# Benchmarking Driving Efficiency using Data Science Techniques applied on Large-Scale Smartphone Data (PhD Summary)

The main **objective** of this PhD is to provide a methodological approach for **driving safety efficiency benchmarking** on a trip and driver basis using data science techniques. It also investigates the way to achieve this by defining a safety efficiency index based on travel and driving behaviour metrics collected from smartphone devices. The driving characteristics of each emerging efficiency group is discussed and the main driving patterns are identified. One of the most significant DEA's weaknesses, i.e. the significant time required for processing large-scale data, is overcome by employing computational geometry techniques. Furthermore, the present doctoral research proposes a methodological framework for identifying the least efficient trips in a database and for estimating the efficient level of metrics that each non-efficient trip should reach to become efficient. Finally, this dissertation's objective is to study the temporal evolution of driving efficiency and identify the main driving patterns and profiles of the driver groups formed.

Literature review revealed that it is significant to study the potential of **benchmarking** driving safety efficiency using **microscopic** driving data collected from **smartphone** devices. This doctoral research attempts to address this certain issue by proposing a methodological framework based on data science techniques for evaluating driving characteristics. An improved DEA model is applied to deal with the analysis of large-scale smartphone data collected while driving. The model developed is incorporating several driving behaviour metrics allowing for the **multi-criteria** analysis of driving efficiency.

The general **methodological** framework applied is illustrated in Figure 1.1. There are two data sources where data are derived from a) a database of drivers who participated in a naturalistic driving experiment in which data where recorded using the **smartphone** device of each participant and b) the **questionnaire** administered to a proportion of the participants. After data are collected, the factors representing driving efficiency in terms of safety are specified based on **literature review** conducted. After it is examined that a) **adequate** data is collected from each participant taken into consideration in this research and b) the driving metrics and distance recorded are proportionally increased and their ratio does not significantly change while monitored kilometres are accumulated, these factors are used as inputs and outputs for the DEA models developed. Consequently, **trip** and **driver efficiency analysis** is implemented per road type following the detailed description given below. The results obtained from the trip efficiency analysis are exploited mainly to reduce processing time for the driver efficiency analysis where the **evolution** of driving efficiency through time is investigated and secondarily to assess the practicability of providing a methodology for less efficient trip identification. The results of driver and driving efficiency evolution investigations are combined to perform cluster analysis on a driver level. For each **driving cluster** that results from this procedure, the typical driving characteristics of the drivers that belong to it are examined and presented.

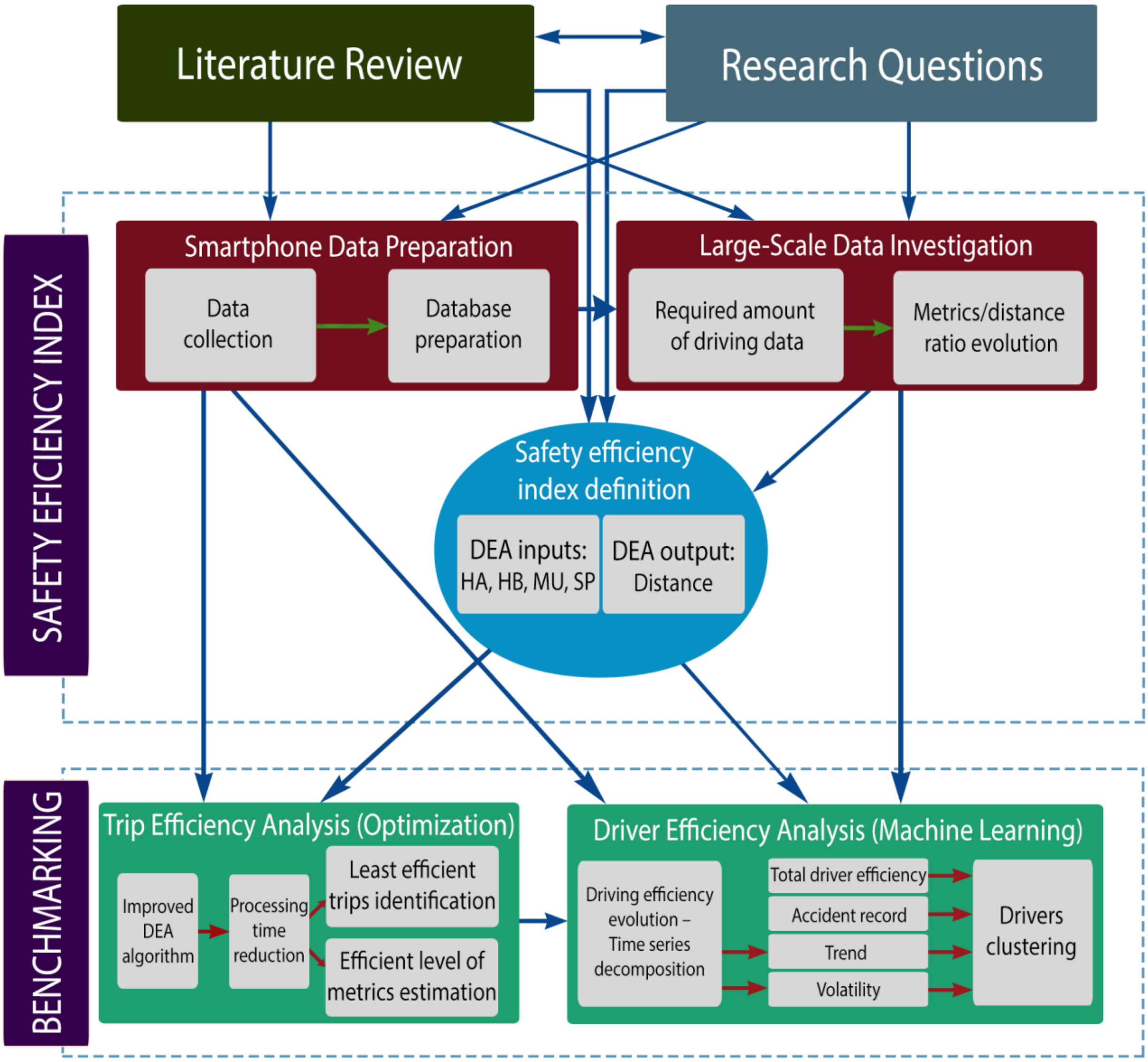
Literature Review
Research Questions
SAFETY EFICIENCY INDEX
Smartphone Data Preparation
Data collection
Database preparation
Large-Scale Data Investigation
Required amount of driving data
Metrics/distance ratio evolution
Safety efficiency index definition
DEA inputs: HA, HB, MU, SP
DEA output: Distance
BENCHMARKING
Trip Efficiency Analysis (Optimization)
Improved DEA algorithm
Processing time reduction
Least efficient trips identification
Efficient level of metrics estimation
Driver Efficiency Analysis (Machine Learning)
Driving efficiency evolution – Time series decomposition
Total driver efficiency
Accident record
Trend
Volatility
Drivers clustering

*Figure 1.1: Graphical representation of the general methodological framework of the present doctoral dissertation.*

To achieve the objectives set above by the present PhD dissertation, the structure of this research consists of six separate methodological steps presented below (Figure 1.1):

Exhaustive **literature review** takes place as a first step, covering an **overview** of road safety and accidents and the fields driving behaviour and risk, driving characteristics, driving efficiency parameters (distraction, aggressiveness, etc.), naturalistic driving experiments, data envelopment analysis methodology, potential improvements on large-scale data analysis and its applications on transport engineering and driving efficiency. The conclusions drawn from the review and the knowledge gap arising assists in setting the research objectives and research hypotheses and generally in setting up the problem.

Based on the **literature review** conducted, it is considered necessary to study driving behaviour on a greater extent and shed more light on the evaluation of driving safety behaviour and the factors influencing it. As we move forward, **UBI** aims to assign insurance premiums to the respective **accident risk** of each individual driver based on travel and driving behavioural characteristics. Therefore, drivers should reduce their annual mileage and improve their driving behaviour. This is because per-mile risk is an unspecified factor that fluctuates over time and therefore although mileage might be reducing, total crash risk can still be increasing. In support of the above, even if per-mile crash risk remains constant and annual mileage is known, total individual crash risk cannot be estimated since it depends on behavioural characteristics that are not currently recorded and considered in UBI. To achieve this, information about driving traits e.g. number of harsh braking and acceleration events, time of driving over the speed limits, road type etc. should be included in driver's evaluation. In other words, risk factor is risk's increase rate that indicates how total individual risk is increased as mileage increases. As a result, it is essential to develop a model that incorporates both distance travelled and the rest of the behavioural characteristics in order to evaluate driving risk. By developing DEA models that take into account these two categories of characteristics, this study aims to examine the applicability of such models.

According to past research, **naturalistic driving experiments** are considered more appropriate for driving behaviour evaluation because behaviour is recorded under normal driving conditions and without any influence from external parameters. Regarding the main drawback of naturalistic driving experiments, driving under normal conditions will be recorded and no bias will appear if drivers are monitored for an appropriate amount of time. On the other hand, it is very important to determine the amount of data required to obtain a complete picture for each driver, where the rate of those metrics described above per km travelled converges to a stable value.

It can be said from the above that the most **significant** human factors recorded by smartphone devices and were found to affect driving risk are mobile phone distraction, speed limit exceedance and the number of harsh braking and acceleration events occurred while driving. It can also be inferred that there are numerous researches that focus on driving behaviour evaluation and mainly on determining the correlation between driving behaviour metrics (speed limit exceedance, number of harsh acceleration/ braking events, mobile phone distraction etc.) either together or separately and accident probability. To the best of the author's knowledge, this doctoral research is the first effort made to estimate and assign a relative **safety efficiency index** to each driver of a sample by exploiting distance travelled and several driving behaviour metrics that result from microscopic driving behaviour data recorded from smartphone devices.

It can be concluded from all the above that it is significant to **benchmark** driving safety efficiency using microscopic driving data collected from smartphone devices. It is showed that DEA has never been used before in driving behaviour research and that driver's efficiency has been studied in a great extent but never by making use of DEA techniques. Therefore, there should be an attempt to address this certain issue by proposing a methodological framework based on **data science** techniques for evaluating driving characteristics. The model that will be developed should incorporate several driving behaviour metrics allowing for the **multi-criteria** analysis of driving efficiency. It is also found important to address the problem of the large computation time required for a DEA algorithm and methodologically speaking, it is momentous to test the effectiveness of the implementation of a DEA and convex-hull algorithm combination in a multiple inputs and outputs settings for large-scale driving data.

The second step of the methodology is data **collection** and **preparation**, which includes a description of the survey design and questionnaire administration and extended description of how the OSeven platform works including the recording, collection, storage, evaluation and visualization process of driving behaviour data using smartphone applications and advanced **machine learning** (ML) algorithms. This innovative large-scale data collection and analysis methodology applied, presents new challenges by gathering large quantities of data for analysis during this research. Furthermore, database is further processed and prepared to be imported in the final data analysis conducted afterwards. This preparation is made using Python programming language, which is suitable for large-scale data analysis.

All aforementioned indicators, which are received directly from the OSeven system, are analysed and filtered to retain only those indicators that will be used as inputs and outputs herein for the DEA problem. Data filtering and DEA improvement algorithms are performed in Python programming language and several scripts are written for this reason. A significant amount of data is recorded using the smartphone application developed by OSeven Telematics. Data used in this research are anonymized before provided by OSeven so that driving behaviour of each participant cannot be connected with any personal information. This is a data exploitation approach that is user-agnostic and therefore less user intrusive. It should also be highlighted at this point that the approach followed in this study aims to **identify driving behaviours** and **patterns** and the factors influencing them and not to explain the causality between behaviour and other factors such as age, gender, occupation etc. or describe the distribution of the driving sample collected. The advantage of such an approach is that behaviours can be studied even in cases where demographic data of a driving sample are not available or cannot be collected.

For the purposes of this doctoral research, a sample of **171 drivers** participated in the designed experiment that endured 7-months and a large database of **49,722 trips** is collected from the database of OSeven. For each individual part of the analysis conducted herein, a part of this database is exploited because of the different requirements of each analysis. The selection made is presented in Table 4.1.

*Table 4.1: Driving sample used in each part of the research*

| | **Sampling time investigation** | **Trip efficiency analysis** | **Driver efficiency analysis** | | | |
|---|---|---|---|---|---|---|
| | | | **data_sample_1** | | **data_sample_2** | |
| | | | **Urban** | **Rural** | **Urban** | **Rural** |
| **Number of drivers** | 171 | 88 | 100 | 100 | 43 | 39 |
| **Number of trips** | 49,722 | 10,088 | 23,000 | 15,000 | 9,890 | 5,850 |

An extended presentation of the **statistical characteristics** of the driving sample used in each of the three types of data analysis are also presented to acquire a clear picture of the sample derived. The whole sample of 171 drivers participated in the designed experiment is used and a large database of 49,722 trips is created. All drivers chosen to be included in this part of the analysis should had driven at least for 10 hours and 40 trips that approximately equals the typical monthly number of trips for a driver assuming that each driver drives 2 trips of 15 minutes a day for 5 working days a week. As for the trip efficiency analysis, a part of the sample of eighty-eight (88) drivers participated in the designed experiment that took place between 28/09/2016 and 05/12/2016 and a large database of 10,088 trips is created.

For the purposes of the driver efficiency analysis, driving data were selected from the initial database of **171 drivers** based on some driver criteria. The first criterion chosen was that all drivers should have travelled at least 50 more trips than the number of trips required so that the total distance per road type is at least equal to the minimum distance found in the previous step of the sample quantification. This criterion is set to ensure that a) inputs are **proportionally** increased to outputs and therefore it is valid to develop a DEA model in each time step of the moving window and in total and that, b) the number of the time series **observations** is satisfying. Of course, this procedure of drivers' selection aims to result to the maximum number of drivers possible.

On the top of that, all drivers should have **positive** mileage on all three types of road network. The third criterion was that drivers with a **zero** sum of input attributes (i.e. harsh acceleration, braking, speed limit violation, mobile phone usage are all equal to zero) should be eliminated from the sample, which is a limitation of DEA. The business equivalent of a zero input could be a factory that is producing a product without making use of any material and/or workforce, which practically cannot occur. This procedure resulted to 100 drivers in urban and rural road type who fulfilled these criteria and were kept for the analysis conducted whereas the rest of the drivers were eliminated from this study. Drivers' elimination resulted to only 18 drivers in highways, which was considered a low number of participants for the analysis to be conducted. The total number of trips that took place by each of the drivers chosen was 230 for urban and 150 for rural roads constructing thus a large database of 23,000 trips in urban and 15,000 in rural. From those drivers, 43 urban and 39 rural drivers have answered the questionnaire administered. Finally, the questionnaire is briefly discussed and its main questions are provided.

The investigation of the **adequate amount** of data to be included in the analysis and the evolution of the metrics/ distance ratio takes place as a next step. This step is essential in order to specify the exact amount of data that should be used in the analysis and is neither deficient nor excessive. A deficient amount of data would lead this research to uncertain or unreasonable results while an excessive amount of data would significantly **increase** required processing **time**.

As for the urban road, HA appears to be the most **critical metric** for the determination of the required sampling distance. It can be noticed that the maximum value of the adequate distance for the relevant metric to converge in the table appears for the percentile range 75-100% of HA. The maximum median distance value is found to be **519km**, which is approximately equal to 75 trips in urban road. Initially, the average distance per trip and consequently the number of required trips that each driver should perform to reach the distance of 519km is calculated. The median value of all users for this variable is estimated to be around 75. This is the length of the moving window used in the driver efficiency analysis to create the time series of driving efficiency in rural road type. The median value is preferred instead of the average value for the same reasons stated above for the determination of the required sampling distance.

As for the rural road, HB and MU appear to be the most **critical metrics** for the determination of the required sampling distance. It can be noticed that the maximum value of the adequate distance to converge for MU in the table appears for the percentile range 0-25% of HA. The maximum median distance value is found to be **579km**, which is approximately equal to 81 trips in rural road. As in urban road analysis, the average distance per trip and consequently the number of required rural trips that each driver should perform to reach the distance of 579km is calculated. The median value of all users for this variable is estimated to be around 81. This is the length of the moving window used in the driver efficiency analysis to create the time series of driving efficiency in rural road type. The median value is preferred instead of the average value for the same reasons stated above for the determination of the required sampling distance.

Finally for highways, HA and HB appear to be the most **critical metrics** for the determination of the required sampling distance. It can be noticed that the maximum value of the adequate distance to converge for MU in the table appears for the percentile range 50-75% of HA. The maximum median distance value is found to be **611km**, which is approximately equal to 106 trips in highways. As in urban road analysis, the average distance per trip and consequently the number of required rural trips that each driver should perform to reach the distance of 611km is calculated. The median value of all users for this variable is estimated to be around 106. This the length of the moving window that should be used in the driver efficiency analysis to create the time series of driving efficiency in rural road type. Unfortunately, this value exceeds the number of trips (100) that are collected for the driver efficiency analysis in highways and therefore this analysis cannot be performed in the specific road type.

It is therefore concluded that the driving efficiency problem can be dealt as a constant returns-to-scale (**CRS**) DEA problem since the required sampling distance is defined so that the sum of all metrics (inputs) recorded for each driver changes **proportionally** to the sum of driving distance (output) in each moving window examined and in total. This step also defines the moving window time step and concludes that the highway road type cannot be included in the analysis because only a short number of participants has been recorded for more than the respective kilometres found.

Taking into account the literature review conducted, the data collected and all the peculiarities of the DEA technique, it is concluded that **safety efficiency index** may be defined using the number of harsh acceleration and braking events, the seconds of mobile usage and the seconds of driving over the speed limits as inputs and the distance travelled as output. This is the **key-step** connecting the “safety efficiency index estimation” and “benchmarking” part of this doctoral research. It constitutes a substantial step for moving forward with the DE analysis, determining the DEA inputs and outputs in such a way to i) be a scientifically sound formulation of the DEA technique and ii) represent driving safety efficiency and therefore the relative driving risk.

**Trip efficiency** analysis is conducted thereafter to determine the best performing technique among those tested and to develop a methodology for identifying the least efficient trips that exist in a certain trip database. Standard DEA, RBE DEA and convex hull DEA are tested and compared on the basis of required processing time. **Convex hull** algorithm combined with DEA outperforms the other two methodologies tested. This is a critical step that enables the reduction in required running time for all consequent steps engaged with DEA modelling. Furthermore, a convex hull DEA algorithm is implemented when both inputs and outputs are more than one. Lastly, a methodological approach is proposed for less efficient trip identification and efficient level of driving metrics estimation based on the safety efficiency index defined above.

**Driver efficiency** analysis is performed to examine the potential of clustering drivers and identify the main driving characteristics of each cluster arose. Based on the safety efficiency index defined in the fourth step, for each driver total driver efficiency for the total recorded period is estimated together with driver efficiency for the time window of each time step examined. The efficiency time-series created is analysed and results are exploited for driver clustering. All driving profiles emerging from each cluster are presented.

As mentioned above, the large-scale driving data were selected from the initial database of 171 drivers based on some criteria. The first criterion chosen was that all drivers should have travelled at least 50 more trips than the number of trips required so as the total distance per road type is securely **higher** than the **minimum distance** found in the previous step of the sample quantification. This procedure of drivers' selection also aims to result to the maximum number of drivers possible. On the top of that, all drivers should have positive mileage on all three types of road network. In addition to that, drivers with a zero sum of **input** attributes (i.e. harsh acceleration, braking, speed limit violation, mobile phone usage are all equal to zero) are eliminated from the sample because this is a DEA limitation. This procedure resulted to 100 drivers in urban and rural road type who met these requirements and were used in the analysis conducted whereas the rest of the drivers were eliminated from this study. Drivers' elimination resulted to only 18 drivers in highways, which was considered a very low number of participants for the analysis to be conducted. The total number of trips that took place by each of the drivers chosen was 230 for urban and 150 for rural roads constructing thus a large database of 23,000 trips in urban and 15,000 in rural. From those drivers, 43 urban and 39 rural drivers has answered the questionnaire administered.

For each of the data_sample_1 and data_sample_2, the median of the attributes of each class arising is shown in Table 5.10 where the models per urban and rural road type are presented based on the inputs that were used in each model. Class 1 drivers are referred to as **most efficient** drivers despite the fact that only drivers with unit efficiency lie on the efficiency frontier; class 2 and 3 drivers are referred to as **weakly efficient** and **non-efficient** drivers.

*Table 5.10: Driving characteristics of the efficiency groups per 100km and per road and sample type*

| Sample type | Road type | No of drivers | Driving characteristics | Efficiency classes | | |
|---|---|---|---|---|---|---|
| | | | | Class 1: 0 - 25 % percentile | Class 2: 25 - 75 % percentile | Class 3: 75 - 100 % percentile |
| data_sample_1 | Urban | 100 | efficiency | 0.22 | 0.36 | 0.61 |
| | | | ha | 21.49 | 11.82 | 8.82 |
| | | | hb | 9.64 | 5.31 | 3.68 |
| | | | mu | 316 | 205 | 141 |
| | | | sp | 1243 | 878 | 355 |
| | Rural | 100 | efficiency | 0.24 | 0.42 | 0.90 |
| | | | ha | 34.11 | 24.06 | 11.30 |
| | | | hb | 14.92 | 9.16 | 5.42 |
| | | | mu | 529 | 419 | 165 |
| | | | sp | 1564 | 1004 | 708 |
| data_sample_2 | Urban | 43 | efficiency | 0.21 | 0.38 | 1.00 |
| | | | ha | 39.26 | 21.71 | 9.98 |
| | | | hb | 16.38 | 8.07 | 4.19 |
| | | | mu | 751 | 553 | 100 |
| | | | sp | 1892 | 965 | 477 |
| | Rural | 39 | efficiency | 0.28 | 0.44 | 1.00 |
| | | | ha | 23.04 | 11.86 | 7.49 |
| | | | hb | 9.28 | 5.21 | 3.16 |
| | | | mu | 316 | 305 | 160 |
| | | | sp | 1423 | 939 | 378 |

As expected, for all road network and sample type models, the median of the attributes is **reducing** while shifting to a class of **higher** efficiency. The difference between classes 1 and 2 is found to be less significant for $mobile_{rural}$ and slightly less significant for $mobile_{urban}$ of both the data_sample_1 and data_sample_2. This result indicates that drivers of both road types (and especially rural road) have similar behaviour in terms of the **mobile** usage and therefore mobile usage is **not** a **critical** factor when measuring driving efficiency using DEA. In other words, the conclusion that can be drawn is that the overall driving safety profile of a less risky driver in urban and rural road **is not** considerably influenced by the driver's mobile usage. A possible explanation of this phenomenon is either the fact that drivers of all classes use the mobile phone approximately the same or DEA's sensitivity to outliers, which means that the model might sometimes be influenced by the extreme values of other inputs or outputs when estimating a DMU's efficiency e.g. low number of speeding or mobile usage seconds. In either case, mobile phone distraction should be examined separately. The main factors influencing shifting from one class to another are also identified and a methodology for estimating the efficient level of driving metrics that each driver should reach to become efficient is proposed. Total efficiency and volatility are also estimated in this step, which will be used in the clustering procedure.

The **evolution** of average driving **efficiency** over time will also be investigated by using different databases, accumulated over different timeframes from the beginning of recording time until the end of each timeframe. The time series that results is studied and decomposed in its main components, **stationarity** and **trend**. The average trend is observed to be approximately the same between the two road types of the data_sample_1 despite the fact that median trend is diverged. This indicates the existence of high outlier trend values in urban road and low outlier trend values in rural road that influence the average trend value. As for the stationarity of the time series, the number of differences required for a time series to become stationary reveal that there are no users in urban road whose driving behaviour is stationary. On the other hand, the relative number in rural roads is low for the data_sample_1 but significantly higher for the data_sample_2.

Using a k-means **machine learning** algorithm, drivers clustering is performed afterwards based on total driving efficiency, volatility, trend, stationarity of the time series arising as well as on the questionnaire data collected from the data_sample_2. The questions concerning the number of driving experience and the number of total accidents to date were the questionnaire data exploited in the clustering approach. These two questions were combined into one variable representing the total number of accidents per 10 years of driving and is presented in this form below. Driving **characteristics** of each cluster arose are analysed and conclusions drawn are presented. To prevent the results from being influenced by the outliers, all variables are normalized before used as inputs in the k-means clustering algorithm. The optimal number of clusters is determined using the elbow method.

*Table 5.19: Qualitative characteristics of the drivers' clusters*

| Sample type | Road type | Cluster | Trend (*10-3) | Volatility | Efficiency | Accidents/ 10 years of driving experience |
|---|---|---|---|---|---|---|
| data_sample_1 | Urban | **1 (typical)** | very low positive | medium - high | low | low - medium |
| | | **2 (unstable)** | medium positive | medium - high | medium | low |
| | | **3 (cautious)** | medium negative | low - medium | medium - high | low |
| | Rural | **1 (typical)** | low positive | medium | low | low - medium |
| | | **2 (unstable)** | high negative | high | medium - high | medium - high |
| | | **3 (cautious)** | high positive | medium - high | high | low |
| data_sample_2 | Urban | **1 (typical)** | very low positive | medium | low | low |
| | | **2 (unstable)** | low - medium | medium | low | high |
| | | **3 (cautious)** | medium negative | low | high | low |
| | Rural | **1 (typical)** | barely no trend | medium - high | low | low |
| | | **2 (unstable)** | low negative | medium | low | high |
| | | **3 (cautious)** | high positive | medium - high | high | low |

Clustering analysis performed resulted to three driving groups, which mainly represent the **average** drivers, the **unstable** drivers and the **cautious** drivers. The main common attribute between all clusters of cautious drivers is the high driving efficiency index and the low value of the accident per year value regardless of whether or not it was included as a factor in the cluster analysis. On the other hand, all clusters of the average drivers feature a high driving efficiency index and an insignificant low positive trend indicating a steadily poor driving behaviour. Finally, the unstable drivers of the second cluster present a medium to high volatility, which is found to be the only common characteristic between them. The rest of the clusters show similar characteristics in all attributes. The results of the clustering procedure are summarized in table 5.19.

Finally, prior **information** on driving **accident** data seems to **affect** only the form of the second cluster of the most **unstable** drivers, which incorporates drivers that are both **less** safety efficient and unstable. This is extremely promising for driving behaviour literature since it implies that it is feasible to study massive anonymous datasets for which no personal data are provided and produce equally significant and not biased outcomes.

This dissertation concludes that the methodological approach for the determination of the required driving data sampling distance depends on the scope of the research methodology that will be applied. In other words, the statistical principles of the methodological approach that will exploit the driving data collected determine the statistical rules that will be specified to estimate the amount of required data.

An equally important conclusion is that the adequate **amount** of driving **data** is decreased as the driving metrics (for all metrics except speeding) increase, at least for rural road and highways. This means that the more **aggressive** a driver becomes, the less monitoring is required to acquire a clear picture about his/ her driving patterns. Results also demonstrated that a **different** type of metric is critical for the determination of the required amount of data that should be recorded in each road type. Additionally, it appears that a different amount of data is necessary to be collected depending on the road type examined.

Results also indicate that the proposed DEA algorithm combined with **convex hull** is performing significantly better for **large-scale** data compared to other existing DEA algorithmic methodologies such as standard and RBE DEA methodologies. Another important contribution of this research is that it suggests a new approach for the assessment of the driving efficiency of a trip. The methodology to estimate a trip's efficiency index, identify its “peers”, and therefore, determine its efficient level of inputs and outputs is provided. Finally, the methodological steps for the identification of the least efficient trips in a database are provided, which would be a valuable finding for a driving recommendation system.

Another important highlight of the analysis conducted for each category is that considerable **differences** exist in driving characteristics between **inefficient** drivers and the classes of **weakly** efficient and **most** efficient drivers with the difference of the two latter to be less significant. On the other hand, mobile usage is not found to be a critical factor in safety efficiency assessment probably because DEA is providing a relative estimation of driving efficiency and at the same time, the difference in the seconds of mobile usage between different classes is not found to be significant. Another very important finding is probably that the shift between efficiency classes is mainly affected by different driving metrics in urban and rural road.

The **temporal** dynamics of driving efficiency are also investigated and the moving **time window** in which each driver is **assessed** is specified. It is shown that despite the fact that drivers retain a steady driving behaviour for a certain period, there exists dynamic major shifts in systematic behaviour within a long-term period. Furthermore, the average trend is observed to be approximately the same between the two road types despite the fact that median trend is differentiated significantly. Finally, studying stationarity demonstrated that three out of four driver groups have similar characteristics and therefore it would not play an important role in the final clustering procedure where this attribute is not included.

Clustering analysis performed resulted to three driving clusters, which mainly represent the **average** drivers, the **unstable** drivers and the **cautious** drivers. The main common attribute between all clusters of cautious drivers is the high driving efficiency index and the low value of the accident per year value regardless of whether or not it was included as a factor in the cluster analysis. On the other hand, all clusters of the average drivers feature a high driving efficiency index and an insignificant low positive trend indicating a steadily poor driving behaviour. Finally, the unstable drivers of the second cluster present a medium to high volatility, which is found to be the only common characteristic between them. Finally, prior information on driving accident data seems to affect only the form of the second cluster of the most unstable drivers, which incorporates drivers that are both less safety efficient and unstable.

This doctoral dissertation contributes towards the understanding of driving safety efficiency benchmarking, and therefore driving risk, using data science techniques applied on **large-scale** data in the form of **travel** and **driving** behaviour metrics collected from each trip and on a driver basis. A new methodological approach is also provided for estimating the efficient level of inputs and outputs that each driver should reach to become efficient in terms of safety. It is also very important that this research recognize the main characteristics of the driving safety efficiency groups arising from the **improved DEA** methodology performed, because this sets the ground for the in-depth study on driving efficiency based on microscopic driving characteristics. Finally, this research studies the time evolution of driving efficiency and reveals the characteristics of the driving profiles arose.

This thesis is also dealing with the problem of **data science** techniques that can be applied in real transportation problems as the one examined, to deal with the problem driving efficiency **benchmarking** using DEA. Consequently, the performance of DEA methodology for large-scale data as well as the potential of applying an improved DEA approach with certain techniques (RBE, Convex Hull) to yield the same optimal solution in less time is examined herein. Moreover, the large-scale driving data collected are investigated through statistical methods in order to specify the certain amount of driving data that should be collected for each driver in each road type. The need for specifying this amount emerges from the fact that collecting either excessive or deficient driving data can be risky because it might lead to excessive computational effort when it comes to large-scale data or to non-significant conclusions, respectively.

The latter approaches combined are the **innovation** of this doctoral research in terms of the large-scale **data handling**. This doctoral research presents how to reduce the **dimensionality** of a problem using large-scale data and draw valuable conclusions from them. This study also provides the methodological steps for estimating the efficient level of metrics for a trip and the approach to identify the least efficient trips in a database.