entire metropolitan area can be seen. The coordinates are divided into a rectangular matrix, where the yellow points represent a higher driving time. As can be inferred, this does not mean that the number of cycles in the area is higher, but the time spent is. This can be related to different factors such as a lower driving speed due to traffic conditions and/or roads infrastructure.

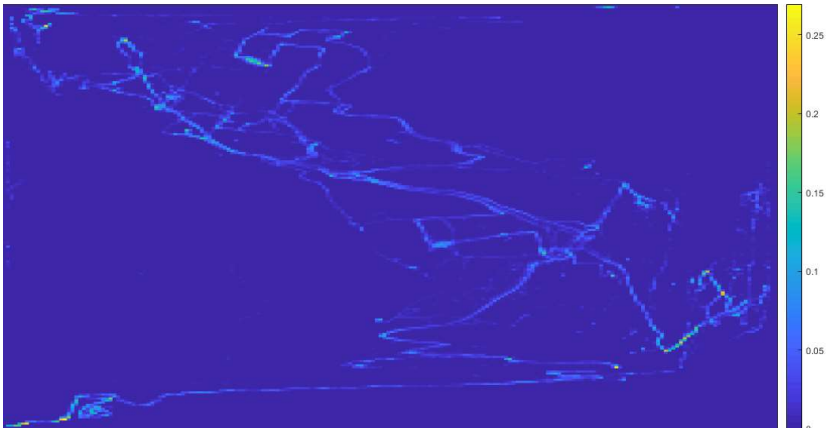

Figure 4.3: Driving time and geographical location of cycles.

Attending the starting/ending locations, the bar graph shown in figure 4.4 indicates the predominance of Guajara, La Laguna, Tabaiba and Salud-La Salle as starting/ending points. This plot was constructed by comparing the starting coordinates provided by the application and the closest designated point from a 50 location (cities/towns) list.

The reader could easily notice that many starting and ending locations do not belong to the municipalities of La Laguna, Santa Cruz, El Rosario and Tegueste, however, due to many factors, such as their proximity to said municipalities and their road infrastructure (type of road, changes of altitude, etc.), they could be also included in this analysis as long as they do not produce substantial changes in the distribution.

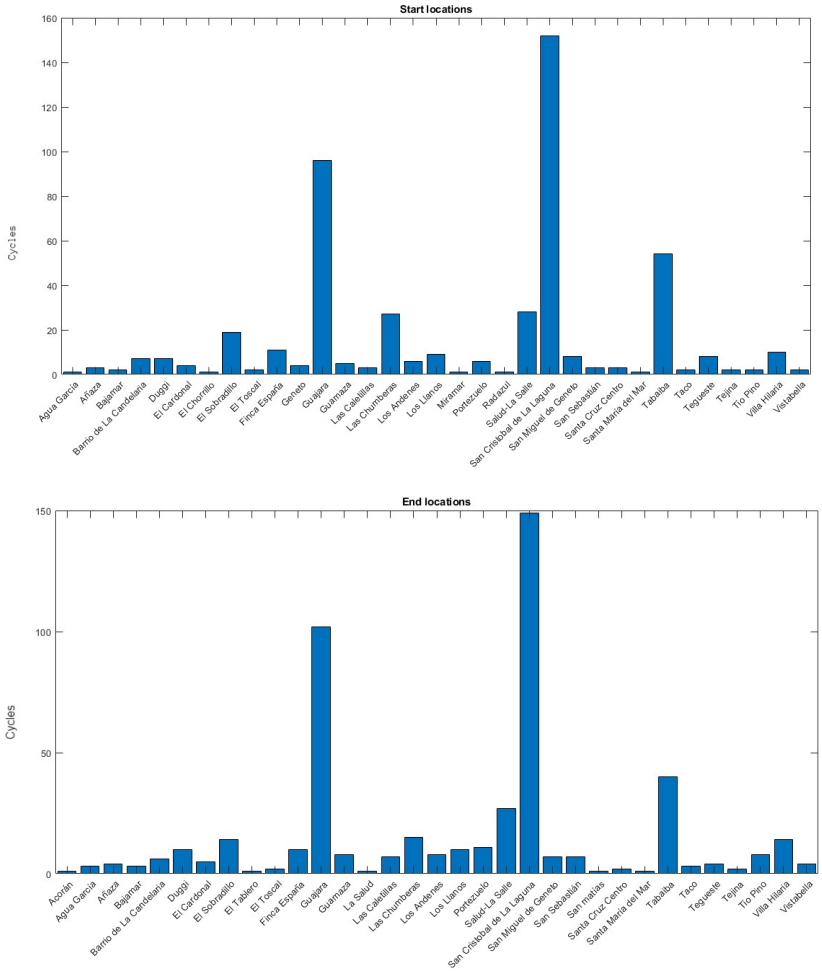

Figure 4.4: Starting and ending locations of driving cycles

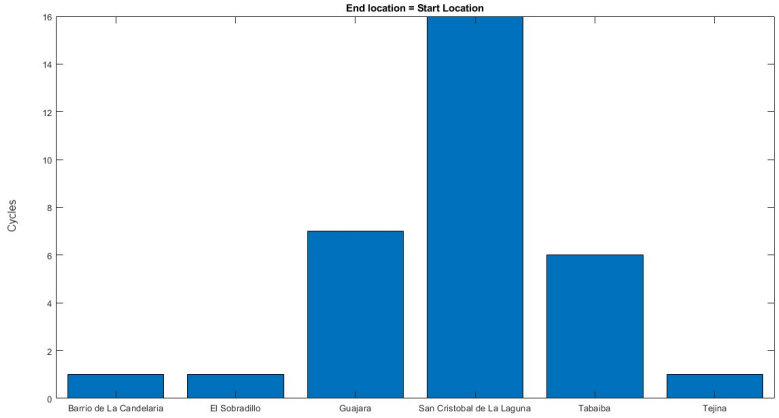

Figure 4.5: Intracity cycles

An intracity cycle is defined as a cycle that has the same starting and ending point. From the 490 analysed cycles, only approximately 30 are intracity cycles, characterized by short distances and shorter time durations (figure 4.5).

To illustrate the representativeness of the different locations covered, the next comparative heatmap with the 20 most frequent locations is shown. As can be seen, most of the cycles have a start/end location in Laguna Centro/Guajara which could imply a non-uniform distribution of the cycles recorded. Additionally, it can be noticed a lack of cycles with start/end location in Santa Cruz. A resume table (with data extracted from the heatmap) is shown to ease the understanding of the relationship between the different locations.

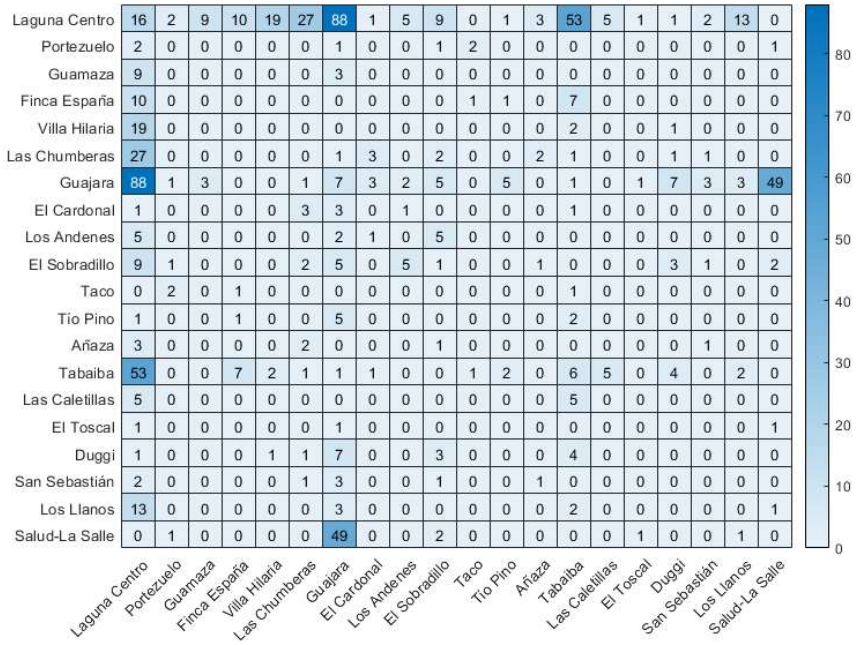

Figure 4.6: Start/end location heatmap.

	Laguna	Santa Cruz Cen	Santa Cruz Sth
Laguna	52%	24%	24%
Santa Cruz Cen	80%	3%	17%
Santa Cruz Sth	74%	16%	10%

Table 4.1: Start/end location. Data set resume table (horizontal)

In order to understand the level of representativeness of the locations involved, the information presented needs to be compared with a mobility study of Tenerife. Hence, table 4.2 exposes the percentage of trips with starting and ending locations in different zones of the Tenerife metropolitan area according to a study conducted by the Council of Tenerife.

	Laguna	Santa Cruz Cen	Santa Cruz Sth	Other
Laguna	57.4%	24.2%	9.7%	9%
Santa Cruz Cen	24.0%	50.4%	13.3%	13%
Santa Cruz Sth	31.0%	32.0%	24.0%	13%

Table 4.2: Percentage of trips (horizontal) in different zones of Tenerife. Source: Cabildo de Tenerife [32].

If the information presented in table 4.1 and 4.2 are compared, it can be seen that the ending location of the cycles that started in La Laguna are correctly distributed, where about half of the cycles stayed in La Laguna, and the other half was equally distributed to Santa Cruz centre and Santa Cruz South. On the other hand, when it comes to Santa Cruz Centre, there is visible a clear difference in the distribution, where most cycles (74%) had La Laguna as destiny point whereas the trips within Santa Cruz only represented 3% of the data set. As can be deduced, the mentioned lack of representation may compromise the final cycles, specifically the urban, where more information would be required.

Residence zone	Population	Location	Vehicles per 1000 people
SC Centro - Anaga	162.263	Santa Cruz de Tenerife	537
SC Sur - El Rosario	57.240	El Rosario	612
		Santa Cruz de Tenerife	
Laguna Centro	108.223	San Cristóbal de La Laguna	535
		Tegueste	
Laguna Norte - Tegueste	37.276	San Cristóbal de La Laguna	595

Tab 4.3: Population and number of vehicles per 1000 people in different zones of Tenerife. Source: Cabildo de Tenerife [32].

Towards the evaluation of the data representativeness, it needs to be considered, alongside the percentage of trips previously exposed, the population and the number of vehicles in each zone. It is shown a higher population in SC Centro and Laguna Centro (doubling the other zones) with an almost constant number of vehicles. Consequently, it can be assumed that the highest representativeness is reached when the starting/ending location takes place in one of the mentioned zones.

Temporal Characteristics

As may be understood, temporal characteristics play an important role in data collection (especially by producing traffic changes). For example, collecting data in peak hours, holidays or when specific incidents happened on the road could produce a variation in the driving features. To study this impact, the following study is performed.

In figure 4.7 a histogram of the cycles collection dates is presented. It can be seen that the data collection took place in a relatively long period of time, on different days. This helps to decrease the impact of specific road situations, weather conditions and holidays. The last two will be studied later more deeply. However, it is remarkable the predominance of December as the most frequent month which could influence the features, especially due to the number of holidays in this month.

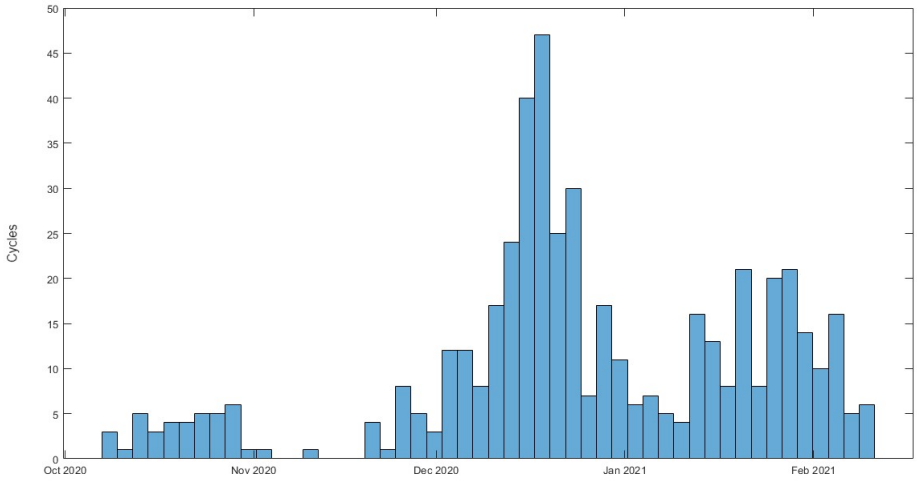

Figure 4.7: Cycles collected per day

In the next bar graph, it can be seen that the number of cycles measured on weekends is lower than those collected on weekdays. On the other hand, the frequency of the number of cycles regarding the weekdays, is uniform, without any relatively predominant day.

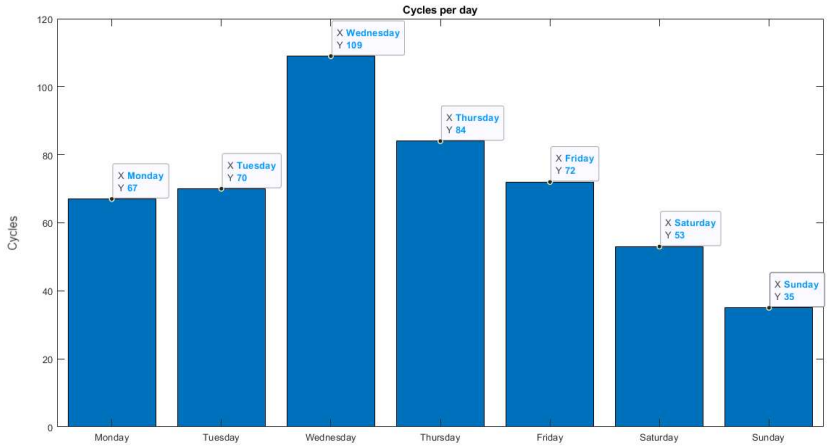

Figure 4.8: Cycles per weekday

To emphasize the impact of weekends, figure 4.9 shows a heatmap comparing the traffic congestion by weekday in Santa Cruz de Tenerife in 2019 [33]. It can be seen that the day with more traffic congestion was Friday, whereas weekends experienced a lower congestion level. However, figure 4.8 shows that only 18% of the cycles were recorded on weekends (Saturdays and Sundays), more than 10 points under an equally distributed data collection (assuming every day contributed a 14.3% to the weekly data collection). It can be said that congestion is not necessarily related to the number of cycles but with their distribution within a specific day, affecting driving features.

	Sun	Mon	Tue	Wed	Thu	Fri	Sat
12:00 AM	5%	0%	1%	0%	1%	2%	4%
	4%	0%	1%	0%	0%	0%	2%
02:00 AM	1%	0%	4%	0%	0%	0%	0%
	0%	0%	0%	0%	0%	0%	0%
04:00 AM	0%	0%	0%	0%	0%	0%	0%
	0%	0%	0%	0%	0%	0%	0%
06:00 AM	0%	0%	0%	0%	0%	0%	0%
	0%	5%	5%	5%	5%	4%	0%
08:00 AM	0%	31%	31%	32%	31%	28%	0%
	1%	37%	36%	37%	34%	30%	3%
10:00 AM	3%	24%	23%	23%	22%	22%	7%
	7%	23%	21%	21%	21%	21%	13%
12:00 PM	12%	26%	25%	25%	24%	26%	18%
	14%	26%	25%	25%	26%	28%	21%
02:00 PM	15%	26%	26%	27%	27%	36%	23%
	10%	31%	33%	33%	33%	57%	19%
04:00 PM	5%	28%	32%	31%	31%	47%	7%
	7%	27%	31%	29%	29%	28%	9%
06:00 PM	10%	33%	36%	33%	36%	25%	14%
	13%	32%	31%	30%	33%	28%	19%
08:00 PM	14%	27%	26%	26%	28%	28%	21%
	11%	17%	19%	19%	21%	25%	17%
10:00 PM	7%	10%	10%	11%	11%	16%	14%
	2%	4%	4%	5%	5%	9%	9%

Figure 4.9: Weekly traffic congestion by the time of the day. Source: [33]

Regarding the driving hours, it is possible to highlight three peak hours: between 7:00 and 8:00, at 14:00 and finally, at 19:00-20:00. This could be explained by the working / school shifts, where the school period takes place between 9 and 14:00 and the working period usually finishes at 19:00. It is highlightable that the last peak hour (19:00) has a lower influence on the traffic conditions.

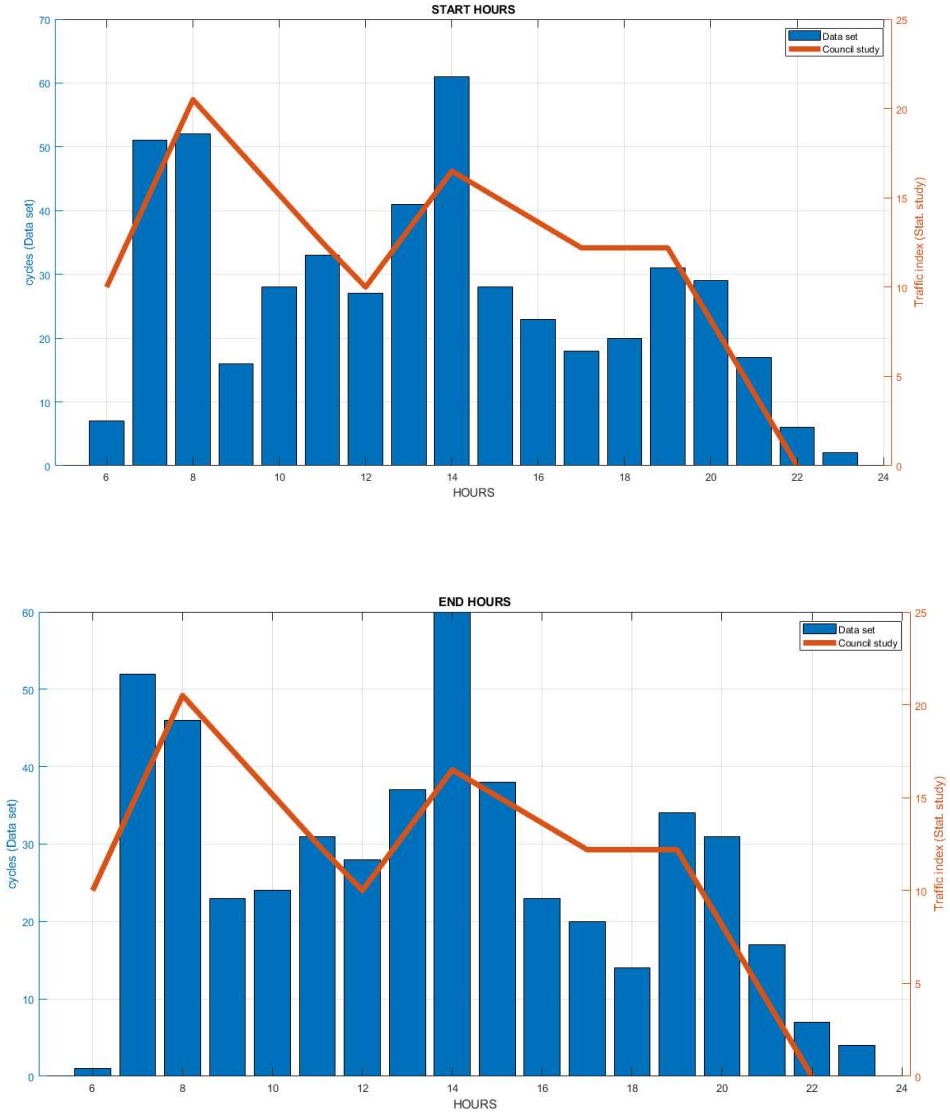

Figure 4.10: Starting and ending hours. Comparison with Tenerife Council study. [32]

To corroborate this, in figure 4.10 is also reflected the distribution of the peak hour in Tenerife, according to a study performed by the Tenerife Council [32], being similar to the one calculated through the driving cycles: higher traffic flow at 8:00 and 14:00.

Finally, after the previously mentioned, it can be said that most of the representativeness of the data set may be compromised due to a non-uniform data collection regarding the days of the week.

Weather Conditions

To evaluate the impact of the weather conditions on the traffic, a group of days is selected (from the weather database *AccuWeather*) and categorized as rainy days within the period exposed in figure 4.7. Those days are employed to compare some features distributions that could be affected by the rain.

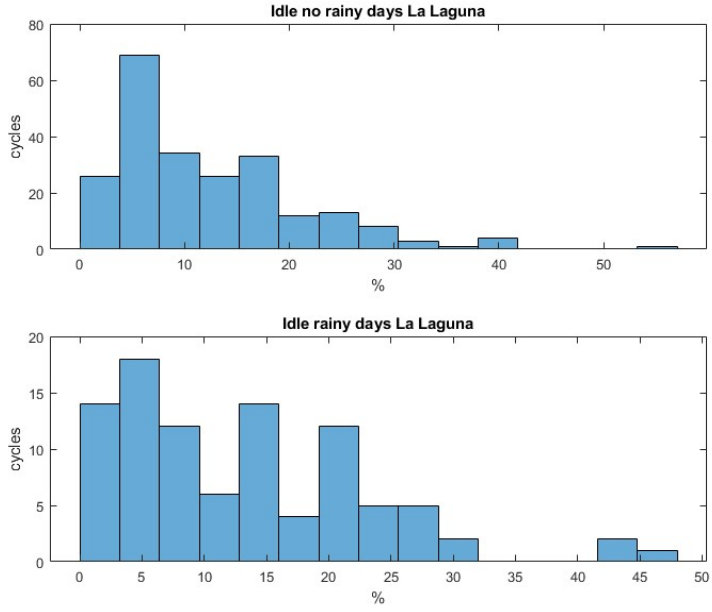

Figure 4.11: Average Idle time (%) and weather conditions

As shown in figure 4.11, in comparison to regular days, the distribution of the idle time (%) during rainy days, in La Laguna, is slightly lower, where it is not possible to get to a clear conclusion due to such a small difference. However, when it comes to RPA, it is possible to highlight lower values on rainy days in the distribution exposed in figure 4.12. This is probably due to cautious drivers trying to generate less aggressive accelerations in wet road conditions.

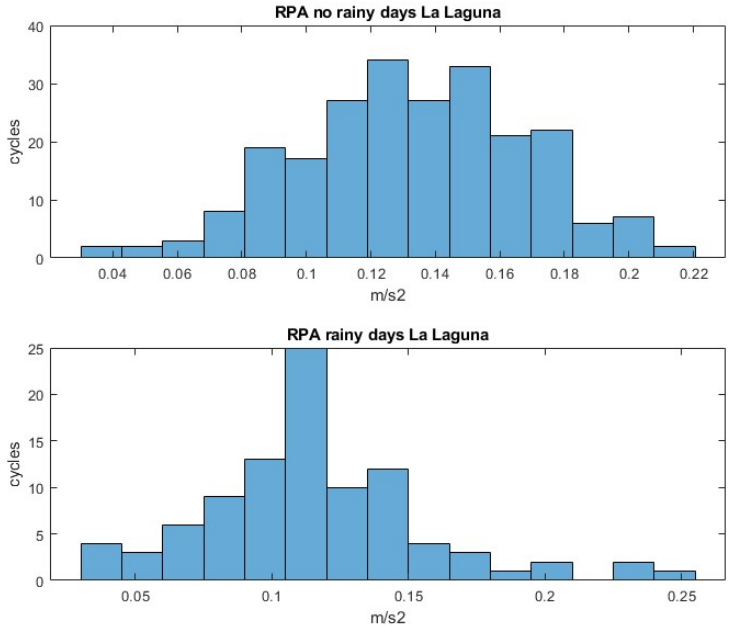

Figure 4.12: Average RPA (m/s2) and weather conditions

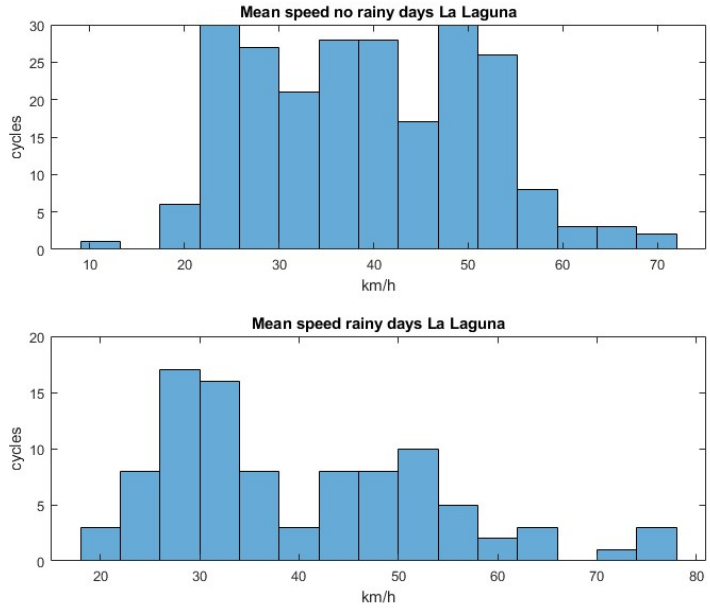

Figure 4.13: Average Driving Speed (km/h) and weather conditions

Finally, as shown in figure 4.13, the average driving speed histogram shows a less uniform distribution on rainy days, reaching the first peak at 28 km/h and a smaller second one at 52 km/h.

After exposing these histograms, it is possible to argue that the weather conditions do not play a significant role in the driving environment in La Laguna, a conclusion that will be extrapolated to other locations in the metropolitan area for this study. However, it would be necessary to conduct a deeper analysis, with more observation points and locations to achieve a general conclusion of the impact of weather conditions on the driving conditions and traffic in Tenerife.

Time and driving distance

When it comes to driving distances, it is shown (figure 4.14 - 4.15) that the average distance is around 8.84 km, with the highest frequency in the range of 3 - 6 km. As it is displayed in the cumulative diagram function, about 80% of the cycles travelled a distance under 10 km. It is highlightable the existence of unusually long cycles (>20 km) that, following the definition given in the second chapter, these cycles may be addressed as outliers.

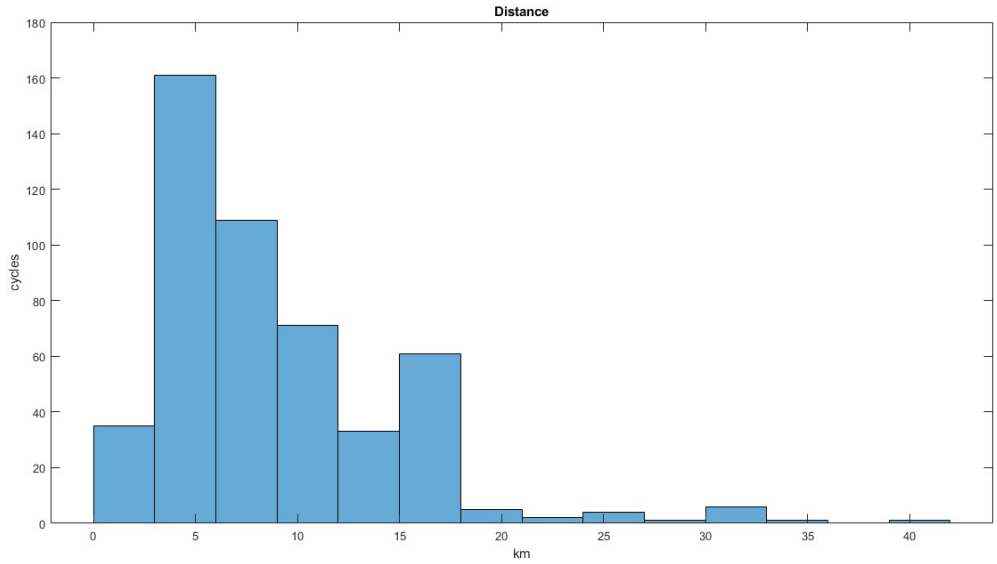

Figure 4.14: Total driving distance (km)

It is shown the presence of two main clusters, since, regarding frequency, the data is grouped forming two peaks: 3-6 km and 15-18 km, this could be due to the existence of at least two different driving profiles derived from different driving environments.

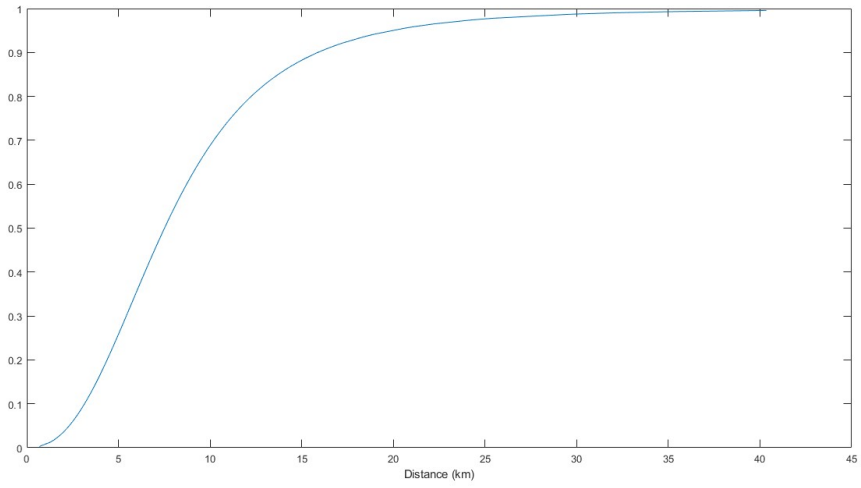

Figure 4.15: Total driving distance CDF (km)

After previously exposed, it is arguable that the driven distance of a representative driving cycle shall be found in a range between 5 km and 15 km.

In order to establish a relationship between the starting/ending locations and the distances travelled, the next heatmap is presented (figure 4.16). Seven locations with the highest frequency are selected to measure the driving distance between them. The most frequent locations (figure 4.5) were San Cristóbal de La Laguna, Guajara and Tabaiba. It can be seen that those cycles associated with Tabaiba have a higher driving distance than the rest of the locations, demonstrating that the said cycles may be treated as outliers. Hence, regarding driving distances, the most representative locations are La Laguna-Guajara, La Laguna-Las Chumberas, La Laguna-El Sobradillo, Finca España- Salud La Salle and El Sobradillo-Finca España.

Distance

	La Laguna	Guajara	Tabaiba	Las Chumberas	Salud- La Salle	El Sobradillo
Guajara	4,8					
Tabaiba	15,5	11,8				
Las Chumberas	4,7	0,8	11,1			
Salud- La Salle	10,7	8,4	13,8	7,9		
El Sobradillo	5,6	3,2	10,2	3,5	10,3	
Finca España	3,3	2	12,8	2,1	4,3	4,7

Figure 4.16: Average driving distances (km) between locations. Source: Google Maps

It is important to mention that the heatmap only shows an approximate distance that could vary depending on the direction (designation of starting and ending points) and route taken, thus, it should not be used as an exact reference.

Regarding the cycle duration measured in seconds, it is displayed, in the following histogram, that cycles with a duration of about 500 seconds (8.3 minutes) have the highest frequency (about 25% of recorded cycles). As it can be inferred, the representative driving cycle needs to be close to the said measure.

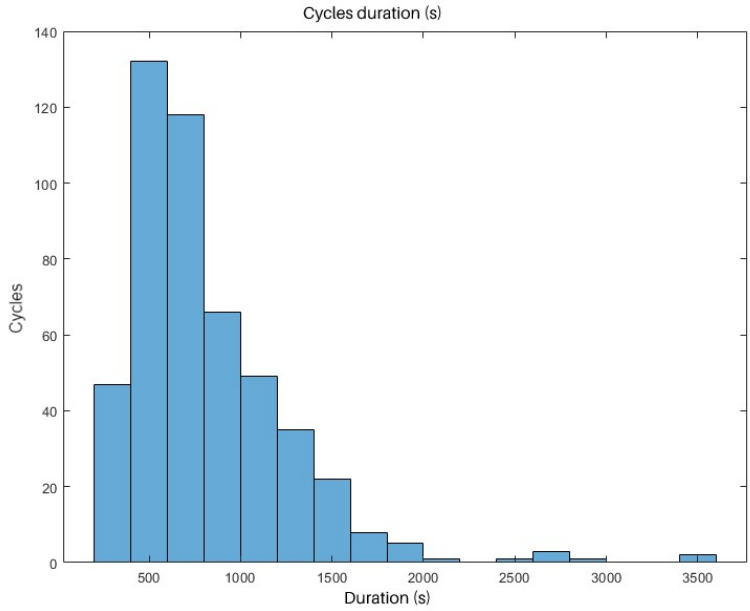

Figure 4.17: Cycles duration (s)

Driving Features

As it was explained in chapter 2, the driving features are defined as the variables that will play a decisive role through the clustering process, allowing the characterization of the groups. Consequently, studying the attributes of the data set features will help to address the final cycle analysis. S

Regarding some of the driving features mentioned in the first chapter of this thesis, it is possible to demonstrate the increment of the average speed once the idle time is removed from the speed vector (from 36.01 to 40.75 km/h). In figure, 4.18 such increment in the distribution is displayed. The importance of calculating the driving speed without considering the instants where the vehicle is not moving (idling) relies on the influence of idle time on this variable. In other words, the average vehicle speed would rest on the percentage of time idling, influencing the independence of the variables.

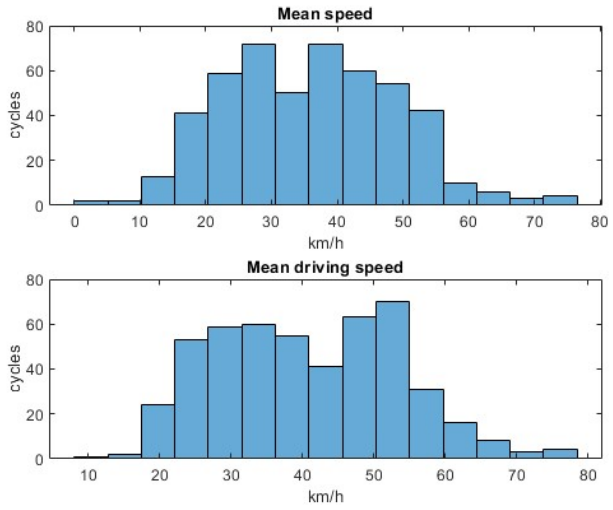

Figure 4.18: Mean speed and Mean driving speed (km/h)

As can be seen in figure 4.18, the mean driving speed histogram indicates the presence of at least two groups, since the shape of the distribution is separated into two main peaks. The first one faces a higher frequency at 31-35 km/h whereas the second one, with a smaller dispersion, takes place at 50-55 km/h. It is important to highlight that the separation of said groups is more notorious once the idle time is removed from the speed vector.

At a first glance, the two main groups could be described as high speed (motorway) and low speed (urban) cycles. However, as explained before, there exist cycles with mixed driving profiles that could affect the clustering.

When it comes to the maximum speed reached in each cycle, contrary to the displayed in the mean driving speed histogram, there is a distribution that follows a chi-square shape with only one peak at the range of 100-110 km/h. This also could be explained by the previous argument, where it was stated that one single driving cycle may be composed of different environments, leading to only representing the maximum speed reached in motorways. Finally, the cycles with a maximum speed under 10 km/h might be treated as outliers.

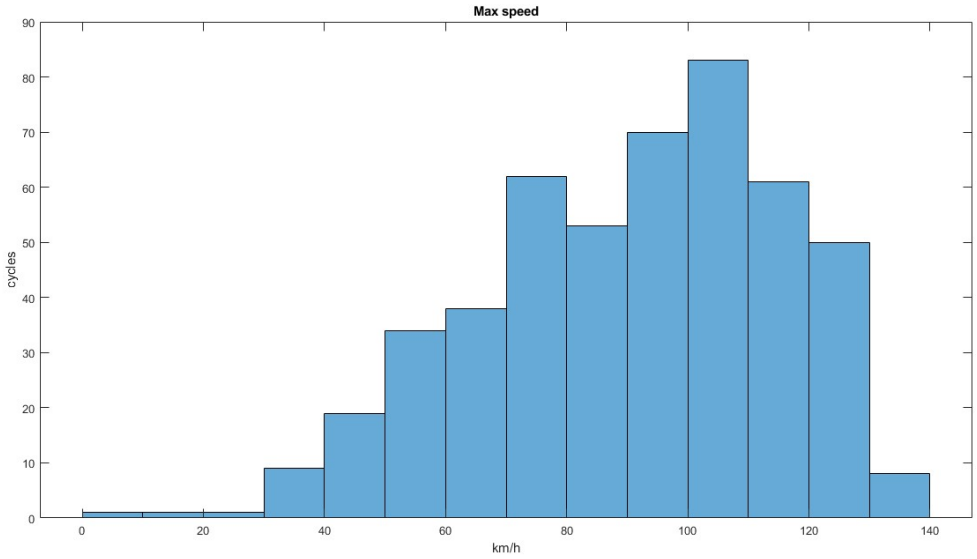

Figure 4.19: Maximum driving speed (km/h)

Regarding the percentage of time idling during the cycles, the histogram presented in figure 4.20 follows a Weibull distribution, reaching a peak at 3-9%. Once again, the distribution is uniform, showing that the cycle might be a compound of different driving profiles, losing representativeness. The mean percentage of time idling is 12.8%.

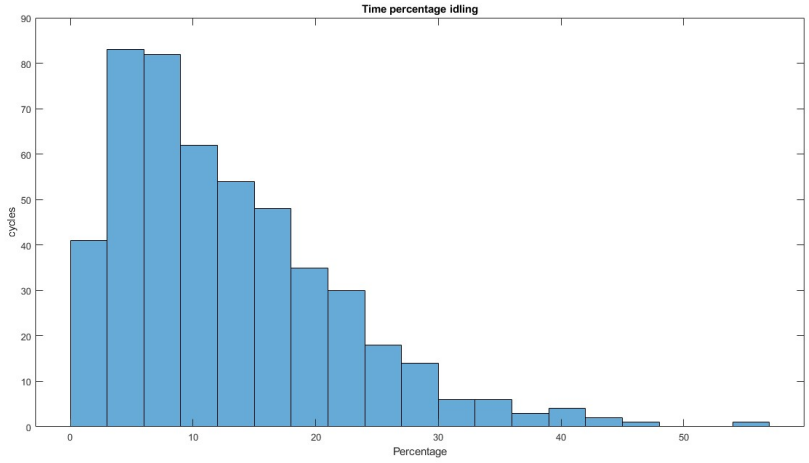

Figure 4.20: Percentage of time idling

Finally, attending acceleration-related features, the distribution found for APA, ANA, RPA and RNA is uniform, without demonstrating the presence of different groups in the histograms.

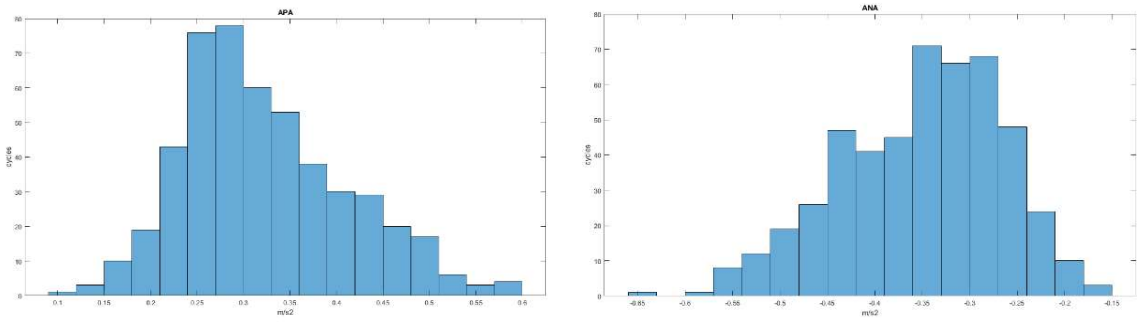

Figure 4.21: APA and ANA (m/s2)

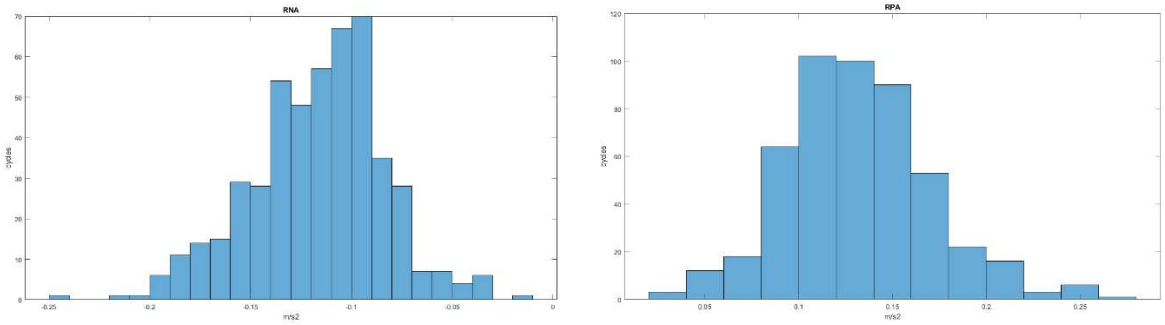

Figure 4.22: RPA and RNA (m/s2)

In relation to acceleration features, it is expected to obtain the highest APA/ANA in urban (low-speed cycles), where the frequency is highest, whereas in motorways it is the lowest since the variations of speed take place less abruptly, resulting in a more uniform driving speed.

On the other hand, RPA and RNA associate the acceleration and driving speed, thus, the highest values will be obtained on extra-urban/rural roads (medium speed/acceleration), and not on urban (high acceleration and low speed) nor motorways (low acceleration and high speed) cycles.

Microtrips Division

After performing the MTs division, it is expected a less uniform distribution in most features, since each MT will fully belong to a specific driving condition, unlike the complete cycles, which can be composed by different conditions.

The maximum speed is selected as an example to illustrate the idea. Suppose an entire driving cycle, exposed to different driving situations, where several types of roads are taken. The maximum driving speed reached will always belong to motorways and principal avenues, underestimating the maximum speed reached in the minor roads and urban sections of the cycle, turning the mentioned variable into a non-useful feature. This is displayed in figure 4.23.e (after the division), where can be highlighted the separation of two main groups after the division in comparison to 4.23.b (before the division).

This change is also remarkable when it comes to the mean driving speed and distance. Regarding the latter, as expected, it can be seen an increment in the frequency of short cycles (under 2 km) since MTs are considered the driving period between idle. Finally, despite the mentioned, there still are MTs with a driving distance over 20 km that may be identified as outliers.

BEFORE DIVISION **AFTER DIVISION**

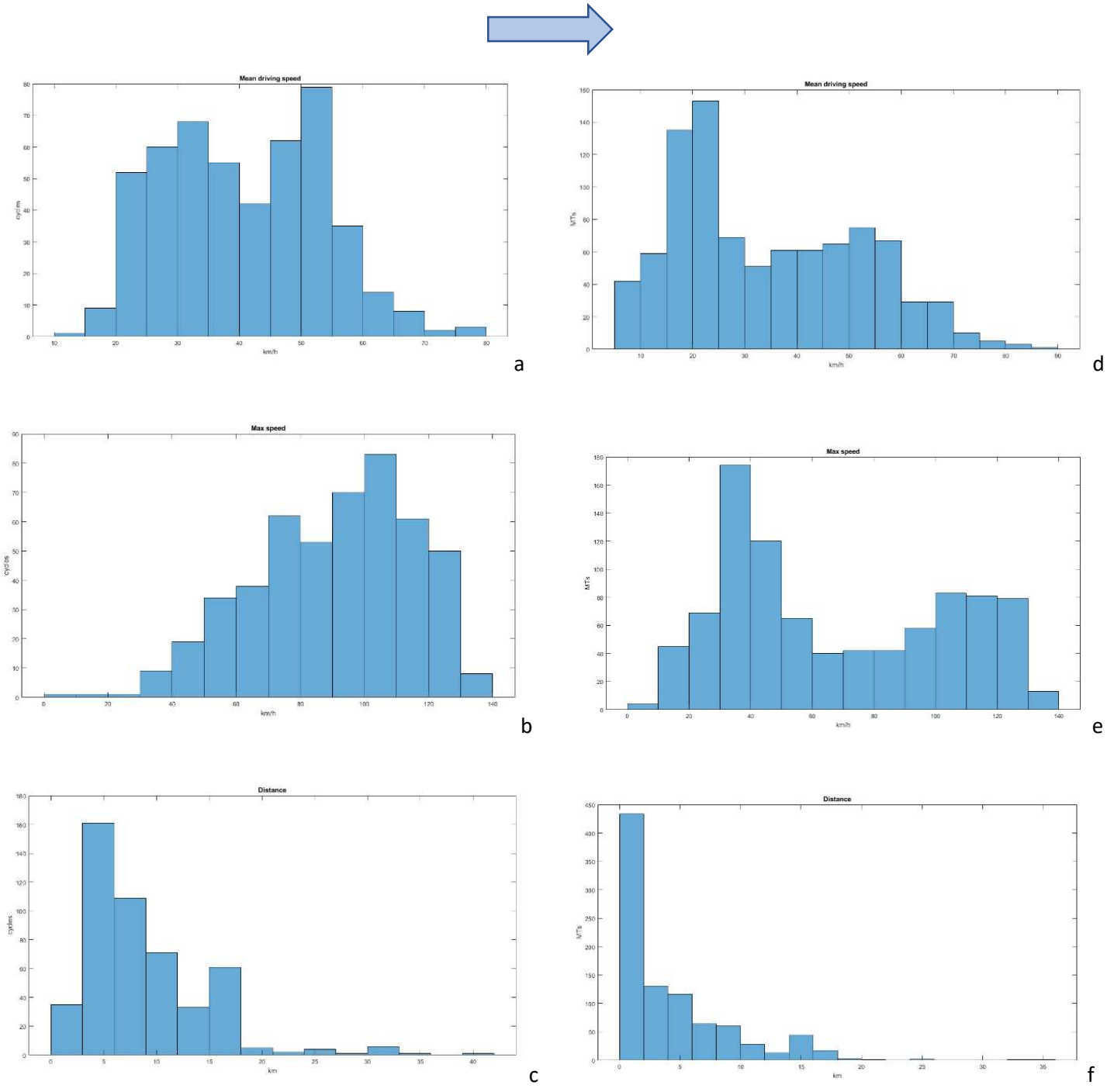

Figure 4.23: a) Average driving speed before division (km/h) b) Maximum speed before division (km/h) c) Distance before division (km) d) Average driving speed after division (km/h) e) Maximum speed after division (km/h) f) Distance after division (km)

In order to address the outliers previously mentioned, it is important to understand where they come from. It can be assumed that, since the outliers are found in the distance, it is possible to also find MTs unusually long.

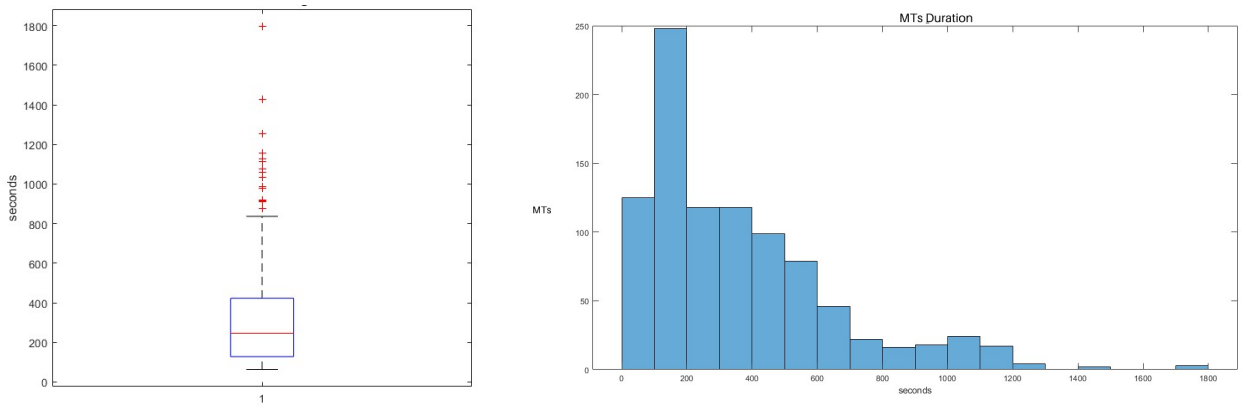

Figure 4.24: MTs duration with outliers (s)

For this case, it will be contemplated the definition of outlier given in chapter 4: Median + 3xMAD (mean absolute deviation). For this case, the median duration is 293 seconds, and the MAD is 220.90. Hence, any MT with a duration over 956 seconds will be considered an outlier and will be removed from the MTs set.

After the outlier removal, it is obtained a more uniform distribution, with the highest frequency between 1.67 and 3.33 minutes of MT duration.

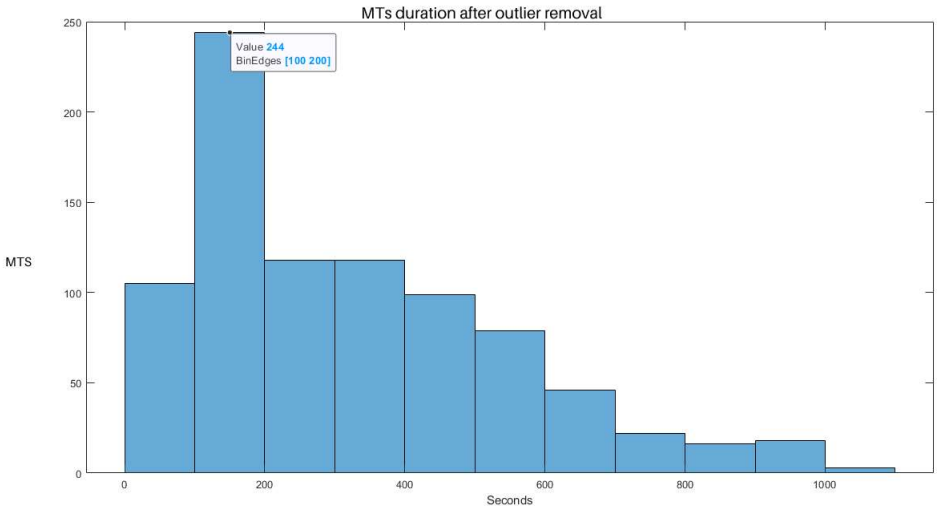

Figure 4.25: MTs duration after outlier removal

As mentioned, the histogram of the duration also changed, supporting the idea of the relationship between unusually long cycles and long distances.

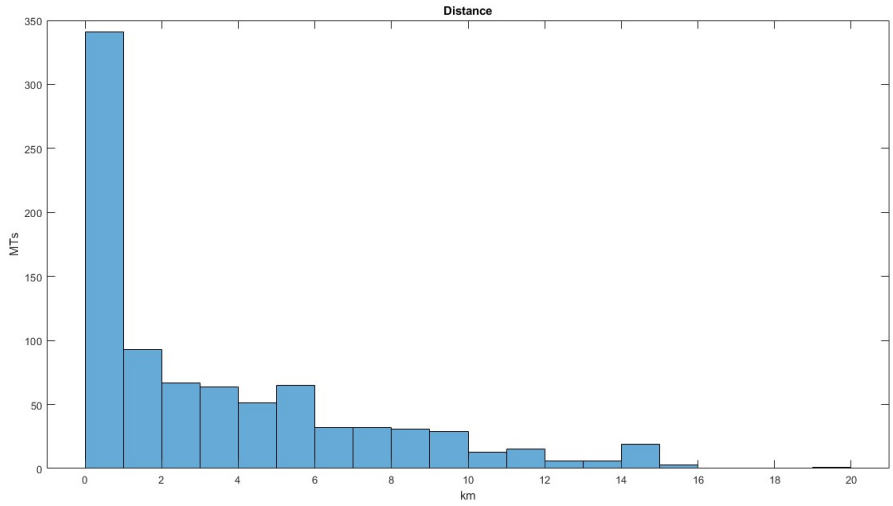

Figure 4.26: Distance after MTs division

It is important to identify the characteristics of the outliers removed, since their elimination may compromise the data set representativeness. Hence, the features of the outliers removed will be studied.

Initially, it can be thought that said outliers belong to long highway cycles, characterized by large distances, long periods of driving time, high velocity, and smaller acceleration. To support this idea and evaluate the impact of the removal of the outliers, figure 4.28 represents the data before and after the removal.

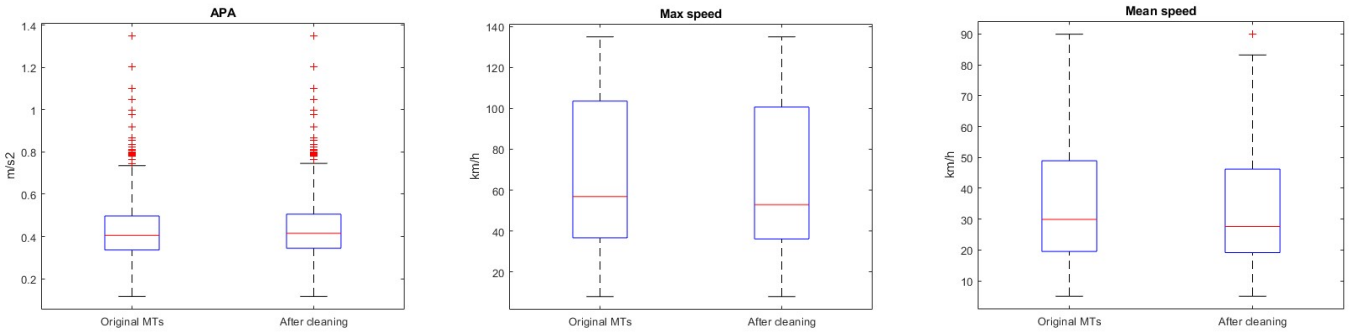

Figure 4.27: Features comparison before and after outliers treatment.

As is illustrated in the previous boxplots, it is noted a decline in maximum speed and mean driving speed. In order to exemplify the impact of the removal of outliers from the data set, the next table is presented. It shows the median of five features before and after the outlier removal while highlighting the median of the outliers.

It can be seen an important reduction in the median MT distance (>17%) since the median distance of the outliers removed was high (16.05 km) in comparison to the data set (2 km).

On the other hand, other features also got altered (i.e., Max speed and mean driving speed), which could mean that removing those outliers could affect the representativeness of the final driving cycle. It is important to recognize that the driving features of the mentioned MTs (outliers) besides distance are within the expected range (no unusual values).

	Median Mean driving speed (km /h)	Median Max speed (km /h)	Median APA (m/s2)	Median RPA (m/s2)	Median distance (km)
Before	29.91	56.88	0.405	0.195	2.35
Outliers	53.19	108.34	0.324	0.157	16.05
After	27.60	52.95	0.415	0.198	2.00
Change	-8.37%	-7.42%	2.41%	1.72%	-17.50%

Table 4.4: Impact of the outlier removal on the data set. Source: own elaboration

As a consequence of the foregoing, the MTs categorized as outliers were not removed from the data set to ensure that the sample representativeness was not affected. However, the aforementioned MTs were not included in the distance-related calculations since it is an independent variable that does not directly affect other features. Nevertheless, it was observed that these outliers belonged to the motorway profile due to their low APA and RPA and high maximum speed and distance. Therefore, these MTs were included in the clustering analysis, despite the possibility of influencing the cluster compactness.

Dimensionality Reduction

As it was mentioned in the second chapter of this study, due to the number of variables, to avoid losing information through the selection of specific features, it will be necessary to perform a dimensionality reduction methodology. The most common algorithms for this task are t-SNE and PCA. Consequently, both algorithms will be studied to select the one that best fits the data set.

Feature Scaling

In chapter 2 it was explained the importance of the feature scaling before performing any dimensionality reduction. In figure 4.28 a boxplot intends to exemplify the effect of normalization on features. In order to illustrate this impact, figure 4.29 is presented, where a MT is shown before and after the normalization. As it is inferred, the relationships between the magnitudes within the MT are the same.

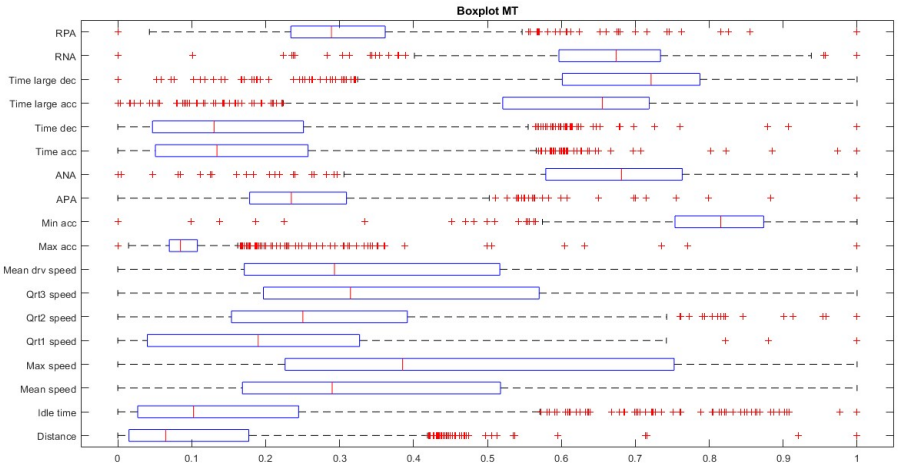

Figure 4.28: Box Plot of normalized features

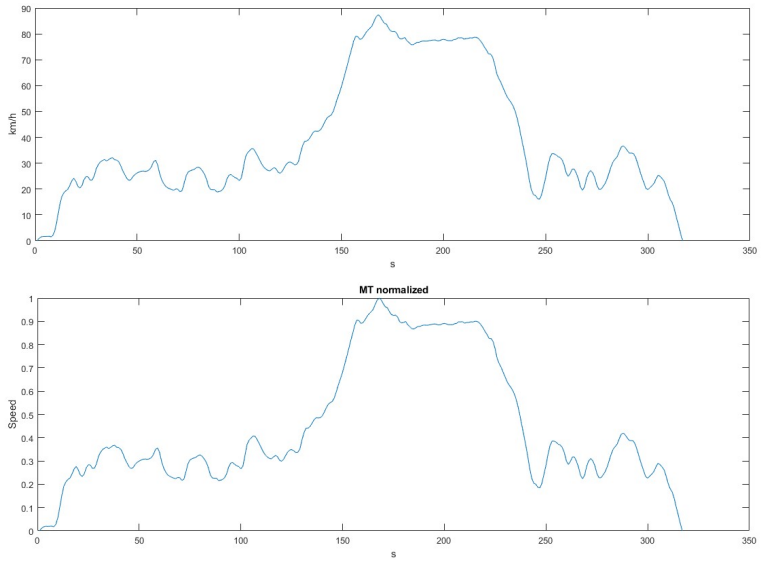

Figure 4.29: MT before and after normalization

Principal Component Analysis

After normalization, following the methodology described in chapter 3, the PCA is executed. The Pareto graph indicates that the employment of 3 principal components (PC) may be correct since they represent about 80% of the data set variability [21]. This is favourable as the result can be plotted in a 3-dimensional graph. The mentioned number of PCs must be determined before executing the algorithm since it is an input parameter.

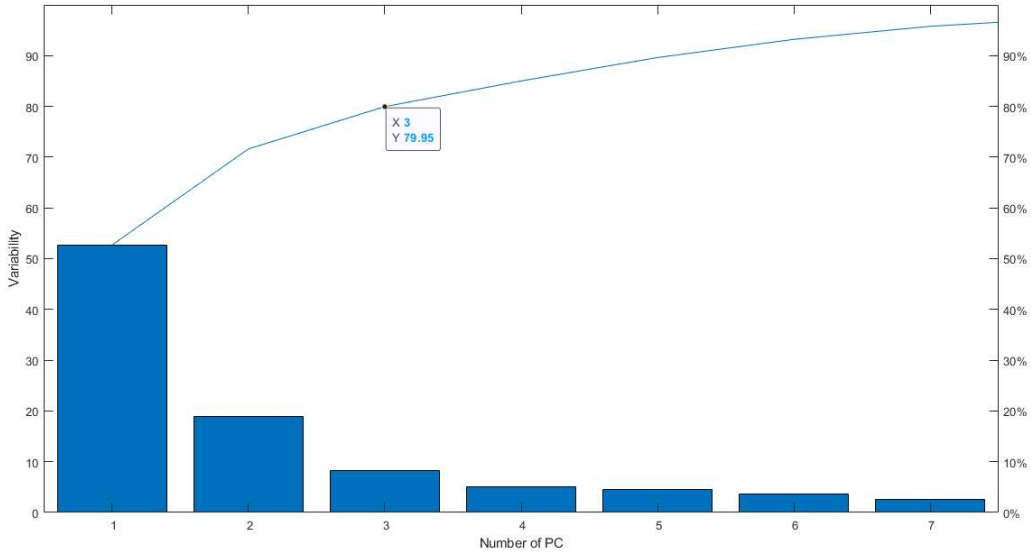

Figure 4.30: Number of PCs needed.

The biplot presented in figure 4.31 intends to represent the influence of the driving features on PCs. As it can be seen, the first component is mainly composed of distance, idle, and speed-related features, such as the speed quartiles and mean driving speed. On the other hand, the second component is mainly influenced by acceleration-related features, such as APA and RPA. Hence, the highest variability will be found on speed-related features, idle, and distance. Finally, the third component (not visible in this plot) will be mainly affected by idle and time accelerating/decelerating.

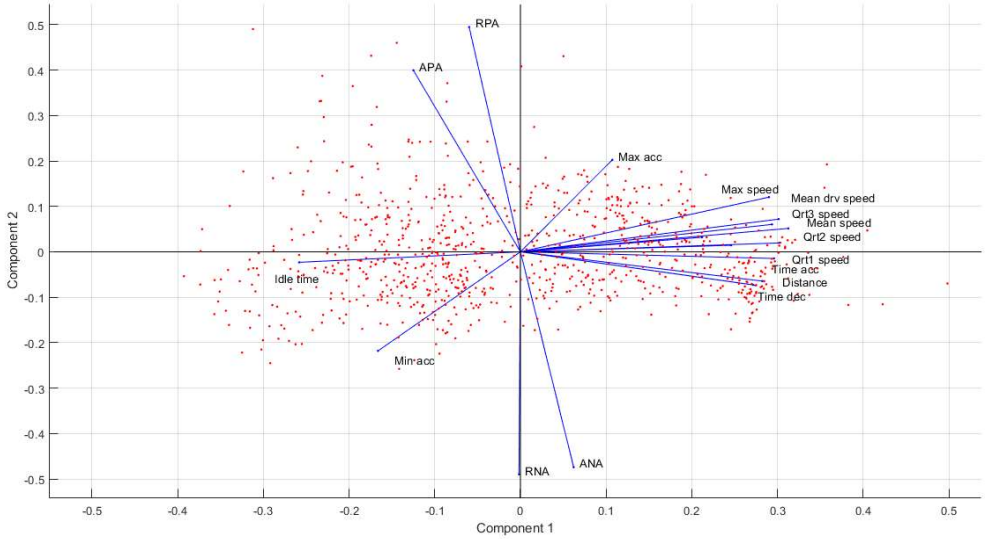

Figure 4.31: Influence of features on PCs.

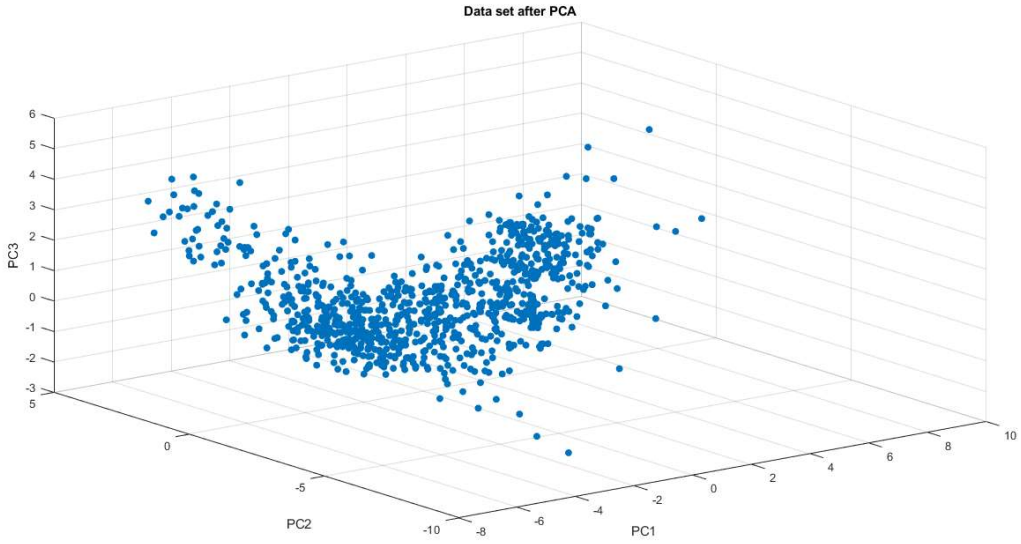

Figure 4.32: Data set after PCA

According to the methodology, after performing the dimensionality reduction algorithm, the reduced data set (Figure 4.32) can be clustered. The algorithms executed are k-means and hierarchical clustering.

Consequently, the performance metrics are employed to evaluate the quality of the clusters obtained which will be addressed through silhouette, Calinski Harabasz and Davies Bouldin coefficients for a different number of clusters (2, 3, and 4).

PCA	**Kmeans**			**HC**		
nº clusters	**2**	**3**	**4**	**2**	**3**	**4**
Silhouette	0.529	0.413	0.414	0.476	0.333	0.333
Calinski Harabasz	579.90	427.30	388.80	473.20	368.90	332.72
Davies Bouldin	1.134	1.353	1.211	1.119	1.303	1.429

Table 4.5: Cluster coefficients by clustering algorithm and number of clusters. PCA.

In this case, on average, the values are higher in 2 clusters using k-means, whereas the worst combination is found in 4 clusters/HC. Hence, the preferred clustering algorithm is k-means. By this, it is remarkable a more uniform distribution where the number of clusters is 2 in comparison to 3.

Figure 4.33 illustrates the silhouette coefficients for the data set when two and three clusters are studied. The aim is to obtain the most uniform distribution (similar widths) and higher silhouettes values. As Expected, the best results are obtained when two clusters are studied.

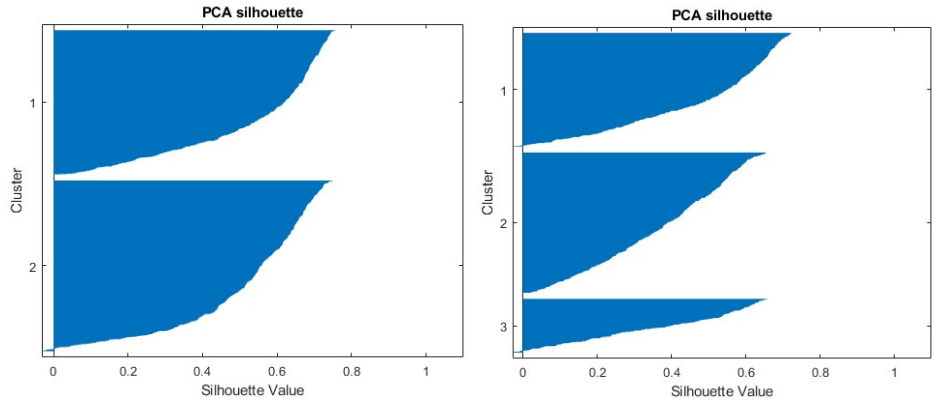

Figure 4.33: Silhouettes 2 and 3 clusters using k-means.

t-Distributed Stochastic Neighbour Embedding

Before executing this algorithm, several input variables need to be known, such as perplexity (mentioned in chapter 2). As it was argued, those variables will vary depending on the data set, hence, it is not possible to know them beforehand.

The Kullback-Leibler divergence (KL), described as the relative entropy earlier, calculates the similarity between the original and reduced data set. As expected, a higher perplexity will derive in a lower KL, which denotes a lower amount of data lost. However, a lower KL will not indicate an optimum perplexity, therefore, the methodology proposed by [23], where a score is designated, takes place.

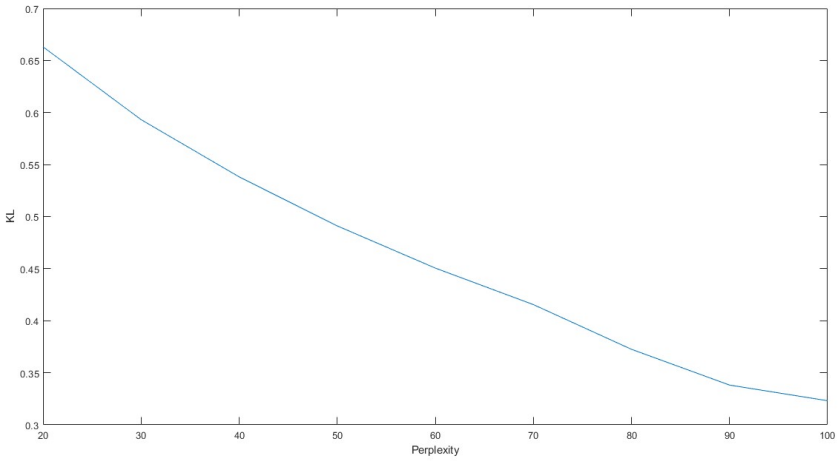

Figure 4.34: KL and perplexity.

Following equation 22, the perplexity will be found by plotting the mentioned score. The best perplexity possible will result in a minimum score, hence, the lowest point in the graph will be selected as ideal perplexity for this data set (figure 4.35).

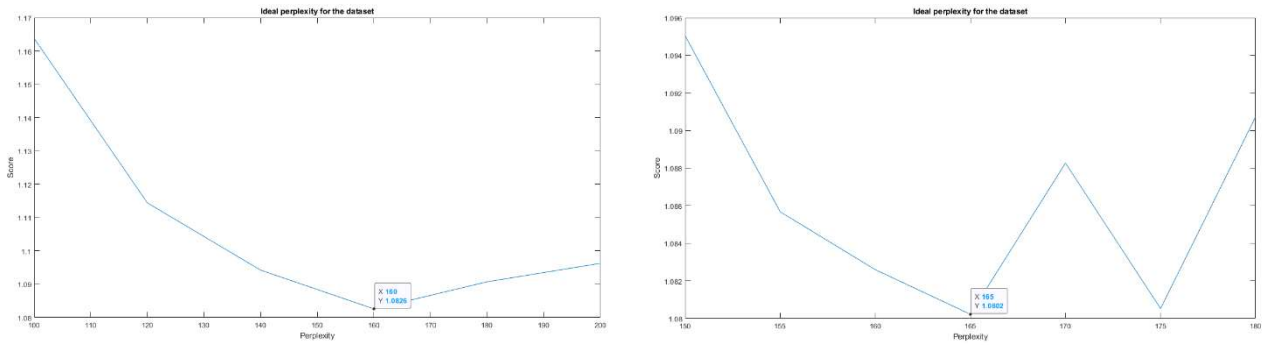

Figure 4.35: Score and perplexity.

As can be seen, the ideal perplexity is found at 165, although the maximum recommended is 50, however, it is considered insufficient due to a high number of local neighbours. In figure 4.36 the data set is shown after the application of t-SNE.

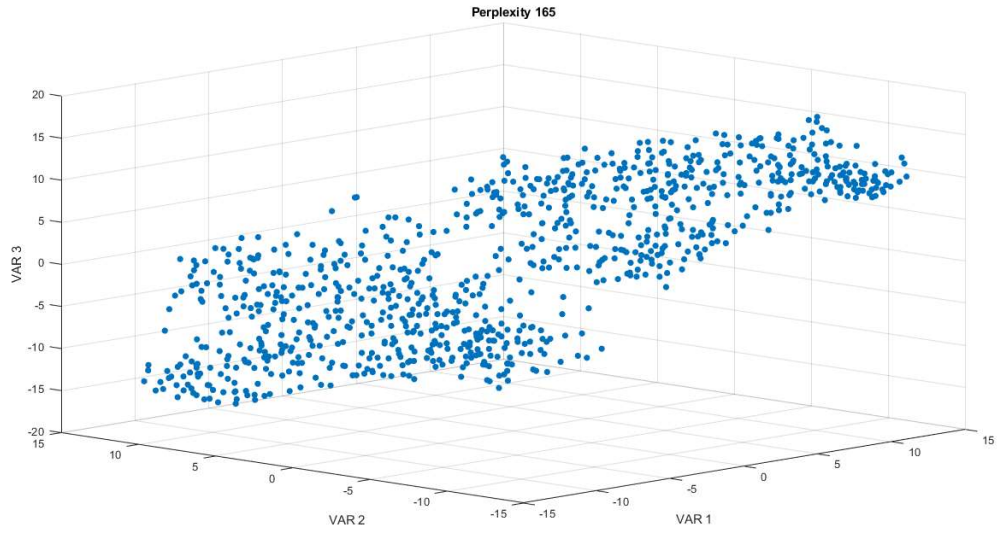

Figure 4.36: Data set after t-SNE

The result after t-SNE is satisfactory since it is visible the existence of at least two well-differentiated clusters. To illustrate the effect of perplexity on the final result, figure 4.37 is shown, where the left represents an insufficient perplexity (30) and the right an excessive one (300). As anticipated, local neighbours will prevail with a lower perplexity and small groups will be formed. On the other hand, if the perplexity is excessive, it is remarkable a lower distinction between groups, decreasing the quality of the clusters.

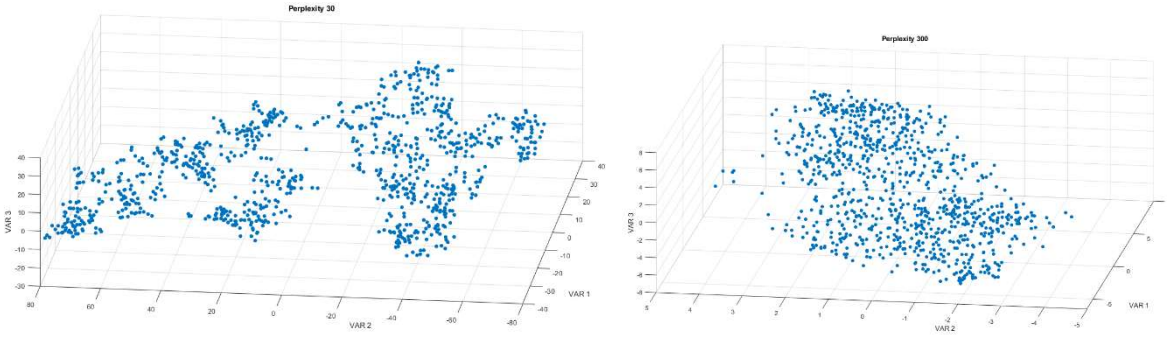

Figure 4.37: Effects of low (left) and high (right) perplexity on the data set.

Finally, as it was done with the PCA, the quality of the clusters is studied through different coefficients (Table 4.6). It can be seen that the optimal number of clusters is 2 obtained through the k-means algorithm.

t-SNE	**Kmeans**			**HC**		
nº clusters	**2**	**3**	**4**	**2**	**3**	**4**
Silhouette	0.766	0.617	0.541	0.760	0.593	0.506
Calinski Harabasz	1933.00	1444.40	1389.19	1906.00	1380.41	1280.20
Davies Bouldin	0.644	0.865	1.031	0.649	0.905	1.073

Table 4.6: Cluster coefficients by clustering algorithm and number of clusters. t-SNE.

After both dimensionality reduction methodologies were studied, it is necessary to select the optimal combination. Based on table 4.5 and 4.6 it is clear that t-SNE offers more accurate clusters in comparison to PCA and, additionally, the k-means algorithm provides a better quality of the groups. Hence, the result of this combination (t-SNE + k-means) will be employed for the next stages of this study.

Regarding the number of clusters, it is clear that the optimal number is 2 for all cases and coefficients. However, as it is known, each cluster represents a driving profile, meaning that, in the case of selecting only two clusters, the final cycle will be composed of only two driving conditions which may be inaccurate.

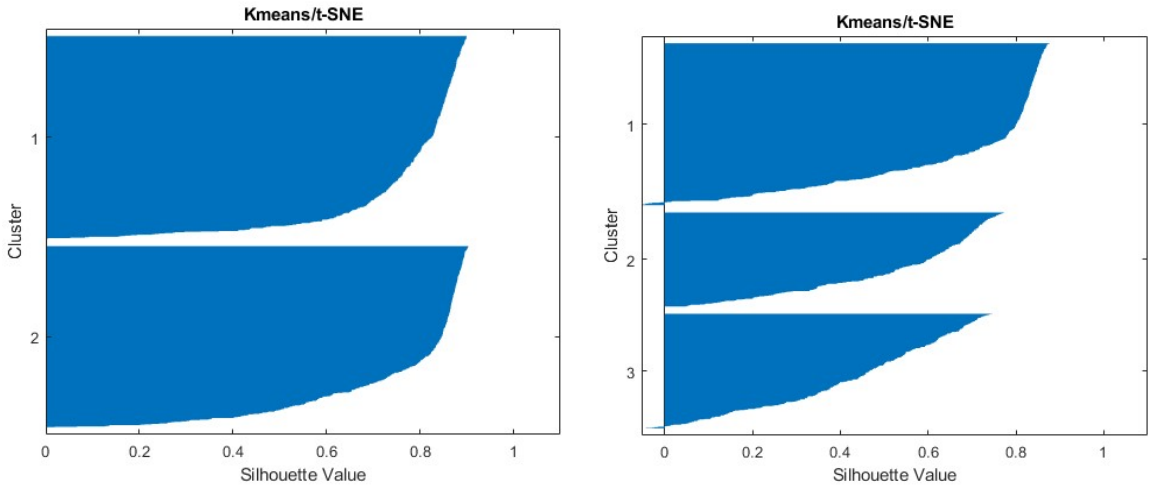

Figure 4.38: Silhouettes of t-SNE and k-means for 2 and 3 clusters.

It can be inferred that the mentioned extra-urban/rural MTs may have features that could be related to high-speed (HS) and low-speed (LS) MTs and, consequently, are separated into those 2 clusters. This may produce a decrease in the maximum speed reached in HS MTs and increase it in LS MTs. Hence, the dispersion of the clusters might be higher.

To address this argument, it would be necessary to compare the results obtained from the clustering algorithm with 2 and 3 clusters (figure 4.39). As expected, when 2 clusters were calculated, it is remarkable the presence of two well-differentiated groups: high speed/low APA and low speed/high APA MTs.

Initially, the data set was almost equally distributed in both profiles: 432 HS MTs and 483 LS MTs for 915 data points (MTs). It is also highlightable the presence of 19 unacceptable outliers in LS, where the maximum speed reached was 107 km/h, which is inappropriate for this driving profile (urban). It is also noteworthy to mention that the outliers addressed before (unusually long MTs) were placed in the HS MTs, which is acceptable.

On the other hand, when three clusters are studied, it is possible to see three profiles: low speed/High APA, medium speed/high APA and high speed/low APA that may represent urban, extra-urban/rural and high speed cycles respectively.

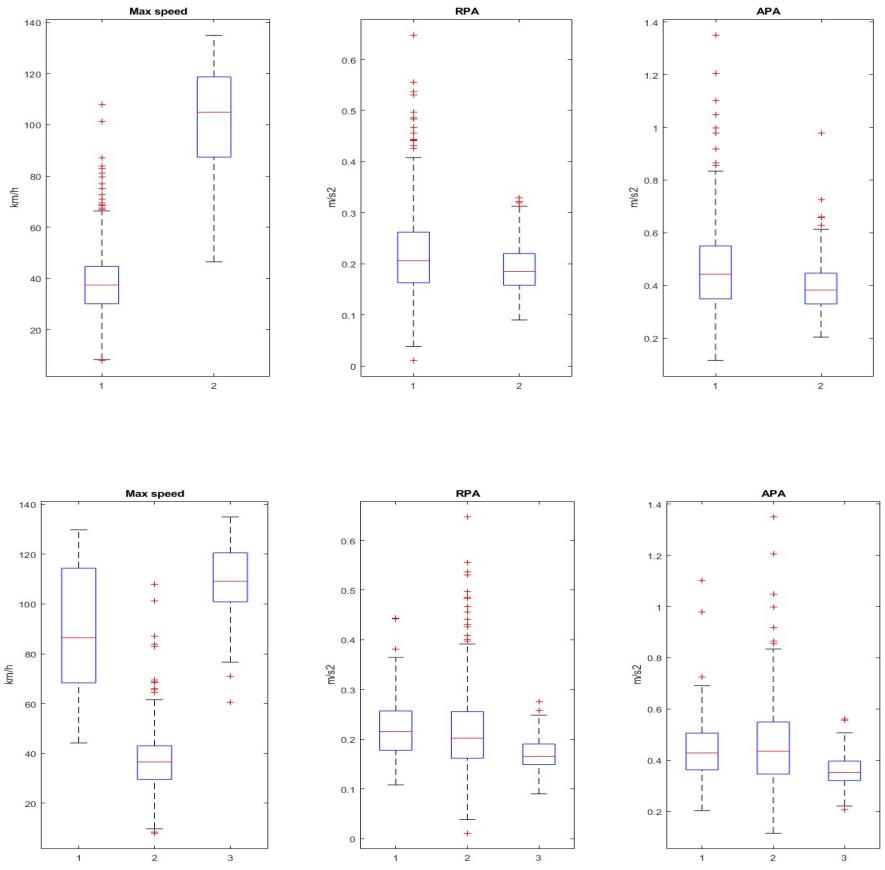

Figure 4.39: Results of clustering 2 (above) and 3 (below) clusters.

The median absolute deviation (MAD) is measured to evaluate the impact of a third cluster addition. If the explanation above is correct, this measure must decrease after the addition of a third cluster.

	2 clusters	3 clusters
High speed	219.35	187.21
Low speed	233.15	236.12

Table 4.8: Median absolute deviation (MAD).

As shown, the MAD decreased by 17% in the HS cluster after adding a third cluster. On the other hand, the MAD of the LS MTs slightly increased. From here, it could be concluded that a third cluster may have a higher impact on the HS cluster. This can be explained by the new data distribution: 210 data points in the HS cluster, 252 in medium speed (MS), and 453 in LS.

When it comes to the driving features, table 4.9 summarizes the changes of some representative variables after the third cluster addition. The HS cluster increased its maximum speed by 4% and reduced the acceleration-related features by 8 and 10% which is favourable.

The impact on the LS cluster is less noticeable, where the maximum speed decreased by 2% (favourable), however, the acceleration-related features slightly decreased by almost 2% (unfavourable).

High speed	Max speed	109.07	km/h	4.00%
	APA	0.352	m/s2	-7.85%
	RPA	0.165	m/s2	-10.33%
Medium speed	Max speed	86.5	km/h	-%
	APA	0.428	m/s2	- %
	RPA	0.218	m/s2	-%
Low speed	Max speed	36.6	km/h	-2.14%
	APA	0.435	m/s2	-1.81%
	RPA	0.202	m/s2	-1.94%

Table 4.9 Impact of third clusters on features.

After this argument, 3 clusters are selected to characterize the driving profiles, despite reaching the highest cluster quality when 2 groups are considered. Hence, figure 4.40 is plotted, where different features are represented. As expected, the high-speed cluster possesses a low RPA, APA, and time idling but the longest distances. The medium-speed cluster is represented by a medium APA and high RPA with medium distances and medium idle time. Finally, the low-speed cluster shows a high APA, medium RPA, high idling time, and shorter distances. Consequently, the latter can be named urban MTs whereas the medium-speed and high-speed MTs may be classified as extra-urban and motorway cycles respectively.

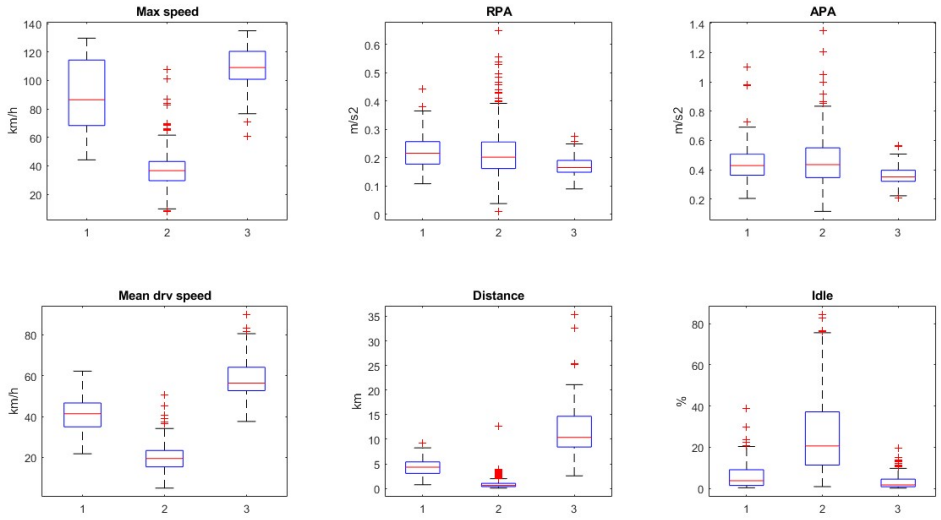

Figure 4.40: Driving features of 3 clusters: 1) medium speed; 2) low speed; 3) high speed.

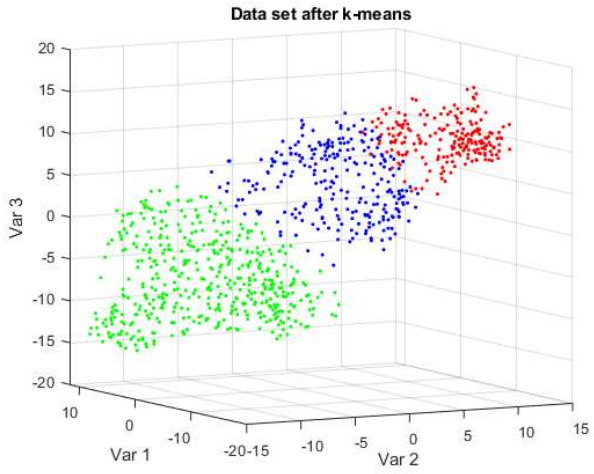

Figure 4.41: Data set after k-means (3 clusters).

Driving Behaviour

In chapter 2 it was expressed the importance of distinguishing between different driving behaviours since it could affect speed and acceleration-related features. For this reason, following the mentioned in [31], the next methodology is followed.

Initially, starting from the stated in the previous reference, it is possible to relate acceleration features to driving behaviour. The main reason to rely on this affirmation is that driving speed is limited and strongly influenced by traffic conditions, road infrastructure, and speed limits. However, the acceleration is not directly influenced by these variables but by driving behaviour and driving profile. Hence, the different behaviours are defined through APA, ANA, RPA and RNA.

As explained in chapter 2, the first step is to calculate cumulative diagram functions (CDF) of the mentioned acceleration-related variables. Hence, 4 different CDFs will be obtained (figure 4.42).

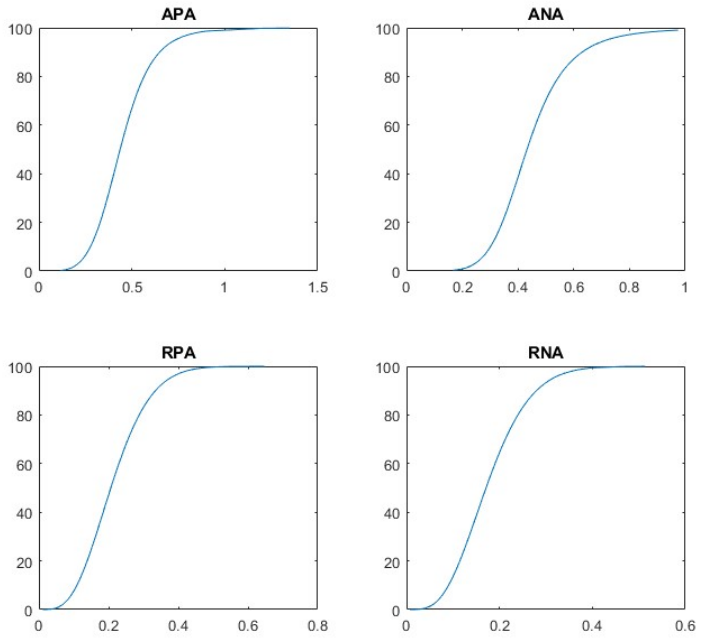

Figure 4.42: CDF of acceleration-related features.

Consequently, it is possible to calculate the average position (score) of every MT on each CDF. This score can be plotted on a final CDF (Figure 4.43) where the driving behaviours will be identified by dividing this graph on 0-1st quartile, 1st-3rd quartile, and 3rd-4th quartile.

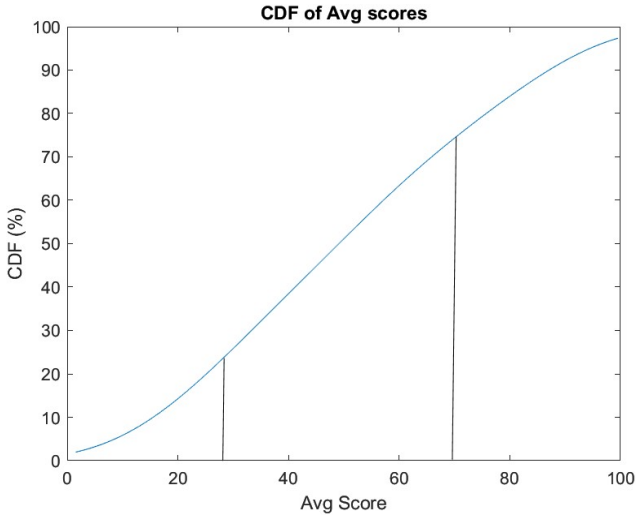

Figure 4.43: CDF of Average score

The MTs under the first quartile (<25%) will be identified as mild drivers, the ones between the 1st and 3rd quartile (>25% -75%>) will be average drivers, and the MTs over the 3rd quartile (>75%) will be addressed as aggressive. As a consequence, half of the MTs are considered average drivers. It is important to highlight that this study does not intend to evaluate the representativeness of the driving behaviour since this would require an exhaustive analysis.

Finally, the result of this division will be 9 groups of MTs that will represent 3 driving behaviours (from the driving behaviour splitting) and 3 different driving profiles (from the clustering algorithm). To illustrate this idea, table 4.10 is shown.

	Drv. Behaviour		
Drv. Profile	Mild	Average	Aggressive
Urban	113	227	113
Extra-Urban	52	106	52
Motorway	63	126	63

Table 4.10: Distribution of MTs by driving behaviour and driving profile.

It can be thought that by splitting the clusters (previously obtained) following acceleration-related features, the driving speed may be affected, which might result in a loss of representativeness. To address this, table 4.11 and figure 4.44 are shown.

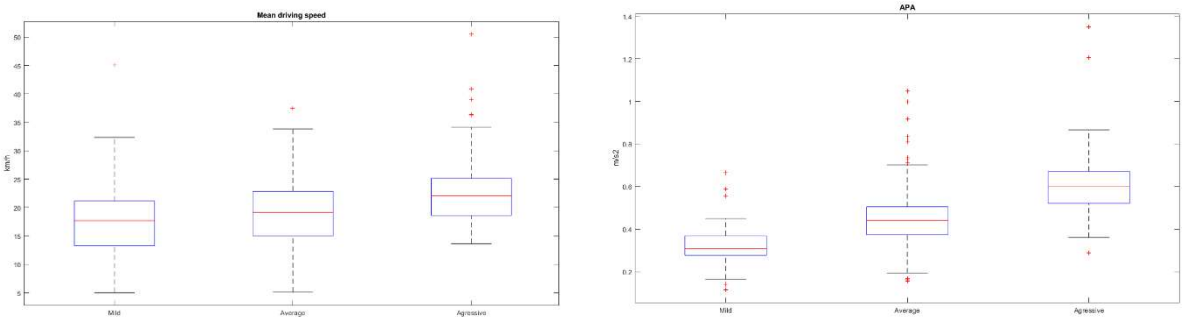

Figure 4.44: boxplot of APA and mean driving speed by driving behaviour for urban MTs.

In this boxplot, it is possible to see how the mean driving speed behaves in comparison to APA after the behaviour division for the urban (LS) cluster. As can be seen, the speed difference (between mild and aggressive) is considerably lower than in APA.

Additionally, table 4.11 illustrates the existing difference between mild and aggressive behaviour for APA and mean driving speed in order to have a general view of their impact.

It is highlightable anew, that the variation of APA is remarkably higher than the one that occurred in the mean driving speed. Also, this variation is lower for urban MTs and higher for motorway MTs. Hence, it can be stated that the higher the speed the lower the distance between the 1st and 3rd quartile for APA. In order to represent this, figure 4.45 is presented.

	APA (m/s2)			Mean driving speed (km/h)		
Drv. Profile	Mild	Aggressive	Variation	Mild	Aggressive	Variation
Urban	0.309	0.602	94.82%	17.68	22.04	24.66%
Extra-urban	0.32	0.546	70.63%	38.52	43.35	12.54%
Motorway	0.296	0.44	48.65%	55.66	57.36	3.05%

Table 4.11: Impact of the driving behaviour division on mean driving speed (median).

In Mild drivers, the variation of the APA is lower for different driving profiles (-4.2% when the driving speed is increased). On the other hand, this variation is higher for aggressive drivers (-26.9%). Hence the changes of APA will be more notorious in aggressive drivers.

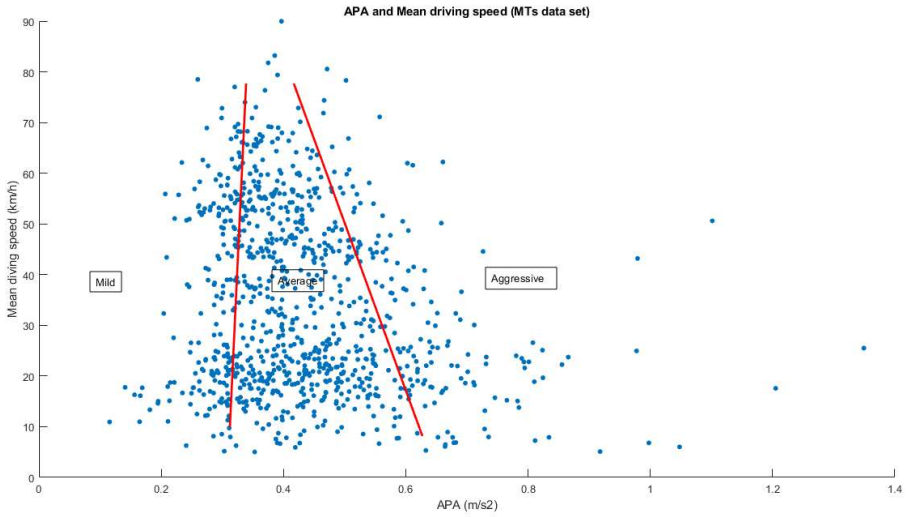

Figure 4.45: Relationship between APA, mean driving speed and driving behaviour.

Finally, according to figure 4.45 (Illustrative/non-real division), it is possible to state that the driving behaviour splitting will have a higher influence on urban cycles due to a higher dispersion of acceleration-related features.

Final Cycle Construction

After the division into groups by driving behaviour, the MTs are ready to be merged into a representative driving cycle. To illustrate the 9 groups, figure 4.46 is shown, where it is possible to observe the 3 different clusters (by colour) and 3 different driving behaviours (by shape).

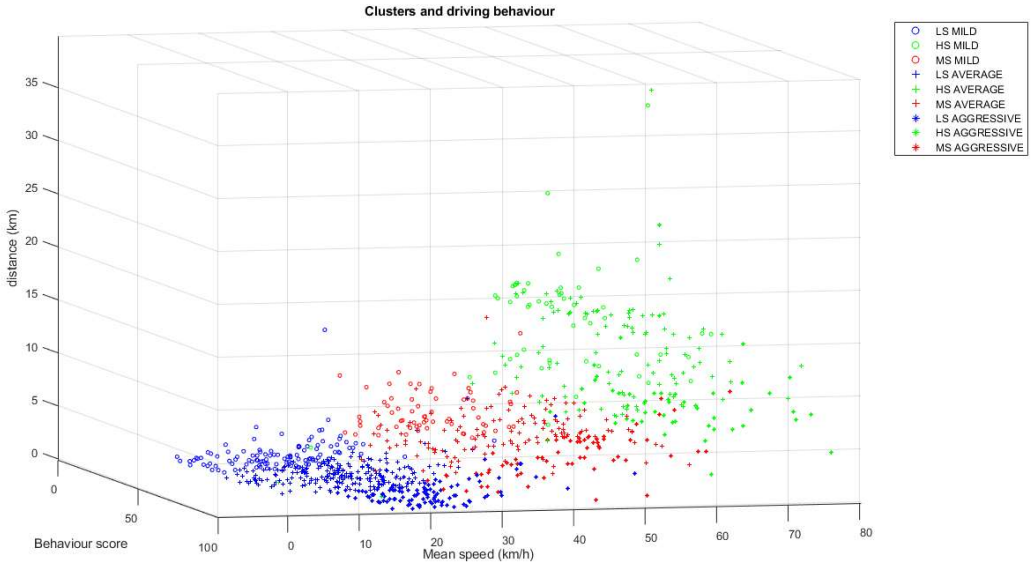

Figure 4.46: Data set after clustering and driving behaviour grouping.

At least one representative MT per group will be selected, hence, 9 MTs will be extracted. As can be inferred, those MTs will be the closest to the group centroid, as it was explained in [8]. It is important to mention that selecting only one MT per group may not be sufficient to satisfy the time distribution. To address this problem, the time of each driving profile in the data set needs to be calculated. To exemplify this, it is possible to have a data set where most MTs have been identified as LS and only a small fraction as HS. It would be inaccurate to only select one MT per cluster. Consequently, the next calculations are performed.

As argued in the previous reference, the share of time of a specific profile in the final cycle will be given by the next expression

$$t_i = \frac{t_{driv}}{t_{overall}} \sum_{j=1}^{n} t_{i,j}$$

Where *ti* is the time of the profile *i* in the final representative cycle, *tdriv* is the estimated duration (time) of the final cycle, *toverall* is the sum of the total duration of the MTs in the data set, and *ti,j* is the total duration of the MTs that belong to the profile *i.*

The total duration of the representative cycle may be determined by the statistical analysis performed (figure 4.17) where, on average, the cycle duration was about 811 seconds (13 min), with the highest frequency at around 500 seconds (8.3 min). The 3rd quartile was located at 1107 seconds (18 min). Hence, according to this analysis, the final cycle duration should be close to the mentioned values.

On the other hand, it was found that the average duration of the HS MTs was 765 seconds (table 4.13). Consequently, if an average HS MT is added to the final cycle, (considering a total duration of 1107 seconds given by the statistical analysis) it would represent about 70% of the final share of time, which is inaccurate.

Driving profile	Time recorded (s)	Percentage
Low speed (LS)	38,022	22.18%
Medium speed (MS)	52,278	30.49%
High speed (HS)	81,161	47.33%
Total	171,461	100.00%

Table 4.12: Share of time of recorded MTs by driving profile (average behaviour).

Driving profile	Average Duration (s)	Percentage
Low speed (LS)	167.50	12.43%
Medium speed (MS)	414.90	30.78%
High speed (HS)	765.67	56.80%
Total	1,348.07	100.00%

Table 4.13: Share of time of average duration by profile (average behaviour).

From this, it can be said that the final cycle must keep the proportions exposed in table 4.12. Consequently, it may be stated that, for this data set, through this methodology, it will not be possible to satisfy both requirements keeping acceptable driving features: proportions of profiles and final duration

In order to keep the proportion of the profiles in the final cycle, according to the exposed in table 4.13, a second LS MT needs to be added (table 4.14) to increase the final duration to over 1500 seconds (25 minutes). This arrangement shows satisfactory results regarding proportion (similar to the ones shown in table 4.12) but undesirable duration.

Driving profile	Average duration (s)	Percentage
Low speed (LS)	335 (167.5 x 2)	22.10%
Medium speed (MS)	414.90	27.37%
High speed (HS)	765.67	50.52%
Total	1,515.57	100.00%

Table 4.14: Share of time of average duration by profile with two LS MTs (average behaviour).

After this, for this study, the duration of the selected MTs was limited to obtain a final cycle of 1200 seconds, which is higher than the desired duration (approximately 811 seconds). As discussed, the representative MTs are the closest ones to the group centroid. However, when the duration is restricted (imposed as a requirement in the selection), the chose MT will no longer be the closest one to the group centroid, therefore decreasing the representativeness of the driving cycles.

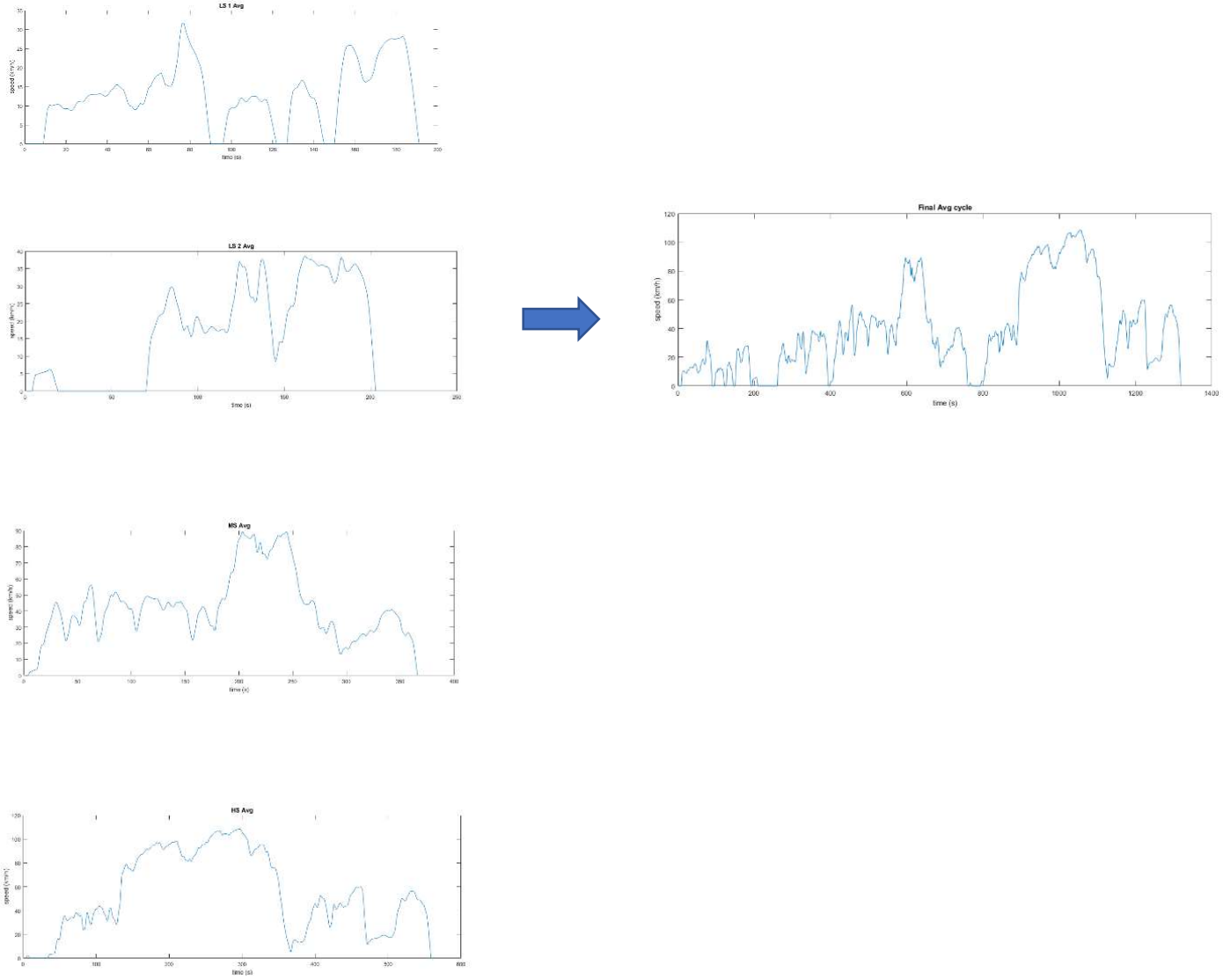

Figure 4.47: Merge of representative average cycle.

Finally, the representative cycles for each driving behaviour are presented (Fig 4.48-50). Following the standardised driving cycles structure, the cycles proposed are arranged from LS to HS (ascending).

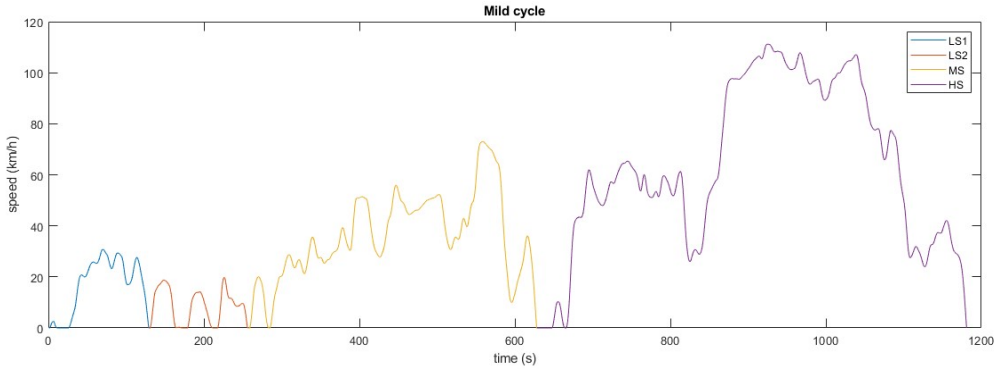

Figure 4.48: Representative mild cycle.

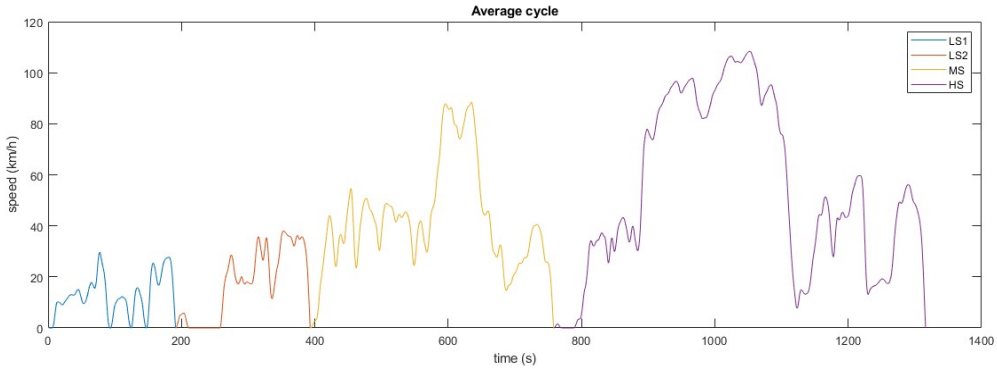

Figure 4.49: Representative average cycle.

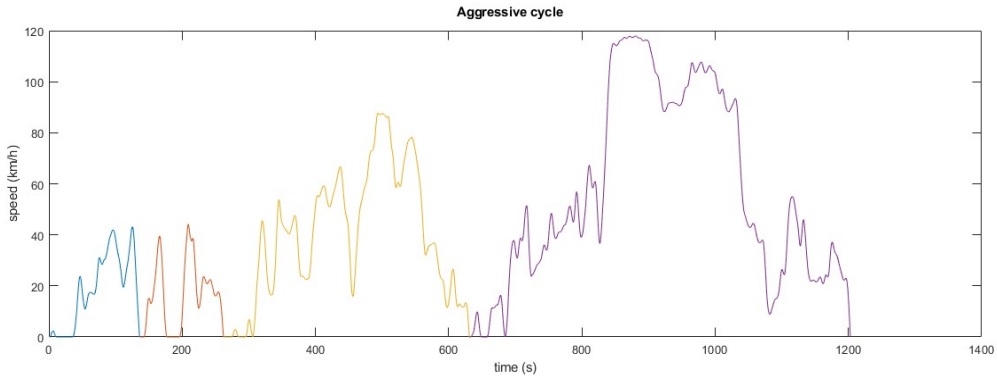

Figure 4.50: Representative aggressive cycle.

CHAPTER 5

CYCLE ANALYSIS

To evaluate the quality of the final cycles it is necessary to compare them with the original data set and with other existing driving cycles (mentioned in the introduction of this study). To illustrate the results obtained, tables 5.1-5.3 shows a comparison of these cycles with the data set. The full tables are in the appendices in case the lector needs further clarification.

In table 5.1 the features of the average representative driving cycle are presented. It can be easily distinguished a motorway profile with high speed and low acceleration, an extra urban segment with high accelerations and medium speeds and, an urban profile with high accelerations and lower speeds.

Average		Mean drv. speed (km/h)	APA (m/s2)	Duration (s)
LS	Rep. cycle	19.355	0.47715	274
	Difference	3.80%	4.43%	38.87%
MS	Rep. cycle	42.78	0.449	366
	Difference	3.18%	4.45%	-13.36%
HS	Rep. cycle	57.00	0.336	559
	Difference	-1.91%	-5.36%	-36.97%
	TOTAL	44.06	0.40	1199.00

Table 5.1: Features of the average representative cycle and its difference with the group average.

Initially, as expected, the acceleration-related features (APA, ANA, RPA, and RNA) of the cycles reached the highest values in aggressive cycles (Table 5.2) and the lowest in mild cycles (Table 5.3). When it comes to speed-related features (mean driving speed and maximum speed), it is also possible to see a similar increase. Additionally, the LS MTs presented a lower idle percentage (appendices 1-3) in comparison to other clusters (for all the behaviours).

Regarding the driving features of the aggressive and mild cycles, it is possible to state that these final cycles also fulfil the requirements (low acceleration/high speed, and high acceleration/low speed).

Aggressive		Mean drv. speed (km/h)	APA (m/s2)	Duration (s)
LS	Rep. cycle	23.63	0.6255	263
	Difference	3.34%	1.84%	39.77%
MS	Rep. cycle	44.43	0.552	370
	Difference	5.04%	-3.99%	3.41%
HS	Rep. cycle	58.25	0.444	573
	Difference	1.69%	3.69%	-6.80%
	TOTAL	46.59	0.53	1,206.00

Table 5.2: Features of the aggressive representative cycle and its difference with the group average.

Mild		Mean drv. speed (km/h)	APA (m/s2)	Duration (s)
LS	Rep. cycle	16.35	0.3399	257
	Difference	-5.14%	5.56%	26.68%
MS	Rep. cycle	37.24	0.3734	373
	Difference	-5.21%	13.50%	-15.74%
HS	Rep. cycle	66.77	0.314	554
	Difference	12.90%	7.01%	-50.00%
	TOTAL	46.23	0.34	1,184.00

Table 5.3: Features of the mild representative cycle and its difference with the group average.

However, the highest discrepancies are reached in the duration of the profiles, where the LS MTs were higher than the group average, whereas the MS and HS MTs were considerably lower. It can be assumed that this is due to the proportions of the different profiles in the final cycle.

As explained in the previous chapter, to reach the highest representativeness, the final cycle should have a similar duration to the original cycles found in the data set (811 s). However, these cycles (initial data set) are, on average, composed of only two driving profiles, where the HS MTs represent the highest share of driving time. As a result, the sum of the representative MTs duration (1200 s) will be well over the third quartile of the initial cycles (1107 s). Additionally, if the duration is restricted (imposed as a requirement when selecting the representative MTs), the characteristic MT will lose representativeness, since it would not be the closest one to the centroid.

To address this problem, it would be necessary to perform an analysis with the objective of determining the ideal duration of the final cycle, reaching a balance between features and distance/duration representativeness as described in figure 5.1. Additionally, a higher number of MTs would be favourable since it may be possible to select a MT closer to the group centroid.

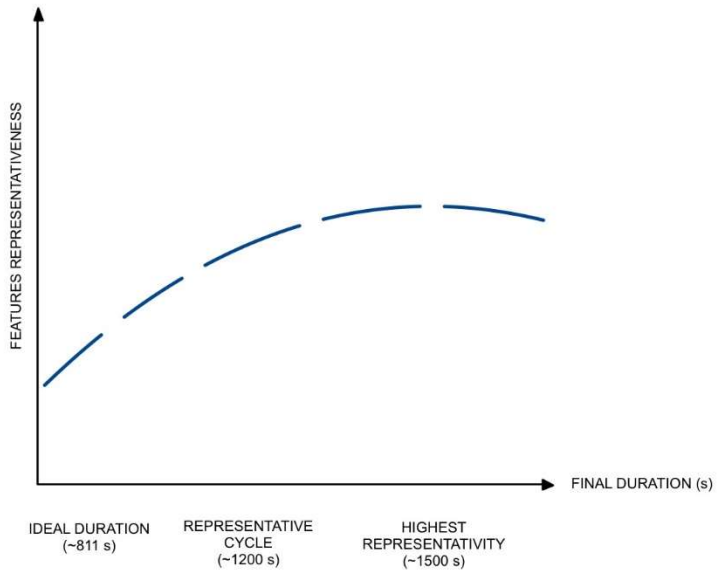

Figure 5.1: Illustrative (non-real) image of the relationship between ideal duration and representativeness.

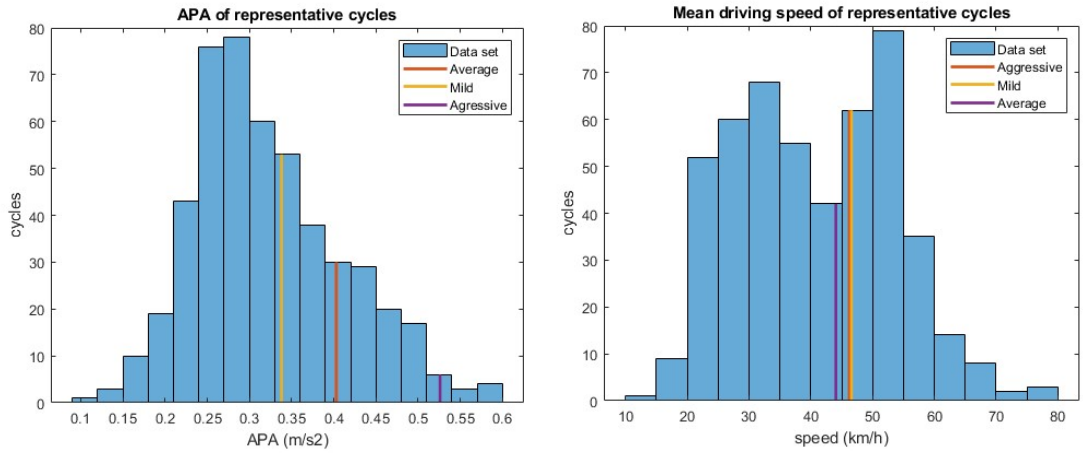

Figure 5.2: Position of representative cycles in the original data set.

As can be seen (figure 5.2), the representative cycles positions (regarding mean driving speed) are close to the data set mean. Concerning APA, the cycles are well over the data set average. This may be due to the duration restrictions since the total driving time (especially idle and cruising) are reduced, increasing the share of time accelerating/decelerating resulting in a higher APA.

The next table describes the final cycles when the duration is not restricted. It is highlightable a higher duration (approximately twice the ideal duration, 1,547 s). However, it can be seen that the results concerning driving features are closer to the group average (better results in comparison to the previous cycles).

Avg. Non-restricted		Mean drv speed (km/h)	APA (m/s2)	Duration (s)
LS	Rep. cycle	19.16	0.4699	336
	Difference	2.82%	2.96%	0.30%
MS	Rep. cycle	37.62	0.398	423
	Difference	-10.10%	-7.79%	1.91%
HS	Rep. cycle	58.42	0.351	788
	Difference	0.56%	-0.85%	2.83%
	TOTAL	43.10	0.39	1,547.00

Table 5.4: Features of the average representative cycle when the duration is not limited.

To illustrate the results when the final duration is restricted in order to obtain a similar duration to the one presented in the statistical analysis, table 5.5 is shown. The final duration is acceptable (786 s), however, it is also remarkable higher differences with the data average features (according to the described in figure 5.1).

Avg. Restricted		Mean drv speed (km/h)	APA (m/s2)	Duration (s)
LS	Rep. cycle	18.29	0.336	174
	Difference	-1.80%	-35.71%	3.74%
MS	Rep. cycle	43.51	0.334	258
	Difference	4.80%	-28.44%	-60.81%
HS	Rep. cycle	67.65	0.361	354
	Difference	14.13%	1.94%	-116.29%
	TOTAL	49.00	0.35	786.00

Table 5.5: Features of the average representative cycle when the duration is limited.

As can be seen (figure 5.3) the durations of the representative cycles are not accurate (when the distance is not limited). The restricted cycle (table 5.5) is represented by a purple line, found close to the data set average. However, the representative cycle and the non-restricted cycles (tables 5.1 and 5.4 respectively) are overly long (red and yellow).

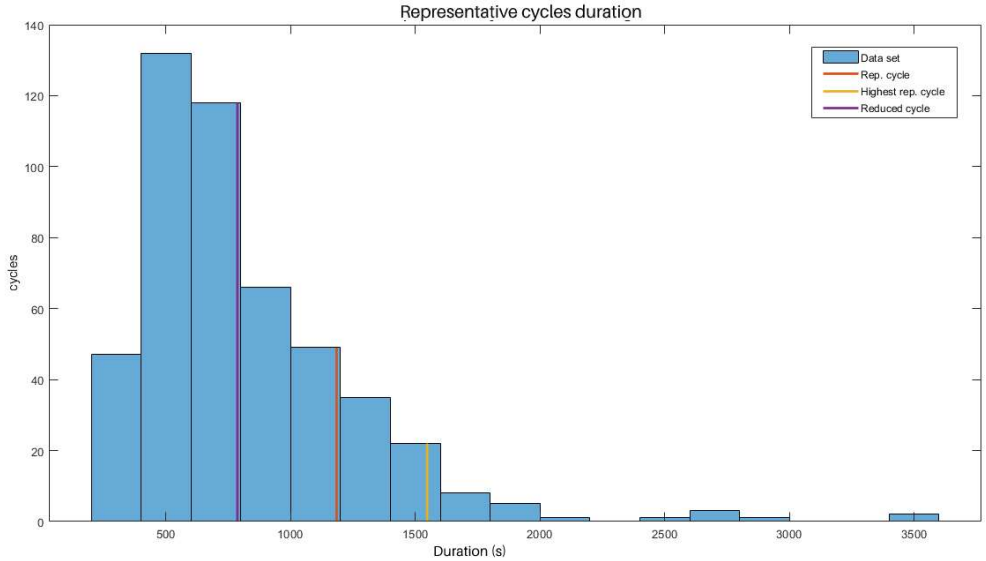

Figure 5.3: Position of representative cycles in the original data set: Duration

Finally, after the cycles presented, it can be inferred that the driving behaviour may be influenced by the distance/duration since longer distances usually relate to longer cruising times, reducing the speed variations. Hence, mild MTs will result in longer cycles whereas aggressive MTs may be related to shorter distances.

COMPARISON WITH STANDARDIZED CYCLES

It is possible to see that the variables exposed in the next table are similar to the ones found on the WLTC and NEDC. It is important to note that this cycle cannot be directly compared to any of the mentioned regulated cycles since they have different structures.

The WLTC class 3 is composed of four driving conditions: low, medium, high, and extra high, whereas class 2 is composed of a low, medium, and high where motorway speeds are not considered.

	NEDC	WLTC	Avg. Rep. Cycle
Duration (s)	1,180	1,800	1,184
Distance (km)	10.97	23.27	13.58
Idle (%)	0.25	0.13	0.0941
Max speed (km/h)	120	131.3	108.11
Mean driving speed (km/h)	34	47	44.06
APA (m/s2)	0.5	0.39	0.4

Table 5.6: Comparison between NEDC WLTP and the average representative cycle.

It can be seen that the distances and accelerations are similar to the standardized cycles, which could denote that, in the case of limiting the cycle duration (to ~811 s), the final cycle may be significantly shorter than the other cycles.

After a thorough research, the final cycle can be compared to the Artemis rural cycle (table 5.7), where the maximum speed and duration are lower and the mean speed higher than the referenced cycles.

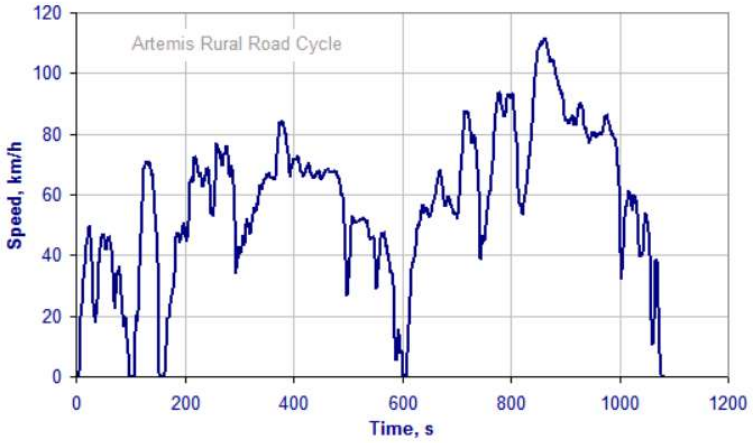

Figure 5.3: Artemis rural road cycle.

Duration (s)	1082
Mean driving speed (km/h)	58.34
Max speed (km/h)	111.09
APA (m/s2)	0.359
RPA (m/s2)	0.153

Table 5.7: Driving features of Artemis Rural road cycle

From this, it can be said that road infrastructure plays an important role in the speeds and distances reached. This can be explained by the sizes of the zones compared (Tenerife metropolitan area and other European cities).

CONCLUSION

This study was aimed to aid in the development of a methodology for obtaining representative driving cycles from a group measured in the Tenerife metropolitan area through machine learning algorithms. It was also examined the impact of various external variables in the driving features, such as location, weather and congestion. Additionally, through statistical analysis, it was possible to address the representativeness of the data set and the most characteristic variables.

The methodology proposed consisted of a selection of representative segments of cycles (Microtrips) through clustering algorithms, followed by the reduction of their dimensionality. This research clearly illustrates that this methodology results in cycles with a high degree of representativeness for the data set, but also raises questions related to appropriate cycle durations. Due to the morphology of the initial cycles, it was not possible to obtain an appropriate cycle duration without reducing the feature representativeness of the data set since the average cycle duration was shorter than the sum of the duration of characteristic Microtrips.

Through the statistical analysis, it was found that the average duration of the cycles was shorter than the standardized international and European cycles. This was mainly attributed to the restricted area that was studied, resulting in shorter distances and cycle durations. The data set obtained also does not represent a realistic transportation situation, where more local cycles in Santa Cruz centre are required.

Regarding driving behaviours, the data set was categorized into groups with similar acceleration-related variables. However, it was observed certain undesired relationships between the acceleration and distance variables that may have affected the aforementioned categorization.

To better understand the implications of these results, future studies should address the selection of an appropriate driving duration. Moreover, it is recommended to collect a higher number of cycles to improve the quality of the final cycle. Finally, in order to obtain a higher degree of representativeness for the studied zone, it is recommended to address the data set according to previous statistical studies to obtain a realistic proportion of locations and driving profiles.

In conclusion, based on the results of this study, a new methodology can be introduced with the aim of reducing the driving distance and cycle duration, maintaining the proportions of the different profiles within it.

This research has also proved the viability of the t-SNE over the PCA as a dimensionality reduction algorithm for the studied data set. Additionally, in accordance with the referenced studies, for these cycles, k-means was the algorithm that presented a higher quality of clusters, assessed through performance metrics.

REFERENCES

[1] EEA European Environment Agency, “Air quality in Europe — 2020 report”, [Online] Available https://www.eea.europa.eu//publications/air-quality-in-europe-2020-report [accessed 14 March 2021]

[2] Environmental Protection Agency of the United States, “Fuel Economy”, [Online] Available https://www.fueleconomy.gov/feg/best-worst.shtml [Accessed 27 March 2021]

[3] Andreas Braun, Wolfgang Rid, “The influence of driving patterns on energy consumption in the electric car driving and the role of regenerative braking”, in *19th EURO Working Group on Transportation Meeting, EWGT2016, 5-7 September 2016*, pp. 175-181.

[4] T.J. Barlow, S. Latham. “Appendix B: Definition of art.kinema parameters” in *A reference book of driving cycles for use in the measurement of road vehicle emissions*, 2009, pp. 17-19.

[5] S. Tsiakmakis, G. Fontaras, C. Cubito, J. Pavlovic “From NEDC to WLTP: effect on the type-approval CO2 emissions of light-duty vehicles”. 2017

[6] Monica Tutuianu , Alessandro Marotta. “Development of a World-wide Worldwide harmonized Light duty driving Test Cycle (WLTC)”. 2013

[7] Xinglong Liu, Fuquan Zhao. ”From NEDC to WLTP: Effect on the Energy Consumption, NEV Credits, and Subsidies Policies of PHEV in the Chinese Market”. 2020

[8] A. Fotouhi, M. Montazeri-Gh. “Tehran driving cycle development using the k-means clustering method”. 2012.

[9] D. Förster, R.B. Inderka. “Data-Driven Identification of Characteristics Real-Driving Cycles Based on k-MeansClustering and Mixed-Integer Optimization”. 2019

[10] J. Liu, X. Wang, A. Khattak “Customizing driving cycles to support vehicle purchase and use decisions: Fuel economy estimation for alternative fuel vehicle users”. 2016

[11] J. Huertas, L. Quirama, M. Giraldo, J. Díaz, “Comparison of driving cycles obtained by the Micro-trips, Markov chains and MWD-CP methods”.

[12] A. Kabra, “Clustering of Driver Data based on Driving Patterns”. Blekinge Institute of Technology Karlskrona, Sweden. 2019

[13] D. W. Scott, “Histogram”. Department of Statistics, Rice University, Houston, TX. 2008

[14] J. C. Watkins “Organizing and Producing Data” in *An Introduction to the Science of Statistics: From Theory to Implementation.* pp. 3-72.

[15] MathWorks, “Centro de ayuda” [online] Available https://es.mathworks.com/help [Accessed on 9 April 2021]

[16] J. Frost. ”How to Identify the Distribution of Your Data” [online] Available https://statisticsbyjim.com/hypothesis-testing/identify-distribution-data/ [Accessed on 10 April 2021]

[17] A. Quesada, R. Lopez, “3 methods to treat outliers in machine learning” [online] Available https://www.neuraldesigner.com/blog/3_methods_to_deal_with_outliers#UnivariateMethod [Accessed on 10 April 2021]

[18] D. Cosineau, S. Chartier. “Outliers detection and treatment: a review”, Université de Montréal, University of Ottawa, Canada. 2010.

[19]M. R. Elliot, N. Stettler, “Using a mixture model for multiple imputation in the presence of outliers: the ‘Healthy for life’ project”, University of Michigan, Ann Arbor, US. 2007

[20] I.T. Jolliffe, “Principal component analysis”, Aberdeen, UK, Springer, 2002. pp. 1-59

[21] F. Zhang, F. Guo and H. Huang, "A research on driving cycle for electric cars in beijing," 2016 Chinese Control and Decision Conference (CCDC), Yinchuan, 2016, pp. 4450-4455.

[22] Van der Maaten, L.J.P.; Hinton, G.E. "Visualizing Data Using t-SNE", Journal of Machine Learning Research. 9: 2579–2605. 2009

[23] Y. Cao, L. Wang, “Automatic Selection of t-SNE Perplexity”, ICML 2017 AutoML Workshop, RBC Research Institute, 2017

[24] F. Nielsen, “Hierarchical clustering” in *Introduction to HCP with MPI for data science,* Springer, ch 9, pp. 221-239

[25] Leonard Kaufman; Peter J. Rousseeuw. *Finding groups in data : An introduction to cluster analysis.* Hoboken, NJ: Wiley-Interscience. p. 87. 1990

[26] Y. Zhang, D. Li, “Cluster Analysis by Variance Ratio Criterion and Firefly Algorithm”, Southeast University, Nanjing, China pp. 689-697, 2013.

[27] M. Halkidi, Y. Batistakis, M. Vazirgiannis, “On Clustering Validation Techniques”, Athens University of Economics & Business, Greece, 2001

[28] Davies, David L. Bouldin, Donald W., "A Cluster Separation Measure". *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 1979.

[29] J. E. Meseguer, C. T. Calafate, J. C. Cano and P. Manzoni, "Assessing the impact of driving behavior on instantaneous fuel consumption," 2015 12th Annual IEEE Consumer Communications and Networking Conference (CCNC), 2015, pp. 443-448, doi: 10.1109/CCNC.2015.7158016.

[30] R. Liessner, M. Dietermann, "Derivation of Real-World Driving Cycles Corresponding to Traffic Situation and Driving Style on the Basis of Markov Models and Cluster Analyses” Daimler AG, Germany.

[31] D. Förster, R. B. Inderka and F. Gauterin, "Data-Driven Identification of Characteristic Real-Driving Cycles Based on k-Means Clustering and Mixed-Integer Optimization," in *IEEE Transactions on Vehicular Technology*, vol. 69, no. 3, pp. 2398-2410, March 2020, doi: 10.1109/TVT.2019.2963272.

[32] Cabildo de Tenerife, “Análisis de la demanda de movilidad” in *Plan Territorial Especial de Ordenación del Transporte de Tenerife.*

[33] Tomtom. 2019. [online] Available at: <https://www.tomtom.com/en_gb/traffic-index/santa-cruz-de-tenerife-traffic/> [Accessed 14 June 2021].

APPENDICES

	AVERAGE CYCLE								
	LS								
	Mean drv speed	Max speed	APA	ANA	RPA	RNA	Duration	Distance	Idle
Cluster	18.62	35.56	0.456	-0.429	0.2013	-0.1666	167.49	0.7546	26.19
Rep. cycle	19,355	40	0.47715	-0.4295	0.212	-0.1815	274	1,077	28.81
Difference	3.80%	11.10%	4.43%	0.12%	5.05%	8.21%	38.87%	29.94%	2.62%
	MS								
	Mean drv speed	Max speed	APA	ANA	RPA	RNA	Duration	Distance	Idle
Cluster	41.42	90.51	0.429	-0.4606	0.218	-0.1998	414.9	4.39	6.44
Rep. cycle	42.78	89.26	0.449	-0.4962	0.2388	-0.222	366	4.31	1.36
Difference	3.18%	-1.40%	4.45%	7.17%	8.71%	10.00%	-13.36%	-1.86%	-5.08%
	HS								
	Mean drv speed	Max speed	APA	ANA	RPA	RNA	Duration	Distance	Idle
Cluster	58.09	108.84	0.354	-0.3924	0.1678	-0.1581	765.67	11.82	2.84
Rep. cycle	57	108.11	0.336	-0.441	0.1711	-0.16	559	8.19	5.18
Difference	-1.91%	-0.68%	-5.36%	11.02%	1.93%	1.19%	-36.97%	-44.32%	2.34%
TOTAL	44.06	108.11	0.403	-0.455	0.201	-0.184	1199	13.58	9.41

Appendix 1: Average Cycle. Partially restricted.

	MILD CYCLE								
	LS								
	Mean drv speed	Max speed	APA	ANA	RPA	RNA	Duration	Distance	Idle
Cluster	17.19	30.54	0.321	-0.326	0.127	-0.106	188.43	1.06	24.98
Rep. cycle	16.35	32.23	0.3399	-0.33135	0.148	-0.1215	257	0.97	22,575
Difference	-5.14%	5.24%	5.56%	1.61%	14.19%	12.76%	26.68%	-9.28%	-2.41%
	MS								
	Mean drv speed	Max speed	APA	ANA	RPA	RNA	Duration	Distance	Idle
Cluster	39.18	74.89	0.323	-0.348	0.155	-0.144	431.7	4.37	6.84
Rep. cycle	37.24	73.31	0.3734	-0.363	0.167	-0.154	373	3.78	3.75
Difference	-5.21%	-2.16%	13.50%	4.13%	7.19%	6.49%	-15.74%	-15.61%	-3.09%
	HS								
	Mean drv speed	Max speed	APA	ANA	RPA		Duration	Distance	Idle
Cluster	58.16	105.54	0.292	-0.318	0.133	-0.127	831	13.11	2.94
Rep. cycle	66.77	111.32	0.314	-0.3376	0.14	-0.134	554	9.78	5.5
Difference	12.90%	5.19%	7.01%	5.81%	5.00%	5.22%	-50.00%	-34.05%	2.56%
TOTAL	46.23	111.32	0.338	-0.344	0.15	-0.137	1184	14.53	8.87

Appendix 2: Mild Cycle. Partially restricted.

	AGGRESSIVE CYCLE								
	LS								
	Mean drv speed	Max speed	APA	ANA	RPA	RNA	Duration	Distance	Idle
Cluster	22.84	44.72	0.614	-0.6235	0.321	-0.264	158.41	0.788	26.68
Rep. cycle	23.63	47.37	0.6255	-0.7005	0.3465	-0.2955	263	12,883	26.65
Difference	3.34%	5.59%	1.84%	10.99%	7.36%	10.66%	39.77%	38.83%	-0.03%
	MS								
	Mean drv speed	Max speed	APA	ANA	RPA	RNA	Duration	Distance	Idle
Cluster	42.19	100.11	0.574	-0.564	0.288	-0.257	357.4	4.08	5.37
Rep. cycle	44.43	89.32	0.552	-0.57	0.279	-0.255	370	4.03	10.27
Difference	5.04%	-12.08%	-3.99%	1.05%	-3.23%	-0.78%	3.41%	-1.24%	4.90%
	HS								
	Mean drv speed	Max speed	APA	ANA	RPA	RNA	Duration	Distance	Idle
Cluster	58.25	116.58	0.444	-0.492	0.211	-0.195	611.96	9.1	4.12
Rep. cycle	59.25	118.45	0.461	-0.489	0.211	-0.197	573	8.87	4.19
Difference	1.69%	1.58%	3.69%	-0.61%	0.00%	1.02%	-6.80%	-2.59%	0.07%
TOTAL	46.59	118.45	0.526	-0.562	0.263	-0.237	1206	14.19	11.18

Appendix 3: Aggressive Cycle. Partially restricted.

		AVG WITH NO DURATION/DISTANCE RESTRICTIONS							
		LS							
	Mean drv speed	Max speed	APA	ANA	RPA	RNA	Duration	Distance	Idle
Cluster	18.62	35.56	0.456	-0.429	0.2013	-0.1666	335	1.5	26.19
Rep. cycle	19.16	37.22	0.4699	-0.43005	0.213555	-0.1755	336	1,337	30.85
Difference	2.82%	4.46%	2.96%	0.24%	5.74%	5.07%	0.30%	-12.19%	4.66%
		MS							
	Mean drv speed	Max speed	APA	ANA	RPA	RNA	Duration	Distance	Idle
Cluster	41.42	90.51	0.429	-0.4606	0.218	-0.1998	414.9	4.39	6.44
Rep. cycle	37.62	89.92	0.398	-0.514	0.221	-0.204	423	4.38	1.18
Difference	-10.10%	-0.66%	-7.79%	10.39%	1.36%	2.06%	1.91%	-0.23%	-5.26%
		HS							
	Mean drv speed	Max speed	APA	ANA	RPA	RNA	Duration	Distance	Idle
Cluster	58.09	108.84	0.354	-0.3924	0.1678	-0.1581	765.67	11.82	2.84
Rep. cycle	58.42	112.7	0.351	-0.371	0.145	-0.137	788	12.46	3.3
Difference	0.56%	3.43%	-0.85%	-5.77%	-15.72%	-15.40%	2.83%	5.14%	0.46%
TOTAL	43.1	112.7	0.393	-0.428	0.184	-0.166	1547	18.18	8.95

Appendix 4: Average Cycle. No restricted.

		AVG WITH LIMITED DURATION/DISTANCE							
					LS				
	Mean drv speed	Max speed	APA	ANA	RPA	RNA	Duration	Distance	Idle
Cluster	18.62	35.56	0.456	-0.429	0.2013	-0.1666	167.5	0.75	26.19
Rep. cycle	18.29	37.4	0.336	-0.401	0.2198	-0.1889	174	0.627	17.24
Difference	-1.80%	4.92%	-35.71%	-6.98%	8.42%	11.81%	3.74%	-19.62%	-8.95%
					MS				
	Mean drv speed	Max speed	APA	ANA	RPA	RNA	Duration	Distance	Idle
Cluster	41.42	90.51	0.429	-0.4606	0.218	-0.1998	414.9	4.39	6.44
Rep. cycle	43.51	87.36	0.334	-0.471	0.17	-0.158	258	2.96	5.42
Difference	4.80%	-3.61%	-28.44%	2.21%	-28.24%	-26.46%	-60.81%	-48.31%	-1.02%
					HS				
	Mean drv speed	Max speed	APA	ANA	RPA	RNA	Duration	Distance	Idle
Cluster	58.09	108.84	0.354	-0.3924	0.1678	-0.1581	765.67	11.82	2.84
Rep. cycle	67.65	103.84	0.361	-0.398	0.156	-0.148	354	6.2	6.78
Difference	14.13%	-4.82%	1.94%	1.41%	-7.56%	-6.82%	-116.29%	-90.65%	3.94%
TOTAL	49	103.84	0.347	-0.421	0.175	-0.16	786	9.79	8.76

Appendix 5: Average Cycle. Restricted.