Temporal Analysis of Driving Efficiency Using Smartphone Data

Dimitrios I. Tselentis[1], Eleni I. Vlahogianni, George Yannis,

National Technical University of Athens
Department of Transportation Planning and Engineering
5, Iroon Polytechniou str., Zografou Campus, GR-15773
Athens, Greece

ABSTRACT

This paper attempts to shed light on the temporal evolution of driving safety efficiency with the aim to acquire insights useful for both driving behavior and road safety improvement. Data exploited herein are collected from a sophisticated platform that uses smartphone device sensors during a naturalistic driving experiment, at which the driving behavior from a sample of two hundred (200) drivers during 7-months is continuously recorded in real time. The main driving behavior analytics taken into consideration for the driving assessment include distance travelled, acceleration, braking, speed and smartphone usage. The analysis is performed using statistical, optimization and machine learning techniques. The driver's safety efficiency index is estimated both in total and in several consecutive time windows to allow for the investigation of safety efficiency evolution in time. Initial data analysis results to the most critical components of microscopic driving behaviour evolution, which are used as inputs in the k-means algorithm to perform the clustering analysis. The main driving characteristics of each cluster are identified and lead to the conclusion that there are three main driving groups of the a) moderate drivers, b) unstable drivers and c) cautious drivers.

Keywords: Driving Behavior, Driving Safety Efficiency, Temporal Evolution, K-means Clustering, Smartphone Data

[1]Corresponding author. Tel.: +30-210-772-2210; fax: +30-210-772-1454.

E-mail address: mailto:dtsel@central.ntua.gr

1. INTRODUCTION

Many studies in driving behavior literature (*1, 2, 3, 4*) have focused on measuring driving safety efficiency. *Driving safety efficiency* in this study refers to the amount of driving events (harsh braking, harsh acceleration, mobile phone usage, speeding) occurred during a certain driving period (*1*). The most efficient drivers are those with the least number of events. Nonetheless, only a few of them have demonstrated that there is a potential in analyzing and evaluating driving behaviour using microscopic driving data (e.g. driving over the speed limits, mobile phone usage and the number of harsh acceleration and braking events occurred while driving) collected from naturalistic driving experiments (*1), (5 - 10*). It is extremely important from a safety perspective to identify the behavioral parameters that influence drivers and, therefore, the probability of getting involved in a crash.

Several studies, thus far, have investigated the microscopic driving factors that could potentially be incorporated in driving assessment models as well as the methodologies for driving behavior data collection and analysis (*1, 11*). Driving behavior is complex; it can be influenced by the mode of transport (*12*), and the type of road (*13*). Due to this inherent complexity, a different optimal driving policy may exist for each driver (*14*). Finally, driving behavior, if controlled, can lead not only to safer roads, but also to healthier cities (*15*).

As for the data collection, literature review revealed that the methodologies most commonly used include driving simulators (*16, 17*), questionnaires (*3*) combined with simulators and naturalistic driving experiments (*5, 18, 19*). Naturalistic experiments provide a wide perspective of understanding typical microscopic travel and driving behaviour (*10*). A naturalistic study can help (*20*): a) determine crash risk, b) study the interaction between road/ traffic conditions and driver's behaviour, c) understand the interaction between car drivers and vulnerable road users, d) specify the relationship between driving pattern and vehicle emissions and fuel consumption, and many other aspects of traffic participation. The most popular devices for monitoring driving measures are recorders related to the car engine (*21, 22*) such as on-board-diagnostics (OBD) devices and smartphones (*23*).

Driving efficiency assessment using microscopic driving parameters is thoroughly studied (*1*). Nevertheless, the evolution of the driving performance in time (in the long term) has not yet been investigated. The temporal characteristics of driving efficiency and especially stationarity, trend and volatility, are of outmost importance when driving efficiency is measured. This is because the average driving efficiency might be representative of the total driving risk only in those cases when driving behavior is not fluctuating in time. This occurs solely when a driver retains a steady behavior that is usually close to the average and is not significantly changing while the driver is being monitored. Therefore, driving efficiency measured might not be representative of a driver's risk in a case of an unstable driver whose behaviour is volatile. In other words, the methodology of average efficiency estimation might not always be able to stand alone. For instance, when comparing the behavior of two drivers of the same average efficiency, it is more likely for the less steady driver (the one with the more volatile behavior) to have a higher crash risk since he features a higher number of less efficient trips.

In order to fill this research gap, this paper presents a methodological framework to study the temporal evolution of measured driver's efficiency with the aim to provide insights on the different driving behavior profiles. Whereas the authors' previous work (*1*) focused on assessing the aggregated driving behavior, this is an effort to reveal more information on how this behavior evolves over time, providing thus more information on the dynamic aspects of driving behavior. In other words, instead of providing only one driving efficiency index for the total monitoring period, monitoring period is divided into several shorter consecutive time periods and an individual driving efficiency index is provided for each of these periods. This process creates a time-series of driving efficiency indices for each driver, the trend and volatility of which is estimated. The latter two characteristics of the time-series, together with the driving efficiency of the total monitoring period, are used as inputs in a clustering algorithm to identify prevailing driver profiles. These profiles constitute an attempt to recognize the several existing driver profiles and their characteristics. The main driving characteristics of each group are presented and important conclusions are drawn regarding the features of each driving group. This methodology could be exploited as a platform's service in order to provide recommendations to drivers on how to improve their driving efficiency and become less risky.

2. METHODS FOR DRIVING EFFICIENCY ANALYSIS

Previous research on efficiency analysis has shown that data envelopment analysis (DEA) is an effective methodology to measure driving efficiency *(24, 25)*. Despite the fact that DEA is mostly used in business, economics, management and health *(26, 27)*, it has also been implemented in transport fields in assessing public transportation system performance *(28)*, as well as traffic safety studies *(29, 30, 31)* where it was proved to be equally useful as in the fields stated above. DEA is a mathematical programming technique with minimal assumptions that determines the efficiency of a Decision-Making Unit (DMU) based on its inputs and outputs and relatively estimates efficiency to the rest of the units involved in the analysis. A Decision-Making Unit (DMU) is "technically efficient" when the amount of outputs produced is maximized for a given amount of inputs, or for a given output the amount of inputs used is minimized *(32)*. Thus, when a DMU is technically efficient, it operates on its production frontier and therefore DMUs lie on the efficiency frontier *(33)*. The first DEA model proposed by *(34)* is the CCR model that assumes a constant returns to scale exhibition in production i.e. outputs are proportionally increased to inputs.

When studying driving behavior, efficiency is defined by the number of driving metrics recorded for a specific period or distance that a driver is being monitored *(1), (6 - 9)*. Therefore, drivers are considered driving units that make decisions about the number of events occurring, the time of mobile phone usage and speed limit violation within a given mileage range. Driving metrics and distance recorded are used as DEA inputs and outputs in order to estimate a driving safety efficiency index. The concept of DEA is to minimize inputs (input-oriented model) or maximize the outputs of a problem (output-oriented model) *(33)*. More specifically in the case study examined herein, a driver should either drive more kilometers maintaining the same number of harsh braking/ accelerating events or reduce the number of harsh braking/accelerating events for the same mileage. The same applies, of course, to the rest of the metrics recorded for each driver. From a road safety perspective, increasing mileage increases the exposure of a driver and consequently crash risk *(11)* and, therefore, an input-oriented (IO) DEA model is developed aiming to minimize inputs (recorded metrics) maintaining the same number of outputs (recorded distance). Previous research also showed that the driving efficiency problem is considered a constant-returns-to-scale (CRS) problem and that the sum of all metrics (inputs) recorded such as the number of harsh acceleration and braking events occurred in each $trip_i$ changes proportionally to the sum of driving distance (output) *(1)*. The mathematical formulation of the DEA problem is thoroughly described in section 3.1. The proposed methodology is applied to a driving sample of 38,000 recorded trips collected from one hundred (100) drivers in urban road and one hundred (100) drivers in rural road.

Models representing driving safety efficiency in urban and rural road types with multiple inputs and outputs are developed. Input and output selection is a critical process in DEA. Taking into consideration *(35, 36)* before applying DEA to a dataset and the specific data used in this study, metrics are used in the form of the number of harsh acceleration/ braking events, seconds driving over the speed limit and seconds used the mobile phone and not as e.g. the percentage of trip duration using the mobile phone. Literature review revealed that all the indicators mentioned before are the most influencing crash risk factors, among those that can be recorded from the smartphone sensors, and therefore they should be included in the models implemented *(1)*, *(5 - 9)*. All indicators along with distance travelled by drivers are provided per road type (urban, rural, highway) e.g.

number of harsh accelerations that occurred in urban road, seconds of speed limit exceedance in rural road etc. Variables used in the analysis along with the input and output combinations performed and their description are illustrated in Table 1. As it appears, one DEA model is developed for each of the two types investigated. The results of this step of the analysis are similar to those found in *(1)* and therefore they are not individually discussed herein.

<Table 1>

In this work, the driver is considered a DMU with an aggregate performance and his driving behavior equivalent to the sum of the driving characteristics for the entire period examined (*1*). For instance, the total distance travelled in rural road is equivalent to the sum of the distance travelled in rural network in each trip_{ij} (where i is the index of driver_i and j the index of trip_j of driver_i) by the specific driver_i (DMU_i). In general, the same applies for all indicators of driver_i, which are calculated aggregately as shown in (1):

$$\mathrm{indicator}_i = \sum_{j=1}^{N_i} \mathrm{indicator}_{ij} \tag{1}$$

recorded $\forall trip_j, j \in (1, N_i)$ that took place by $driver_i$. Since every driver is treated as DMU, the linear program constructed has 101 variables (λ_i, θ_B), equal to the number of drivers plus the efficiency of $driver_0$. The number of constraints on the other hand is equal to the sum of a) the number of inputs ($\theta_B * \chi_0 - \mathrm{X} * \lambda \geq 0$), b) the number of outputs ($Y * \lambda \geq y_0$) and c) the number of drivers ($\lambda_i \geq 0$). The DEA procedure described in (*1*) is followed separately for each model as described in Table 1.

3. TEMPORAL FEATURES OF DRIVING EFFICIENCY

3.1 Driving Safety Efficiency Estimation - Mathematical formulation

For the sake of simplicity, it is noted that from now on DMUs will be referred as drivers. In order to evaluate the driving efficiency of Driver_0 and assuming a sample of N drivers, let X and Y represent the set of inputs and outputs respectively, for the rest of the drivers' sample. In other words, $X=\{x_1,x_2,...,x_i\}$ and $Y=\{y_1,y_2,...,y_i\}$ where $i \in [1, N-1]$. The input-oriented CCR model evaluates the efficiency of Driver_0 by solving the linear program *(33)* presented below. Considering each driver as a DMU and taking into account the principles of DEA *(34)*, the mathematical formulation for the specific driving efficiency problem examined herein is:

$$\min(\text{Driving\_Efficiency}_0)$$

Subject to the following constraints:

$$\text{Driving\_Efficiency}_0 * x_0 - X * \lambda \geq 0$$

$$\text{Y} * \lambda \geq y_0 \qquad (2)$$

$$\lambda_i \geq 0 \forall \lambda_i \in \lambda$$

where λ_i is the weight coefficient for each Driver_i that is an element of set λ, X is the set of Inputs (number of harsh acceleration/braking events, seconds of mobile phone usage, seconds driving over the speed limits), Y is the set of Outputs (distance travelled) and $\text{Driving\_Efficiency}_0$ is a scalar representing the efficiency of reference driver (the one with Driver ID=0) i.e. Driver_0. Apparently, the use of the sets in the constraints indicates the creation of (N-1) inequalities when "building" the constraints of the linear problem i.e. 1+3*(N-1)=3*N-2 constraints. The rationality behind these constraints is to ensure that, compared to the rest of the sample, there could not be any other X, Y combination leading to a higher efficiency than that of the driver being evaluated. The set of λ estimated from the linear program is positive only for those drivers who act as peers to the driver being evaluated and is used afterwards to estimate the efficient level of inputs for the inefficient drivers (driving efficiency <1) that each driver should reach to become efficient. The objective function of DEA is min $\text{Driving\_Efficiency}_i$ i.e. to determine the minimum efficiency of driver_i that satisfies the above conditions. This formulation denotes that the target of this linear problem is to find the minimum driving efficiency that satisfies all the above constraints and not to minimize driving efficiency itself. To benchmark the efficiency of each and every driver in the database, this linear programming (LP) problem should be solved for each driver_i of the sample or in other words, N times in total. Finally, in order to create the time-series of the driving safety efficiency for each driver, this methodology is applied on a sliding time window.

3.2 Driver's Behaviour Volatility Measure

Observing how each driver alters everyday behaviour and whether or not there is a stability in his driving behaviour is crucial. To this end, the natural logarithm of the ratio of the performance of two consecutive time steps is estimated (*37*). This corresponds to the improvement or impairment of overall driving efficiency respectively, which changes according to the way he/she is driving over time. Let $E_{t,i}$ be the efficiency of a driver at the time step t, $E_t = (1,\dots, n)$, where t = (1,2,.., n) the number of time steps and I = (1,…, N) the ID of each driver. The improvement/ impairment ratio per time step is given by (3):

$$r_{t,i} = \ln\left(\frac{E_{t,i}}{E_{t-1,i}}\right) \tag{3}$$

This indicator has a positive value when driver's overall driving efficiency is improving in the next time step and a negative value when driver's overall driving efficiency is deteriorating. It is highlighted that equation (3) provides a measurement of the improvement/ impairment of the driving efficiency per time step and not the effect of two consecutive harsh events (braking/ acceleration etc.). In other words, it shows whether driving efficiency index that is estimated for each trip is improved (i.e. increased) after each trip compared to the previous one. The equation estimates the natural logarithm of the ratio of the current to the previous driving efficiency index. The total driver's behaviour volatility measure is estimated as the standard deviation of the improvement/ impairment ratio, in order to detect the instability in driver's safety efficiency evolution as shown in (4):

$$\text{Driver's behaviour volatility} = \sqrt{\frac{\sum_{t=1}^{n} (r_{t,i} - \bar{r}_i)^2}{n-1}} \tag{4}$$

where $r_{t,i}$ is the gain/loss per trip corresponds to driver i, $\bar{r}$ the average of gain/loss ratio and n the number of trips.

3.3 Driving Efficiency Time Series Analysis

Two of the most important concepts of time series analysis are used herein in order to create the rest of the critical driving behavior components that will be used as inputs in the clustering procedure.

3.3.1 Stationarity

Stationarity is a very important concept in time series analysis because time series models may apply only to stationary series. Therefore, before proceeding with time series modelling, no trend must be identified in the time series. Assuming the existence of a time series Y_t, where t is the observation period, the time series is strictly stationary if the joint probability distributions of $(Y_{t_1}, Y_{t_2}, \ldots, Y_{t_n})$ and $(Y_{t_{1+L}}, Y_{t_{2+L}}, \ldots, Y_{t_{n+L}})$ are the same for all $t_1, t_2, \ldots, t_n$ and L (length of seasonality). This implies that the joint distribution of $(Y_{t_1}, Y_{t_2}, \ldots, Y_{t_n})$ is time invariant, a strong condition that requires verification in practice (*38*).

When both the mean of Y_t, and the covariance between Y_{t_n}, and Y_{t_n+L} are time invariant (for an arbitrary L) weak stationarity applies, which is a weaker notion of stationarity. For $n=1$, the univariate distribution of Y_{t_1} is the same to that of Y_{t_1+L}. Accordingly, $E(Y_t)=E(Y_{t+L})$ and $VAR(Y_t)=VAR(Y_{t+L})$ implying that the mean μ and variance σ^2 of the time series are constant over time (*39*). For $n=2$, the joint probability distributions of (Y_{t_1}, Y_{t_1}) and $(Y_{t_{1+L}}, Y_{t_{1+L}})$ are the same and their covariances are equal (*40*) such that

$$\mathrm{COV}\left(\mathrm{Y}_{t_1}, \mathrm{Y}_{t_2}\right)=\mathrm{COV}\left(\mathrm{Y}_{t_1+L}, \mathrm{Y}_{t_2+L}\right) \tag{5}$$

When a time series is not stationary, stationarity is usually obtained through first-order differencing in transportation research *(40)* i.e. Y_t is $\mathrm{Z}_t=\mathrm{Y}_t-\mathrm{Y}_{t-1}$. In this case, the original series Y_t are called unit-root non-stationary. Several tests for non-stationarity (unit-root tests) have been proposed in literature *(41)* with the most satisfactory among them to be the Dickey–Fuller tests. The null hypothesis of this test is that the time series Y_t, is non-stationary and it requires at least first-order differencing to become stationary whereas the alternative is that the time series is already stationary.

3.3.2 Trend

The trend of a time series can be defined as the “long-term” movement in a time series without calendar related and irregular effects, and is a reflection of the underlying level. It may appear as a result of an alteration in several factors that affect the time series e.g. price inflation, general economic changes and population growth. It is necessary thus in time series analysis to estimate the magnitude of this change in time. There are five main methodological approaches to quantify the trend of a time series namely, least squares linear regression (*42*), Theil-Sen regression (*43*), Delta approach (*44*), The Hodrick-Prescott Trend approach (*45*) and the Rotemberg Trend approach (*46*).

Literature review revealed that if there is no prior knowledge of the time series characteristics, then the simplest function to fit is probably a straight line with the data plotted vertically and values of time (t = 1, 2, 3, ...) plotted horizontally *(47)*. This is the case when solving an online driving efficiency problem; there is no prior knowledge of the time series characteristics of each driver because a) these attributes are not known when a new driver is included in the analysis and b) even if they were, they are affected by the rest of the driving sample since efficiency is relatively estimated in each time step. Additionally, it is generally preferable to use the same approach for all drivers for the sake of simplicity and, therefore, this study makes use of the least squares linear regression methodology for the determination of the time series trend. It should be noted though, that the optimal choice would be to investigate the trend estimation approach that fits each time series best, but this is beyond the scope of this paper.

3.4 K-means Clustering Analysis

K-means clustering is a type of unsupervised learning, which aims to find the optimum way to group given data, with the number of groups represented by the variable k that is given as input. The algorithm iteratively assigns each data point to one of the k groups on the basis of the features provided. Data points are grouped based on the similarity of their features. The results of the k-means clustering algorithm are the centroids of the k clusters, which can be used to label new data and the labels for the training data (each data point is assigned to a single cluster). Clustering allows finding and analysing the groups that were formed naturally, instead of defining groups prior to looking at the data. The section below describes how the number of groups can be determined. The centroid of each cluster is a collection of feature values that defines resulting groups. When the centroid feature weights are examined, the kind of group each cluster represents can be qualitatively interpreted.

An iterative refinement is used by the k-means clustering algorithm to produce the final result. The inputs used by the algorithm are the number k of clusters to be created and the data to be clustered. Data are a collection of features for each data point or in other words each data point has several attributes that account for its features. Initially, the algorithm randomly assigns the k centroids, which are randomly either generated or selected from the dataset. Afterwards, the algorithm iterates between two steps, data assignment and centroid update. During the first step, each data point is assigned to the nearest existing centroid in this step, based on the squared Euclidean distance while during the second, the centroids are re-estimated by taking the mean of all data points assigned to that centroid's cluster.

The algorithm iterates between the two steps until one of the stopping criteria is met or in other words, no data point changes cluster and the sum of the distances is minimized or the maximum number of iterations set is reached. These algorithmic steps are guaranteed to converge to a result. The result though is likely to be a local optimum and not the optimum solution and therefore a better outcome might be reached by assessing more than one algorithm run with randomized starting centroids (*48*).

The algorithmic steps presented above result to k clusters for a particular pre-determined k. To estimate the optimum number of clusters arising, it is necessary to run the algorithm for a range of k values and compare the results. In general, there is not a methodology to determine the exact value of k, but the techniques presented below can be used to obtain an accurate estimate. One of the most commonly used metrics for comparing results across different values of k is the mean distance between cluster centroids and the data points assigned to each one of them. Increasing the number of clusters will always lead to a reduction of this distance, to the extreme of reaching zero when the number k is equal to the number of data points. Therefore, the minimization of this metric cannot be used as the sole target. Instead of that, the mean distance to the centroid is plotted as a function of the number of clusters k and the "elbow point," which appears at the point where the rate of decrease sharply shifts, can be used to estimate k.

4. EXPERIMENTAL DATA COLLECTION

OSeven Telematics (*49*) has developed an integrated platform for the recording, transmission, storage, evaluation and visualization of driving behaviour data using a smartphone application, statistical and advanced machine learning (ML) algorithms. Recorded data come from various smartphone sensors (including the accelerometer, the gyroscope, the GPS and the compass) and data fusion algorithms provided by Android and iOS. The application transmits all data to a central database after the end of each trip via Wi-Fi or cellular network. Data are stored in the cloud server for central processing and data reduction and thereafter processed using big data mining techniques and machine learning algorithms. This results to outlier detection and removal, data cleaning, smoothening and filtering, identification of repeating patterns, detection of aggressive behavior of the driver in the form of harsh events, mobile phone distraction, travel mode identification, identification of speed limit exceedance as well.

On a high-level, harsh events (braking and acceleration) are detected by identifying the relatively higher absolute values of acceleration (positive acceleration for harsh acceleration and negative acceleration for harsh braking) during each trip. As for the mobile phone usage, it is detected through the identification of the driving time during which unusual movement or rotation of the mobile device is recorded inside the vehicle. The steps described above for data processing are exclusively performed by OSeven Telematics and no raw data processing was implemented in this study (e.g. converting data from the gyroscope and accelerometer to harsh braking events). Therefore, they do not constitute part of this study. Unfortunately, more details on the data processing steps cannot be provided since they are intellectual property of the company. The main risk exposure indicators are:

- total distance travelled (between the start point and the end point of the trip),
- driving duration (time difference between the trip start time and trip end time),
- type(s) of the road network used (urban, rural, highway), and
- time of the day driving (morning peak/ rest of the day).

It is clarified that at the present study, those motorways with a speed limit of 90km/h and higher are taken into account as highways. The above are combined with other data sources (e.g. google maps service). The main driving behavior indicators are:

- speeding (percentage of time driving over the speed limits),
- mobile phone use while driving (identifying any movement of the smartphone device),
- number and severity of harsh events:
    - harsh braking,
    - harsh acceleration and
    - harsh cornering.

The descriptive statistics of the cumulative values of the driving metrics used in this study are shown in table 2.

<Table 2>

Aggregated data are analyzed and filtered to retain only those indicators that will be used as inputs for the analysis conducted in this research. This analysis is performed in Python programming language and more specifically, using the python packages of Pandas and Numpy for numeric calculations and transformations, Scipy that features Quickhull algorithm and pulp for linear programming problem construction and solving.

A significant amount of data is recorded using the platform of OSeven Telematics described above. The application records a random sample of drivers in terms of gender, age, education etc. in their everyday driving activities that have agreed to be recorded. The drivers are advised to use the application during their everyday activities and to drive normally as if they were not recorded. The application is used for a long period of time, e.g. more than 3-6 months, in order for the drivers to get used to being recorded and avoid monitoring a biased behaviour, e.g. a more cautious behaviour, because of the fact that they know they are being monitored. Drivers may place their mobile devices anywhere in their vehicle such as on a seat, the mobile phone holder or the cup holder etc. or store it in a personal item such as their pocket, pursue, bag etc. and sensor data are processed after collected to identify the characteristics of the vehicles' movement such as headway, longitudinal and lateral acceleration, speed.

The data are stored in the OSeven backend system using advanced encryption and data security techniques, in compliance with the national laws and EU Directives for the protection of personal data. The APIs (Application Programming Interfaces) used support user authentication and encryption to prevent unauthorized data access. The data used in this research were derived from the OSeven database and provided by the company for the purposes of this study. Data are anonymized before provided by OSeven so that driving behavior of each participant cannot be connected with any personal information. This is a data exploitation approach that is user-agnostic and therefore not user intrusive. The approach followed in this study aims to identify the main driving patterns without having any prior knowledge on the demographics of the sample collected.

The advantage of such an approach is that behaviors can be studied and drivers can be categorized based on their driving behavior even in cases where demographic data of a driving sample are not available or cannot be collected.

To achieve the goals of this research, large-scale driving data of 38,000 trips are randomly selected from the OSeven database of which 23,000 trips took place in urban road by one hundred (100) drivers and 15,000 trips took place in rural road by one hundred (100) drivers. All drivers were passenger vehicle drivers and trips took place between August 2016 and April 2017. In order to create time series of the same length for all participants, 230 urban and 150 rural trips are collected for each driver. In order to acquire a reliable measure for analyzing driving patterns and changes in drivers' behavior over time, driver's sample size is specified based on (*50, 13*). Moreover, drivers selected have a zero input attributes (i.e. zero harsh acceleration, braking, speed limit violation, mobile phone usage) since this is a limitation of DEA. This is because the business equivalent of a zero input driver is a factory that produces a product without making use of any material and/or workforce, which is practically impossible.

5. IMPLEMENTATION AND RESULTS

5.1 Components of the Efficiency Time Series

Driver's efficiency is estimated for each time step of a sliding time window, following exactly the same procedure described above in the estimation of total driving efficiency. This allows for studying the evolution of the average driving efficiency over different timeframes from the beginning of the recording time until the end of each timeframe. The length of the time window is estimated using specific statistical tests that identify the convergence of the driving analytics of each driver to a certain behavior (*13*). Results on the specific dataset indicate that this time window is 75 and 82 trips for urban and rural road types respectively.

The estimation of the driving efficiency in a sliding time window results to a sequence of driving efficiency indexes from which, the volatility measure of the time series is estimated. Table 3 illustrates the results of the volatility analysis performed, which indicate that although there is a higher range of volatility in rural road type, the average is approximately the same in both road and sample types. Based on driving volatility's definition, it is inferred that when it is equal to zero, a driver demonstrates a solid performance throughout monitoring. As a result, drivers with steady unit efficiency exist only in rural road since the minimum value of volatility is found to be higher than zero in urban road. On the other hand, the rural road sample includes users with a more alterable behaviour, which is evident from maximum volatility that is twice as users in the urban road sample.

<Table 3>

Time series is afterwards decomposed to acquire trend and stationarity using the methodological approach described above. It is observed in Table 3 that the average trend is approximately the same between the two road types 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 influences the average trend value. The negative skewness value of both road types of the anonymous sample indicates that the distribution's left tail is longer, compared to the right, whereas the positive skewness value of the non-anonymous sample testifies the opposite. Regarding the number of differences required for a time series to become stationary, zero urban road users and five rural road users with a stationary driving behaviour were found. The number of required differences for a time series to become stationary is equal to one for the vast majority of users and therefore, this variable is not illustrated in Table 3 and not used in the clustering procedure.

Figures 1 and 2 illustrate the driving efficiency time series of seven random drivers in the urban and rural sample respectively. Each color in these figures represents the evolution of the driving efficiency time-series of a different driver ID and therefore, a legend would not add an extra value to the figure and this is why it is intentionally omitted. The observed time series fluctuation is indicative to the existence of different driving patterns. It is observed in both figures that drivers who are the least efficient in total, also appear to be the least volatile among the rest. The most efficient drivers also appear to be less volatile but not as much as the latter. On the other hand, medium efficiency drivers are the most volatile among the drivers' sample. This is probably because in order to maintain either a very high or low efficiency level, efficiency cannot fluctuate that much because it will approximate the average efficiency. All these observations are also confirmed by the results illustrated below, where driving patterns identified are quantitatively discussed. Nonetheless, this does not affect the entire picture of the time series since the index values of all the other time series are relatively estimated at that time point or period.

<Figure 1>

It is also evident that there are some common local minimum and maximum points for most of the drivers, which is attributed to existing efficiency outliers at these time points or periods. These actually represent a time point or period when the most efficient drivers increases or decreases their driving metrics significantly and since efficiency is benchmarked, this results in an equally significant drop or increment in the efficiency of the other drivers. This might not have an impact on drivers' efficient index, as it still remains equal to one, and therefore it is not always visualized in figures 1 and 2.

<Figure 2>

5.2 Clusters' Driving Characteristics

The k-means algorithm is applied to cluster drivers based on the total driving efficiency, volatility, and the trend of the time series. To prevent the results from being influenced by the outliers, all variables are normalized before used as inputs. Driving characteristics of each cluster arose are presented and discussed in Table 4.

The optimal number of clusters is determined using the elbow method. In most cases, the elbow appears to exist at k = 3 or 4 indicating that the optimal number of clusters is either 3 or 4. After several clustering tests performed using both (3 and 4), it was found that some of the clusters formed in most cases included a significantly low number of users (e.g. 2) when k is set to 4 and therefore the results obtained were not representative. As a result, the number k of clusters is set to 3, which is rational considering the sample size used in this study.

The macroscopic characteristics of the urban sample's clusters that resulted from the analysis are illustrated in Table 4. Cluster 1 presents a very low positive trend compared to the rest of the clusters formed showing thus a slight tendency of these drivers to improve their driving behaviour. The volatility of their behaviour also seems to be medium to high and therefore an instability exists in behaviour. Drivers of this cluster also feature an average low total efficiency value, which shows a poor average behaviour. All the above along with the high number of drivers included in the specific cluster, lead to the conclusion that this cluster mainly represent the moderate driver. As for cluster 2, it features a medium positive efficiency trend indicating an overall improvement trend. This positive instability is also confirmed by the medium to high volatility presented. This cluster's drivers also have a medium average efficiency index which along with all the aforementioned demonstrate that this cluster is comprised from unstable drivers with less risky behaviour and a constant trend of improvement. Drivers of cluster 3 present a medium negative trend and a low to medium behavioral volatility. They also feature a medium to very high average driving efficiency confirmed by the low crash frequency. Consequently, this cluster includes the most cautious drivers of the sample and the negative trend is probably because of the fact that it is extremely rare for a driver to be highly efficient and steadily improved at the same time. It can be named as the cluster of cautious drivers.

Table 4 also shows the macroscopic characteristics of the rural sample's clusters arising from the analysis performed. Cluster 1 presents a low positive trend compared to the rest of the clusters formed showing thus a slight tendency of these drivers to improve their driving behaviour. A medium behavioral volatility also appears and as a result, an instable behaviour exists. Drivers of this cluster also feature an average low total efficiency value, which shows a poor average behaviour. All the above along with the high number of drivers included in the specific cluster, lead to the conclusion that this cluster mainly represents the moderate driver. In general, this cluster is very similar to cluster 1 obtained from the cluster analysis performed in urban roads. Drivers of cluster 2 present a high negative trend, a high behavioral volatility and a medium to high average driving efficiency and are generally unstable. Consequently, a significant deterioration of their behaviour is present while being monitored, which is the most important difference observed between the clustering results obtained from the analyses of the two road types.

As for cluster 3, it features a high positive efficiency trend indicating an overall improvement trend. This positive instability is also confirmed by the medium to high volatility that is shown. This cluster's drivers also have a high average efficiency index which along with all the aforementioned demonstrate that this cluster is comprised of cautious drivers with less risky behaviour and a constant trend of improvement. This cluster is also similar to the third cluster that results from the analysis of the anonymous sample in urban roads.

<Table 4>

6. DISCUSSION

This paper provides a structured approach to investigate the evolution of driving efficiency in time, aiming to draw conclusions on the different existing driving patterns. Driving efficiency estimation is based on a methodology introduced in previous research using DEA (*1*). Based on driving analytics of a sample of two hundred (200) drivers during 7-months, the analysis of the efficiency time series arising revealed that although there is a higher range of volatility in rural road type, the average driving efficiency is approximately the same in both road and sample types. The average trend is observed to be approximately the same between the two road types despite the fact that median trend is diverged. Finally, stationarity is not included in the final clustering procedure since all driver groups have similar characteristics and therefore this characteristic would not play an important role. All the above lead to the conclusion that when driving efficiency is benchmarked using DEA, the sample should be assessed on a regular basis to identify any alterations made in the efficiency frontier, which will result in a change in the ranking of the drivers. As a result, drivers should be continuously monitored and re-evaluated to capture these shifts and provide personalized advice on how their behaviour could be improved in the future.

A potential is identified in this study for classifying drivers' sample based on macroscopic temporal driving characteristics. The clustering analysis performed resulted to three main driving groups of the a) moderate drivers, b) unstable drivers and c) cautious drivers. The highlights of the analysis conducted for each category indicated considerable differences in driving characteristics between cautious drivers and the other two clusters in terms of driving performance. As mentioned in the introduction, the previous work of the authors (*1*) focused on the assessment of the aggregated driving behavior and therefore, the classification of the drivers as most efficient, weakly efficient, and non-efficient was made without taking into account how driving behavior evolves over time. The results of this research provide more information on the dynamic aspects of driving behavior, which are now taken into consideration in the driving behavior analysis. In other words, the quantified efficiency found in the previous work of the authors was incorporated in this study as a parameter for driver profiling. It becomes apparent that the results of the two studies are not directly connected and consequently, they cannot be compared. However, as mentioned in the results section, the main common attribute between all clusters of cautious drivers is the high driving efficiency index, which probably indicates that most of the drivers found in the class of the "most efficient" drivers can also be found in the class of "cautious" drivers.

Apart from the high driving efficiency index, the main common attribute between all clusters of the cautious drivers is the low value of the crash frequency per year value regardless of the fact that it was not included as a factor in the cluster analysis. On the other hand, all clusters of the moderate 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 regarding all attributes. It is highlighted that no statistical testing for the results obtained is undertaken since both DEA and k-means clustering are neither a statistical/ econometric method for forecasting future observations of a variable/time-series nor a supervised method in general, and therefore no validation technique could be applied (e.g. cross-validation, testing different datasets etc.).

In a real case scenario, drivers could be monitored for a certain period to analyze and evaluate their driving behavior. Thus, the riskiest driving traits that significantly influence crash probability would be recognized. Those results can potentially feed a platform's service and provide feedback and recommendations to drivers on their driving characteristics that need further improvement to become less risky. To this end, gamification policies based on this approach such as competitions, learning goals and awards could contribute to this scope. The results of this research could also be exploited in order to create innovative insurance pricing schemes that will be based on driving characteristics (e.g. Pay-How-You-Drive driving insurance schemes) and not mainly on demographics.

The main driving characteristics of the clusters that result from the analysis such as mobile usage, speed limit violation and number of harsh events should be further analyzed in the future to acquire a clearer picture on the dominant driving patterns that exist. Future research should also focus on larger drivers' samples with a representative sample collected from the entire population or from many countries so that more generalized conclusions can be drawn. It is a fact that models become more representative of the average characteristics of each cluster as more trips and drivers are aggregated. On the top of that, it would be beneficial to collect the crash record of the participants and include it in the clustering procedure in order to check if results arising are also representative of the individual driving risk. Finally, more driving metrics influencing crash risk should be used and test whether or not driving behavior models are improved.

ACKNOWLEDGEMENT

The authors would like to thank OSeven Telematics, for providing all necessary data to accomplish this study.

REFERENCES

1. Tselentis, D. I., Vlahogianni, E. I., & Yannis, G. (2019). Driving safety efficiency benchmarking using smartphone data. Transportation research part C: emerging technologies, 109, 343-357.
2. Matthews, Gerald, Timothy J. Sparkes, and Helen M. Bygrave. Attentional overload, stress, and simulate driving performance. Human Performance Vol. 9, No. 1 1996, pp. 77-101. http://dx.doi.org/10.1207/s15327043hup0901_5
3. Matthews, Gerald, Lisa Dorn, Thomas W. Hoyes, D. Roy Davies, A. Ian Glendon, and Ray G. Taylor. Driver stress and performance on a driving simulator. Human Factors Vol. 40, No. 1, 1998, pp. 136-149. https://doi.org/10.1518/001872098779480569
4. Young, Mark S., Stewart A. Birrell, and Neville A. Stanton. Safe driving in a green world: A review of driver performance benchmarks and technologies to support 'smart'driving. Applied ergonomics Vol. 42, No. 4, 2011, pp. 533-539. https://doi.org/10.1016/j.apergo.2010.08.012
5. Toledo, Tomer, Oren Musicant, and Tsippy Lotan. In-vehicle data recorders for monitoring and feedback on drivers' behavior. Transportation Research Part C: Emerging Technologies Vol. 16, No. 3, 2008, pp. 320-331. https://doi.org/10.1016/j.trc.2008.01.001
6. Eboli, L., Mazzulla, G., & Pungillo, G. (2016). Combining speed and acceleration to define car users' safe or unsafe driving behaviour. Transportation research part C: emerging technologies, 68, 113-125.
7. Eboli, L., Mazzulla, G., & Pungillo, G. (2017). How to define the accident risk level of

car drivers by combining objective and subjective measures of driving style. Transportation research part F: traffic psychology and behaviour, 49, 29-38.

8. Guo, F., & Fang, Y. (2013). Individual driver risk assessment using naturalistic driving data. Accident Analysis & Prevention, 61, 3-9.
9. Dingus, T. A., Guo, F., Lee, S., Antin, J. F., Perez, M., Buchanan-King, M., & Hankey, J. (2016). Driver crash risk factors and prevalence evaluation using naturalistic driving data. Proceedings of the National Academy of Sciences, 113(10), 2636-2641.
10. Mantouka, E., Barmpounakis, E., Vlahogianni, E., & Golias, J. (2020). Smartphone Sensing for Understanding Driving Behavior: Current Practice and Challenges. International Journal of Transportation Science and Technology.
11. Tselentis, Dimitrios I., George Yannis, and Eleni I. Vlahogianni. Innovative motor insurance schemes: a review of current practices and emerging challenges. Accident Analysis & Prevention Vol. 98, 2017, pp. 139-148. https://doi.org/10.1016/j.aap.2016.10.006
12. Barmpounakis, E., & Vlahogianni, E. I. (2020). Powered Two-Wheeler detection using crowdsourced smartphone data. IEEE Transactions on Intelligent Vehicles.
13. Stavrakaki, A.-M.; Tselentis, D.I.; Barmpounakis, E.; Vlahogianni, E.I.; Yannis, G. Estimating the Necessary Amount of Driving Data for Assessing Driving Behavior. Sensors 2020, 20(9):2600.
14. Vlachogiannis, Dimitris M., Eleni I. Vlahogianni, and John Golias (2020). "A Reinforcement Learning Model for Personalized Driving Policies Identification." International Journal of Transportation Science and Technology
15. Adamidis, F. K., Mantouka, E. G., & Vlahogianni, E. I. (2020). Effects of controlling aggressive driving behavior on network-wide traffic flow and emissions. International Journal of Transportation Science and Technology.
16. Desmond, Paula, Peter Hancock, and Janelle Monette. Fatigue and automation-induced impairments in simulated driving performance. Transportation Research Record: Journal of the Transportation Research Board No. 1628, 1998, pp. 8-14. https://doi.org/10.3141/1628-02
17. Lenné, Michael G., Thomas J. Triggs, and Jennifer R. Redman. Time of day variations in driving performance. Accident Analysis & Prevention Vol. 29, No. 4, 1997, pp. 431-437. https://doi.org/10.1016/S0001-4575(97)00022-5
18. Birrell, Stewart A., Mark Fowkes, and Paul A. Jennings. Effect of using an in-vehicle smart driving aid on real-world driver performance. IEEE Transactions on Intelligent Transportation Systems Vol. 15, No. 4, 2014, pp. 1801-1810. https://doi.org/10.1109/TITS.2014.2328357
19. Hanowski, R. J., Wierwille, W. W., & Dingus, T. A. (2003). An on-road study to investigate fatigue in local/short haul trucking. Accident Analysis & Prevention, 35(2), 153-160.
20. Regan, M. A., Williamson, A., Grzebieta, R., & Tao, L. (2012, August). Naturalistic driving studies: literature review and planning for the Australian naturalistic driving study. In Australasian college of road safety conference 2012, Sydney, New South Wales, Australia.
21. Zaldivar, Jorge, Carlos T. Calafate, Juan Carlos Cano, and Pietro Manzoni. Providing

accident detection in vehicular networks through OBD-II devices and Android-based smartphones. In Local Computer Networks (LCN), 2011 IEEE 36th Conference on, pp. 813-819. IEEE, 2011. https://doi.org/10.1109/LCN.2011.6115556

22. Backer-Grøndahl, Agathe, and Fridulv Sagberg. Driving and telephoning: Relative accident risk when using hand-held and hands-free mobile phones. Safety Science Vol. 49, No. 2, 2011, pp. 324-330. https://doi.org/10.1016/j.ssci.2010.09.009

23. Vlahogianni, Eleni I., and Emmanouil N. Barmpounakis. Driving analytics using smartphones: Algorithms, comparisons and challenges. Transportation Research Part C: Emerging Technologies Vol. 79, 2017, pp. 196-206. https://doi.org/10.1016/j.trc.2017.03.014

24. R. Shone, Applications in Intermediate Microeconomics, Martin Robertson, Oxford, 1981.

25. Emrouznejad, A., Parker, B. R., & Tavares, G. Evaluation of research in efficiency and productivity: A survey and analysis of the first 30 years of scholarly literature in DEA. Socio-economic planning sciences, Vol. 42, No 3, 2008, pp. 151-157. https://doi.org/10.1016/j.seps.2007.07.002

26. Cook, W. D., & Seiford, L. M. Data envelopment analysis (DEA)–Thirty years on. European journal of operational research, Vol. 192, No. 1, 2009, pp. 1-17. https://doi.org/10.1016/j.ejor.2008.01.032

27. Hollingsworth, B., Dawson, P. J., & Maniadakis, N. Efficiency measurement of health care: a review of non-parametric methods and applications. Health care management science, Vol 2, No. 3, 1999, pp. 161-172. https://doi.org/10.1023/A:1019087828488

28. Karlaftis, M.G., Gleason, J.M., Barnum, D.T. (2013) Bibliography of Urban Transit Data Envelopment Analysis (DEA) Publications. Available at SSRN: http://ssrn.com/abstract=1350583, http://dx.doi.org/10.2139/ssrn.1350583

29. Egilmez, G., & McAvoy, D. Benchmarking road safety of US states: A DEA-based Malmquist productivity index approach. Accident Analysis & Prevention, Vol. 53, 2013, pp. 55-64. https://doi.org/10.1016/j.aap.2012.12.038

30. Alper, D., Sinuany-Stern, Z., & Shinar, D. Evaluating the efficiency of local municipalities in providing traffic safety using the Data Envelopment Analysis. Accident Analysis & Prevention, Vol. 78, 2015, pp. 39-50. https://doi.org/10.1016/j.aap.2015.02.014

31. Alper, D., Sinuany-Stern, Z., & Shinar, D. (2015). Evaluating the efficiency of local municipalities in providing traffic safety using the Data Envelopment Analysis. Accident Analysis & Prevention, 78, 39-50.

32. M.J. Farrell, The measurement of productive efficiency, J. Royal Statistical Society Series A Vol. 120, No. 3, 1957, pp. 253–281. https://doi.org/10.2307/2343100

33. Ramanathan, Ramu, ed. An introduction to data envelopment analysis: a tool for performance measurement. Sage, 2003.

34. Charnes, A., Cooper, W. W., & Rhodes, E. Measuring the efficiency of decision making units. European journal of operational research, Vol. 2, No. 6, 1978, pp. 429-444. https://doi.org/10.1016/0377-2217(78)90138-8

35. Dyson RG, Allen R, Camanho AS, Podinovski VV, Sarrico CS, Shale EA. Pitfalls and protocols in DEA. European Journal of Operational Research, Vol. 132, Issue 2, 2001,

pp. 245–259. https://doi.org/10.1016/S0377-2217(00)00149-1

36. Cooper WW, Seiford LM, Tone K. Data envelopment analysis: a comprehensive Text with models, applications, references and DEA-Solver Software, Springer Science & Business Media, 2006.
37. Mantouka, E. G., Barmpounakis, E. N., & Vlahogianni, E. I. (2019). Identifying driving safety profiles from smartphone data using unsupervised learning. Safety Science, 119, 84-90.
38. Ray, B. K., & Tsay, R. S. (2002). Bayesian methods for change-point detection in long-range dependent processes. Journal of Time Series Analysis, 23(6), 687-705.
39. Shumway, R. H., & Stoffer, D. S. (2000). Time series analysis and its applications. Studies In Informatics And Control, 9(4), 375-376.
40. Washington, S., Karlaftis, M. G., & Mannering, F. L. (2003). Statistical and econometric methods for transportation data analysis (pp. 241-255). Boca Raton, FL: Chapman & Hall/CRC.
41. Engle, R. F., & Granger, C. W. (1987). Co-integration and error correction: representation, estimation, and testing. Econometrica: journal of the Econometric Society, 251-276.
42. Cameron, S. (2005). "Making Regression Analysis More Useful, II". Econometrics. Maidenhead: McGraw Hill Higher Education. pp. 171–198. ISBN 0077104285.
43. Pranab Kumar Sen (1968) Estimates of the Regression Coefficient Based on Kendall's Tau, Journal of the American Statistical Association, 63:324, 1379-1389 http://dx.doi.org/10.1080/01621459.1968.10480934
44. Kendall, M., & Stuart, A. (1977). The advanced theory of statistics. Vol. 1: Distribution theory. In ats.
45. Hodrick, R. J. and E. C. Prescott. (1997), Postwar U.S. Business Cycles: An Empirical Investigation, Journal of Money Credit and Banking, Vol. 29(1), 1-16.
46. Rotemberg, J. J. (1999), A Heuristic Method for Extracting Smooth Trends from Economic Time Series, National Bureau of Economic Research Working Paper No. 7439.
47. Mills, T. C. (2003). Modelling trends and cycles in economic time series (Vol. 10). Basingstoke: Palgrave Macmillan.
48. Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. Journal of the Royal Statistical Society. Series C (Applied Statistics), 28(1), 100-108.
49. OSeven Telematics. Available online: https://www.oseven.io/
50. Shichrur, R., Sarid, A., & Ratzon, N. Z. (2014). Determining the sampling time frame for in-vehicle data recorder measurement in assessing drivers. Transportation research part C: emerging technologies, 42, 99-106.

Table 1: Variables Recorded During the Experiment

Model	Type	Name	Short description
1	Input	ha_{urban}	number of harsh acceleration events in urban road
	Input	hb_{urban}	number of harsh braking events in urban road
	Input	$mobile_{urban}$	total seconds of mobile phone usages in urban road
	Input	$speeding_{urban}$	total seconds of speed limit violation in urban road
	Output	$distance_{urban}$	total distance driven in urban road
2	Input	ha_{rural}	number of harsh acceleration events in rural road
	Input	hb_{rural}	number of harsh braking events in rural road
	Input	$mobile_{rural}$	total seconds of mobile phone usage in rural road
	Input	$speeding_{rural}$	total seconds of speed limit violation in rural road
	Output	$distance_{rural}$	total distance driven in rural road

Table 2: Descriptive statistics of the cumulative per driver values of the variables recorded

	Distance (km)	HA	HB	Mobile (secs)	Speeding (secs)
			Urban		
Min	288	31	18	902	3971
Max	5224	2033	1033	84677	81640
Average	2238.6	633.8	226.7	18109.7	26134.3
Standard Deviation	1061.8	485.4	184.5	17547.4	15164.7
Median	2166.5	477.5	150.5	11403	23621.5
Kurtosis	0.5	1	5.3	3.7	1.9
Skewness	0.7	1.2	1.9	1.8	1.1
			Rural		
Min	391	45	14	39	1802
Max	7209	1566	1213	65623	68405
Average	2274.6	342.5	173	10735.7	21214.3
Standard Deviation	1363.4	308.4	198.3	14155.8	14862.7
Median	2165.5	235	108	6420.5	19547
Kurtosis	3.2	3.8	15.5	7.5	0.7
Skewness	1.5	1.8	3.6	2.8	1
			Highway		
Min	180	2	1	37	23
Max	6932	185	98	27979	48367
Average	1512.7	30.3	16.4	2841.8	9126.4
Standard Deviation	1209.5	35.2	18	4384.8	10204.4
Median	1147.5	16	9.5	1604.5	5357.5
Kurtosis	6	8.4	7.4	19	4.1
Skewness	2	2.8	2.4	3.8	1.9

Table 3: Descriptive Statistics of the Driving Efficiency Volatility and Trend of the Drivers' Sample.

Sample type	**Volatility**		**Trend (*10e^{-3})**	
Road type	**Urban**	**Rural**	**Urban**	**Rural**
Min	0.022	0.000	-4.56	-8.79
Max	0.152	0.379	4.09	8.46
Average	0.119	0.111	0.68	0.66
Standard Deviation	0.021	0.055	1.25	2.69
Median	0.123	0.095	0.51	0.80
Kurtosis	7.245	6.393	3.820	3.696
Skewness	-2.388	2.102	-0.222	-0.550

Table 4: Macroscopic Characteristics of the Drivers' Clusters.

Road type	Cluster	Statistical metric	Trend (*10e^{-3})	Volatility	Efficiency index	Number of drivers
Urban	Cluster 1 (moderate drivers)	**Min**	-1.045	0.066	0.122	79
		Max	1.686	0.152	0.725	
		Average	0.516	0.123	0.340	
		Standard Deviation	0.534	0.013	0.108	
		Median	0.486	0.124	0.328	
		Kurtosis	0.303	4.969	0.944	
		Skewness	-0.123	-1.438	0.713	
	Cluster 2 (unstable drivers)	**Min**	2.032	0.066	0.448	13
		Max	4.085	0.141	1.000	
		Average	3.006	0.119	0.673	
		Standard Deviation	0.628	0.022	0.206	
		Median	3.067	0.125	0.608	
		Kurtosis	0.334	-1.815	-2.281	
		Skewness	0.209	-1.278	0.732	
	Cluster 3 (cautious drivers)	**Min**	-4.557	0.022	0.367	8
		Max	0.322	0.122	1.000	
		Average	-1.512	0.080	0.746	
		Standard Deviation	1.530	0.038	0.263	
		Median	-0.937	0.090	0.813	
		Kurtosis	-1.027	0.925	-1.154	
		Skewness	-1.053	-0.385	-0.237	
Rural	Cluster 1 (moderate drivers)	**Min**	-1.987	0.048	0.127	72
		Max	3.375	0.228	0.664	
		Average	0.764	0.099	0.363	
		Standard Deviation	1.040	0.035	0.120	
		Median	0.778	0.091	0.356	
		Kurtosis	-0.639	2.144	-1.806	
		Skewness	-0.252	1.437	0.410	
	Cluster 2 (unstable drivers)	**Min**	-8.785	0.072	0.323	12
		Max	-1.545	0.379	1.000	
		Average	-4.288	0.155	0.716	
		Standard Deviation	2.530	0.088	0.246	
		Median	-3.811	0.125	0.685	
		Kurtosis	0.412	2.323	-0.250	
		Skewness	-0.824	1.490	-0.042	
	Cluster 3 (cautious drivers)	**Min**	0.000	0.000	0.483	16
		Max	8.455	0.306	1.000	
		Average	3.904	0.133	0.847	
		Standard Deviation	2.573	0.072	0.160	
		Median	4.295	0.115	0.880	
		Kurtosis	-0.712	1.167	-0.268	
		Skewness	0.398	0.789	-0.802	

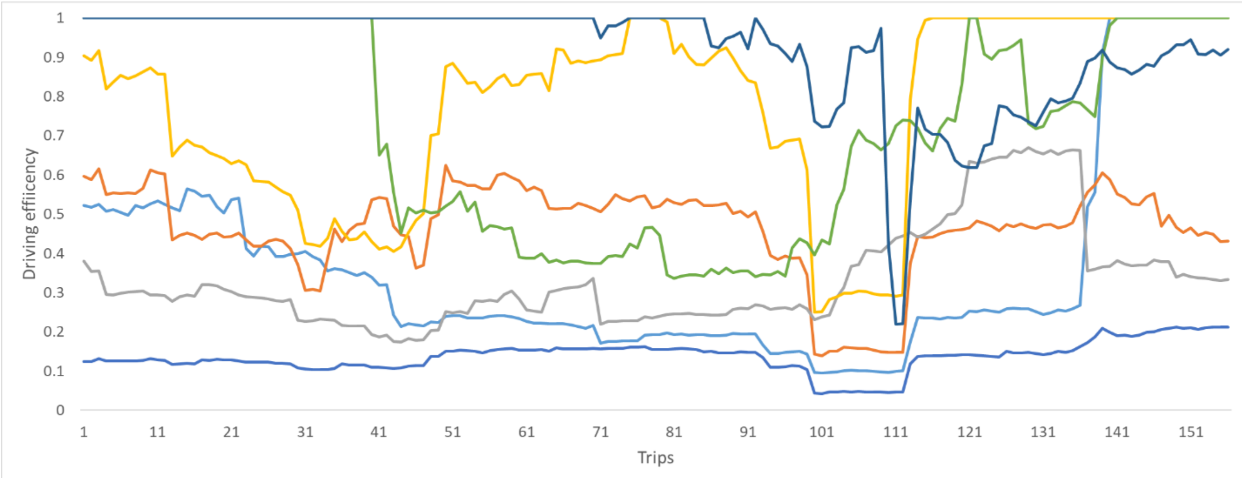

Figure 1: Efficiency time series of 7 random drivers of the anonymous urban sample.

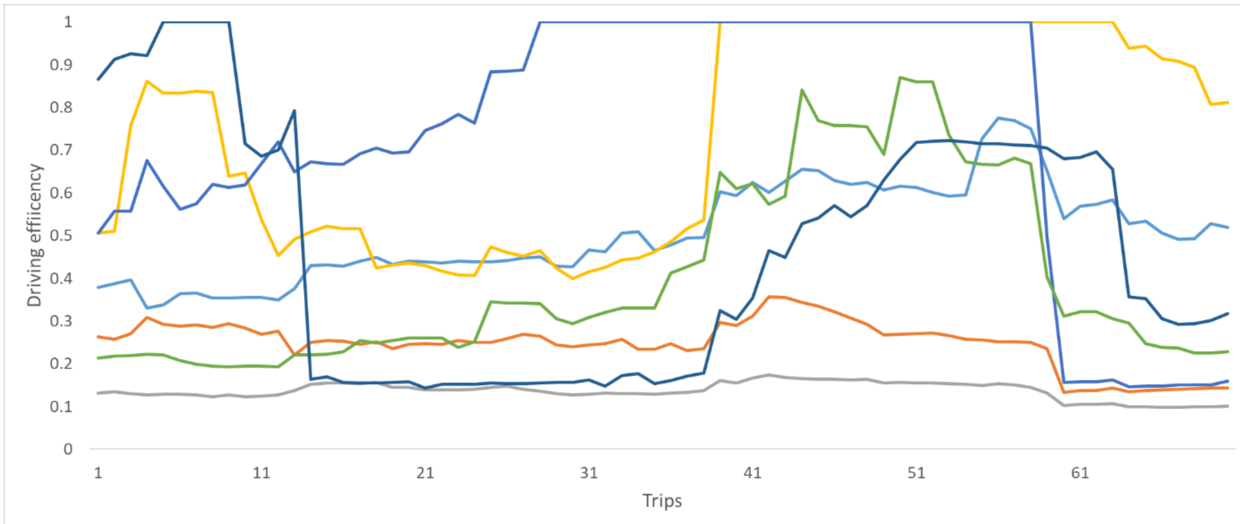

Figure 2: Efficiency time series of 7 random drivers of the rural sample.