Benchmarking Driving Efficiency using Data Science Techniques applied on Large-Scale Smartphone Data

Ph.D. Dissertation

Prepared by

Dimitrios I. Tselentis

Supervisor: Professor George Yannis
Co - Supervisor: Professor Nectarios Koziris
Co - Supervisor: Assistant Professor Eleni I. Vlahogianni

Athens, September 2018

To my family, Giannis, Sofia and Panagiotis

Acknowledgements

First, I would like to thank Professor George Yannis for the continuous support and guidance he provided on many levels. Our cooperation has been exceptional and his contribution in order for this research to be accomplished was extremely significant throughout the entire period. Working with Professor Yannis taught me how to cultivate my engineering skills and be problem-solver, creative and very well organized. Without his contribution, this doctoral research would have never been carried out and I will always be grateful to him for this.

I would also like to express my gratitude to Assistant Professor Vlahogianni and Professor Koziris whose indications and recommendations have been crucial from the very beginning of this research. I cannot thank enough Assistant Professor Vlahogianni with whom I have been working since the early years of my career and she has always been stimulating my analytical thinking and providing me with productive feedback on my work. I would equally like to thank Professor Koziris who significantly assisted by providing critical comments on this research.

Moreover, I am sincerely thankful for the advices and recommendations received from the remaining members of the examination committee, Professor John Golias, Professor Andreas Loizos, Professor Constantinos Antoniou and Associate Professor Nikolaos Geroliminis whose input has been vital for the finalization of this thesis.

I would also like to acknowledge the support of OSeven Telematics, who has provided all necessary data to carry out this study. The OSeven team has been working many years towards a safer road environment and the fact that the company has shown great trust and support to this dissertation from its very first research steps, is honourable for me.

Furthermore, a sincere gratitude is expressed to the entire research team of the Department of Transportation Planning and Engineering and more specifically to Eleonora, Alexandra, Panagiotis, Akis, Dimos, Katerina, Apostolis, Foteini, Areti, Manos, Elena, Manos and Katerina for the very productive conversations we have had over the years (scientific or not). It is always a great pleasure to work and be part of such a team of beautiful and creative minds that provides stimulation for further development.

At this point, I could not forget to thank all my friends that have been always very supportive throughout my career. A huge and sincere thank you is specially attributed to the group of "Yugoi", who always keep standing by each other in easy and difficult times.

I would also like to thank everyone who has passed through my life. Either they know it or not, every and each of them has individually influenced and formed who I am today and the way I think.

Above all, I thank my family for teaching me to be curious, persistent and passionate as well as to dedicate myself to everything I love. Since the very beginning, they always provided me a free, healthy and balanced environment to grow up in and become myself and this is why this doctoral dissertation is specially dedicated to them.

Contact details

To everyone who is willing to learn more about this PhD dissertation, please feel free to contact me using the following contact details:

Dimitrios I. Tselentis

Professional mobile phone: (+30) 697 8068955

Professional e-mail: dtsel@central.ntua.gr

Personal e-mail: tselentisdimitrios@gmail.com

Table of contents

Table of tables

Table of figures

Abstract

The main objective of this PhD is to provide a methodological approach for benchmarking driving efficiency in terms of safety on a trip and driver basis using data science techniques. The methodological approach is based on the definition of a safety efficiency index based on the Data Envelopment Analysis (DEA) theory and is related to macroscopic behavioral driving characteristics such as the number of sudden accelerations / decelerations, mobile phone usage time, and time exceeding the speed limit. In this dissertation, machine learning models are also developed to identify the different driving profiles that exist based on the temporal evolution of driving efficiency. The proposed methodological approach is applied on a real-world driving dataset collected from smartphones, which is analyzed using statistical methods to determine the amount of driving data to be used in the analysis. The results show that the optimized convex hull - DEA algorithm gives the exact solution in significantly less time than the classic DEA approaches. Furthermore, the methodology allows for the identification of the least efficient trips in a database as well as the efficient level of driving metrics of a trip to make it more efficient in terms of safety. Further clustering of the drivers based on the temporal evolution of driving performance leads to the identification of three main driver groups, the typical driver, the unstable driver and the less dangerous driver. Results indicate that having a prior knowledge on user's accident history solely affects the composition of the second cluster of the most volatile drivers, which incorporates drivers that are less efficient and unstable in terms of safety. It is shown that mobile phone use is not a critical factor in determining the safety efficiency of a driver since slight differences are found with regards to this characteristic between drivers of different efficiency classes. Furthermore, it is shown that a different driving data sampling is required for each a) road type, b) driving characteristic and c) driving aggressiveness level to collect enough data to obtain a clear picture of a driver's behaviour and perform DE analysis. Results could be exploited to provide personalized feedback to drivers on their total driving efficiency and its evolution in order to improve and reduce accident risk.

Σύνοψη

Ο κύριος στόχος της παρούσας διδακτορικής διατριβής είναι η ανάπτυξη μιας ολοκληρωμένης μεθοδολογικής προσέγγισης για τη συγκριτική αξιολόγηση της οδηγικής επίδοσης, όσον αφορά την οδική ασφάλεια, τόσο σε επίπεδο διαδρομής, όσο και οδηγού, με τη χρήση τεχνικών της επιστήμης δεδομένων. Η μεθοδολογική προσέγγιση στηρίζεται στον καθορισμό ενός δείκτη επίδοσης που βασίζεται στη θεωρία της Περιβάλλουσας Ανάλυσης Δεδομένων (Data Envelopment Analysis - DEA) και σχετίζεται με μακροσκοπικά συμπεριφοριστικά χαρακτηριστικά οδήγησης, όπως ο αριθμός των απότομων επιταχύνσεων/ επιβραδύνσεων, ο χρόνος χρήσης του κινητού τηλεφώνου και ο χρόνος υπέρβασης του ορίου ταχύτητας. Ακόμα, αναπτύσσονται μοντέλα μηχανικής μάθησης για τον προσδιορισμό διακριτών προφίλ οδήγησης που βασίζονται στη χρονική εξέλιξη της οδηγικής επίδοσης. Η προτεινόμενη μεθοδολογική προσέγγιση εφαρμόζεται σε πραγματικά δεδομένα οδήγησης ευρείας κλίμακας που συλλέγονται από έξυπνες συσκευές κινητών τηλεφώνων (smartphones), τα οποία αναλύονται μέσω στατιστικών μεθόδων για τον προσδιορισμό της απαιτούμενης ποσότητας δεδομένων οδήγησης που θα χρησιμοποιηθούν στην ανάλυση. Τα αποτελέσματα δείχνουν ότι ο βελτιστοποιημένος αλγόριθμος convex hull – DEA δίνει εξίσου ακριβή και ταχύτερα αποτελέσματα σε σχέση με τις κλασικές προσεγγίσεις της DEA. Ακόμα, η μεθοδολογία επιτρέπει τον προσδιορισμό των λιγότερο αποδοτικών ταξιδιών σε μια βάση δεδομένων καθώς και το αποδοτικό επίπεδο οδηγικών στοιχείων ενός ταξιδιού για να καταστεί αποδοτικότερη από την άποψη της ασφάλειας. Η περαιτέρω ομαδοποίηση των οδηγών με βάση της απόδοσή τους σε βάθος χρόνου οδηγεί στον εντοπισμό τριών ομάδων οδηγών, αυτή του μέσου οδηγού, του ασταθή οδηγού και του λιγότερο επικίνδυνου οδηγού. Τα αποτελέσματα δείχνουν ότι η εκ των προτέρων γνώση σχετικά με το ιστορικό ατυχημάτων του χρήστη φαίνεται να επηρεάζουν μόνο τη σύσταση της δεύτερης συστάδας των πιο ασταθών οδηγών, η οποία ενσωματώνει τους οδηγούς που είναι λιγότερο αποδοτικοί και ασταθής ως προς την ασφάλεια. Φαίνεται επίσης ότι η χρήση κινητών τηλεφώνων δεν αποτελεί κρίσιμο παράγοντα για τον καθορισμό της επίδοσης της ασφάλειας ενός οδηγού, καθώς διαπιστώθηκαν μικρές διαφορές σε σχέση με αυτό το χαρακτηριστική οδήγησης μεταξύ οδηγών διαφορετικών κατηγοριών επίδοσης. Επιπλέον, δείχνεται ότι απαιτείται μια διαφορετική δειγματοληψίας δεδομένων οδήγησης για κάθε α) οδικό τύπο, β) χαρακτηριστικό οδήγησης και γ) οδηγική επιθετικότητα για να συγκεντρωθούν αρκετά δεδομένα και να αποκτηθεί μια σαφής εικόνα της οδηγικής συμπεριφοράς και να εκτελεστεί ανάλυση με χρήση DEA. Τα αποτελέσματα θα μπορούσαν να αξιοποιηθούν για την παροχή εξατομικευμένης ανατροφοδότησης στους οδηγούς σχετικά με τη συνολική τους οδηγική επίδοση και την εξέλιξή της, προκειμένου να βελτιωθεί και να μειωθεί ο κίνδυνος ατυχήματος.

Ελληνική Περίληψη Διδακτορικής Διατριβής

Ο κύριος στόχος της Διδακτορικής Διατριβής είναι η ανάπτυξη μιας μεθοδολογίας για τη συγκριτική αξιολόγηση της οδηγικής επίδοσης ως προς την ασφάλεια σε επίπεδο ταξιδιού και οδηγού με τη χρήση τεχνικών της επιστήμης δεδομένων. Προκειμένου να επιτευχθεί αυτό, καθορίζεται ένας δείκτης επίδοσης της οδήγησης ως προς την ασφάλεια βασισμένος στις μετρήσεις των χαρακτηριστικών του ταξιδιού και της οδηγικής συμπεριφοράς που συλλέγονται από συσκευές έξυπνων κινητών τηλεφώνων. Αναλύονται τα χαρακτηριστικά οδήγησης κάθε ομάδας επίδοσης που προκύπτει και εντοπίζονται τα κύρια μοτίβα οδήγησης. Μία από τις σημαντικότερες αδυναμίες της μεθόδου της Περιβάλλουσας Ανάλυσης Δεδομένων (Data Envelopment Analysis - DEA), είναι ο σημαντικός χρόνος που απαιτείται για την επεξεργασία δεδομένων μεγάλης κλίμακας, η οποία ξεπερνιέται χρησιμοποιώντας τεχνικές υπολογιστικής γεωμετρίας. Επιπλέον, η παρούσα διδακτορική έρευνα προτείνει ένα μεθοδολογικό πλαίσιο για τον εντοπισμό των ταξιδιών ενός δείγματος που έχουν την χαμηλότερη επίδοση καθώς επίσης και για την εκτίμηση του επιπέδου των οδηγικών χαρακτηριστικών που πρέπει να φτάσει κάθε ταξίδι με χαμηλή επίδοση για να καταστεί αποδοτικό. Τέλος, ο στόχος αυτής της διατριβής είναι να μελετήσει τη χρονική εξέλιξη της οδηγικής επίδοσης και να προσδιορίσει τα κύρια μοτίβα οδήγησης και τα προφίλ των ομάδων οδηγών που σχηματίζονται.

Η ανασκόπηση της βιβλιογραφίας αποκάλυψε ότι είναι σημαντικό να μελετηθεί η δυνατότητα για συγκριτική αξιολόγηση της οδηγικής επίδοσης ως προς την ασφάλεια χρησιμοποιώντας μικροσκοπικά δεδομένα οδήγησης που συλλέγονται από συσκευές έξυπνων κινητών τηλεφώνων. Το συγκεκριμένο ζήτημα επιχειρείται να αντιμετωπιστεί προτείνοντας ένα μεθοδολογικό πλαίσιο που βασίζεται σε τεχνικές επιστήμης δεδομένων για την αξιολόγηση των χαρακτηριστικών οδήγησης. Χρησιμοποιείται ένα βελτιωμένο μοντέλο περιβάλλουσας ανάλυσης δεδομένων (DEA) για την ανάλυση των δεδομένων ευρείας κλίμακας που συλλέγονται από έξυπνα κινητά τηλέφωνα κατά την οδήγηση. Το μοντέλο που αναπτύχθηκε ενσωματώνει διάφορες μετρήσεις συμπεριφοράς οδήγησης που επιτρέπουν την ανάλυση πολλαπλών κριτηρίων σχετικά με την απόδοση οδήγησης.

Το γενικό μεθοδολογικό πλαίσιο που εφαρμόζεται απεικονίζεται στο Σχήμα 1. Υπάρχουν δύο πηγές δεδομένων: α) μια βάση δεδομένων με οδηγούς που συμμετείχαν σε ένα πείραμα φυσικής οδήγησης (naturalistic driving) όπου τα δεδομένα καταγράφηκαν χρησιμοποιώντας το έξυπνο τηλέφωνο κάθε συμμετέχοντα και β) το ερωτηματολόγιο που συμπληρώθηκε από ένα μέρος των συμμετεχόντων. Μετά τη συλλογή των δεδομένων, καθορίζονται οι παράγοντες που αντιπροσωπεύουν την οδηγική επίδοση όσον αφορά στην ασφάλεια με βάση την ανασκόπηση της βιβλιογραφίας. Αφού εξεταστεί ότι α) συλλέγονται επαρκή στοιχεία από κάθε συμμετέχοντα που λαμβάνονται υπόψη στην έρευνα αυτή και β) οι τιμές των χαρακτηριστικών οδήγησης και η συνολική απόσταση που καταγράφηκαν αυξάνονται αναλογικά καθώς και ότι ο λόγος τους δεν μεταβάλλεται σημαντικά όσο τα χιλιόμετρα αυξάνονται, αυτοί οι παράγοντες χρησιμοποιούνται ως δεδομένα εισόδου και δεμένα εξόδου για τα μοντέλα DEA που αναπτύσσονται. Η ανάλυση της επίδοσης του ταξιδιού και του οδηγού εφαρμόζεται ανά τύπο οδού

ακολουθώντας την λεπτομερή περιγραφή που δίνεται παρακάτω. Τα αποτελέσματα που προκύπτουν από την ανάλυση της επίδοσης του ταξιδιού αξιοποιούνται κυρίως για να μειωθεί ο χρόνος επεξεργασίας για την ανάλυση της απόδοσης του οδηγού, όπου διερευνάται η εξέλιξη της επίδοσης της οδήγησης μέσα στο χρόνο και δευτερευόντως για να αξιολογηθεί η πρακτικότητα της παροχής μεθοδολογίας για αναγνώριση των ταξιδιών χαμηλότερης επίδοσης. Τα αποτελέσματα των ερευνών εξέλιξης του οδηγού και της οδηγικής επίδοσης συνδυάζονται για την εκτέλεση ανάλυσης συσταδοποίησης σε επίπεδο οδηγού. Για κάθε ομάδα οδήγησης που προκύπτει από αυτή τη διαδικασία, εξετάζονται και παρουσιάζονται τα τυπικά χαρακτηριστικά οδήγησης των οδηγών που ανήκουν σε αυτήν.

Για την επίτευξη των στόχων που θέτει η παρούσα διδακτορική διατριβή, η δομή αυτής της έρευνας αποτελείται από έξι επιμέρους μεθοδολογικά βήματα που παρουσιάζονται ως ακολούθως (Σχήμα 1):
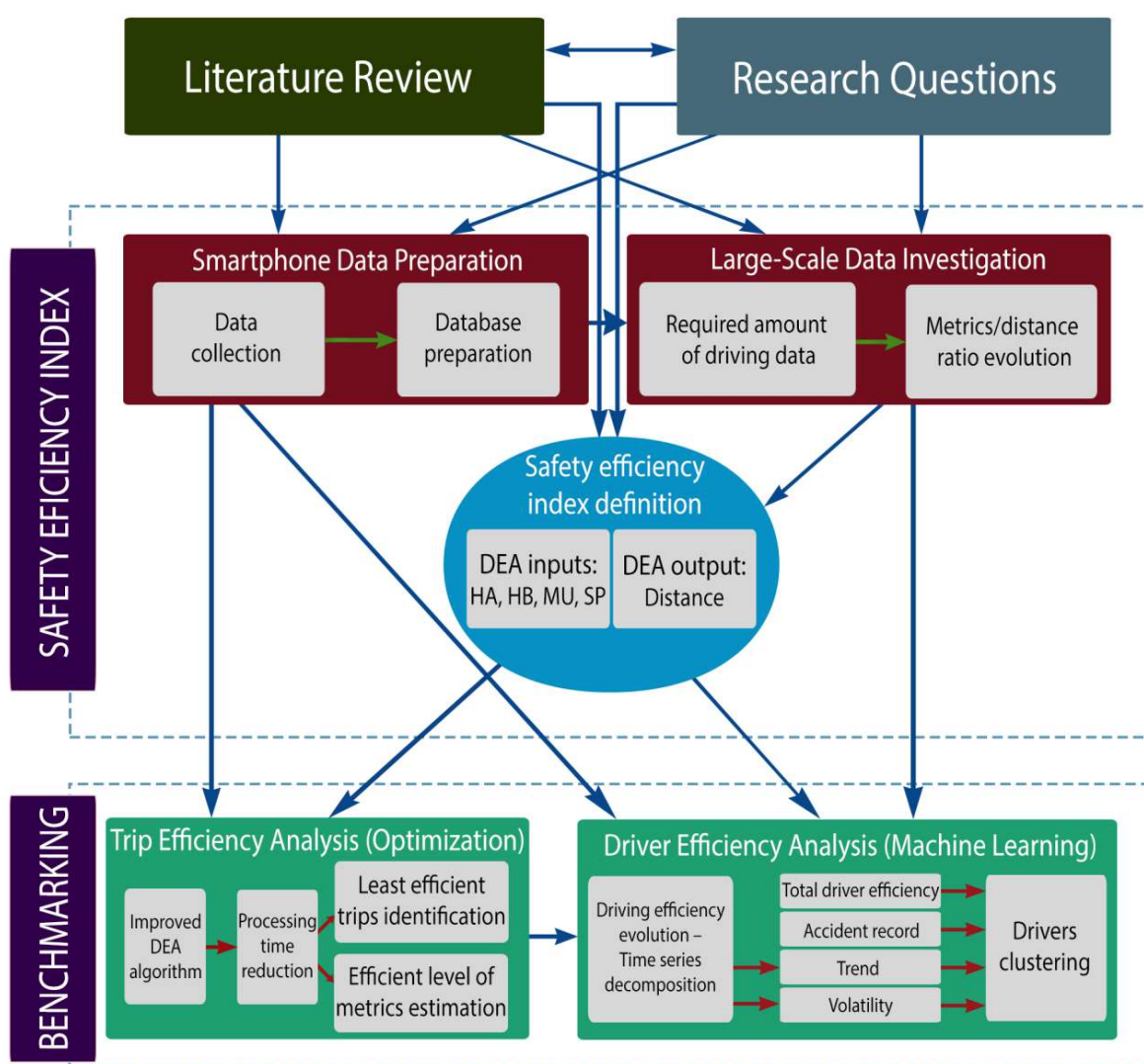
Σχήμα 1: Γραφική αναπαράσταση του πλαισίου της μεθοδολογικής προσέγγισης της παρούσας διδακτορικής διατριβής..

Η αναλυτική ανασκόπηση της βιβλιογραφίας λαμβάνει χώρα ως πρώτο βήμα, καλύπτοντας μια συνολική εικόνα της οδικής ασφάλειας και ατυχημάτων και τους τομείς οδηγικής συμπεριφοράς και κινδύνου, τα χαρακτηριστικά οδήγησης, παράμετροι οδηγικής επίδοσης (απόσπαση της προσοχής, επιθετικότητα, κλπ), πειράματα φυσικής οδήγησης, μεθοδολογία περιβάλλουσας ανάλυσης δεδομένων (DEA), τρόποι του χρόνου υπολογισμού της ανάλυσης δεδομένων μεγάλης κλίμακας και τις εφαρμογές της στις μεταφορές και την οδηγική επίδοση. Τα συμπεράσματα που προέκυψαν από την ανασκόπηση και το κενό γνώσης που προκύπτει συμβάλλουν στον καθορισμό των ερευνητικών στόχων και των ερευνητικών υποθέσεων και γενικότερα στη δημιουργία του προβλήματος.

Από την ανασκόπηση της βιβλιογραφίας προκύπτει ότι είναι απαραίτητο να μελετηθεί περεταίρω η συμπεριφορά οδήγησης και να διερευνηθεί περισσότερο η αξιολόγηση της συμπεριφοράς της οδήγησης και οι παράγοντες που την επηρεάζουν. Με την πάροδο των ετών, φαίνεται ότι η ασφάλιση με βάση τη χρήση (UBI) σκοπεύει να τιμολογεί ασφάλιστρα με βάση τον κίνδυνο ατυχήματος του κάθε οδηγού βάσει των χαρακτηριστικών της οδηγικής συμπεριφοράς και της διαδρομής. Ως εκ τούτου, οι οδηγοί θα πρέπει να μειώσουν τα ετήσια χιλιόμετρα και να βελτιώσουν την οδηγική τους συμπεριφορά. Αυτό συμβαίνει επειδή η οδηγική επικινδυνότητα ανά χιλιόμετρο είναι ένας απρόβλεπτος παράγοντας που κυμαίνεται με την πάροδο του χρόνου και επομένως ακόμη και αν τα χιλιόμετρα μειωθούν, ο συνολικός κίνδυνος συντριβής μπορεί ακόμα να αυξηθεί. Προς επίρρωση αυτού, ακόμα και αν ο κίνδυνος ατυχήματος ανά χιλιόμετρο παραμένει σταθερός και η ετήσια χιλιομετρική διαδρομή είναι γνωστή, ο συνολικός ατομικός κίνδυνος ατυχήματος δε μπορεί να εκτιμηθεί, δεδομένου ότι εξαρτάται από χαρακτηριστικά συμπεριφοράς που δεν καταγράφονται και λαμβάνονται υπόψη επί του παρόντος στο UBI. Για να επιτευχθεί αυτό, πληροφορίες σχετικά με τα χαρακτηριστικά της οδήγησης όπως π.χ. ο αριθμός των απότομων συμβάντων επιβράδυνσης και επιτάχυνσης, ο χρόνος οδήγησης πάνω από τα όρια ταχύτητας, ο τύπος οδού κ.λπ. πρέπει να περιλαμβάνονται κατά την αξιολόγηση του οδηγού. Με άλλα λόγια, ο συντελεστής επικινδυνότητας είναι ο ρυθμός αύξησης του κινδύνου, ο οποίος υποδεικνύει τον τρόπο με τον οποίο ο συνολικός ατομικός κίνδυνος αυξάνεται καθώς αυξάνεται η απόσταση. Ως αποτέλεσμα, είναι απαραίτητο να αναπτυχθεί ένα μοντέλο που να ενσωματώνει τόσο την απόσταση που διανύθηκε όσο και τα υπόλοιπα χαρακτηριστικά συμπεριφοράς, προκειμένου να αξιολογηθεί ο συνολικός κίνδυνος κατά την οδήγηση. Με την ανάπτυξη των μοντέλων DEA που λαμβάνουν υπόψη αυτές τις δύο κατηγορίες χαρακτηριστικών, η παρούσα μελέτη στοχεύει να εξετάσει τη δυνατότητα εφαρμογής τέτοιων μοντέλων.

Σύμφωνα με παλαιότερες έρευνες, τα πειράματα οδήγησης σε πραγματικές συνθήκες οδήγησης θεωρούνται πιο κατάλληλα για την αξιολόγηση συμπεριφοράς οδήγησης επειδή η συμπεριφορά καταγράφεται υπό κανονικές συνθήκες οδήγησης και χωρίς οποιαδήποτε επιρροή από εξωτερικές παραμέτρους. Το κύριο μειονέκτημα των πειραμάτων οδήγησης σε πραγματικές συνθήκες είναι ότι οι οδηγοί θα πρέπει να παρακολουθούνται για ένα επαρκές χρονικό διάστημα έτσι ώστε τα χαρακτηριστικά της οδήγησης υπό κανονικές συνθήκες που θα καταγραφούν να μην έχουν καμία προκατάληψη. Από την άλλη πλευρά, βρίσκεται να είναι πολύ σημαντικό να προσδιοριστεί η ποσότητα των δεδομένων που απαιτείται να συλλεχθεί για την απόκτηση

της πλήρους εικόνας για κάθε οδηγό, όπου ο ρυθμός των μετρήσεων που περιγράφηκαν παραπάνω ανά χιλιόμετρο συγκλίνει σε μια σταθερή (και μέση) τιμή.

Μπορούμε να πούμε από τα παραπάνω ότι οι σημαντικότεροι ανθρώπινοι παράγοντες που καταγράφηκαν από τις συσκευές smartphone και βρέθηκαν να επηρεάζουν τον κίνδυνο οδήγησης είναι η απόσπαση της προσοχής των κινητών τηλεφώνων, η υπέρβαση του ορίου ταχύτητας και ο αριθμός των απότομων επιβραδύνσεων και επιταχύνσεων που συνέβησαν κατά την οδήγηση. Μπορεί επίσης να συναχθεί ότι υπάρχουν πολυάριθμες έρευνες που επικεντρώνονται στην αξιολόγηση της συμπεριφοράς οδήγησης και κυρίως στον προσδιορισμό της συσχέτισης μεταξύ μετρήσεων συμπεριφοράς οδήγησης (υπέρβαση ορίου ταχύτητας, αριθμός σκληρών επιταχύνσεων / φρεναρίσματος, απόσπαση κινητού τηλεφώνου κ.λπ.) και της πιθανότητας ατυχήματος. Σύμφωνα με όσα είναι γνωστά ως σήμερα, αυτή η διδακτορική έρευνα είναι η πρώτη προσπάθεια για την εκτίμηση και την εκχώρηση ενός σχετικού δείκτη αποτελεσματικότητας ασφάλειας σε κάθε οδηγό ενός δείγματος αξιοποιώντας την απόσταση που διανύθηκε και μερικές μετρήσεις συμπεριφοράς οδήγησης που προκύπτουν από τα δεδομένα μικροσκοπικής συμπεριφοράς οδήγησης που καταγράφονται από συσκευές smartphone.

Από τα παραπάνω συμπεραίνεται ότι είναι σημαντικό να συγκριθεί η αποτελεσματικότητα της οδικής ασφάλειας χρησιμοποιώντας μικροσκοπικά δεδομένα οδήγησης που συλλέγονται από συσκευές smartphone. Αποδεικνύεται ότι η DEA δεν έχει χρησιμοποιηθεί ποτέ προηγουμένως στην έρευνα συμπεριφοράς οδήγησης και ότι η αποτελεσματικότητα του οδηγού έχει μελετηθεί σε μεγάλο βαθμό αλλά ποτέ με την χρήση των τεχνικών της DEA. Ως εκ τούτου, θα πρέπει να επιχειρηθεί να αντιμετωπιστεί αυτό το συγκεκριμένο ζήτημα προτείνοντας ένα μεθοδολογικό πλαίσιο που βασίζεται σε τεχνικές της επιστήμης δεδομένων για την αξιολόγηση των χαρακτηριστικών οδήγησης. Το μοντέλο που θα αναπτυχθεί θα πρέπει να ενσωματώνει διάφορες μετρήσεις συμπεριφοράς οδήγησης που επιτρέπουν την ανάλυση πολλαπλών κριτηρίων της απόδοσης της οδήγησης. Θεωρείται επίσης σημαντικό να αντιμετωπιστεί το πρόβλημα του μεγάλου χρόνου υπολογισμού που απαιτείται για να «τρέξει» ένας αλγόριθμος DEA και από μεθοδολογικής απόψεως είναι σημαντικό να ελεγχθεί η αποτελεσματικότητα της εφαρμογής ενός συνδυασμού του αλγορίθμου της DEA και του αλγορίθμου Convex Hull σε ρυθμίσεις πολλαπλών δεδομένων εισόδου και εξόδου για μεγάλης κλίμακας δεδομένα οδήγησης.

Το δεύτερο βήμα της μεθοδολογίας είναι η συλλογή και προετοιμασία δεδομένων, η οποία περιλαμβάνει τον σχεδιασμό της έρευνας και του ερωτηματολογίου καθώς και περιγραφή του τρόπου λειτουργίας της πλατφόρμας της OSeven, συμπεριλαμβανομένης της διαδικασίας καταγραφής, συλλογής, αποθήκευσης, αξιολόγησης και απεικόνισης των δεδομένων συμπεριφοράς οδήγησης με τη χρήση εφαρμογών έξυπνων κινητών τηλεφώνων και προηγμένων αλγορίθμων μηχανικής μάθησης (Machine Learning - ML). Μια σημαντική ποσότητα δεδομένων καταγράφεται χρησιμοποιώντας την εφαρμογή έξυπνων κινητών τηλεφώνων που αναπτύχθηκε από την OSeven Telematics. Αυτή η καινοτόμος μέθοδος συλλογής και ανάλυσης δεδομένων μεγάλης κλίμακας που εφαρμόζεται, παρουσιάζει νέες προκλήσεις συγκεντρώνοντας μεγάλες ποσότητες δεδομένων για ανάλυση κατά τη διάρκεια αυτής της έρευνας. Επιπλέον, η βάση δεδομένων επεξεργάζεται περαιτέρω και προετοιμάζεται για εισαγωγή προκειμένου να

διεξαχθεί η τελική ανάλυση δεδομένων στη συνέχεια. Αυτή η προετοιμασία γίνεται χρησιμοποιώντας τη γλώσσα προγραμματισμού Python, η οποία είναι κατάλληλη για ανάλυση δεδομένων μεγάλης κλίμακας.

Όλοι οι προαναφερθέντες δείκτες, που λαμβάνονται απευθείας από το σύστημα της OSeven, αναλύονται και φιλτράρονται για να διατηρηθούν μόνο εκείνοι οι δείκτες που θα χρησιμοποιηθούν εδώ ως δεδομένα εισόδου και εξόδου για το πρόβλημα της DEA. Τα φίλτρα δεδομένων και οι αλγόριθμοι βελτίωσης DEA εκτελούνται στη γλώσσα προγραμματισμού Python και για αυτό το λόγο έχουν γραφτεί πολλά σενάρια. Μια σημαντική ποσότητα δεδομένων καταγράφεται χρησιμοποιώντας την εφαρμογή smartphone που αναπτύχθηκε από την OSeven Telematics. Τα δεδομένα που χρησιμοποιούνται σε αυτήν την έρευνα παρέχονται ανώνυμα από την OSeven, έτσι ώστε η συμπεριφορά οδήγησης του κάθε συμμετέχοντα να μην μπορεί να συνδεθεί με οποιαδήποτε προσωπική πληροφορία. Πρόκειται για μια προσέγγιση αξιοποίησης δεδομένων που είναι αγνωστικιστική ως προς τον χρήστη και ως εκ τούτου λιγότερο παρεμβατική ως προς εκείνον. Θα πρέπει επίσης να επισημανθεί σε αυτό το σημείο ότι η προσέγγιση που ακολουθήθηκε σε αυτή τη μελέτη στοχεύει στον εντοπισμό των οδηγικών συμπεριφορών και μοτίβων και των παραγόντων που τους επηρεάζουν και όχι στην εξήγηση της αιτιότητας μεταξύ της συμπεριφοράς και άλλων παραγόντων όπως ηλικία, φύλο, κατανομή του οδηγού δείγματος που συλλέχθηκε. Το πλεονέκτημα μιας τέτοιας προσέγγισης είναι ότι οι συμπεριφορές μπορούν να μελετηθούν ακόμη και σε περιπτώσεις όπου τα δημογραφικά στοιχεία και το ιστορικό ατυχημάτων ενός δείγματος οδήγησης δεν είναι διαθέσιμα ή δεν μπορούν να συλλεχθούν.

Στο πλαίσιο αυτής της διδακτορικής έρευνας, συλλέχθηκε δείγμα από 171 οδηγούς σε διάρκεια 7 μηνών και αναπτύχθηκε μια μεγάλη βάση δεδομένων με 49,722 ταξίδια από τη βάση δεδομένων της OSeven. Για κάθε επιμέρους μέρος της ανάλυσης, αξιοποιήθηκαν διαφορετικά υποσύνολα αυτής της βάσης δεδομένων που επιλέχθηκαν με βάση τις απαιτήσεις κάθε ανάλυσης. Η επιλογές δείγματος για κάθε ανάλυση παρουσιάζονται στον Πίνακα 1.

Πίνακας 1: Οδηγικά δείγματα που χρησιμοποιήθηκαν σε κάθε μέρος της παρούσας έρευνας

	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

Μια εκτεταμένη περιγραφή των στατιστικών χαρακτηριστικών του οδηγού δείγματος που χρησιμοποιείται σε κάθε έναν από τους τρεις τύπους ανάλυσης δεδομένων παρουσιάζεται επίσης για να αποκτηθεί μια σαφής εικόνα του αποκτώμενου δείγματος. Χρησιμοποιήθηκε το σύνολο των 171 οδηγών που συμμετείχαν στο σχεδιασμένο πείραμα και δημιουργήθηκε μια μεγάλη βάση δεδομένων με 49.722 ταξίδια. Όλοι οι οδηγοί που επιλέχθηκαν να συμπεριληφθούν σε αυτό το τμήμα της ανάλυσης θα έπρεπε να έχουν οδηγήσει τουλάχιστον για 10 ώρες και 40 ταξίδια που αντιστοιχούν κατά

προσέγγιση στον τυπικό μηνιαίο αριθμό ταξιδιών για έναν οδηγό υποθέτοντας ότι κάθε οδηγός οδηγεί 2 ταξίδια 15 λεπτά ημερησίως για 5 ημέρες την εβδομάδα. Όσον αφορά την ανάλυση της απόδοσης του ταξιδιού, ένα μέρος του δείγματος ογδόντα οκτώ (88) οδηγών συμμετείχε στο σχεδιασμένο πείραμα που πραγματοποιήθηκε μεταξύ 28/09/2016 και 05/12/2016 και δημιουργήθηκε μια μεγάλη βάση δεδομένων με 10.088 ταξίδια.

Για τους σκοπούς της ανάλυσης αποτελεσματικότητας του οδηγού, τα δεδομένα οδήγησης επιλέχθηκαν από την αρχική βάση δεδομένων των 171 οδηγών βάσει ορισμένων κριτηρίων οδηγού. Το πρώτο κριτήριο που επιλέχθηκε ήταν ότι όλοι οι οδηγοί έπρεπε να έχουν πραγματοποιήσει τουλάχιστον 50 επιπλέον διαδρομές από τον αριθμό των απαιτούμενων διαδρομών ώστε η συνολική απόσταση ανά τύπο δρόμου να είναι τουλάχιστον ίση με την ελάχιστη απόσταση που βρέθηκε στο προηγούμενο στάδιο της ποσοτικής εκτίμησης του δείγματος. Αυτό το κριτήριο έχει οριστεί για να διασφαλιστεί ότι α) τα δεδομένα είσοδού αυξάνονται αναλογικά με τα δεδομένα εξόδου και ως εκ τούτου να είναι έγκυρη η ανάπτυξη ενός μοντέλου DEA σε κάθε βήμα του κινούμενου παραθύρου αλλά και συνολικά και ότι β) ο αριθμός των παρατηρήσεων της χρονοσειράς ικανοποιητικός. Φυσικά, αυτή η διαδικασία επιλογής οδηγών έχει ως στόχο να οδηγήσει στον μέγιστο δυνατό αριθμό οδηγών.

Επιπρόσθετα, όλοι οι οδηγοί θα πρέπει να έχουν διανύσει κάποια απόσταση και στους τρεις τύπους οδικού δικτύου. Το τρίτο κριτήριο ήταν ότι οι οδηγοί με μηδενικό άθροισμα των δεδομένων εισόδου (δηλαδή η απότομη επιτάχυνση, επιβράδυνση, παραβίαση ορίου ταχύτητας και χρήση κινητού τηλεφώνου να μην είναι όλα μηδενικά) πρέπει να εξαλειφθούν από το δείγμα, γεγονός που αποτελεί περιορισμό της DEA. Το επιχειρηματικό ισοδύναμο μηδενικών δεδομένων εισόδου θα μπορούσε να είναι ένα εργοστάσιο που παράγει ένα προϊόν χωρίς να χρησιμοποιεί οποιοδήποτε υλικό ή / και εργατικό δυναμικό, το οποίο πρακτικά δεν μπορεί να συμβεί. Η διαδικασία αυτή οδήγησε σε 100 οδηγούς αστικού οδικού δικτύου και αστικού οδικού δικτύου ταχείας κυκλοφορίας που πληρούσαν τα κριτήρια αυτά και διατηρήθηκαν για την ανάλυση που διενεργήθηκε ενώ οι υπόλοιποι οδηγοί εξαλείφθηκαν από το οδηγικό δείγμα που χρησιμοποιήθηκε. Η κατάργηση των οδηγών οδήγησε σε μόνο 18 οδηγούς σε αυτοκινητόδρομους, ο οποίος θεωρείται ιδιαίτερα χαμηλός αριθμός συμμετεχόντων για την ανάλυση που πρέπει να διεξαχθεί. Ο συνολικός αριθμός των διαδρομών που πραγματοποιήθηκαν από καθέναν από τους οδηγούς που επιλέχθηκαν ήταν 230 για αστικούς και 150 για αγροτικούς δρόμους που δημιούργησαν έτσι μια μεγάλη βάση δεδομένων 23.000 ταξίδια σε αστικές οδούς και 15.000 σε αστικές οδούς ταχείας κυκλοφορίας. Από τους οδηγούς που απάντησαν στο ερωτηματολόγιο που δόθηκε, 43 είχαν οδηγήσει σε αστικές οδούς και 39 σε αστικές οδούς ταχείας κυκλοφορίας. Οι ερωτήσεις του ερωτηματολόγιου παρέχονται και συζητούνται εν συντομία παρακάτω.

Η διερεύνηση της επαρκούς ποσότητας δεδομένων που θα συμπεριληφθεί στην ανάλυση και η εξέλιξη του λόγου μετρήσεων οδήγησης και απόστασης λαμβάνει χώρα ως επόμενο βήμα. Αυτό το βήμα είναι απαραίτητο για να προσδιοριστεί το ακριβές ποσό των δεδομένων που πρέπει να χρησιμοποιούνται στην ανάλυση ώστε να μην είναι ούτε ανεπαρκές ούτε υπερβολικό. Ένα ανεπαρκές ποσό δεδομένων θα οδηγούσε την έρευνα αυτή σε αβέβαια ή παράλογα αποτελέσματα ενώ ένα υπερβολικό ποσό δεδομένων θα αύξανε σημαντικά τον απαιτούμενο χρόνο επεξεργασίας.

Όσον αφορά τον αστικό τύπο οδού, ο αριθμός των απότομων επιταχύνσεων φαίνεται να είναι η πιο κρίσιμη μέτρηση για τον προσδιορισμό της απαιτούμενης απόστασης δειγματοληψίας. Μπορεί να παρατηρηθεί επίσης ότι η μέγιστη τιμή της κατάλληλης απόστασης για τη σύγκλιση της σχετικής μέτρησης στον πίνακα εμφανίζεται για το εκατοστημόριο του ποσοστού 75-100% του αριθμού των απότομων επιταχύνσεων. Η μέγιστη μέση τιμή απόστασης βρίσκεται 519 χιλιόμετρα, η οποία είναι περίπου ίση με 75 ταξίδια σε αστικό τύπο οδού. Αρχικά, υπολογίζεται η μέση απόσταση ανά διαδρομή και κατά συνέπεια ο αριθμός των απαιτούμενων διαδρομών που πρέπει να πραγματοποιήσει κάθε οδηγός για να φτάσει στην απόσταση 519km. Η διάμεση τιμή όλων των χρηστών για αυτή τη μεταβλητή υπολογίζεται ότι είναι περίπου 75. Αυτό είναι το μήκος του κινούμενου παραθύρου που χρησιμοποιείται στην ανάλυση αποτελεσματικότητας του οδηγού για να δημιουργηθούν οι χρονοσειρές απόδοσης οδήγησης σε αστικές οδούς ταχείας κυκλοφορίας. Η διάμεση τιμή προτιμάται αντί της μέσης τιμής για τους ίδιους λόγους που αναφέρονται παραπάνω για τον προσδιορισμό της απαιτούμενης απόστασης δειγματοληψίας.

Όσον αφορά τον αγροτικό δρόμο, ο αριθμός των απότομων επιβραδύνσεων και η χρήση κινητού τηλεφώνου φαίνεται να είναι οι πιο σημαντικές μετρήσεις για τον προσδιορισμό της απαιτούμενης απόστασης δειγματοληψίας. Μπορεί να παρατηρηθεί ότι η μέγιστη τιμή της κατάλληλης απόστασης προς σύγκλιση για το MU στον πίνακα εμφανίζεται για το εκατοστημόριο της κλίμακας 0-25% του αριθού των απότομων επιταχύνσεων. Η μέγιστη διάμεση τιμή της απόστασης υπολογίζεται ότι είναι 579 χιλιόμετρα, η οποία είναι περίπου ίση με 81 ταξίδια σε αστικές οδούς ταχείας κυκλοφορίας. Όπως και στην ανάλυση των αστικών οδών, υπολογίζεται η μέση απόσταση ανά ταξίδι και κατά συνέπεια ο αριθμός των απαιτούμενων διαδρομών σε αστικές οδούς ταχείας κυκλοφορίας που πρέπει να εκτελέσει κάθε οδηγός για να φθάσει την απόσταση των 579 χιλιομέτρων. Η διάμεση τιμή όλων των χρηστών για αυτή τη μεταβλητή εκτιμάται σε περίπου 81. Αυτό είναι το μήκος του κινούμενου παραθύρου που χρησιμοποιείται στην ανάλυση αποτελεσματικότητας του οδηγού για να δημιουργηθούν οι χρονοσειρές απόδοσης οδήγησης σε αστικές οδούς ταχείας κυκλοφορίας. Η διάμεση τιμή προτιμάται αντί της μέσης τιμής για τους ίδιους λόγους που αναφέρονται παραπάνω για τον προσδιορισμό της απαιτούμενης απόστασης δειγματοληψίας.

Τέλος για τους αυτοκινητοδρόμους, ο αριθμός των απότομων επιταχύνσεων και επιβραδύνσεων φαίνεται να είναι οι πιο κρίσιμες μετρήσεις για τον προσδιορισμό της απαιτούμενης απόστασης δειγματοληψίας. Μπορεί να παρατηρηθεί ότι η μέγιστη τιμή της κατάλληλης απόστασης προς σύγκλιση για την χρήση κινητού τηλεφώνου στον πίνακα εμφανίζεται για το εκατοστημόριο του πεδίου τιμών 50-75% του HA. Η μέγιστη διάμεση απόσταση υπολογίζεται ότι είναι 611 χιλιόμετρα, η οποία είναι περίπου ίση με 106 διαδρομές σε αυτοκινητόδρομους. Όπως και στην ανάλυση των αστικών οδών, υπολογίζεται η μέση απόσταση ανά διαδρομή και κατά συνέπεια ο αριθμός των απαιτούμενων διαδρομών σε αστικές οδούς ταχείας κυκλοφορίας που πρέπει να εκτελέσει κάθε οδηγός για να φθάσει την απόσταση των 611χλμ. Η διάμεση τιμή όλων των χρηστών για αυτήν τη μεταβλητή υπολογίζεται ότι είναι περίπου 106. Αυτό το μήκος του κινούμενου παραθύρου που πρέπει να χρησιμοποιηθεί στην ανάλυση αποτελεσματικότητας του οδηγού για να δημιουργηθούν οι χρονοσειρές απόδοσης οδήγησης σε αγροτικό δρόμο. Δυστυχώς, αυτή η τιμή υπερβαίνει τον αριθμό των ταξιδιών

(100) που συλλέγονται για την ανάλυση αποτελεσματικότητας του οδηγού σε αυτοκινητόδρομους και ως εκ τούτου η ανάλυση αυτή δεν μπορεί να πραγματοποιηθεί στον συγκεκριμένο τύπο οδού.

Η έρευνα αυτή καταλήγει στο συμπέρασμα ότι το πρόβλημα της οδηγικής επίδοσης μπορεί να αντιμετωπιστεί ως ένα πρόβλημα σταθερής απόδοσης (CRS) DEA, δεδομένου ότι η απαιτούμενη απόσταση δειγματοληψίας ορίζεται έτσι ώστε το άθροισμα όλων των μετρήσεων (δεδομένων εισόδου) που καταγράφονται για κάθε οδηγό να αλλάζει αναλογικά το άθροισμα της απόστασης οδήγησης (δεδομένων εξόδου) σε κάθε κινούμενο παράθυρο που εξετάστηκε αλλά και συνολικά. Αυτό το βήμα ορίζει επίσης το χρονικό βήμα του κινούμενου παραθύρου και καταλήγει στο συμπέρασμα ότι ο τύπος οδικού άξονα δεν μπορεί να συμπεριληφθεί στην ανάλυση επειδή μόνο ένας μικρός αριθμός συμμετεχόντων έχει καταγραφεί για περισσότερα από τα αντίστοιχα χιλιόμετρα που βρέθηκαν.

Λαμβάνοντας υπόψη την ανασκόπηση της βιβλιογραφίας, τα δεδομένα που συγκεντρώθηκαν και όλες τις ιδιαιτερότητες της τεχνικής DEA, συμπεραίνεται ότι ο δείκτης αποτελεσματικότητας της ασφάλειας μπορεί να οριστεί χρησιμοποιώντας τον αριθμό των απότομων επιταχύνσεων και επιβραδύνσεων, τα δευτερόλεπτα χρήσης του κινητού και τα δευτερόλεπτα οδήγησης πέρα από τα όρια ταχύτητας ως δεδομένα εισόδου και την απόσταση που διανύθηκε ως δεδομένο εξόδου. Αυτό είναι το βήμα-κλειδί που συνδέει την αξιολόγηση του δείκτη αποτελεσματικότητας της ασφάλειας και συγκριτικής αξιολόγησης αυτής της διδακτορικής έρευνας. Αποτελεί ένα σημαντικό βήμα που πρέπει να προηγηθεί της ανάλυσης DEA, καθορίζοντας τα δεδομένα εισόδου και εξόδου της DEA με τέτοιο τρόπο ώστε i) να είναι μια επιστημονικά σωστή διατύπωση της τεχνικής DEA και ii) να αντιπροσωπεύει την αποτελεσματικότητα της οδικής ασφάλειας και συνεπώς την σχετική επικινδυνότητα κατά την οδήγηση.

Στη συνέχεια διεξάγεται ανάλυση απόδοσης σε επίπεδο διαδρομής για τον προσδιορισμό της τεχνικής με την καλύτερη επίδοση από άποψη απαιτούμενου χρόνου από αυτές που δοκιμάστηκαν καθώς επίσης και για την ανάπτυξη της μεθοδολογίας για τον εντοπισμό των λιγότερο αποτελεσματικών διαδρομών που υπάρχουν σε μια συγκεκριμένη βάση δεδομένων ταξιδιού. Τα πρότυπα DEA, RBE DEA και Convex Hull DEA που δοκιμάζονται, συγκρίνονται με βάση τον απαιτούμενο χρόνο επεξεργασίας. Ο αλγόριθμος Convex Hull σε συνδυασμό με την DEA υπερέχει των άλλων δύο μεθοδολογιών που εξετάστηκαν. Αυτό είναι ένα κρίσιμο βήμα που επιτρέπει τη μείωση του απαιτούμενου χρόνου εκτέλεσης για όλα τα επακόλουθα βήματα που εμπλέκονται με τη χρήση της DEA. Τέλος, προτείνεται μια μεθοδολογική προσέγγιση για την λιγότερο αποτελεσματική ταυτοποίηση του ταξιδιού και την αποτελεσματική εκτίμηση των μετρήσεων οδήγησης βάσει του δείκτη αποτελεσματικότητας ασφαλείας που ορίζεται παραπάνω.

Η ανάλυση επίδοσης οδηγού πραγματοποιείται για να εξεταστεί η δυνατότητα της συσταδοποίησης των οδηγών και να εντοπιστούν τα κύρια χαρακτηριστικά οδήγησης κάθε ομάδας. Με βάση το δείκτη επίδοσης της ασφαλείας που ορίζεται, υπολογίζεται για κάθε οδηγό η συνολική επίδοση για τη συνολική περίοδο, μαζί με την επίδοση του οδηγού για το χρονικό παράθυρο κάθε εξεταζόμενου βήματος. Αναλύεται η χρονοσειρά επίδοσης που δημιουργείται και τα αποτελέσματα αξιοποιούνται για την ομαδοποίηση των οδηγών.

Όπως αναφέρθηκε παραπάνω, τα μεγάλης κλίμακας δεδομένα οδήγησης επιλέχθηκαν από την αρχική βάση δεδομένων των 171 οδηγών βάσει ορισμένων κριτηρίων. Το πρώτο κριτήριο που επιλέχθηκε ήταν ότι όλοι οι οδηγοί έπρεπε να έχουν πραγματοποιήσει τουλάχιστον 50 επιπλέον διαδρομές από τον αριθμό των διαδρομών που απαιτούνται, έτσι ώστε η συνολική απόσταση ανά τύπο δρόμου να είναι ασφαλώς υψηλότερη από την ελάχιστη απόσταση που διαπιστώθηκε στο προηγούμενο στάδιο της ποσοτικής εκτίμησης του δείγματος. Αυτή η διαδικασία επιλογής οδηγών στοχεύει επίσης στην επίτευξη του μέγιστου δυνατού αριθμού οδηγών. Επιπρόσθετα, όλοι οι οδηγοί θα πρέπει να έχουν θετική απόσταση σε χιλιόμετρα και στους τρεις τύπους οδικού δικτύου. Εκτός αυτού, οι οδηγοί με μηδενικό άθροισμα των χαρακτηριστικών εισόδου (δηλαδή η απότομη επιτάχυνση, επιβράδυνση, παραβίαση ορίου ταχύτητας, χρήση κινητού τηλεφώνου να είναι όλα μηδενικά) εξαλείφονται από το δείγμα, επειδή πρόκειται για περιορισμό της DEA. Η διαδικασία αυτή οδήγησε σε 100 οδηγούς σε αστικό οδικό δίκτυο και αστικό οδικό δίκτυο ταχείας κυκλοφορίας που πληρούσαν αυτές τις απαιτήσεις και χρησιμοποιήθηκαν στην ανάλυση που πραγματοποιήθηκε ενώ οι υπόλοιποι οδηγοί εξαλείφθηκαν από τη μελέτη αυτή. Η απαλοιφή των οδηγών οδήγησε σε μόνο 18 οδηγούς σε αυτοκινητόδρομους, ο οποίος θεωρήθηκε πολύ μικρός αριθμός συμμετεχόντων για την ανάλυση που πρέπει να διεξαχθεί. Ο συνολικός αριθμός των ταξιδιών που πραγματοποιήθηκαν από καθέναν από τους οδηγούς που επιλέχθηκαν ήταν 230 για αστικό οδικό δίκτυο και 150 για αστικό οδικό δίκτυο ταχείας κυκλοφορίας που δημιούργησαν έτσι μια μεγάλη βάση δεδομένων 23.000 ταξίδια σε αστικό δίκτυο και 15.000 σε αστικό οδικό δίκτυο ταχείας κυκλοφορίας. Από αυτούς τους οδηγούς, απάντησαν στο ερωτηματολόγιο που τους δόθηκε οι 43 που είχαν πραγματοποιήσει διαδρομές σε αστικό οδικό δίκτυο και οι 39 που είχαν πραγματοποιήσει διαδρομές σε αστικό οδικό δίκτυο ταχείας κυκλοφορίας.

Για καθένα από τα 2 δείγματα που χρησιμοποιήθηκαν για την ανάλυση σε επίπεδο οδηγού, η μέση τιμή των χαρακτηριστικών κάθε κατηγορίας εμφανίζεται στον Πίνακα 2 όπου παρουσιάζονται τα μοντέλα ανά αστικό οδικό δίκτυο και αστικό οδικό δίκτυο ταχείας κυκλοφορίας με βάση τα δεδομένα εισόδου που χρησιμοποιήθηκαν σε κάθε μοντέλο. Οι οδηγοί της κλάσης 1 αναφέρονται ως οι πιο αποδοτικοί οδηγοί, παρά το γεγονός ότι μόνο οι οδηγοί με αποδοτικότητα μονάδας βρίσκονται στο όριο αποτελεσματικότητας. οι οδηγοί κατηγορίας 2 και 3 αναφέρονται ως ελαφρώς αποδοτικοί και μη αποδοτικοί οδηγοί.

Πίνακας 2: Οδηγικά χαρακτηριστικά για τις κλάσεις απόδοσης των ομάδων ανά 100 χλμ σε κάθε τύπο οδικού δικτύου και δείγματος

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

Όπως αναμενόταν, για όλα τα μοντέλα οδικών δικτύων και δειγμάτων, το μέσο των χαρακτηριστικών μειώνεται ενώ μετατοπίζεται σε μια κατηγορία υψηλότερης απόδοσης. Η διαφορά μεταξύ των κλάσεων 1 και 2 διαπιστώθηκε ότι είναι λιγότερο σημαντική για το $mobile_{rural}$ και ελαφρώς λιγότερο σημαντική για το $mobile_{urban}$ τόσο για το δείγμα_δεδομένων_1 και το δείγμα_δεδομένων_2. Αυτό το αποτέλεσμα δείχνει ότι οι οδηγοί και των δύο τύπων δρόμων (και ιδιαίτερα του αστικού οδικού δικτύου ταχείας κυκλοφορίας) έχουν παρόμοια συμπεριφορά όσον αφορά τη χρήση κινητών τηλεφώνων και ως εκ τούτου η χρήση κινητού τηλεφώνου δεν αποτελεί κρίσιμο παράγοντα για τη μέτρηση της αποτελεσματικότητας οδήγησης με χρήση DEA. Με άλλα λόγια, το συμπέρασμα που μπορεί να συναχθεί είναι ότι το συνολικό προφίλ ασφάλειας οδήγησης ενός λιγότερο επικίνδυνου οδηγού στο αστικό οδικό δίκτυο και στο αστικό οδικό δίκτυο ταχείας κυκλοφορίας δεν επηρεάζεται σημαντικά από τη χρήση του κινητού τηλεφώνου του οδηγού. Μία πιθανή εξήγηση αυτού του φαινομένου είναι είτε το γεγονός ότι οι οδηγοί όλων των κλάσεων χρησιμοποιούν το κινητό τηλέφωνο περίπου το ίδιο είτε ότι επειδή η DEA είναι ευαίσθητη στις ακραίες τιμές, τα μοντέλα που αναπτύσσονται μπορεί μερικές φορές να επηρεάζονται από τις ακραίες τιμές άλλων δεδομένων εισόδου ή εξόδου κατά την εκτίμηση των παραμέτρων του κάθε DMU π.χ. χαμηλό αριθμό ταχυτήτων ή δευτερόλεπτα χρήσης κινητής τηλεφωνίας. Και στις δύο περιπτώσεις όμως, η διάσπαση του κινητού τηλεφώνου πρέπει να εξεταστεί χωριστά. Οι κύριοι παράγοντες που επηρεάζουν τη μετατόπιση από τη μια κατηγορία στην άλλη προσδιορίζονται επίσης και προτείνεται μια μεθοδολογία για την εκτίμηση του αποτελεσματικού επιπέδου των οδηγικών μετρικών που πρέπει να φτάσει κάθε οδηγός για να καταστεί αποτελεσματική.

Η συνολική επίδοση και η μεταβλητότητα εκτιμώνται επίσης σε αυτό το στάδιο, το οποίο θα χρησιμοποιηθεί στη διαδικασία συσταδοποίησης.

Παρατηρώντας τη μετατόπιση μεταξύ των κατηγοριών απόδοσης σε αστικό δρόμο, η μετάβαση από την κλάση 1 σε 2 επηρεάζεται κυρίως από τον αριθμό των απότομων επιταχύνσεων και επιβραδύνσεων και ελάχιστα από τη χρήση κινητού τηλεφώνου, ενώ ο χρόνος οδήγησης πάνω από τα όρια ταχύτητας παίζει σημαντικότερο ρόλο από την κλάση 2 έως 3. Όσον αφορά το αστικό οδικό δίκτυο ταχείας κυκλοφορίας, η ταχύτητα επηρεάζει κυρίως τη μετάβαση από την κλάση 1 σε 2, ενώ ο αριθμός των απότομων επιταχύνσεων και η χρήση κινητής τηλεφωνίας επηρεάζουν κυρίως τη μετάβαση από την κλάση 2 στην 3. Η αξιοσημείωτη διαφορά στο data_sample_2 είναι ότι στην κλάση 1 και 2 των οδηγών είναι υψηλότερη και υψηλότερη από την κατηγορία 1 και 3 των οδηγών και ελαφρώς υψηλότερη στην κατηγορία 2 των οδηγών. Αυτή η διαφορά μπορεί να οφείλεται στον μικρότερο αριθμό συμμετεχόντων στο data_sample_2 και επομένως πρέπει να αξιοποιηθεί ένα μεγαλύτερο δείγμα για περαιτέρω διερεύνηση.

Όσον αφορά τη μετατόπιση μεταξύ κλάσεων απόδοσης στην αστική οδό, η μετάβαση από την κλάση 1 σε 2 επηρεάζεται κυρίως από τον αριθμό των απότομων επιταχύνσεων και επιβραδύνσεων και ελάχιστα από το χρόνο οδήγησης πάνω από τα όρια ταχύτητας, ενώ η χρήση κινητού τηλεφώνου παίζει σημαντικότερο ρόλο από την κλάση 2 έως 3. Ο αριθμός των απότομων επιταχύνσεων και επιβραδύνσεων είναι επίσης οι παράγοντες που επηρεάζουν περισσότερο τις μετατοπίσεις μεταξύ κλάσεων 1 και 2 στον αστικό τύπο οδού ταχείας κυκλοφορίας, ενώ η μετάβαση από την κλάση 2 σε 3 επηρεάζεται κυρίως από τη χρήση κινητών τηλεφώνων. Ο συντελεστής ταχύτητας φαίνεται να επηρεάζει εξίσου και τις δύο μετατοπίσεις από την κλάση 1 έως 2 και από την κλάση 2 έως 3. Είναι πολύ σημαντικό να προσδιοριστούν αυτές οι παράμετροι προκειμένου να παρασχεθούν σημαντικές στοχευμένες μακροσκοπικές συστάσεις για περαιτέρω βελτίωση σε μια συγκεκριμένη ομάδα οδηγών, η αποτελεσματικότητα είναι γνωστή.

Το σχήμα 2 είναι ένα ιστόγραμμα που δείχνει τον αριθμό των οδηγών του δείγματος_οδηγών_1 που εμπίπτουν σε κάθε εύρος αποτελεσματικότητας σε αστικό οδικό δίκτυο και αστικό οδικό δίκτυο ταχείας κυκλοφορίας. Είναι εμφανές ότι η απόδοση οδήγησης των περισσότερων οδηγών είναι συγκεντρωμένη στο εύρος απόδοσης 0,2-0,6 και στους δύο οδικούς τύπους και ότι μόνο ένα μικρό ποσοστό των οδηγών εμπίπτει στην ανώτερη και κατώτερη κλίμακα. Αυτό πιθανότατα οφείλεται στο γεγονός ότι οι πιο αποδοτικοί οδηγοί έχουν σημαντικά υψηλότερη απόδοση και κατά συνέπεια η απόδοση των υπόλοιπων οδηγών εκτιμάται ότι είναι σχετικά χαμηλότερη. Από την άλλη πλευρά, δεδομένου ότι η απόδοση της οδήγησης εκτιμάται σχετικά, μπορεί να συναχθεί το συμπέρασμα ότι οι οδηγοί του εύρους απόδοσης 0,4-0,6 είναι πιθανό να αποτελέσουν οδηγούς μέσου κινδύνου. Όσον αφορά το εύρος απόδοσης 0,2-0,4, η ομάδα αυτή πρέπει να διερευνηθεί περαιτέρω για να γίνει διάκριση μεταξύ μέσου και υψηλού κινδύνου οδηγών.
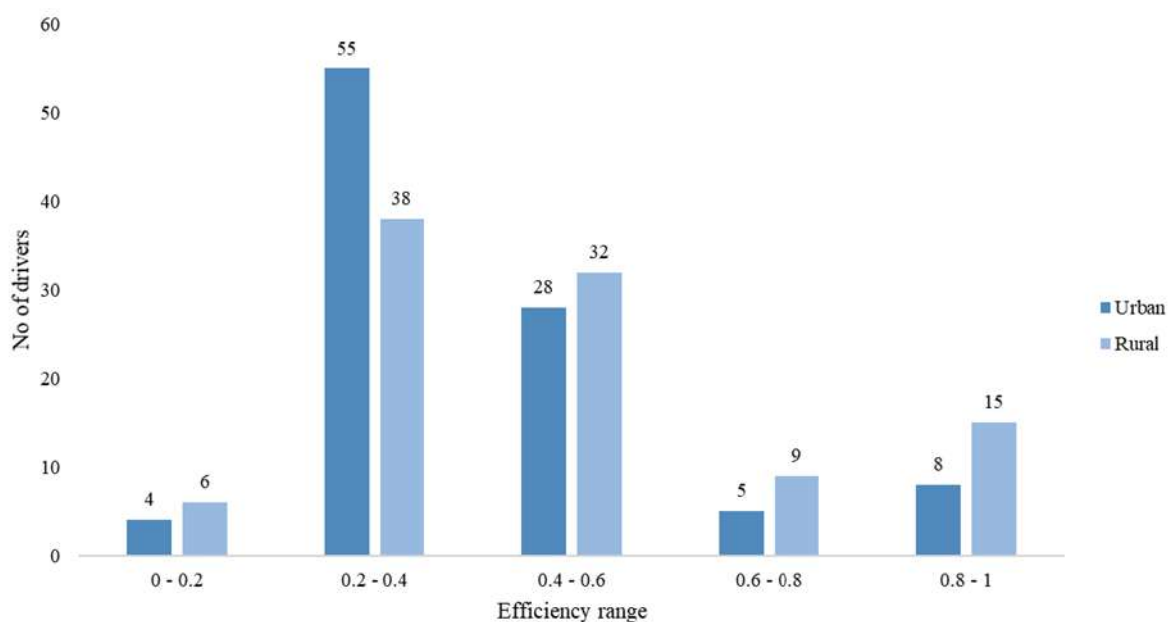
Σχήμα 2: Ο αριθμός των οδηγών του δείγματος 1 που βρίσκονται σε κάθε εύρος αποδοτικότητας για αστικό οδικό δίκτυο και αστικό οδικό δίκτυο ταχείας κυκλοφορίας

Το Σχήμα 3 είναι ένα ιστόγραμμα που δείχνει τον αριθμό των οδηγών του δείγματος_οδηγών_2 που εμπίπτουν σε κάθε περιοχή απόδοσης σε αστικό οδικό δίκτυο και αστικό οδικό δίκτυο ταχείας κυκλοφορίας. Προφανώς, τα σχήματα 2 και 3 έχουν παρόμοια χαρακτηριστικά όσον αφορά τη κατανομή των οδηγών. Και πάλι, οι περισσότεροι οδηγοί είναι συγκεντρωμένοι στο εύρος απόδοσης 0,2-0,6 και στους δύο οδικούς τύπους και μόνο ένα μικρό ποσοστό των οδηγών εμπίπτει στην ανώτερη και κατώτερη σειρά. Η διαφορά που παρατηρήθηκε στο σχήμα 3 είναι ότι υπάρχει ένας μεγαλύτερος αριθμός οδηγών στην περιοχή 0.8-1 πιθανώς λόγω της μειωμένης αναλογίας του αθροίσματος των εισόδων και εξόδων προς τον αριθμό των συμμετεχόντων, μία παράμετρο που αναφέρθηκε προηγουμένως λεπτομερώς. Γενικά, η εξέταση μιας διαφορετικής κατηγοριοποίησης (π.χ. τάξη 1: 0-33,3%, κλάση 2: 33,3-66,6%, κλάση 3: 66,6-100%) προτείνεται για έρευνα στο μέλλον.
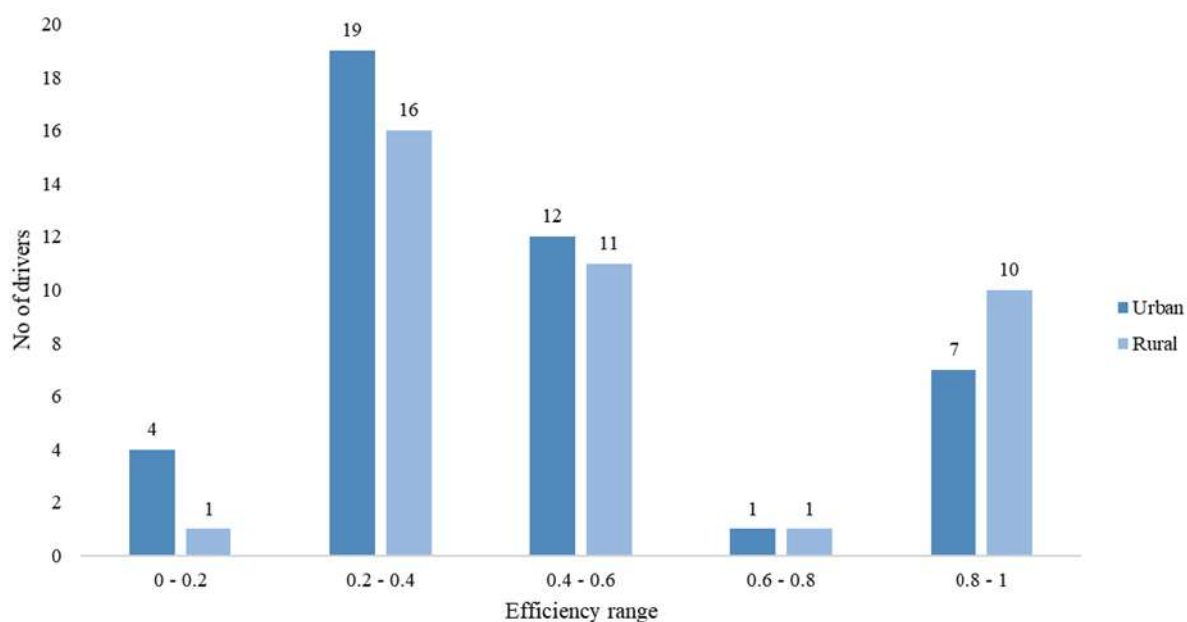
Σχήμα 3: Ο αριθμός των οδηγών του δείγματος 2 που βρίσκονται σε κάθε εύρος αποδοτικότητας για αστικό οδικό δίκτυο και αστικό οδικό δίκτυο ταχείας κυκλοφορίας

Ο Πίνακας 3 δείχνει τα λάμδα και την θήτα για τους πρώτους δώδεκα (έντεκα μη αποδοτικούς οδηγούς συν έναν αποτελεσματικό οδηγό) οδηγούς των μοντέλων DEA στον αστικό δρόμο, όπου αντιπροσωπεύει το συντελεστή λάμδα του αποδοτικού που λειτουργεί ως κριτής για τον οδηγό που εξετάζεται κάθε φορά. Για λόγους συντομίας, δεν παρουσιάζονται όλα τα λάμδα και οι υπολογισμοί. Για παράδειγμα, στην πρώτη σειρά του πίνακα όπου αναπτύσσεται το μοντέλο της DEA, η τιμή του συντελεστή θήτα είναι 0.581 (λιγότερο αποτελεσματική) και οι συντελεστές λάμδα, είναι ίσοι με 0.52, 0.14, 0.06 αντίστοιχα. Το αποδοτικό επίπεδο εισόδων για το μπορεί να υπολογιστεί ως το άθροισμα των προϊόντων των λάμδα και των τιμών των δεδομένων εισόδου του καθενός από τους οδηγούς που έχουν αναγνωριστεί ως κριτές, ενώ για να βρεθεί το αποτελεσματικό επίπεδο των δεδομένων εξόδου για τον ίδιο οδηγό, κάθε τιμή εξόδου θα πρέπει να διαιρείται με την τιμή θήτα.

28

Πίνακας 3: Λάμδα, θήτα, πραγματικό και αποδοτικό επίπεδο των οδηγικών χαρακτηριστικών (απόσταση(χλμ), απότομες επιταχύνσεις, απότομες επιβραδύνσεις, υπέρβαση του ορίου ταχύτητας (δ), χρήση κινητού τηλεφώνου (δ)), σε αστικό τύπο οδού για τους πρώτους 12 οδηγούς (DMUs)

Driver No	Real level of metrics					Theta	Lamdas of peers: Driver No				Efficient level of metrics				
	distance$_{urban}$	ha$_{urban}$	hb$_{urban}$	speeding$_{urban}$	mobile$_{urban}$		12	34	40	42	distance$_{urban}$	ha$_{urban}$	hb$_{urban}$	speeding$_{urban}$	mobile$_{urban}$
1	1868	326	134	21712	9954	0.581	-	0.52	0.14	0.06	3214.3	159.8	77.9	12617.9	5784.7
2	2456	574	85	27049	13974	0.696	-	0.19	0.40	0.20	3526.5	154.8	59.2	18838.2	9732.1
3	1634	709	509	15888	42817	0.391	0.79	-	0.20	-	4182.6	277.0	78.1	6206.9	10434.2
4	2219	233	181	27052	12421	0.637	-	0.23	0.31	0.18	3481.0	148.5	59.1	17244.6	7917.9
5	4223	1088	309	37825	29581	0.613	0.30	1.16	0.47	-	6887.3	417.6	189.5	23192.7	18137.9
6	2773	652	251	25829	25887	0.529	0.60	0.51	0.30	-	5245.3	344.7	128.8	13654.8	13685.4
7	2086	265	149	20880	14036	0.619	-	0.52	0.28	0.04	3371.9	163.9	79.2	12917.2	8683.2
8	1789	460	323	17185	5960	0.656	-	0.75	-	0.01	2728.0	184.3	98.1	11269.8	3908.5
9	1630	184	82	30801	8955	0.528	-	-	0.21	0.22	3085.0	97.2	30.8	14784.6	4731.5
10	808	266	95	9985	6562	0.443	0.11	0.24	0.04	-	1824.6	97.2	42.1	4421.6	2905.8
11	3012	913	152	21604	43562	0.585	0.72	-	0.83	-	5149.6	308.3	88.9	12636.1	23107.5
12	1462	329	92	5074	7781	1.000	1.00	-	-	-	1462.0	329.0	92.0	5074.0	7781.0

* Inputs = ['ha$_{urban}$', 'hb$_{urban}$', 'speeding$_{urban}$', 'mobile$_{urban}$'], Output = ['distance$_{urban}$']

Η εξέλιξη της μέσης απόδοσης οδήγησης με την πάροδο του χρόνου διερευνάται επίσης με τη χρήση διαφορετικών βάσεων δεδομένων που συγκεντρώνονται σε διαφορετικά χρονικά πλαίσια από την αρχή του χρόνου καταγραφής έως το τέλος. Η χρονοσειρά των αποτελεσμάτων μελετάται και αποσυντίθεται στα κύρια συστατικά της, την στασιμότητα και την τάση. Η μέση τάση παρατηρείται ότι είναι περίπου η ίδια μεταξύ των δύο τύπων δρόμων του δείγματος_οδηγών_1 παρά το γεγονός ότι η διάμεση τιμή της τάσης αποκλίνει. Αυτό υποδηλώνει την ύπαρξη υψηλών τιμών απομειωμένων τάσεων σε αστικές οδούς και χαμηλών τιμών εκτός τάσης στον αστικό τύπο δικτύου ταχείας κυκλοφορίας που επηρεάζουν τη μέση τιμή της τάσης. Όσον αφορά την στασιμότητα των χρονοσειρών, ο αριθμός των διαφορών που απαιτούνται για να γίνει στάσιμη μια χρονοσειρά αποκαλύπτει ότι δεν υπάρχουν χρήστες σε αστικές οδούς των οποίων η συμπεριφορά οδήγησης να είναι στάσιμη. Από την άλλη πλευρά, ο σχετικός αριθμός σε αστικές οδούς ταχείας κυκλοφορίας είναι χαμηλός για το δείγμα_οδηγών_1 αλλά σημαντικά υψηλότερο για το δείγμα_οδηγών_2.

Χρησιμοποιώντας έναν αλγόριθμο μηχανικής μάθησης k-mean, η συσταδοποίηση των οδηγών πραγματοποιείται στη συνέχεια βάσει της συνολικής απόδοσης οδήγησης, της μεταβλητότητας, της τάσης, της στασιμότητας των χροσειρών που προκύπτουν, καθώς και των δεδομένων ερωτηματολογίου που συλλέγονται από το δείγμα_οδηγών _2. Τα ερωτήματα σχετικά με τον αριθμό των ετών της οδηγικής εμπειρίας και τον αριθμό των συνολικών ατυχημάτων μέχρι σήμερα ήταν τα δεδομένα ερωτηματολογίου που ελήφθησαν υπόψη στην προσέγγιση της συσταδοποίησης. Αυτά τα δύο ερωτήματα συνδυάστηκαν σε μία μεταβλητή που αντιπροσωπεύει το συνολικό αριθμό ατυχημάτων ανά δεκαετία οδήγησης. Τα χαρακτηριστικά οδήγησης κάθε συστάδας αναλύονται και παρουσιάζονται τα συμπεράσματα. Προκειμένου να αποφευχθεί η επίδραση των αποτελεσμάτων από τις ακραίες τιμές, όλες οι μεταβλητές κανονικοποιούνται πριν χρησιμοποιηθούν ως είσοδοι στον αλγόριθμο ομαδοποίησης k-mean. Ο βέλτιστος αριθμός συστάδων προσδιορίζεται χρησιμοποιώντας τη μέθοδο του αγκώνα.

Πίνακας 4: Ποιοτικά χαρακτηριστικά των συστάδων οδηγών

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

Η ανάλυση συσταδοποίησης οδήγησε σε τρεις ομάδες οδήγησης, οι οποίες αντιπροσωπεύουν κυρίως τους τυπικούς οδηγούς, τους ασταθείς οδηγούς και τους επιφυλακτικούς οδηγούς. Το κύριο κοινό χαρακτηριστικό μεταξύ όλων των ομάδων προσεκτικών οδηγών είναι ο υψηλός δείκτης απόδοσης της οδήγησης και η χαμηλή τιμή της αξίας του ατυχήματος ανά έτος ανεξάρτητα από το αν συμπεριλήφθηκε ή όχι ως παράγοντας στην ανάλυση συσταδοποίησης. Από την άλλη πλευρά, όλες οι συστάδες των μέσων οδηγών χαρακτηρίζονται από υψηλό δείκτη οδηγικής επίδοσης και χαμηλή θετική τάση που δείχνει μια σταθερά κακή οδηγική συμπεριφορά. Τέλος, οι ασταθείς οδηγοί του δεύτερης συστάδας παρουσιάζουν μεσαία έως υψηλή μεταβλητότητα, η οποία θεωρείται ότι είναι το μόνο κοινό χαρακτηριστικό μεταξύ τους. Τα υπόλοιπα σύνολα παρουσιάζουν παρόμοια χαρακτηριστικά σε όλα τα χαρακτηριστικά. Τα αποτελέσματα της διαδικασίας συσταδοποίησης συνοψίζονται στον Πίνακα 3.

Τέλος, η πληροφορία σχετικά με τα ατυχήματα κατά την οδήγηση φαίνεται να επηρεάζουν μόνο τη μορφή της δεύτερης συστάδας των πιο ασταθών οδηγών, η οποία ενσωματώνει οδηγούς με χαμηλότερη επίδοση και ασταθείς. Αυτό είναι εξαιρετικά ελπιδοφόρο για τη βιβλιογραφία συμπεριφοράς οδήγησης, καθώς υποδηλώνει ότι είναι εφικτό να μελετηθούν μαζικά ανώνυμα σύνολα δεδομένων για τα οποία δεν παρέχονται προσωπικά δεδομένα και παράγονται εξίσου σημαντικά και όχι προκατειλημμένα αποτελέσματα.

Αυτή η διατριβή καταλήγει στο συμπέρασμα ότι η μεθοδολογική προσέγγιση για τον προσδιορισμό της απαιτούμενης απόστασης δειγματοληψίας οδήγησης εξαρτάται από το πεδίο εφαρμογής της ερευνητικής μεθοδολογίας που θα εφαρμοστεί. Με άλλα λόγια, οι στατιστικές αρχές της μεθοδολογικής προσέγγισης που θα αξιοποιήσει τα δεδομένα οδήγησης που συλλέγονται, καθορίζουν τους στατιστικούς κανόνες που θα χρησιμοποιηθούν για την εκτίμηση της ποσότητας των απαιτούμενων δεδομένων.

Ένα εξίσου σημαντικό συμπέρασμα είναι ότι η επαρκής ποσότητα δεδομένων οδήγησης μειώνεται καθώς αυξάνονται οι μετρήσεις οδήγησης (για όλες τις μετρήσεις εκτός από την ταχύτητα), τουλάχιστον για αστικούς δρόμους ταχείας κυκλοφορίας και αυτοκινητόδρομους. Αυτό σημαίνει ότι όσο πιο επιθετικός γίνεται ο οδηγός, τόσο λιγότερη παρακολούθηση απαιτείται για να αποκτηθεί μια σαφής εικόνα για τα πρότυπα οδήγησης του.

Ένα εξίσου σημαντικό συμπέρασμα είναι ότι η επαρκής ποσότητα δεδομένων οδήγησης μειώνεται καθώς αυξάνονται οι μετρήσεις οδήγησης (για όλες τις μετρήσεις εκτός από την ταχύτητα), τουλάχιστον για αγροτικούς δρόμους και αυτοκινητόδρομους. Αυτό σημαίνει ότι όσο πιο επιθετικός γίνεται ο οδηγός, τόσο λιγότερη παρακολούθηση απαιτείται για να αποκτηθεί μια σαφής εικόνα για τα πρότυπα οδήγησης του. Τα αποτελέσματα κατέδειξαν επίσης ότι ένας διαφορετικός τύπος χαρακτηριστικών οδήγησης είναι κρίσιμος για τον προσδιορισμό της απαιτούμενης ποσότητας δεδομένων που πρέπει να καταγράφονται σε κάθε τύπο οδού. Επιπλέον, φαίνεται ότι πρέπει να συλλεχθεί ένα διαφορετικό ποσό δεδομένων ανάλογα με τον τύπο οδού που εξετάστηκε.

Τα αποτελέσματα δείχνουν επίσης ότι ο προτεινόμενος αλγόριθμος DEA σε συνδυασμό με το Convex Hull έχει σημαντικά καλύτερη απόδοση για δεδομένα μεγάλης κλίμακας σε σύγκριση με άλλες υπάρχουσες αλγοριθμικές μεθοδολογίες DEA όπως οι μεθοδολογίες Standard και RBE DEA. Μια άλλη σημαντική συμβολή αυτής της έρευνας είναι ότι προτείνει μια νέα προσέγγιση για την αξιολόγηση της αποτελεσματικότητας οδήγησης ενός ταξιδιού. Η μεθοδολογία παρέχει τη δυνατότητα για την εκτίμηση του δείκτη αποτελεσματικότητας του ταξιδιού, τον προσδιορισμό των "κριτών" του, και ως εκ τούτου, καθορίζει το αποτελεσματικό επίπεδο δεδομένων εισόδου και εξόδου. Τέλος, παρέχονται τα μεθοδολογικά βήματα για την ταυτοποίηση των λιγότερο αποτελεσματικών ταξιδιών σε μια βάση δεδομένων, κάτι που θα ήταν πολύτιμο εύρημα για ένα σύστημα οδηγιών προς το χρήστη.

Ένα άλλο σημαντικό χαρακτηριστικό της ανάλυσης που διενεργήθηκε για κάθε κατηγορία είναι ότι υπάρχουν σημαντικές διαφορές στα χαρακτηριστικά οδήγησης μεταξύ των κατηγοριών των αναποτελεσματικών οδηγών και κατηγοριών ελαφρώς αποδοτικών και αποδοτικότερων οδηγών, με τη διαφορά των δύο τελευταίων να είναι λιγότερο σημαντικές. Από την άλλη πλευρά, η χρήση κινητής τηλεφωνίας δεν θεωρείται κρίσιμος παράγοντας στην εκτίμηση της αποτελεσματικότητας της ασφάλειας, πιθανότατα επειδή η DEA παρέχει σχετική εκτίμηση της αποδοτικότητας οδήγησης και ταυτόχρονα δεν διαπιστώθηκε σημαντική διαφορά στην χρήση κινητών τηλεφώνων μεταξύ διαφορετικών κατηγοριών. Ένα άλλο πολύ σημαντικό εύρημα είναι ότι η μετατόπιση μεταξύ κατηγοριών αποδοτικότητας επηρεάζεται κυρίως από διαφορετικά χαρακτηριστικά οδήγησης σε αστικό τύπο δικτύου και αστικό τύπο δικτύου ταχείας κυκλοφορίας.

Διερευνάται επίσης η χρονική εξέλιξη της αποτελεσματικότητας οδήγησης και καθορίζεται το χρονικό παράθυρο κίνησης στο οποίο αξιολογείται κάθε οδηγός. Αποδεικνύεται ότι παρά το γεγονός ότι οι οδηγοί διατηρούν μια σταθερή οδηγική συμπεριφορά για μια ορισμένη χρονική περίοδο, υπάρχουν σημαντικές δυναμικές μεταβολές στη συστηματική συμπεριφορά σε μια μακροπρόθεσμη περίοδο. Επιπλέον, παρατηρείται περίπου η ίδια μέση τάση μεταξύ των δύο οδικών τύπων παρά το γεγονός ότι η μέση τάση διαφοροποιείται σημαντικά. Τέλος, η μελέτη της στασιμότητας έδειξε ότι τρεις στις

τέσσερις ομάδες οδηγών έχουν παρόμοια χαρακτηριστικά και ως εκ τούτου δεν θα διαδραματίσει σημαντικό ρόλο στην τελική διαδικασία συσταδοποίησης όπου αυτό το χαρακτηριστικό δεν λαμβάνεται υπόψη.

Η ανάλυση συσταδοποίησης που πραγματοποιήθηκε οδήγησε σε τρεις συστάδες οδήγησης, οι οποίες αντιπροσωπεύουν κυρίως τους τυπικούς οδηγούς, τους ασταθείς οδηγούς και τους επιφυλακτικούς οδηγούς. Το κύριο κοινό χαρακτηριστικό μεταξύ όλων των ομάδων προσεκτικών οδηγών είναι ο υψηλός δείκτης απόδοσης της οδήγησης και η χαμηλή τιμή της αξίας του ατυχήματος ανά έτος ανεξάρτητα από το αν συμπεριλήφθηκε ή όχι ως παράγοντας στην ανάλυση συσταδοποίησης. Από την άλλη πλευρά, όλες οι συστάδες των μέσων οδηγών χαρακτηρίζονται από υψηλό δείκτη αποτελεσματικότητας οδήγησης και ασήμαντη χαμηλή θετική τάση που δείχνει μια σταθερά κακή οδηγική συμπεριφορά. Τέλος, οι ασταθείς οδηγοί της δεύτερης συστάδας παρουσιάζουν μεσαία έως υψηλή μεταβλητότητα, η οποία θεωρείται ότι είναι το μόνο κοινό χαρακτηριστικό μεταξύ τους. Τέλος, οι προηγούμενες πληροφορίες σχετικά με τα δεδομένα σχετικά με τα ατυχήματα κατά την οδήγηση φαίνεται να επηρεάζουν μόνο τη μορφή της δεύτερης συστάδας των πιο ασταθών οδηγών, η οποία ενσωματώνει οδηγούς που είναι λιγότερο αποτελεσματικοί και ασταθείς ως προς στην ασφάλεια.

Αυτή η διδακτορική διατριβή συμβάλλει στην κατανόηση της συγκριτικής αξιολόγησης της αποτελεσματικότητας της οδήγησης και συνεπώς στην κατανόηση της οδηγικής επικινδυνότητας, χρησιμοποιώντας τεχνικές επιστήμης δεδομένων που εφαρμόζονται σε μεγάλης κλίμακας δεδομένα με τη μορφή μετρήσεων διαδρομής και οδηγικής συμπεριφοράς που συλλέγονται από κάθε διαδρομή σε επίπεδο οδηγού. Μια νέα μεθοδολογική προσέγγιση παρέχεται επίσης για την εκτίμηση του αποτελεσματικού επιπέδου δεδομένων εισόδου και εξόδου που πρέπει να φτάσει κάθε οδηγός ώστε να καταστεί αποδοτικός από πλευράς ασφάλειας. Είναι επίσης πολύ σημαντικό ότι αυτή η έρευνα αναγνωρίζει τα κύρια χαρακτηριστικά των ομάδων απόδοσης της οδικής ασφάλειας που προκύπτουν από τη βελτιωμένη μεθοδολογία DEA που χρησιμοποιείται, διότι αυτό θέτει το έδαφος για την εις βάθος μελέτη της αποτελεσματικότητας οδήγησης με βάση μικροσκοπικά χαρακτηριστικά οδήγησης. Τέλος, η έρευνα αυτή μελετά τη χρονική εξέλιξη της οδηγικής επίδοσης και αποκαλύπτει τα χαρακτηριστικά των προφίλ οδήγησης που παρατηρούνται.

Αυτή η εργασία ασχολείται επίσης με το πρόβλημα των τεχνικών επιστήμης των δεδομένων που μπορούν να εφαρμοστούν σε πραγματικά προβλήματα μεταφορών όπως αυτό που εξετάστηκε, για να αντιμετωπιστεί το πρόβλημα που οδήγησε στη συγκριτική αξιολόγηση της αποτελεσματικότητας χρησιμοποιώντας την DEA. Συνεπώς, εξετάζονται εδώ οι χρονικές επιδόσεις της μεθοδολογίας DEA για μεγάλης κλίμακας δεδομένα καθώς και η δυνατότητα εφαρμογής βελτιωμένης προσέγγισης DEA με ορισμένες τεχνικές (RBE, Convex Hull) για την επίτευξη της ίδιας βέλτιστης λύσης σε λιγότερο χρόνο. Επιπλέον, τα μεγάλης κλίμακας δεδομένα οδήγησης που συλλέγονται διερευνώνται μέσω στατιστικών μεθόδων, προκειμένου να προσδιοριστεί το συγκεκριμένο ποσό δεδομένων που πρέπει να συλλέγονται για κάθε οδηγό σε κάθε τύπο οδού. Η ανάγκη προσδιορισμού αυτού του ποσού προκύπτει από το γεγονός ότι η συλλογή υπερβολικών ή ελλιπών δεδομένων οδήγησης μπορεί να είναι επικίνδυνη, διότι μπορεί να οδηγήσει σε υπερβολική υπολογιστική προσπάθεια όταν πρόκειται για δεδομένα μεγάλης κλίμακας ή για μη σημαντικά συμπεράσματα αντίστοιχα.

Ο συνδυασμός των δύο τελευταίων προσεγγίσεων συνδυάζονται με την καινοτομία αυτής της διδακτορικής έρευνας όσον αφορά τη διαχείριση δεδομένων μεγάλης κλίμακας. Αυτή η διδακτορική έρευνα παρουσιάζει έναν τρόπο μείωσης της διάστασης ενός προβλήματος χρησιμοποιώντας δεδομένα μεγάλης κλίμακας και αντλεί πολύτιμα συμπεράσματα από αυτά. Η μελέτη αυτή παρέχει επίσης μεθοδολογικά βήματα για την εκτίμηση του αποτελεσματικού επιπέδου μετρήσεων για μια διαδρομή και της προσέγγισης για τον εντοπισμό των λιγότερο αποδοτικών διαδρομών σε μια βάση δεδομένων.

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.

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 benchmarking of a trip's driving efficiency. 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 benchmarking 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 **benchmarked** 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.

Chapter 1: Introduction

1.1) Overview

Road safety is a complex scientific field with far reaching implications to societies and industry field and has systematically attracted the interest of researchers and practitioners. Road accidents impose serious problems to society in terms of human costs, economic costs, property damage costs and medical costs. It is one of the most important concerns in modern societies around the globe since they result to approximately 1.25 million fatalities (around 3,400 per day) and between 20 to 50 million non-fatal injuries per year (WHO 2015). The impact of these injuries and fatalities on the families and the communities in which these people lived and worked is immeasurable. Road traffic crashes are extremely costly for most countries at a national level and particularly to developing economies ranging from 1–2 % for low- and middle-income countries, estimated at over US$ 100 billion a year (Jacobs 2000), to 3% of their gross domestic product (WHO 2015). Luckily, a steadily decreasing trend in road traffic casualties is noticed during the last few years. Nonetheless, the number of road fatalities in several countries and in Greece remains in unacceptably high levels which demonstrates the need for greater efforts and improvements in all aspects of road safety including driving behaviour and performance (OECD 2013).

According to World Health Organization (WHO 2015), road accidents are estimated to be the tenth leading cause of fatalities globally and the leading cause of fatalities among people aged between 15 and 29 years old (nearly 400 000 young people per year). It is also a fact that people aged between 15 and 44 years account for 48% of global road traffic deaths. From a socioeconomic perspective, more than 90% of these fatalities and injuries occur in low- and middle-income countries and the highest rates are found in Africa and the Middle East region. However, even within high-income countries, people from lower socioeconomic backgrounds are more likely to be involved in road traffic crashes. These are probably the effects of rapid motorization which has not been accompanied sufficiently by improved road safety strategies and operations. While many strategies could be applied to reduce the need for road travel, and thus the exposure to crash risk, in practice the combination of increased road transportation and better and continuous forms of communication may be detrimental to the global road safety picture (WHO 2013). In terms of sex, males are more likely to be involved in road traffic crashes than females. Approximately 73% of all road traffic deaths occur among young males under the age of 25 years who are almost 3 times as likely to be killed in a road traffic crash as young females.

A significant number of risk factors that affect the probability of participating in a road traffic accident have been identified in literature. It is a matter of great importance to limit the number of these risk factors in order to succeed in the efforts of reducing road traffic injuries. Among others, the most important risk factors recognized in literature (WHO 2015, Elvik 2004, Pedden et al. 2004) are human factors (speeding, distracted driving, driving under the influence of alcohol and other psychoactive substances etc.), unsafe

road infrastructure, unsafe vehicles and inadequate law enforcement of traffic laws. Human factors are considered to be one of the main causes of road traffic fatalities and injuries every year and therefore it is highly important to study how these factors can affect traffic risk. As a result, it would be valuable to quantify the influence of driving behaviour on crash risk at least relatively to the rest of the drivers' population.

Regarding mobile phone usage while driving, literature has shown that it has a significant effect on driver behaviour. Cell phone use causes drivers to have higher variation in accelerator pedal position, drive more slowly with more variation in speed and report a higher level of workload (Md Mazharul and Washington 2015) regardless of conversation difficulty level. Drivers tend to select larger vehicle spacing (Nilsson 1982), and longer time headways (Mohammad et al. 2015) suggesting possible risk compensatory behavior (Md Mazharul and Washington 2015, Törnros and Bolling 2006). Furthermore, the participants' reaction times (Patten et al., 2004) increase significantly when conversing, but no benefit of hands-free units over handheld units on rural roads/motorways were found (Handel et al., 2014, Yannis et al., 2014). Thus, with regards to mobile telephones, the content of the conversation was far more important for driving and driver distraction.

Speeding is also recognized as one of the most important factors in driving risk since it influences the accident probability (e.g. decreased reaction distance, loss of control) as well as the crash impact. According to (OECD, 2006) speeding has been a contributory factor in 10% of the total accidents and more than 30% in fatal accidents. According to (Andersson and Nilsson 1997, Nilsson 1982) the probability of a crash involving an injury is proportional to the square of the speed, the probability of a serious crash is proportional to the cube of the speed and the probability of a fatal crash is related to the fourth power of the speed. Moreover, (Nilsson 2004) depicts the relationship between speed and driving risk via an exponential curve, showing that the driving risk is not proportional to the speed.

Harsh acceleration, harsh breaking and harsh cornering events are three significant indicators for driving risk assessment (Derick and Trivedi 2011, Bonsall et al. 2005) especially for evaluating driving aggressiveness. This is because they are strongly correlated with unsafe distance from adjacent vehicles, possible near misses, lack of concentration, increased reaction time, poor driving judgement or low level of experience and involvement in situations of high risk (e.g. marginal takeovers). The correlation between HA and HB events with driving risk has been highlighted in the scientific papers published by (Tselentis et al., 2017, Bonsall et al. 2005) and it has been widely recognized by the insurance and telematics industry.

Measuring driving efficiency has been the focus of many studies in driving behaviour literature in the past (Gerald et al. 1996, Gerald et al. 1998, Young et al. 2011). From a road safety perspective, it is extremely significant to identify the parameters that influence driving behaviour and therefore traffic risk. To achieve this, it is extremely important to study driving behaviour on a microscopic level a fact that necessitates the driver behaviour recording. Several studies have been carried out regarding mobile phone usage distraction and methodologies for collecting and analysing (Tselentis et al. 2017) driving behaviour data. The most common methodology applied to date, include driving simulators (Desmond et al. 1998, Lenné et al. 1997), questionnaires (Gerald et al. 1998)

combined with simulators and naturalistic driving experiments (Toledo et al. 2008, Birell et al. 2014).

The most common method for monitoring driving measures necessary for behaviour evaluation include recorders that relate to the car engine (Zaldivar et al 2011, Backer-Grøndahl et al. 2011) and smartphones (Vlahogianni and Barmpounakis 2017). Generally, there are several emerging methods exploited for collecting naturalistic driving data based on in-vehicle sensors. The advantage of these sensors is the light and relatively low-cost equipment as well as the new possibilities offered by information and communication technologies for data transmission and processing, compared to traditional "heavy" vehicle instrumentation of early naturalistic driving experiments. This emerging telematics field is tested for a number of applications including traffic management, accident detection and emergency response, monitoring of fuel consumption and emissions, monitoring of hybrid electric vehicles, monitoring of professional drivers etc. (Zaldivar et al. 2011, Yang et al. 2013). It is used in innovative motor insurance schemes (UBI - usage based insurance) on pricing users based on distance travelled or driving behaviour by the OBD unit (Tselentis et al. 2017). Finally, using smartphones as recording devices is becoming popular during the last decade because of the precision of their installed sensors and the fact that they are considered to be the most affordable solution thus far.

As it appears, this rapid technological progress along with the increasingly penetration rate by drivers (e.g. Smartphones), provide unprecedented opportunities to accurately monitor and analyse driving behaviour. First results from related applications (Tselentis et al. 2017, Theofilatos et al. 2017, Araújo et al. 2012, Enev et al. 2016) have confirmed the efficiency and usefulness of such big data collection schemes. Nevertheless, the exact amount of the necessary driving data that should be collected for each driver in driving behaviour assessment is not determined yet. Since both small and big data samples lurk the risk of leading to doubtable results, by acquiring a sample either biased or computationally expensive to analyse, it is a matter of great importance to specify exactly how much driving data should be recorded from each participant in the experiment.

In order to benchmark driving efficiency based on different type of driving metrics, a data envelopment analysis (DEA) approach is proposed in this study. DEA is a technique of mathematical programming problem with minimal assumptions that determines a unit's efficiency based on its inputs and outputs, and compares it to other units involved in the analysis. It is a data-oriented methodology that effects performance evaluations and other conclusions drawn from the analysis directly from the observed data. The efficiency of a Decision Making Unit (DMU) is comparatively measured and analysed relatively to the rest of the DMUs considering that all DMUs lay on or below the efficiency frontier. DEA has become one of the most popular fields in operations research, with applications involving a wide range of context (Thanassoulis 2001). It has been applied in great extent in literature (Cook & Seiford 2009, Emrouznejad et al. 2008) to measure and compare the productivity performance of a group of DMUs. Martić et al. (2009) presented the ample possibilities for using DEA for evaluating among others the performance of banks, schools, university departments, farming estates, hospitals and social institutions, military services and entire economic systems. DEA has also been successfully implemented in

transport fields in assessing public transportation system performance (Karlaftis et al. 2013) and traffic safety studies (Egilmez & McAvoy 2013, Alper et al. 2015) but never before in driving behaviour research.

One of the most important issues arising is that linear programming (LP) techniques, such as DEA requires a significant amount of time to perform on large-scale data. Many suggestions including Reduced Basis Entry (RBE) and Early Identification of Efficient DMUs (EIE) have been made thus far to reduce the running time of DEA with some of them performing notably better (Dulá, 2008; Barr & Durchholz, 1997; Ali, 1993; Dulá & López, 2009). Something also never addressed before is the multiple inputs and multiple outputs DEA problem for large-scale data.

It can be concluded from all the above that it is significant to study the potential of measuring 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. From many perspectives, the contribution of this thesis is deemed to be significant since a) it provides a methodology for driving safety efficiency assessment, based on microscopic driving characteristics b) it provides a methodology for least efficient trips identification and c) the results of this thesis can potentially be advantageous towards providing personalized feedback to drivers on how to enhance their driving behaviour. It should be highlighted at this point that from now on in the present dissertation, the term of driving safety efficiency will be mentioned as driving efficiency for brevity purposes.

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.

1.2) Methodological steps

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):

A) Exhaustive literature review takes place 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 the problem up.

B) Data collection and preparation which includes a description of the survey design and questionnaire administration and the procedure of data collection and transmission from the smartphone application, the storage of data to the central database of the platform developed by OSeven and the whole data processing procedure. Furthermore, database is further processed and prepared to be imported in the final DE analysis that is conducted afterwards. This preparation is made using Python programming language, which is suitable for large-scale data analysis.

C) Investigation of the adequate amount of data to be included in the analysis and the evolution of the metrics/ distance ratio. 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.

D) Safety efficiency index is defined taking into account, literature review conducted, data collected and all the peculiarities of the DEA technique. 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.

E) Trip efficiency analysis is conducted 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.

F) 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 step D, 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 time-series created is analysed and results are exploited for driver clustering. All driving profiles emerging from each cluster are presented.
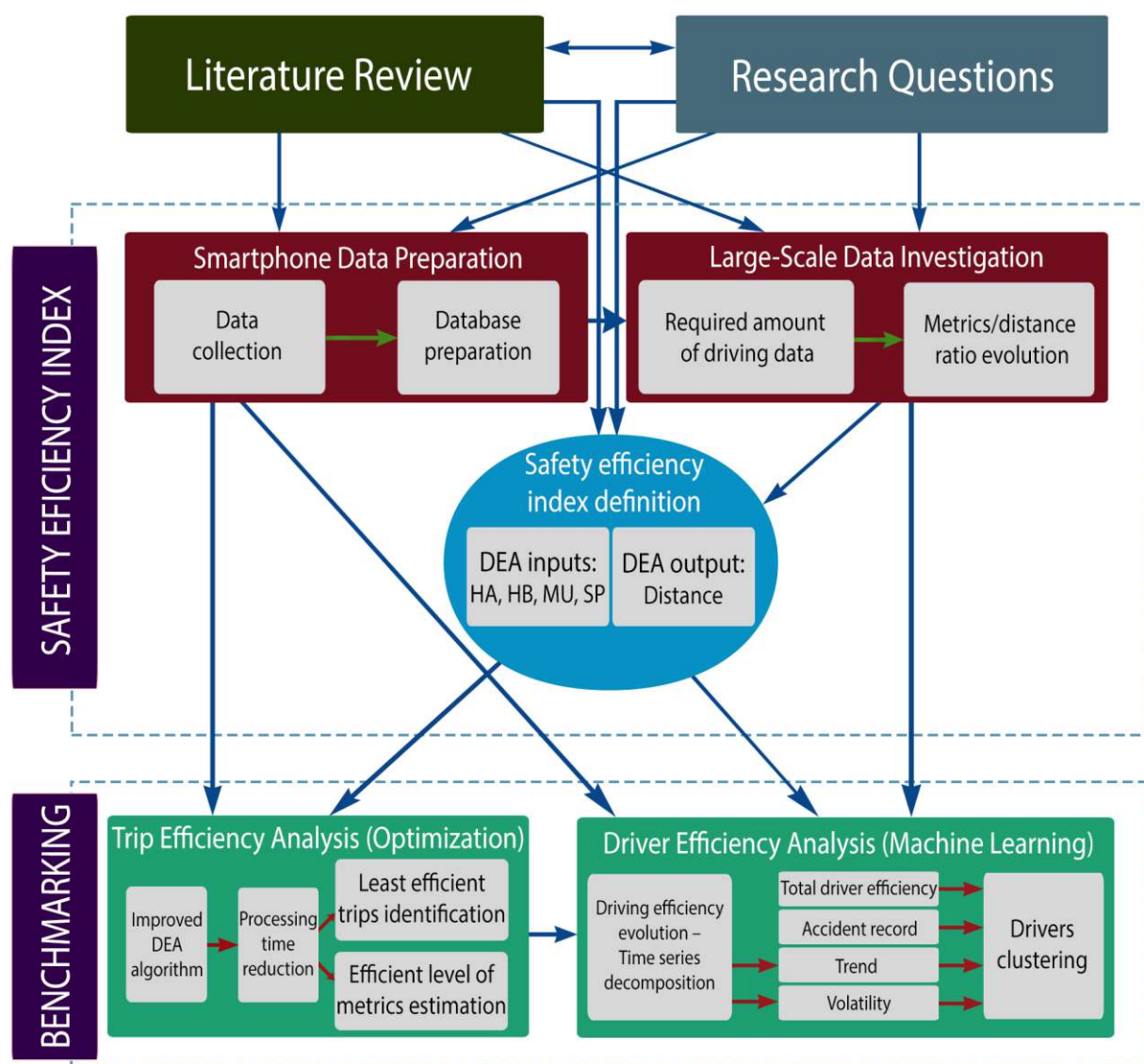
Figure 1.1: Graphical representation of the general methodological framework of the present doctoral dissertation.

The above graphical representation of the methodological approach is used to provide a broader and comprehensible picture of the workflow that takes place and results to the better understanding on how to analyse driving efficiency, using data science techniques for large-scale data. Further details on the methodological background and implementation of the techniques applied in this thesis are presented in each of the following sections.

1.3) Structure of the dissertation

Chapter 1 serves as an introductory part of this dissertation that aims to assist the reader in adapting smoothly with the specific research field that might not be familiar with. It provides a general description of the road safety sector, descriptive statistics on road

accidents and an overview of the microscopic risk factors that influence driving behaviour. The methodological steps followed to accomplish this research are also introduced. Finally, the contribution of the present doctoral dissertation is presented in details to document the reason why this research was necessary to be conducted.

Chapter 2 constitutes the main part of the entire literature review and consists of several parts. This part includes an overview of road safety literature, driving behaviour and risk and a review of driving characteristics and their correlation with driving efficiency and performance. Several implementations of DEA methodology in Economics, Business, Transportation Engineering and Driving efficiency are also presented. The state-of-the-art is given afterwards on how driving data are collected from naturalistic driving experiments. The existing knowledge gap in literature is identified allowing for conclusions to be drawn with respect to methodological and statistical limitations of existing studies and setting the key research questions for the present doctoral research. Finally, the research hypotheses and objectives on which this dissertation is based are specified.

All data sources exploited in this research are explicitly referred in Chapter 3. The procedure of the experiment i.e. participants choice, naturalistic driving experiment data recording, transmission, data processing and storage from the smartphone devices to the central database is described in detail in this chapter. More specifically, a) a general description of the whole procedure of collecting data from the smartphone application, transmitting and stored to a central database of the OSeven platform and how these data are processed is provided and b) the way that large-scale smartphone data and questionnaire data are initially collected is presented. Details regarding the large database's construction and processing before being exploited in the second phase of the analysis using Python as well as some sample characteristics are also provided. In the last section of this chapter, the descriptive statistics of the driving sample collected are illustrated.

Chapter 4 presents the methodological approach of the research conducted. The structure of the methodological approach followed to achieve the objectives specified is described in this chapter including the processing procedure and the analysis conducted using large-scale data collected through the smartphone devices. The theoretical background of every statistical, parametric, non-parametric, linear programming, etc. method used in this thesis is presented in details in this chapter. Finally, the key methodological research questions are formulated.

The results of the application of the methodology proposed to achieve the objectives of this PhD thesis are presented in Chapter 5. As a first step of the modelling application, the database is prepared to be used as input for the analysis. An investigation of the metrics/ distance ratio evolution takes place thereafter to verify the potential usability of the data collected in the model. As a next step, it is examined how much data should be used in the forthcoming analysis. Thereupon, it is explained which parameters will be used in the DEA model and with what input/output combination formulation and the safety efficiency index is estimated. Trip efficiency analysis is performed afterwards to study the potential of applying and providing an improved DEA model on the large-scale data collected as well as a methodology for identifying the less efficient trips. Next step is

driver efficiency analysis that is conducted to estimate efficiency on a driver basis given the recorded metrics. Analysis of the driving efficiency evolution is performed thereupon using time-series analysis and statistics. Finally, drivers are clustered through machine learning techniques to identify the attributes of the main driving profiles arising.

Chapter 6 includes the conclusions and a critical synthesis of the results which answers all research questions raised at the commencement of the present PhD dissertation and suggestions for future research steps on studying driving efficiency and driving behaviour in general are made. Finally, the impact and scientific contributions of this research are highlighted.

1.4) Contribution of this dissertation

This doctoral dissertation aims to contribute in the field of driving safety **efficiency** measurement, which is correlated with driving risk, using data science techniques applied on **large-scale** data. Therefore, the way to perform driving efficiency benchmarking based on microscopic driving metrics collected from smartphones will be studied. It is equally important to investigate the evolution of driving efficiency over time and the main characteristics of the efficiency **groups** and drivers' clusters arose. To achieve this, optimization and machine learning techniques will be combined.

Findings will constitute a significant **contribution** towards the better understanding of the existing **driving patterns**, which is extremely significant for the literature of driving behaviour. This pattern recognition may assist in the identification of a driver's pattern, which would assist in the provision of more representative and personalized information on how to improve driving behaviour.

This thesis will also cope 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, as the potential of applying an **improved DEA** methodology to yield the same optimal solution in less time is examined herein. Moreover, the **large-scale** driving data collected will be investigated to specify the certain amount of driving data that should be collected. These two approaches discussed constitute the innovation of this doctoral research in terms of the large-scale data handling.

This study will attempt to make use of **driving data** collected using a **smartphone** application, which is an approach that is becoming popular nowadays for collecting data from naturalistic driving experiments.

The **practical value** of this study is that results could be exploited to provide **feedback** to drivers on their total driving efficiency and its **evolution**. This constitutes a very significant contribution since driving safety efficiency **measurement** is correlated with driving risk, which means that this can potentially affect the accident probability of a driver, which can potentially lead to a reduction in the total number of accidents.

Chapter 2: Literature Review

2.1) Road safety

A very complicated scientific field of transportation research that has attracted huge efforts by researchers and practitioners is road safety. Since accidents impose serious problems to society in terms of human, economic, property damage and medical costs, it constitutes a major concern in the transportation industry. Accidents on a global level resulted to approximately 1.25 million fatalities (around 3,400 per day) in 2013 (WHO, 2015) (Fig. 2.1), 25 thousands in Europe and 824 in Greece (EL.STAT. 2016) as well as between 20 to 50 million non-fatal injuries per year (WHO 2015) worldwide. These injuries and fatalities' effect on families and communities where these people live and work is not something measurable. Lower income families are suffer more by both direct (e.g. medical) and indirect costs (e.g. lost wages) that result from these injuries. Road traffic crashes' cost is extremely high for most countries at a national level and particularly to developing economies ranging from 1-2 % for low- and middle-income countries, estimated at over US$ 100 billion a year (Jacobs 2000), to 3% of their gross domestic product (WHO 2015).
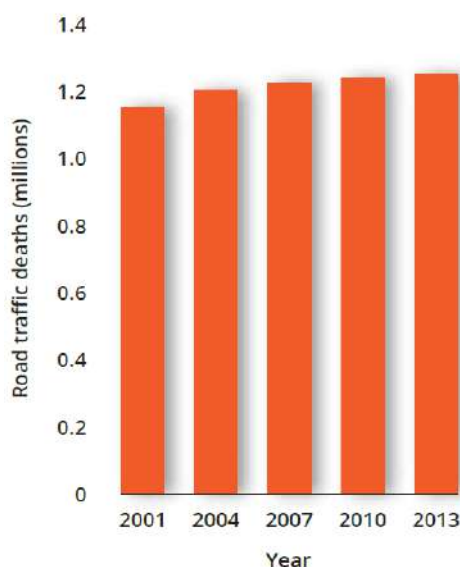
Figure 2.1: Number of road traffic deaths, worldwide, 2013

In terms of fatalities around the globe, road accidents (WHO 2015) are the eighth leading cause in all age categories, with a similar impact to that caused by many communicable diseases, such as malaria (Lazano et al., 2012) and the leading cause of fatalities for people between 15 and 29 years old (approximately 400,000 people/ year) (Fig. 2.2). Current trends observed in the number of road fatalities suggest that road traffic deaths will become the fifth leading cause of death by 2030 unless urgent action is taken (Global status report 2009). It should also be mentioned that people between 15-44 years old

account for the 48% of road crash fatalities globally. From a socioeconomic perspective, more than 90% of these fatalities and injuries take place in low and middle-income countries and the highest rates are found to be in Africa and the Middle East region. Nonetheless, people from lower socioeconomic backgrounds in high-income countries are more likely to be involved in road traffic accidents. This is likely to be the impact of rapid motorization, which is not sufficiently accompanied by improved road safety strategies and operations. According to the same report (WHO 2015), eighty-eight countries (with almost 1.6 billion residents) reduced the number of roads fatalities between 2007 - 2010, showing that improvements are possible, and that there is potential in saving many more lives if countries take further actions. Nevertheless, the fact that 87
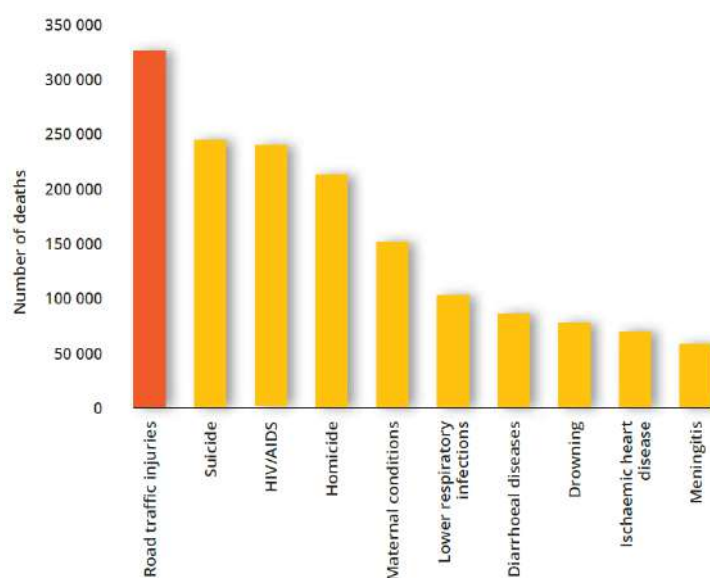
Figure 2.2: Top ten causes of death among people aged 15–29 years, 2012

countries saw increases in the numbers of road traffic fatalities over the same period is also concerning. Regarding sex, it is more likely for males to be involved in road traffic crashes than females. More than three-quarters of all road traffic deaths are among young males under the age of 25 years who are approximately 3 times more likely to participate in a fatal road traffic crash than young females.

A steadily decreasing trend in road traffic casualties is fortunately observed during the last few years. Nonetheless, the number of road fatalities remains in unacceptably high levels in several countries including Greece, a fact that demonstrates the need for greater efforts, improvements and precautionary measures in all road safety aspects including driving behaviour and performance (OECD 2013). Despite the above mentioned fact, the number of fatalities in road accidents in most countries including Greece is still unacceptably high and therefore the need for increased efforts with respect to better driving behaviour and road safety is highlighted (WHO, 2015). National road safety policy formulation, aggregate and disaggregate level of road safety monitoring, speeding interventions for infrastructure improvement in interurban and urban road network, information and education campaigns, road users retraining and upgrade of the

requirements for a driver's license should be included by the necessary recommendations for the improvement of the road safety level (Golias et al., 1997).

Road traffic safety is dealing with the necessary methods and precautionary measures for reducing a road network user's risk of being seriously or fatally injured. The road network users' categories include pedestrians, cyclists, motorists, their passengers, and passengers of on-road public transport, mainly buses and trams. The difference between the old and the new road safety paradigm is that the latter focuses upon the prevention of serious injuries and death crashes in spite of human fallibility (this is considered to be a best-practice road safety strategy) in contrast with the old that is focusing on simply reducing road crashes assuming road user's compliance with traffic regulations. The goal of current safe road design is to provide a road environment that ensures vehicle speeds are within the human tolerances for serious injury and casualty in case of conflict points existence (International Transport Forum, 2008).

With a focus on road accidents, a number of factors is identified to affect the probability of a road crash and therefore, limiting the exposure of these risk factors is a matter of great significance in order to reduce the effects of road traffic accidents. For instance, there are numerous scientific researches demonstrating that the increased risk of road traffic fatalities and injuries results from human factors, such as inappropriate speed, alcohol or mobile phone use (Alm and Nilsson 1994, Agathe and Sagberg, 2001), non-use of seat-belts, child restraints or motorcycle helmets and driver distraction (Elvik et al., 2004). Human factor is proved to be the basic cause of road accidents in a percentage of 65-95% (Sabey and Taylor, 1980; Salmon et al., 2017; Treat, 1980). The rest of the factors that have an impact on accident probability include road environment (pavement, road signs, weather conditions, road design etc.) and vehicles (equipment and maintenance, damage etc.) as well as combinations of all three contributory factors.

Human factors research investigates people's interaction with various aspects of the social environment as well as the way to make them safer, healthier, and more efficient. This interdisciplinary field of research has a wide scope of application, spanning road safety, healthcare delivery, physical, cognitive, and technological systems. Within the context of road safety, human factors research focuses on the understanding of a driver's role in the safe operation of his or her vehicle. There are various factors contributing to the way a person behaves when on driver's seat including environment, psychology, and vehicle design. Human factors research ultimate goal is to set out these factors, determine their influence on driver performance and propose road or vehicle design modifications for driving risk reduction and driving performance improvement. Vehicle safety features are part of a safe-driving system that includes human factors: the various ways drivers interact with those features help with the determination of both driver's safety and safety features effectiveness.

On the other hand, the list of human factors affecting crash risk is not limited to risky driving behaviour such as driving over the speed limits and distraction originating from mobile phone use. Age, driving experience etc. can have an impact on safety features performance. Drivers are a vital part of the road safety system, so factors that affect drivers are accordingly affecting road safety in general. There are several factors such as physiology, psychology, knowledge, culture, traffic laws and regulations, driver's

experience and temper, and brain health on which driving traits, quality, and performance of a driver depend on. Driving characteristics can also be classified by skills and styles into prudent/ aggressive, stable/ unstable, conflict risk avoidance/ risk prone, skilfulness/ non-skilfulness, and law-abiding/ frequent violation driving.

A nation-wide survey (TIRF, 2012) revealed that several drivers admit that if they owned a vehicle fully equipped with modern safety features, they would probably engage with unsafe driving practices such as distracted driving and speed limit exceedance. This presents how human factors may affect total safety benefits for drivers that arise as a consequence of using a vehicle with safety features. Nevertheless, many human factors are not obvious and a lack of familiarity with how safety features work may lead to a slight but negative influence on driver's adaptability on these safety features and therefore on the potential benefits.

In order to assess driving behaviour, several types of driving behaviour experiments exist - naturalistic driving experiments, driving simulator experiments, on road experiments, in-depth accident investigations and surveys on opinion and stated behaviour, just to name a few. Focusing on naturalistic driving experiments, they allow for the examination of a range of driving performance measures in a completely natural, realistic and safe driving environment. However, during these experiments, drivers might be influenced by the fact that they are aware that they are being monitored, which can potentially influence the validity of the results obtained. Despite this fact, the present doctoral research assumes that the influence is roughly the same on all drivers and therefore the results obtained are not significantly affected since the analysis conducted is relative. In general, despite of these limitations, naturalistic driving experiments are an increasingly popular experimental design quickly adopted from many researchers to accurately measure and analyse driving characteristic in natural driving conditions, where originally observed, such as speed limit exceedance, distraction etc. and numerous studies have been conducted, particularly in the last decade (Tselentis et al., 2018).

2.2) Driving behaviour analysis and benchmarking

Measuring driving efficiency has been the focus of many studies in driving behaviour literature in the past (Gerald et al. 1996, Gerald et al. 1998, Young et al. 2011). From a road safety perspective, it is extremely significant to identify the parameters that influence driving behaviour and therefore traffic risk. To achieve this, it is extremely important to study driving behaviour on a microscopic level and therefore necessary to investigate on the factors that mainly influence it. Several studies have been carried out so far on these factors such as mobile phone usage distraction, speed limit exceedance and the number of harsh manoeuvers occurring.

2.2.1) Risk factors

There is a significant number of risk factors identified in literature, which affect accident probability. Previous studies showed that regardless of the type of measurement

technology, speed and hard braking are associated with higher accident rates (Klauer et al., 2008, Simons-Morton et al., 2009, Kloeden et al., 2001, Aarts and van Schagen, 2006). It is therefore extremely important to limit these risk factors in order to succeed in the effort of reducing the number and severity of road traffic injuries. The most important risk factors recognized in literature (WHO 2015, Elvik 2004, Pedden et al. 2004) are human factors (speeding, distracted driving, driving under the influence of alcohol and other psychoactive substances etc.), unsafe road infrastructure, unsafe vehicles and inadequate law enforcement of traffic laws. Among them, human factors are likely to be the most important cause of road traffic fatalities and injuries every year and therefore the importance of studying how these factors can affect traffic risk is high. Consequently, it is valuable to quantify the influence of driver's behaviour on crash risk at least relatively to the rest of the drivers' population.

Driving behaviour includes many different aspects e.g. adjustability to traffic conditions, vehicle operation and driver's intention and action such as acceleration and deceleration, etc., decision-making, smartphone and navigation systems usage, eating, drinking, talking to other passengers, applying cosmetics, looking at the external environment. Vehicle state is mainly comprised of vehicle position, speed, acceleration, and steering angle and rate (single vehicle trajectory), as well as distance headway, time headway, and time to collision.

Driver distraction

Driver distraction is one of the most important human factors that influence road accident probability. In general, driving distraction is defined as "a diversion of driver's attention by focusing on an object, person, task or event not related to driving, which reduces driver's awareness, decision making ability and/or performance, leading to an increased risk of corrective actions, near-crashes, or crashes" (Regan et al., 2008). More specifically, driver distraction involves a secondary task that distracts driver attention from the primary driving task (Donmez et al., 2006; Sheridan, 2004) and may include four different types: physical distraction, visual distraction, auditory distraction and cognitive distraction. A distracting activity involves one, or more of these. The act of operating a hand-held cell phone for example, may involve all four types of distraction.

There are many types of distractions that can lead to impaired driving (WHO 2015). The distraction caused by mobile phones is a growing concern for road safety. Drivers using mobile phones are approximately 4 times more likely to be involved in a crash than drivers that are not using. Conversing while driving slows (Patten et al., 2004) reaction times (notably braking reaction time, but also reaction to traffic signals), and makes it difficult to keep in the correct lane and to keep the correct following distances. Hands-free phones are not much safer than hand-held phone sets on rural roads/motorways (Handel et al., 2014, Yannis et al., 2014) and texting considerably increases the risk of a crash. With regards to mobile phones, the content of the conversation was far more important for driving and driver distraction.

Physical distraction takes place when a driver has to use one or both hands to use the mobile phone instead of concentrating on the physical tasks that driving requires (Young

et al., 2007). Visual distraction occurs when drivers' eyes are either totally off the road or on the road but failing to see because another activity takes place simultaneously (talking over the telephone). The use of mobile phones that display visual information (e.g. texting, watching videos) while driving are further distracting drivers' visual attention (Dragutinovic and Twisk, 2005). Auditory distraction takes place when a driver's attention is disturbed by external sounds not related with the driving procedure such as conversation, telephone ringing and music. Cognitive distraction occurs when two mental tasks are performed simultaneously and involves lapses in attention and judgment. Example given, conversation competes with driving demands, listening can reduce activity in the part of the brain associated with driving and the extent of the negative effects of mobile phone use while driving depends on the complexity of both cell phone conversations and of driving situation.

Driver distraction is part of the broader category of driver inattention. The presence of a specific event or activity that triggers the distraction distinguishes distracted driving from inattentive driving (Regan et al., 2005). On the other hand, very few definitions of driver inattention exist in the literature such as driver distraction that vary in meaning. Driver inattention has been defined (Lee et al., 2008) as "diminished attention to activities critical for safe driving in the absence of a competing activity". Other definitions (Regan et al., 2005) of driver inattention and driver distraction are "insufficient or no attention to activities critical for safe driving" and "is just one form of driver inattention, with the explicit characteristic of the presence of a competing activity" respectively.

Driving distraction factors can be subdivided into external and internal (in-vehicle). Those occurring inside the vehicle seem to have greater effect on driver's behaviour and safety (Horberry et al., 2006). While some studies (Stutts et al., 2001) report that external distraction factors are less than 30% of the total distraction factors, other specify that they account for less than 10% (Sagberg, 2001; MacEvoy et al., 2007).

Patel et al., (2008) examined perceived qualitative characteristics of 14 different driving distraction factors. Participants were asked to rank a list of distracting factors according to certain criteria. Table 2.4 illustrates the average perceived risk ratings for each of the 14 factors.

As shown, the highest perceived risk factors are those associated with mobile phone usage, followed by the 'looking at a map or book' and 'grooming' factors. On the other hand, the lowest perceived risk ratings were associated with 'listening to music', 'talking to passengers' and 'looking at road signs'. It is noticeable that advertising signs and landscape are both a significant perceived external distraction factor.

Table 2.1: Perceived risk associated with driver distraction (Patel et al., 2008)

Driver Distraction Hazard	Risk rating	Lower limit	Upper limit
Listening to music	3.3	1.2	4.8
Talking to passengers	3.8	2.0	5.0
Looking for/at road signs	4.2	3.0	6.0
Satellite navigator use	4.6	3.0	6.0
Hands-free kit use	4.7	3.0	6.0
Looking at Landscape	5.2	3.0	7.0
Adjusting device	5.3	4.0	7.0
Smoking	5.3	3.0	7.0
Looking at advertising sign	5.7	4.0	8.0
Eating or drinking	6.3	5.3	8.0
Looking for object	7.4	6.0	9.0
Grooming/make-up	8.5	8.0	10.0
Looking at a map or book	8.5	8.0	10.0
Mobile phone use	**8.6**	8.0	10.0

The actual relative importance of different distraction factors was studied in more depth in the reports of the 100-Car naturalistic driving study carried out in the USA. Table 2.5 shows results on the odds ratio or in other words the accident probability when engaging in various secondary distracting tasks over when driving without any distracting task (statistically significant results are in bold). As a result, a significant odds ratio indicates an important increase in accident risk when engaged with such a distracting activity. Results suggest that crash probability of "reaching for a moving object" is more than eight times higher when compared to just driving, followed by "reading' and 'applying make-up" where risk is increased by more than 3 times. Finally, mobile phone use is increasing accident risk by 2.8 times.

Table 2.2: Odds ratio for secondary task (NHTSA, 2008)

Type of Secondary Task	Odds Ratio
Reaching for a moving object	8.82
Insect in vehicle	6.37
Reading	3.38
Applying makeup	3.13
Dialling hand-held device	2.79
Inserting/retrieving CD	2.25
Eating	1.57
Reaching for non-moving object	1.38
Talking/listening to a handle-held device	1.29
Drinking from open container	1.03
Other personal hygiene	0.70
Adjusting the radio	0.50
Passenger in adjacent seat	0.50
Passenger in rear seat	0.39
Child in rear seat	0.33

As for the influence of distraction on driving performance, a fundamental question is how and how much drivers are capable of self-regulating their driving behaviour in order to

compensate their driving inattention. As it appears, this issue has not been thoroughly examined thus far. This is probably because research has mainly focused on the identification of the impairment of driving performance as a result of distraction activities (Haigney et al., 2001). Nonetheless, the engagement with non-driving tasks is not necessarily indicative of driving performance impairment, and research (Poysti et al., 2005) suggests that most drivers do engage in a range of conscious and unconscious compensatory behaviours (speed reduction etc.) in order to maintain an adequate level of safe driving.

Compensatory behaviour (Alm and Nilsson, 1994; Lamble et al., 2002) might take place from a strategic (e.g., preferring not to use a mobile phone during driving) to an operational level (e.g., reducing speed). In order to moderate risk exposure, the choice of not engaging into a distracting task while driving can be the option at the highest level. Research (Alm and Nilsson, 1994; Lamble et al., 2002) showed that mobile phone usage impairs driving performance of elder drivers more than that of younger drivers and that this results in a compensatory behaviour at the highest level; a significant portion of older drivers choose not use a mobile phone during driving.

Research showed that at an operational level, drivers endeavour to reduce workload and moderate their risk exposure when there is interaction with in-vehicle devices. This is done through speed reduction (Alm and Nilsson, 1990; Burns et al., 2002; Haigney et al., 2001), headway distance increase (Strayer and Drews, 2004; Strayer et al., 2003), adjusting the relative amount of attention given to driving and non-driving tasks to respond to road environment changes (Brookhuis et al., 1991) and accepting a temporary degradation in certain driving tasks (Brookhuis et al., 1991; Harbluk et al., 2002). Mobile phone use results in higher variation of the throttle position, lower speed with higher variation and higher workload level (Md Mazharul and Washington 2015) regardless of the difficulty level of the conversation. Drivers tend to select larger headways (Nilsson 1982) and longer time-headways (Mohammad et al. 2015) suggesting possible risk compensatory behaviour (Md Mazharul and Washington 2015, Törnros and Bolling 2006).

Speeding

Speeding is also recognized (WHO, 2015) as one of the most important factors in driving risk analysis that influences accident probability (e.g. decreased reaction distance, loss of control) and crash severity. While, at an individual level, the perceived risk is low, the societal risk is high and usually not well understood. The higher the absolute speed, the higher the crash risk. This is probably because the driver needs a specific amount of time to react to unpredictable events. Therefore, the faster a vehicle is moving, the longer it moves before a reaction. In general, at high speeds the time to react to changes in the environment is shorter, manoeuvrability is harder and the stopping distance is larger. Apart from that, the larger the speed differences, the higher the accident probability. High speed differences increase the number of potentially conflicting situations. For instance, the probability of a rear-end collision with a slower car in front and a head-on collision when overtaking a slower car is increased because of the high speed difference.

The relationship between speed and accident probability is less direct and more complicated to quantify than the relationship between speed and accident severity. There are several studies (Kloeden et al., 1997, 2001, 2002) that relate crash risk to individual driver's travel speed or the difference between it and the mean speed of traffic. One important factor that determines to what extent driving speed affects accident probability is the layout and design speed of a road. There are some roads that can cope with higher driving speeds than others, without influencing accident probability much. Traffic volume and traffic composition are also important factors.

The severity of a crash follows from the laws of physics. At higher speeds, the energy released during a crash increases with the square of the speed and the change of speed experienced by those involved directly or indirectly in the crash, increase with speed. Crash risk increase because when speed increases, there is a shorter time to react to environment's changes and the capability to maneuver the vehicle effectively is reduced. Drivers need approximately one second (reaction time) on average to react to an unforeseen event and choose how to respond when in traffic. The higher the driving speed, the longer the distance covered during the reaction time and before a response is initiated, reducing thus the opportunity to avoid a crash.

Excessive speed is a major problem in all motorised countries. A research (Elvik, 2011) for Norway showed that if all drivers were driving below speed limits, there would be 20% less fatalities. Speeding has been (OECD, 2006) a contributory factor in 10% of the total accidents and more than 30% in fatal accidents. According to (Andersson and Nilsson 1997, Nilsson 1982) the probability of a crash involving an injury is proportional to the square of the speed, the probability of a serious crash is proportional to the cube of the speed and the probability of a fatal crash is related to the fourth power of the speed. Moreover, (Nilsson 2004) depicts the relationship between speed and driving risk via an exponential curve, showing that the driving risk is not proportional to the speed. For instance, a 1 km/h increase in average vehicle speed results in a 3% increase in the frequency of accidents involving injuries and a 4–5% increase in the incidence of fatal crashes. The probability of a vehicle collision with an adult pedestrian to be fatal is less than 20% if the vehicle's speed is less than 50 km/h and almost 60% if it is 80 km/h.

The estimates in figure 2.4 are aligned with older research indicating (Rosen et al 2009, Kröyer et al., 2014) that fatality rate is about 4-5 times higher in collisions between a vehicle and a pedestrian at 50 km/h compared to 30 km/h. More than 50 km/h is not acceptable in areas where motorised vehicles and vulnerable road users might mix and share the same space. In those cases, e.g. in residential areas, a limit of 30 km/h is imposed.
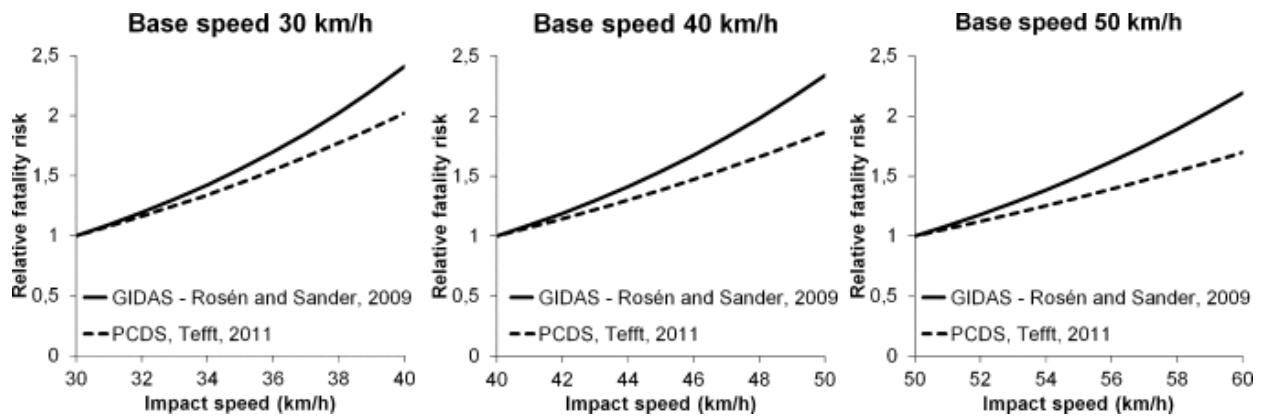
Figure 2.3: The upper part of the relative fatality risk curves for base speeds 30 km/h, 40 km/h and 50 km/h. (Source: Kröyer et al., (2014))

There are many empirical studies (Nilsson, 2004) that assess the influence of a change in average speed on a specific road on the number and severity of crashes on that road. Most scientifically sound studies are comparing the average speed and the number of crashes before and after a speed management intervention, e.g. a speed limit change or the introduction of speed enforcement. Results should be compared with a similar group of roads that were not affected by the speed management measure but can be taken to have been affected similarly by all other concurrent changes. This is to ensure that there are no other factors other than speed change that can explain a change in crash frequency (e.g. a change in traffic volume or an anti-speeding publicity campaign).

Figure 2.4 illustrates the relationships between change in average speed and crashes. As it appears, a 10% increase in average speed will approximately lead to a 20% increase in all injury crashes, a 30% increase in fatal and serious crashes and a 40% increase in fatal crashes.
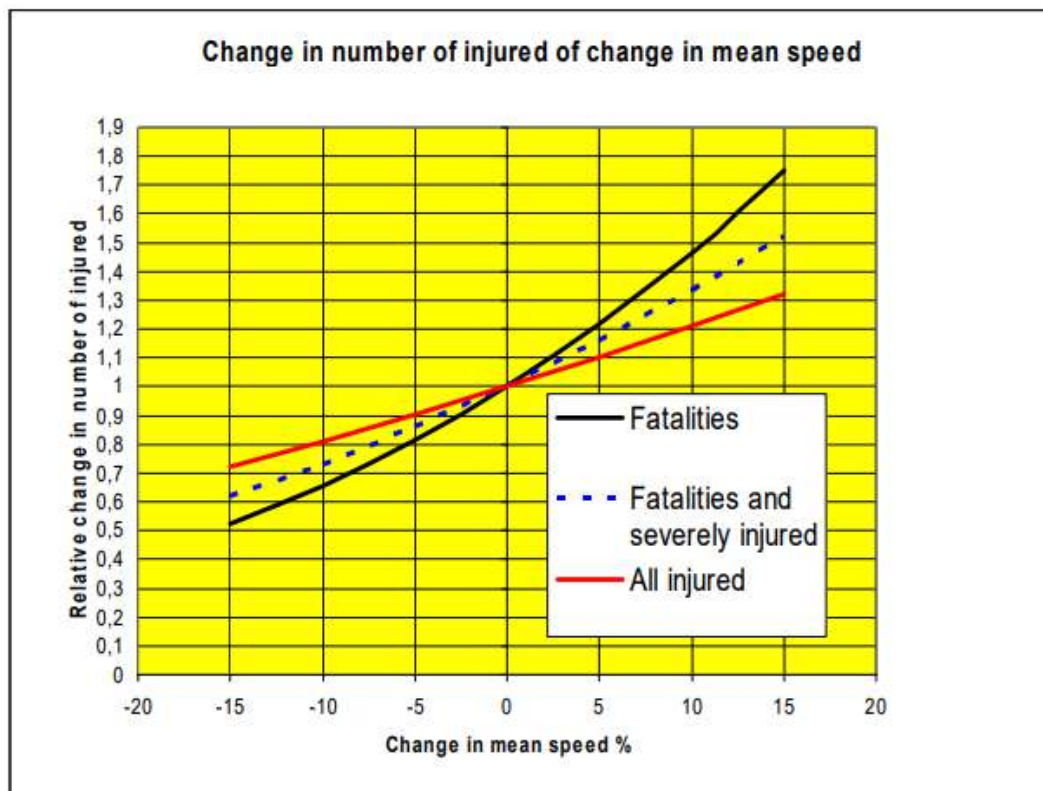
Figure 2.4: Illustration of the power model and the relationship between percentage change in speed and the percentage change in crashes (Source: Nilsson (2014))

Another strong theory is that speed variance is also related to accident frequency. Nonetheless, relative studies were not entirely conclusive and it is not clear how to interpret the relationship found. Speed differences are defined as the speed range over a 24 hours period of recording in most studies. This means that apart from speed differences between vehicles at a particular moment, measured speed differences also include peak and off-peak periods' differences and related traffic volume differences.

However, the use of loop detector data from highways during the last years has made it possible to study the effects of speed variance in a much more robust and well-controlled manner. If stored in s suitable format, data from loop detectors can be used to simulate traffic in great detail. It then becomes possible to determine if the period immediately before a crash occurred was characterized by a larger speed variance than other periods.

A review (Elvik, 2014) of thirteen studies evaluating the speed variance effects on crash rates, based on loop detector data showed that although almost all found that a large speed variance increased crash risk, the numerical estimates of effect varied greatly. The fact that each study employed a different method made it impossible to synthesize their results by means of meta-analysis. It is evident that increased speed variance increases crash risk. Nevertheless, most of these crashes are probably property-damage-only crashes. Dense traffic featured by frequent and sudden speed changes is associated with a particularly high risk of crashes.

Harsh maneuvers

When modeling the kinetic energy of an accident a very close relation to forces is observed. The forces in traffic depend mainly on speed change, the accelerations and the decelerations, and not on the speed itself. Accidents are a subset of the events that include unforeseen deceleration events and the higher the speed, the harsher are the deceleration events and the higher is the probability of injuries and/or fatalities among the involved road users. In specific accident type investigation, those forces applied on the human body are of a great interest, related to the driving speed level of the motor vehicle.

Some studies (OECD, 2006, Elvik et al, 2009) mention that apart from the average speed, other factors, such as the frequency of accelerations, can sometimes be more important. Harsh acceleration, harsh breaking and harsh cornering events are three highly significant indicators for driving risk assessment, and risk level correlation and classification (Johnson and Trivedi 2011, Bonsall et al. 2005, Gündüz et al., 2018, Tselentis et al., 2017) especially when driving aggressiveness is evaluated. This is because they are strongly correlated with unsafe distance from adjacent vehicles, possible near misses, lack of concentration, increased reaction time, poor driving judgement or low level of experience and involvement in situations of high risk (e.g. marginal takeovers). The correlation between HA and HB events with driving risk has been highlighted in the scientific papers published by (Tselentis et al., 2017, Bonsall et al. 2005) and it has been widely recognized by the insurance and telematics industry.

Alcohol and other psychoactive substances

Driving under the influence of alcohol and any psychoactive substance or drug increases accident probability and may result in death or serious injuries (WHO, 2015). When drink-driving takes place, crash risk remains at low levels when blood alcohol concentration (BAC) is low and increases significantly when the driver's BAC is higher than 0.04 g/dl.

On the other hand, when drug-driving takes place, accident risk is increased to differing degrees depending on the psychoactive drug used. For example, fatal crash risk among those who have made use of amphetamines is about 5 times the risk of someone who has not.

Motorcycle helmets, seat-belts and child restraints

Wearing a motorcycle helmet correctly can reduce the fatal crash risk by approximately 40% and the risk of severe injury crash risk by over 70% (WHO, 2015). Wearing a seat-belt reduces fatal accident risk among front-seat passengers by 40–50% and of rear-seat passengers by between 25–75%. When child restraints are correctly installed and used, they can reduce fatalities among infants by approximately 70% and casualties among small children by between 54% - 80%.

2.2.2) Methods for quantifying safety efficiency in transportation

The measurement of driving performance as well as the determination of which factors and how much they are influencing it has been studied a lot in the past. Matthews et al. (1996) examined whether driver stress is related to performance impairment because stress-prone drivers are vulnerable to overload of attentional resources. The driving sample comprised from eighty young drivers who participated at the same time in a simulated drive and a grammatical reasoning task, which was presented either visually or auditorily, with random priority assigned to the two tasks. Results indicated that driving performance is characterized by effort's adaptive mobilization to keep up with changing task demands. The analysis conducted showed that stressed drivers adapted quite efficiently to high demand levels, but they are probably at performance impairment risk when the task requires relatively little active control.

The potential improvement of driving performance using a smart driving system (that provides safety and fuel-efficient driving advice in real time) in on-road experiments was assessed by Birrell et al. (2014). Forty participants participated in the experiment using an instrumented vehicle over a 50-min mixed-route driving scenario. The two different conditions adopted are to control the vehicle with and without smart driving feedback offered to the driver via a smartphone device. The study resulted that a 4.1% improvement in fuel efficiency can be achieved when using the smart driving aid (with no increase in travel time or average speed reduction). There were also observed significant changes in safety behaviour, namely an increase in mean headway to 2.3 s and an almost threefold reduction in time spent traveling closer than 1.5 s to the vehicle in front.

Rakauskas et al. (2014) studied the influence of mobile phone usage on driving performance as well as the relationship between the distraction resulting and the difficulty level of a phone conversation. A driving simulator experiment is conducted to determine how and how much several levels of cell phone conversations influence driving performance. Mobile phone use resulted in higher variation in accelerator pedal position, average speed reduction and higher speed variation, and a higher level of workload regardless of conversation difficulty level. Drivers are likely to endure higher workloads or set reduced performance goals to cope with the additional phone conversations stress.

In order to compare the effects of fatigue, performances and sleepiness in real-life driving and driving in a simulator, Philip et al. (2005) conducted a cross-over study that involved real driving (1200 km) or simulated driving after controlled habitual sleep (8 hours) or restricted sleep (2 hours). Twelve healthy young men (range 19-24 years) participated who were free of sleep disorders and measures collected included self-rated fatigue and sleepiness, simple reaction time before and after each session, the number of inappropriate line crossings from the driving simulator and from video-recordings of real driving. Results indicated that line crossings were more frequent in the driving simulator than in real driving and were increased by sleep deprivation in both conditions. Reaction times (10% slowest) were found to be slower during simulated driving and sleep deprivation. It was also found that the sleepiness scores were higher in the driving simulator and in the sleep restricted condition. Fatigue increased over time and with sleep deprivation but it was similar in both driving conditions.

Lenné et al. (1997) examined the time of day variation in driving ability as a possible cause of accidents. Eleven male subjects participated in a driving simulator experiment for 30 minutes, six times of day. The instructions given to subjects were to maintain a stable position in the left-hand lane, drive at a constant speed of 80 km/hour and perform a secondary reaction time task. The mood of each subject was measured at the beginning and end of every session. Driving performance was measured on the basis of the mean and standard deviation of lateral position and speed. It appeared that speed's mean and standard deviation for curved and straight segments as well as reaction time varied significantly across the day. Performance was impaired more at 06:00 and 02:00 hours, with improvements in driving performance between 10:00 and 22:00 hours and an early afternoon dip. Results indicated that driving performance is subject to time of day variations. The fact that impairments in driving performance in the late evening and early morning are of a similar magnitude to those occurring in the early afternoon is also very important.

It can be inferred from the literature review conducted that there is a plethora of researches studying the assessment of driving behaviour that are mainly focusing 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 is the first effort made to estimate and assign a relative safety efficiency index to each driver of a sample by exploiting several driving behaviour metrics that result from microscopic driving behaviour data recorded from smartphone devices.

Linear programming for efficiency measurement

The terms "efficiency" and "productivity" are widely used in economics and refers to the optimal way a production unit can make use of its available resources *(*Shone, 1981*)*. More specifically *(*Farrell, 1957*)*, 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. Thus, when a DMU is technically efficient, it operates on its production frontier and therefore DMUs lie on the efficiency frontier. Based on the assumptions stated below, drivers in this study are considered as DMUs and DEA applicability on the field of driver's assessment based on microscopic behavioural characteristics is examined.

DEA is therefore a mathematical programming technique with minimal assumptions that determines a unit's efficiency based on its inputs and outputs, and compares it to other units involved in the analysis. A data-oriented methodology that effects performance evaluations and other conclusions drawn from the analysis directly from the observed data. The efficiency of a DMU is comparatively measured and analysed relatively to the rest of the DMUs considering that all DMUs lay on or below the efficiency frontier. No assumption is required about functional form (e.g. a regression equation, a production function, etc.) or the statistical distribution of data sample and as a result DEA is classified as a non-parametric method. It is a frontier analysis, a process of extremities, not driven by central tendencies in contrast to all statistical procedures. Each DMU is analysed

separately and the real and optimal performance that can be achieved for each unit is estimated.

Efficiency can be defined as the ratio of input and output in a theoretical scenario of units that have a single input and output but in a real case scenario where typical organizational unit have multiple and incommensurate inputs and outputs a more scientific approach is needed. DEA is an approach for efficiency and productivity analysis of production units with multiple inputs to produce multiple outputs mostly used thus far in business, economics, management and health. The rationale for using DEA is its applicability to the multiple input–output nature of DMUs provision and the simplicity of the assumptions underlying the method. It is a methodology of several different interactive approaches and models used for the assessment of the relative efficiency of DMU and for the assessment of the efficiency frontier. It assists in drawing important conclusions on operational management of the efficient and inefficient units.

DEA has become one of the most popular fields in operations research, with applications involving a wide range of context (Thanassoulis, 2001). It has been applied in great extent in literature (Cook & Seiford, 2009, Emrouznejad et al., 2008, Hollingsworth et al., 1999) to measure and compare the productivity performance of a group of DMUs. It is one of the most popular fields in operations research (Emrouznejad et al., 2008, Seiford, 1997) to say the least. (Martić et al., 2009) presented the ample possibilities for using DEA for evaluating among others the performance of banks, schools, university departments, farming estates, hospitals and social institutions, military services and entire economic systems. Since the introduction of CCR model (Charnes et al., 1978) in 1978, the number of publications where DEA is implemented has exponentially grown.

DEA has also been implemented in transport fields in assessing public transportation system performance (Karlaftis et al., 2013), as well as traffic safety studies (Egilmez & McAvoy, 2013, Alper et al., 2015) where it was proved to be equally useful as in the fields stated above. DEA approaches have been applied so far to benchmark road safety on a country (Shen et al., 2011, Shen et al., 2012, Hermans et al., 2009, Wegman & Oppe, 2010,), state (Egilmez & McAvoy, 2013) and municipality level (Alper et al., 2015). It has also been used to prioritize highway accident sites (Cook et al., 2001) as well as to investigate target achievements and identify best traffic safety practices (Odeck, 2006). Finally, other studies considered combining road safety information in a performance index (Hermans et al., 2008) or creating more basic safety indicators (policy performance, final road safety outcomes, intermediate outcomes and background characteristics of countries) to relatively evaluate road safety among different countries (Gitelman et al., 2010).

Nonetheless, DEA has never been used before in driving behaviour research. Driver's efficiency has been studied in a great extent but never by making use of DEA techniques. This research proposes a methodological framework to address the issue of measuring driver's efficiency and categorize the drivers of the sample used in three groups i.e. non-efficient, weakly efficient, most efficient. The main characteristics of each group are presented in order to draw important conclusions on the features of each driving group and provide recommendations for drivers on how to improve their driving efficiency. For the purposes of this study, drivers will be considered as DMUs, which is deemed to be

rational since a driver is a unit that makes decisions for a given mileage range about the number of events occurring and the time of mobile phone usage and speed limit violation. Driving attributes (metrics and distance recorded) will be considered the inputs and outputs of the DEA program. More details on the structure of the DEA formulation implemented are given below.

The general idea of DEA is to minimize inputs (input-oriented model) given a specific production outcome or maximize the outputs of a problem (output-oriented model) give a specific set of inputs. More specifically in the case study examined herein, a driver should either drive more kilometres maintaining the same number of harsh braking or accelerating events or reduce the number of harsh braking/accelerating events given a specific driven mileage. The same applies to the rest of the metrics recorded for each driver. From a road safety perspective, increasing mileage increases crash risk (Tselentis et al., 2017) and, therefore, an input-oriented DEA program is being developed aiming to minimize inputs (recorded metrics) maintaining the same number of outputs (recorded distance). Although a trip cannot literally behave as a DMU, it can be evaluated as a DMU and, therefore, it will be considered as such for the purpose of this research considering trip attributes as inputs and outputs of the DEA program. This is deemed to be a correct assumption on a trip basis since a) all variables used are continuous quantitative variables as those used in previous DEA studies (Cook & Seiford, 2009, Hollingsworth 1999, Karlaftis et al., 2013, Egilmez & McAvoy, 2013) and b) a driver should reduce his mileage (Tselentis et al., 2017) and the frequency of some of his driving characteristics such as harsh acceleration and braking, mobile phone usage and speeding (Tselentis et al., 2017, Aarts & Van Schagen, 2006, Young & Regan 2007). The proposed methodology is applied to a real-life case study of 23,000 and 15,000 trips recorded in urban and rural road respectively collected from one hundred drivers in each road type.

Data envelopment analysis – Main principles

Data envelopment Analysis (DEA) is a technique that has been exhaustively applied in literature (Cook & Seiford, 2009, Emrouznejad et al., 2007, Hollingsworth et al., 1999) to evaluate the efficiency of Decision-Making Unit (DMU) mainly in scientific fields such as economics, management and health. It has become one of the most popular fields in operations research, with applications involving a wide range of context (Thanassoulis 2001). It has been applied in great extent in literature (Cook & Seiford 2009, Emrouznejad et al. 2008) to measure and compare the productivity performance of a group of DMUs. Martić et al. (2009) presented the ample possibilities for using DEA for evaluating among others the performance of banks, schools, university departments, farming estates, hospitals and social institutions, military services and entire economic systems. DEA has also been succesfully implemented in transport fields in assessing public transportation system performance (Karlaftis et al. 2013) and traffic safety studies (Shen et al., 2011, Egilmez & McAvoy 2013, Alper et al. 2015) but never before in driving behaviour research. In this study, a data envelopment analysis (DEA) approach is proposed to benchmark driving efficiency based on different type of driving metrics.

The accuracy of the estimated performance measures depends on the use of appropriate and well-specified models, the inclusion of relevant inputs and outputs, and the use of

accurate data. Several pitfalls in DEA are discussed in literature (Dyson et al. 2001). It is mentioned that when input and output variables simultaneously are formulated as percentiles and/or ratios (e.g. profit per employee, and returns on investment), and raw data (e.g. revenues, assets, employees, profits), the efficiency score might be miscalculated. An example of such a potential problem in a DEA analysis is also demonstrated (Cooper et al., 2007). The choice of an appropriate model is also an important methodological issue. Different approaches have advantages and disadvantages and the choice of the most appropriate estimation method should depend on the type of organizations under investigation, the perspective taken, and the quality of available data. DEA is a non-parametric method and does not impose a functional form on the production frontier and hence can accommodate wide-ranging behaviour. However, measurement errors can bias results and DEA may be best employed in applications having relatively small potential measurement errors. A further line of enquiry is the impact on efficiency scores of sample size, and of more advanced DEA techniques, which allow for the ranking of efficient, as well as in-efficient units.

Rapid technological progress, especially in telematics and big data analytics, along with the increase in the information technologies' penetration and use by drivers (e.g. smartphones), provide unprecedented opportunities to accurately monitor and collect large-scale data on driving behaviour. On the other hand, linear programming techniques such as DEA are performing relatively fast on small-scale databases but much slower when it comes to large-scale data. So far, it hasn't been necessary to analyse large databases using optimization techniques but with the innovative data collection methods emerging for collecting data it becomes necessary for stakeholders, policy-makers and scientists to come up with a practical solution. As a result, it becomes even more necessary for scientists to come across a practical solution to tackle the problem of analysing large databases using optimization techniques to support policy-makers and stakeholders (Vlahogianni 2015).

This thesis is dealing with the problem of optimization techniques that can be applied in real transportation problems to deal with the problem of driving efficiency benchmarking using DEA. It addresses the multiple inputs and multiple outputs DEA problem for large-scale data by proposing an algorithmic modification of DEA based on computational geometry and its effectiveness is tested. The approach is based on the "quickhull" algorithm, which is the computational geometry approach of frame recognition, i.e. the convex hull, to allow for extreme points identification before applying the standard DEA approach to the whole sample. The proposed approach is evaluated in terms of computation time, compared to other proposed approaches i.e. the RBE technique and the standard DEA procedure. The proposed methodology is applied to a real-life case study of 10,088 recorded trips collected from Eighty-eight (88) drivers.

Improvements on DEA application on large-scale

One of the most important issues arising when working with linear programming (LP) techniques such as DEA is that it requires a considerable amount of time to perform on large-scale data. Nonetheless, it is feasible to achieve enhancements and improvements for the standard DEA approach. Apart from all well-known approaches outside DEA for

improving LP performance (e.g., multiple pricing, product forms, hot starts, etc.), there are techniques that exploit DEA LPs' special structure. Many suggestions are made so far to reduce DEA running time with the two best known among them, in terms of processing time, to be Reduced Basis Entry (RBE) and Early Identification of Efficient DMUs (EIE) (Dulá, 2008; Barr & Durchholz, 1997; Ali, 1993; Dulá & López, 2009). Both approaches are a result of the DEA LPs formulation, which defines the DEA LP optimal basis based only on the data points of efficient DMUs. The efficiency frontier consists only of those extreme data points (DMUs) whose efficiency equals to one. The inefficient DMUs do not play any role in defining the optimal solution and therefore their presence or absence from the LP coefficient matrix makes no difference. In other words, presence or absence of less efficient DMUs from the LP coefficient matrix has no effect on defining the optimal solution and, therefore, it is preferable to be absent in order to reduce the required running time for each LP. Ideally, for each LP that is required to be solved in order to identify the efficiency index and peers of every DMU m, it is recommended to omit all DMUs apart from the efficient ones and DMU m.

On this basis, the concept of RBE is to leave out from the subsequent (next iterations) LP formulations every DMU found to be inefficient. The implementation tuning of this idea is easy since LPs are iteratively formulated and solved thereafter. By systematically applying this approach, it gradually leads to a reduction of the size of the LPs and consequently to the running time of each iteration and of the total solution. EIE is based on the concept that a DMU is deemed to be efficient if its constraint is an equality at optimality in a multiplier form or its variable appears in a basis of an optimal solution of an envelopment LP. Literature (Dulá, 2008) shows that EIE has less impact on improving performance than RBE. This is because in large data sets, the number of efficient DMUs is small compared to the number of data points. Furthermore, a relatively small subset of the efficient DMUs seem to have a preponderant presence in optimal bases. These two techniques were tested (Barr and Durchholz, 1997) together and a significant impact on reducing computation times is reported. It is also concluded (Bougnol et al., 2005) that RBE has the most impact on improvement.

There are several alternative procedures for reducing DEA processing time. First of all, a combination of RBE and data partitioning schemes can be implementation (Barr & Durchholz, 1997). The main concept is based on the principle that if a DMU is inefficient within a subset of DMUs, it will be inefficient compared to any superset of the DMUs. An application consists of creating several dataset partitions of uniformly sized data blocks and independently solving the DEA LPs using the standard procedure to identify the inefficient DMUs within these blocks. This becomes an effective scheme for reducing processing time since it is a procedure that can be repeated each time with a new set of data blocks consisting of DMUs with unknown status, until a final single block with efficient DMUs is created. All DMUs are evaluated and scored in a second phase using LPs composed of the DMU that were not eliminated in the first phase. A significantly smaller LP than would otherwise be used in the standard approach is likely to arise after the first phase and so on. It is quite possible though, that more than n LPs will have to be solved. It was observed that the performance of a "hierarchical decomposition" schemes such as the one applied is mainly affected by the size of the initial and intermediary data blocks formed. This is an issue that requires experimental tuning before the

implementation. Tests performed in this research, indicated that substantial time reductions were achieved once these parameters are identified.

Additionally, another implementation found in literature is the adaptation of procedures mainly used in computational geometry to solve the finite polyhedral frame problem, which is a version of the convex hull problem. This is the problem of identifying the extreme points of a finitely generated polyhedral set. The frame problem is similar to the problem of efficient DMUs identification in DEA (Dulá and López, 2006). The extreme point identification algorithm builds the exterior points set by identifying a superset that includes the efficient DMUs. Thereafter, the number of LPs solved applying the standard approach is equal to the number of inefficient DMUs while the final dimension of the LPs is equal to the total number of efficient DMUs in the data set plus one (the inefficient DMU examined in each iteration).

Another option for reducing computation time for DEA is pre-processing ideas several of which have been implemented in literature (Dulá & López, 2009, Dulá & Hickman, 1997). In general, when the status of one or more DMUs is determined without having to solve the entire LP problem then less time is needed and a pre-processor is deemed as effective. DMUs classification is exactly what pre-processors are used for. It is not required in most cases to solve an LP for that DMU since advanced classification of an efficient DMU takes place with an inexpensive pre-processor. Especially when estimating the efficiency score of a DMU is not the goal of a research, the entire LP solution is obviated for inefficient DMUs since they are classified by the pre-processor. Therefore, a pre-processor can be proved extremely valuable in terms of cost and effectiveness. If efficiency score is required, a pre-processor could also be exploited to reduce the cost of inefficient DMUs identification by achieving fast classification. As described above, if a DMU is inefficient, it can be omitted from the solution of the LPs that should be solved for scoring and benchmarking the entire inefficient DMUs set. As an example, in variable-returns-to-scale (VRS) model, a pre-processor could be simple sorting to identify efficient DMUs. A sorting is equivalent to the translation of a hyperplane that is parallel to one of the axes in the attribute space. This suggests other types of pre-processors based on translating and rotating hyperplanes. Pre-processors based on translating or rotating hyperplanes involve only inner products. More details on sorting in DEA can be found in (Dulá & Hickman, 1997). Maximum and minimum attribute values of outputs and inputs of DMUs respectively of the entire data set are also likely to represent efficient DMUs.

Other approaches that yield significant results towards the reduction of the required time when applying DEA are the data partitioning approaches. The basic concept of these approaches is that RBE methodologies are combined with data partitioning schemes. The fundamental idea lies on the principle that if a DMU m is found to be inefficient within a set E of DMUs, DMU m will be inefficient within any superset that includes set E. To apply this, the main dataset is partitioned into equally sized subsets and DEA LPs are solved using RBE techniques for every DMU in each of those subsets. When this procedure is over, all inefficient DMUs are identified and compose the new sets of DMUs which still are subsets of the initial database. By applying the same approach repetitively in every superset created by efficient DMUs of each subset, a final superset comprising of the efficient DMUs of the initial dataset is composed. As a second phase of the procedure, a DEA LP is solved for each of the inefficient DMU m using only all DMUs that

were evaluated as efficient plus DMU m in each iteration. As described above, the advantage that this approach offers is that computation time is significantly reduced since the LP comprises of a considerable lower number of variables; especially when the number of efficient DMUs (density) is low.

The LPs arising are estimated to be much smaller than would otherwise be used in the standard approach. Nevertheless, depending on how data is partitioned, the number of efficient DMUs and a few other parameters, there is a slight chance that total required time will be more than that required if RBE without data partitioning was applied. In schemes like these, called "hierarchical decomposition" schemes, performance is affected by the size and density of the initial and intermediary sets of DMUs. According to (Barr & Durchholz, 1997), optimizing this procedure requires experimental tuning and results of the same research indicate a considerable improvement in computation time.

Computational geometry procedures are another idea that has been extensively used as an improvement to the DEA procedure to solve the problem of the exterior points frame. The most common version of frame determination is the convex hull problem that tackles the problem of identifying the extreme points of a finitely generated polyhedral set. Convex hull algorithm builds the frame consisting of the extreme points (the efficient DMUs) as well as the vertices and line segments that joins them two by two (without intersecting the interior of the polyhedron). In general, the convex hull of a finite point set S is the set of all convex combinations of its points. Convex hull's equivalent problem in DEA is the one of identifying solely the efficient DMUs (Dulá & López, 2006) without having to solve the entire LP for each of the DMUs in the set. As described above, the benefit of computational geometry methods such as the convex hull approach is that the size of the new LPs to be solved are considerably smaller and the number of the LP's variable is represented by the total number of efficient DMUs in the dataset plus one, which is the DMU examined in each iteration.

2.3) Naturalistic driving experiments

In this section, an extended literature review is carried out regarding all available experiment types of assessing driving behaviour. More specifically, benefits and limitations are presented regarding naturalistic driving experiments, driving simulator experiments, on road experiments, in-depth accident investigations and surveys on opinion and stated behaviour. In the end, a comparative assessment of experiments for the assessment driver behaviour is taking place.

Until recently, the high cost of real-time driving data recording systems, data programs, cloud computing services, the inability to accumulate and exploit massive data bases (big data) for transport and traffic management purposes (De Romph, 2013, Lee, 2014), as well as the low penetration rate of Smartphones and social networks, made it extremely hard to collect and manage real-time data and, therefore, to study the relation between driving behaviour and travel behaviour and the probability of crash involvement. Research has indicated that barriers like those mentioned above can be overcome when consumers are given an incentive such as a monetary prize (Reese and Pash, 2009);

this, along with informing drivers through personalised feedback about their speeding are also effective at encouraging drivers to reduce their (mainly speeding) driving behaviour (Ellison, 2015a). It is shown (Elvik, 2014) that the highest rates in speeding reduction by incentive schemes is around 60-80% while the respective percentage for mileage reduction is 0-10%.

Thus, the main challenge road safety entities and policy-makers are facing at the moment is the wide provision of information on the social benefits that could arise from an implementation of such a policy. As a matter of fact, high level of interest has been observed among users who were given a medium value financial incentive of $88 per 6-months period to reduce their mileage. Consumers stated that lower Insurance premiums is among the strongest incentive for them and that a mileage-based Insurance could probably lead them to ultimately consider car sharing or even using public transportation (Reese and Pash, 2009).

2.3.1) Driving data collection

Nowadays, it is feasible to collect high quality real-time data in an efficient way in order to model individual and total crash risk. Several studies have been carried out using innovative methodologies for collecting and analysing (Tselentis et al. 2017) driving behaviour data. The most common method for monitoring driving measures necessary for behaviour evaluation include recorders that relate to the car engine (Zaldivar et al 2011, Backer-Grøndahl et al. 2011) and smartphones (Vlahogianni and Barmpounakis 2017). Generally, there are several emerging methods exploited for collecting naturalistic driving data based on in-vehicle sensors.

The advantage of these sensors is the light and relatively low-cost equipment as well as the new possibilities offered by information and communication technologies for data transmission and processing, compared to traditional "heavy" vehicle instrumentation of early naturalistic driving experiments. This emerging telematics field is tested for a number of applications including traffic management, accident detection and emergency response, monitoring of fuel consumption and emissions, monitoring of hybrid electric vehicles, monitoring of professional drivers etc. (Zaldivar et al. 2011, Yang et al. 2013). It is used in innovative motor insurance schemes (UBI - usage based insurance) on pricing users based on distance travelled or driving behaviour by the OBD unit (Tselentis et al. 2017). Finally, using smartphones as recording devices is becoming popular during the last decade because of the precision of their installed sensors and the fact that they are considered to be the most affordable solution thus far.

On-road experiments

The most common methodologies applied to date, include driving simulators (Paula et al. 1998, Lenné et al. 1997), questionnaires (Gerald et al. 1998) combined with simulators and naturalistic driving experiments (Toledo et al. 2008, Birell et al. 2014). In terms of the driving data collection process, data in most studies are recorded either by the vehicle's OBD (Jensen et al., 2011) or user's smartphone (Handel et al., 2014) and transmitted to

a central database for central processing and analysis (Boquete et al., 2010, Iqbal and Lim, 2006). This allows the development of special indicators for evaluating driver's risk travel and driving behaviour.

In some studies, there exists an on board platform inside the vehicle which acquires and processes data obtained from the GPS, the EOBD system and a mobile-telephone use detection circuit (Boquete et al., 2010). Data are transmitted to a control centre (CC) via a mobile telephone connection, where the risk reflected by each vehicle to the insurance company is estimated. The system uses mobile telephony connection to transmit data between the On-board system (OS) and the CC. Vehicle function data (such as number of seatbelts fastened) are captured from the EOBD system, vehicle position-speed data from the GPS and driver mobile-telephone use data from a detector circuit (RF energy scavenging) are ultimately acquired by the OS. Before transmitting it to the CC, data captured by the OS are processed and stored by a high-performance microcontroller that exists inside the core of the OS.

Other studies also incorporate light or weather sensors that interact via a communications channel (infrared or Bluetooth) with the on-board computing unit reporting a numerical value (Iqbal and Lim, 2006). Position, speed and time are continuously recorded by the GPS receiver and transmitted to the central computing unit. Finally, Barmpounakis et al. (2016) conclude that, since a few technological obstacles that exist nowadays are overtaken, these systems can also be exploited for real time traffic monitoring. Other methods include extraction of vehicular trajectories from video recordings using a trajectory extraction system to collect vehicle traffic data (Barmpounakis et al., 2015). Although this is also not available for real time traffic monitoring it is very likely to be used for this purpose in the near future.

In other on-road experimental studies, an instrumented vehicle is equipped with instrumentation that records a variety of driving aspects (Rizzo et al., 2002) (Figure 2.6). The equipment includes GPS, video-cameras, sensors, accelerometers, computers, and radar and video lane tracking systems. Researchers designing on-road experiments attempt to gain in depth information regarding the factors influencing road crash risk. Experiments are conducted by trained experts from a wide spectrum of disciplines to collect the greatest amount of useful information possible, in order to answer key research questions arising (Wadley et al., 2009; Bowers et al., 2013; Okonkwo, 2009).

On-road driving evaluations are generally considered a widely applied method for assessing driving behaviour since a large number of different variables can be recorded and evaluated. Nevertheless, on-road studies are criticized when data are not collected over a longer time period and in more naturalistic settings as it should be. Another methodological concern is that naturalistic studies typically have at least one researcher present, who gives navigation directions and sometimes a second researcher is also present to observe driving behaviour. This, along with the cost of the equipment, increases (Ball and Ackerman, 2011) the total cost of the experiment excessively.
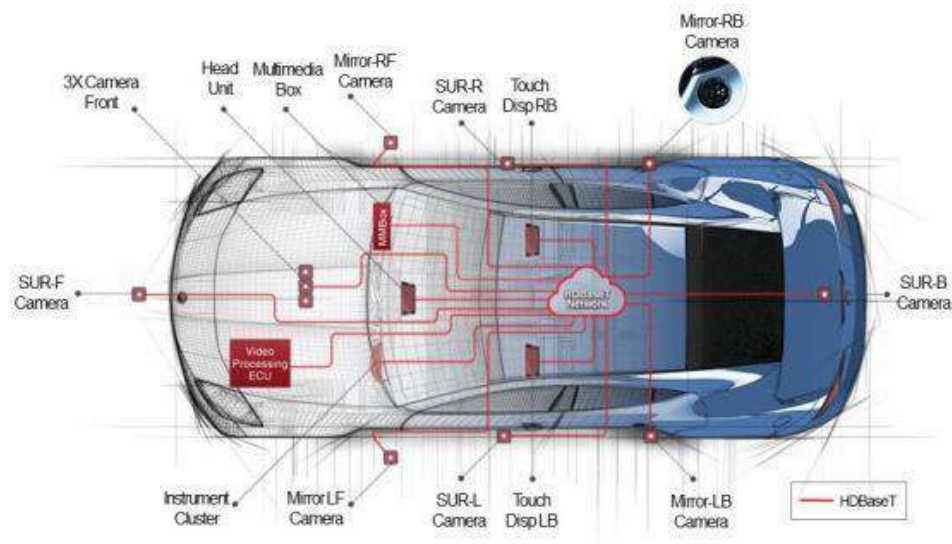
Figure 2.5: An instrumented vehicle used for on-road studies

Naturalistic driving experiments

Naturalistic driving experiments is a relatively new research method used to observe and collect data of the everyday driving behaviour in natural driving conditions. To this end, equipment installed in participants' own vehicles, record manoeuvres occurred, driver's in-vehicle physical motions (e.g. eye, head, hand manoeuvres) and external driving conditions (Figure 2.7). In some naturalistic driving studies (SWOV, 2010), participants are not given any specific instructions and no intervention is made, to let drivers drive the
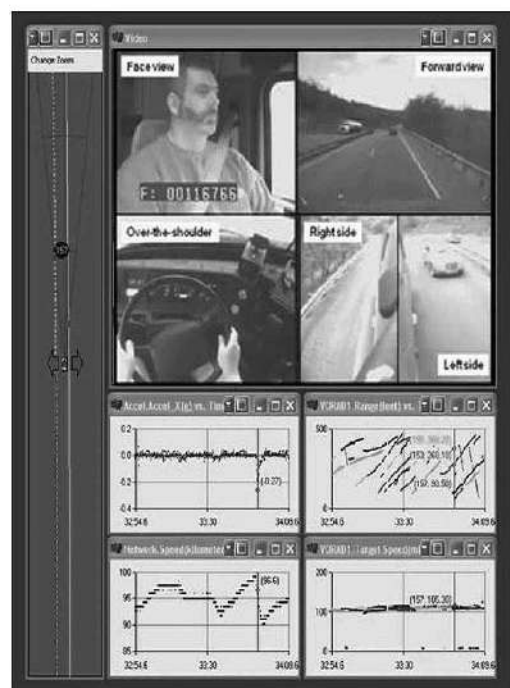
Figure 2.6: Naturalistic driving data collection (FHWA-HRT-12-040, 2012)

way they would normally do in their own car. This provides very interesting information about the relationship between driver, road, vehicle, weather and traffic conditions, both under normal driving conditions and in the case of unforeseen events or incidents.

Naturalistic experiments provide a wide perspective of understanding normal microscopic travel and driving behaviour in normal conditions. It is extremely important for participants to be involved in an experiment that no experimenter is present, there are no experimental interventions and does not include events or incidents that participants have prior knowledge or can guess and act for. Moreover, there should not be a possibility for participants to observe or predict conflicts, near crashes or even actual crashes in real time without potential biases of post-hoc reports. Finally, a naturalistic study (Regan et al., 2012) can help a) determine accident 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.

Nonetheless, an important disadvantage of naturalistic driving studies is that there is no experimental control of the various variables that can potentially affect the road user's behaviour. This is probably because these studies usually result in specifying a correlation between variables examined and driver's behaviour and not an unambiguous causal relationship. Moreover, in real driving conditions traffic incidents are extremely rare and not as usual as they appear to be in a naturalistic driving experiment scenario. A general assumption is that participants in a naturalistic study drive normally because after a while they forget that they are being recorded. Despite the fact that there are strong indications that this is what actually happens, strict scientific proof is still lacking. Finally, since drivers in the experiment are participating on a voluntary basis, the observed behaviour may not always be representative of the whole population. It cannot be inferred that there is no self-selection bias and that volunteers do not significantly differ from non-participants. However, the direction and the approximate size of such a bias can be established and taken into account by using carefully designed background questionnaires (Van Schagen et al., 2011). The advantage of the present research is that since benchmarking is performed, the safety efficiency index of each participant is relatively estimated and therefore conclusions drawn partially take into account existing bias. Nonetheless, the more representative a driving sample is, the more accurate the results.

Driving simulator experiments

Driving simulators are used for the assessment of several driving performance measures in a controlled, relatively realistic and safe driving environment (Figure 2.8). Nonetheless, there are many kinds of driving simulators with different characteristics, which can affect their realism and therefore the validity of the results obtained.

Driving simulators have a number of advantages and disadvantages (Blana & Golias, 2002). First, the fact that a safe environment is provided for the examination of various driving scenarios keeps the driver safe without having real crash risk but on the other hand, this gives the driver a tolerance for driving in a more risky way than in a real case scenario. As for the type and difficulty of driving tasks, these can be precisely specified, and any potentially confounding variables, such as weather, can be eliminated or controlled. On the contrary, these variables cannot be controlled in on-road experiments but conditions are more realistic. Conducting an experiment on a driving simulator, particularly a high-fidelity simulator, can be very expensive because of the high installation cost, while it might be significantly less for an on-road driving experiment especially for smaller-scale experiments. In general, data collected from a driving simulator include also the effects of adjusting on how to use the simulator and may also include the effects of being monitored by the experimenter. Simulator sickness is another problem encountered when conducting simulator experiments and appears particularly when older drivers participate (Papantoniou et al., 2013). Finally, the data volume collected from simulators is relatively small compared to an on-road naturalistic driving experiment where a vast amount of data can be collected.

Figure 2.7: Driving simulator experiment

In-depth accident investigation

Trained experts from multiple disciplines conduct in-depth accident investigations to collect on-site information useful to describe the causes of an accidents or a collision. These studies aim to reveal in-depth information on the facts that led to an accident by describing the accident process and determining appropriate countermeasures.

Additionally, in-depth data shed light on injury prevention research by identifying the injury outcomes in different impact scenarios, including vulnerable road users, and how the interaction between different vehicle types affects injury outcome. These data have also

been utilized for future research and development as a tool to come up with innovative ideas and evaluate the expected effectiveness of innovative safety systems.

The basic disadvantage of in-depth accident investigation data is that there is insufficient reconstruction evidence in each case investigated and the long time required for the final investigation conclusions to be drawn (Hill et al., 2012).

Surveys on opinion and stated behaviour

A reference questionnaire based on a list of selected topics is built in stated behaviour surveys and a representative sample of population is interviewed. In order to answer the research questions raised, the survey approach can employ a range of methods e.g. postal questionnaires, face-to-face interviews, and telephone interviews.

Surveys produce data on the basis of real-world observations, which allows for the investigation of new situations, outside the current set of experiences. Furthermore, a wide spectrum of many people or events is covered, which means that it is more likely to obtain more representative data that will lead to more representative results that can be generalized more easily. A large amount of data in a short time frame and for a fairly low cost can be obtained through surveys, making it easier to plan and deliver desired results.

The main disadvantage is that questions are often hypothetical and the actual behaviour cannot be observed, while data produced are likely to lack from important details or depth on the topic that is investigated (Kelley et al., 2003).

Experiments overview

As it appears in Table 2.6, each driving behaviour assessment method may have different advantages and limitations.

Table 2.3: Comparative assessment of experiments.

Experiment type	Method / tools	Advantages	Limitations
On-road	Instrumented vehicle	Large degree of control over the variables, examination of driver competency	Data collection for a short period, in response to selected interventions, high cost
Naturalistic driving	Systems installed in participants' own vehicles	Understanding normal traffic, observation of conflicts	No experimental control of variables, traffic incidents are very rare
Driving simulator	Driving simulator	Safe environment, greater experimental control, large range of test conditions	learning effect, simulator sickness, very expensive
In-depth accident investigation	Trained experts investigate the causes of an actual accident	Identification of the factors contributing to an accident, research into injury prevention	Insufficient reconstruction evidence, long time period
Surveys on opinion and stated behaviour	Questionnaire	investigate new situations, large amount of data in a short time, low cost	Hypothetical questions, data lack details, self-reported data

The selection of the appropriate methodology for assessing driving safety performance should be carried out based on the research questions of the assessment and the

objectives of the research conducted, the time frame, the infrastructure, available resources etc. To this end, all experiment types should also carefully follow the basic experimental design principles, allowing for reliable data analysis. The selection of the appropriate and relevant driving performance measures, the application of appropriate analysis techniques, and the reliability and validity of the analysis are also very important analysis challenges that should be addressed when assessing driving performance.

Simulators (Pavlou et al., 2016) and questionnaire surveys (Vardaki & Karlaftis, 2011) are usually employed to assess the influence of various human factors on driving performance, especially in older drivers, yet they suffer from the known limitations of self-reported information. Naturalistic driving experiments are considered to be more appropriate for the assessment of driving behaviour (Tselentis et al., 2017, Yannis et al., 2017, Tselentis et al., 2018, Papadimitriou et al., 2018). This is because behaviour is recorded under normal driving conditions and without any influence from external parameters such as the presence of an experimenter, prior knowledge or possibility for participants to observe or predict conflicts, near crashes or even actual crashes in real time without potential biases on the recording. Furthermore, if drivers are monitored for an appropriate amount of time, driving under normal conditions will be recorded and no bias will appear because of the fact that drivers are aware that they are being recorded.

2.3.2) Driving metrics - Adequate amount

As it appears, this rapid technological progress along with the increasingly penetration rate by drivers (e.g. Smartphones), provide unprecedented opportunities to accurately monitor and analyse driving behaviour. First results from related applications (Tselentis et al. 2017, Theofilatos et al. 2017, Araújo et al. 2012, Enev et al. 2016) have confirmed the efficiency and usefulness of such big data collection schemes. Nevertheless, the exact amount of the necessary driving data that should be collected for each driver in driving behaviour assessment is not determined yet. Since both small and big data samples lurk the risk of leading to doubtable results, by acquiring a sample either biased or computationally expensive to analyse, it is a matter of great importance to specify exactly how much driving data should be recorded from each participant in the experiment.

Literature review conducted, revealed that there is not enough research addressing the required amount of data, in terms of recording period, to identify driving patterns or evaluate the improvements achieved by intervention programs over time (Musicant et al., 2011). An aspect close but indirectly related to this particular issue has been studied since the 1960s (Perkins, Harris, 1968; NCHRP, 1999), in an effort to answer the research question of how long a site should be observed to obtain reliable estimates of conflict rates (Robertson et al., 1994; Parker and Zegeer, 1988). Reviewing the recent literature on driving behaviour evaluation using the IVDRs, it was found that there is no unanimity regarding the required driving data measurements for driving assessment. Many studies using IVDR, conclude to a variable range that includes 80 h (Musicant et al., 2007), 400 h (Neale et al., 2002), and 2107 h (Musicant et al., 2011) per driver. It should be highlighted though, that the scope of each research as well as the limitation of

each study (driving conditions (driving at night, peak hours etc.), type of road, drivers (professional or not) etc.).

There is only a handful of studies (Shichrur et al., 2014) that introduce methods for determining the required sampling time-frame from an IVDR in order to analyze and assess driving behaviour. Shichrur et al., (2014) installed IVDRs in the vehicle of each of the 64 taxi drivers participated, recording detailed information about unforeseen events occurred during the trip with regards to vehicle's position, speed, vertical and horizontal acceleration, and maneuvers. This research concludes that i) collecting a sample of more than 300 hours per driver is more likely to result in a relatively stable and reliable measure of driver's average event rate, ii) sampling less than 100 driving hours per driver is not likely to result in a reliable measure and iii) sampling between 100–300 h may also result in a stable measure but it is less recommended.

It is therefore deemed to be valuable to determine the amount of data that is necessary to be recorded, in order to obtain a complete picture for each driver, where the rate of those metrics described above per km travelled converges to a stable value. One of the objectives of this dissertation is also to examine and quantify the need for driving data collection in driving performance assessments based on data collected through smartphone devices.

2.4) Industrial and operator perspective

When studying driving efficiency measurement, it is also very important to take into account and highlight the impact that such a study will have on several different aspects. In this study, this is done by studying driving efficiency measurement from an industrial and operator perspective. In order to obtain a valuable insight of the correlation between driving behaviour and crash risk it is extremely useful to study the evolution of conventional to usage-based insurance (UBI) that addresses the real life problem of assigning insurance premiums to the respective accident risk.

UBI schemes

Each road network user is charged a lump sum according to the current pricing policy of motor insurance companies around the world. This has been considered for long unfair and inefficient (Butler et al., 1988) since drivers with similar demographical characteristics (age, gender, etc.) pay approximately the same premiums regardless of the distance they drive each year and their driving characteristics. This approach is often compared (Bordoff and Noel, 2008) to an unlimited food policy restaurant that charges a fixed price per person, a fact that encourages people to eat more. Respectively, the conventional insurance pricing policy neither discourages drivers from driving more kilometres annually nor punishes risky driving behaviour and, on the other hand, it does not reward prudent driving behaviour. Most of all, this implies increased number of crashes, congestion conditions, carbon emissions, local pollution and dependence from oil. Drivers with lower annual mileage and safer driving behaviour are literally forced by the

unfair conventional pricing system to subsidize the insurance costs for drivers who drive more kilometres per year and in a more risky manner. Apart from that, research findings indicate that people with lower income drive fewer kilometres, which allows to draw the conclusion that most existing policies promote social inequities (Litman, 2002).

It should be highlighted that within this part of the review, driver's strategic choices (at real-time or not) regarding type of road network used and time of the day driving in order to fulfil travel needs arose will be referred as driver's travel behaviour. Choices made are directly linked to driver's exposure and therefore accident risk, through distance driven, road network type chosen and the related traffic conditions, the time of the day chosen to drive and the related weather conditions. Insurance charging systems based on travel behaviour are often called pay-as-you-drive (PAYD) UBI schemes. Furthermore, driver's operational choices at real time regarding handling of the vehicle within the existing traffic conditions will be referred as driver's driving behaviour. These choices are also directly linked to the probability of getting involved in a traffic accident, based on the way someone's driving, e.g. by driving over the speed limit, harsh braking/ accelerating/cornering event occurrence, distraction generated by mobile phone usage, etc. Insurance charging systems based on driving behaviour are often called pay-how-you-drive (PHUD) UBI schemes.

Generally, a probability of crash involvement based on driving behaviour could be assigned to each driver (Tselentis et al., 2017). If all drivers are charged a lump sum, it is assumed that crash probability is approximately equal across the entire drivers population. Evidently, this is not a user optimum and socially equitable approach, since drivers with lower crash risk are imposed to "subsidize" those with higher risk. In other words, safer drivers are imposed to "purchase" higher probability of crash risk than the actual existing, unlike risky drivers who "purchase" less.

There are potentially significant effects on safety from an innovative insurance policy depending on its design (Zantema et al., 2008). Since different driving styles could be sorted on a high to low risk scale (Sagberg et al., 2015) and create thus a safety scoring scale, differentiating premiums to reflect safety, more specifically by charging higher fees for unsafe road categories and night-time driving, most effectively and apply it to all drivers is a feasible solution. The insurance policy based on vehicle use (UBI) includes PAYD and PHYD systems. As mentioned above, premiums are charged based on total travel behaviour characteristics such as mileage and road network used in PAYD while in PHYD, this is based on measuring parameters such as speed, harsh acceleration/ braking etc., i.e. on individual driving behaviour. The automotive diagnostic systems, OBD (On-Board Diagnostics), installed in the vehicle and/ or the drivers' smartphone is used as the main data source to collect the aforementioned parameters, sending all necessary information in a central database via mobile network.

If PAYD schemes were to be implemented in the Netherlands, total number of crashes could be reduced more than 5% leading to 60 less fatalities and 1000 less injuries each year (Zantema et al., 2008). Research in other countries outside Europe (Reese and Pash, 2009) on differentiating premiums indicates the same percentage of 5% mileage reduction on average although driving during low and medium risk hours was only significantly reduced.

UBI data collection

Until recently, it was extremely difficult to collect and manage real-time data and therefore to study the relationship between driving behaviour and travel behaviour and the probability of crash involvement, due to the high cost of real-time driving data recording systems, data programs, cloud computing services, the inability to accumulate and exploit massive databases (big data) for transport and traffic management purposes (De Romph, 2013, Lee, 2014) and the low penetration rate of Smartphones and social networks. Research (Reese and Pash, 2009) so far showed that these barriers can be overcome when consumers are given an incentive such as a monetary prize, which along with providing personalised feedback to drivers about their speeding are extremely effective (Ellison, 2015a) at encouraging them to drive (mainly in terms of speeding) safer. According to literature (Elvik, 2014), the highest reduction rates achieved by incentive schemes in speeding are around 60-80% and 0-10% for mileage reduction.

As a result, the wide provision of information on the social benefits that could arise from an implementation of such a policy is the main present challenge road safety entities and policy-makers are facing. It is a fact that users who were given a medium value financial incentive of $88 per 6-months period to reduce their mileage, noted a high level of interest. A mileage-based insurance could probably lead users to consider car sharing or using public transportation (Reese and Pash, 2009) and lower insurance premiums is proved to be among the strongest incentive for switching to such a policy.

High quality real-time data can be collected in an efficient way in order to model both individual and total crash risk. In most studies data are recorded either by vehicle's OBD (Jensen et al., 2011) or user's smartphone (Handel et al., 2014) and transmitted to a central database for processing and analysis (Boquete et al., 2010, Iqbal and Lim, 2006). This allows for the development of special indicators to estimate driver's travel (PAYD) and driving (PHYD) behaviour.

Data obtained from GPS, EOBD system and mobile-telephone use detection circuit (Boquete et al., 2010) are usually acquired and processed from an in-vehicle device. Data are transmitted via a mobile telephone connection to a control centre (CC), where individual crash risk for each vehicle is estimated. Mobile telephone connection is used for data transmission between the on-board system (OS) and the CC. The EOBD system, the GPS and a detector circuit (RF energy scavenging) respectively captures function data of the vehicle (e.g. number of seatbelts fastened), vehicle position-speed data and driver mobile-telephone use data, all of which are ultimately acquired by the OS. Before transmitted to the CC, data captured by the OS are processed and stored by a high-performance microcontroller that exists inside the core of the OS.

There are also other studies in literature (Iqbal and Lim, 2006) that incorporate light or weather sensors which interact via a communications channel (infrared or Bluetooth) with the on-board computing unit and report a numerical value. The GPS receiver continuously records and transmits all information regarding position, speed and time to the central computing unit.

These systems can also be exploited for real time traffic monitoring (Barmpounakis et al., 2016) since a few technological obstacles that exist nowadays are overtaken. Extraction

of vehicular trajectories from video recordings using a trajectory extraction system is also used to collect vehicle traffic data (Barmpounakis et al., 2015). This is not available for real time traffic monitoring at present but it is very likely to be used for this purpose in the near future.

Table 2.1 summarizes the main methods of several the telematics manufacturer to transmit collected data to the CC. Transmission methods include USB cable connection with the OBD and the CC, GPRS/CDMA network, wirelessly from a Bluetooth device built-in the OBD or micro-SD card. The installation cost is moderate whereas the monthly/yearly fees vary from a $0 to $19 charge every month after the first year of installation.

Table 2.4: Manufacturers providing Telematic recording devices of driving characteristics.

Manufacturer	Data recorded: Distance, speed, time	Method of transmission	Installation cost	Monthly/yearly fee
CarChipFleetPro	Distance, time, acceleration, speed, GPS location, fuel, Engine speed	USB cable/port (customer loaded)	$149 (plus a $395 charge for software, one per fleet) Can also be used wirelessly with a $200 base unit	None
Sky-meter	time, distance, place, speed, acceleration of all driving, and the location and time of all parking	GPRS/CDMA (other protocols available at extra charge)	$50 - $250 activation fee	$5 per month plus 5%–8% of monthly premium (depending on volume)
OnStar	Distance, speed, time, (incl. other features)	Automatic through GPS S	First year free for new GM cars (only available for GM)	$18.95 per month after one year
Freematics	Speed, distance, time, location, acceleration, engine RPM	Built-in Bluetooth Low Energy and SPP module for wireless data communication or via microSD card (32GB)	99$ (Plus $30 for GPS module, plus $10 for MEMS MPU-9150 (9-axis) module, plus $10 for DUO BLE-BT 2.1 and plus 5$ for 32GB microSD)	None
Progressive (MyRate Device)	Distance, speed, time, location, acceleration, trip frequency	Wirelessly	None but $75 fee if not timely returned at end of policy	Varies

Risk factors used in UBI

The indicators recorded by each device refer to travel and driving behavioural characteristics - distance, time, location and speed, acceleration/deceleration, seatbelt use (www.skymetercorp.com, www.carchip.com etc.) to name a few. Apart from these, there are manufacturers that measure additional information such as location and parking duration (www.skymetercorp.com). This information is processed afterwards based on rating information provided by the insurer, to generate the individual risk factors of interest for each user.

Per minute or mile (or km) travelled charge is used so far by some insurers which can be modified based on driver's driving record, vehicle type owned, the class of road, time of

the day driving, the riskiness of the historical behaviour or the riskiness of the current behaviour. Others also charge for parking (www.skymetercorp.com) per hour at high-risk locations (e.g., on street, in mall) but this is beyond the scope of this research.

In general, the main driving indicators mainly used thus far in literature for estimating an individual's driving risk are shown in the Table 2.2 below:

Table 2.5: Risk indicators classification.

Travel behaviour	Driving behaviour
Total distance driven by the user (the higher the mileage the higher the risk)	Speeding expressed either as a percentage of kilometres/time driving over the speed limit or a percentage of speeding
Road network type (increased crash frequency in the cities, increased crash severity outside)	Harsh braking
Risky hours driving (increased crash frequency during a particular hours range)	Harsh acceleration
Trip frequency (a driver is more likely to cause a crash during an infrequent trip)	Harsh cornering
Vehicle type	Seatbelt use
Weather conditions	Mobile phone use

PHYD concept is not yet thoroughly examined and much less implemented than PAYD concept. Nevertheless, it is worth mentioning that only a handful of studies include behavioural characteristics in their models. Thus far, there is only one insurance company exploiting behavioural information to assess driving behaviour and estimating their charges (https://www.progressive.com/auto/snapshot/).

On a research level, there are several indicators both for travel behaviour (vehicle maintenance condition, safety rating of the vehicle from the IIHS (Insurance Institute for Highway Safety)) and driving behaviour (harsh cornering, alcohol use, ecological driving etc.) that affect crash risk as well but are not yet incorporated in risk modelling. Eco-driving for instance, is a factor considerably significant for crash risk estimation (Haworth and Symmons, 2001). According to the manufacturer's specifications, conclusions can be drawn about how someone's driving (aggressively, over the speed limits etc.) if fuel consumption estimated by the manufacturer is compared to the real fuel consumption recorded. Furthermore, the simultaneous existence of two driving traits such as excessive speeding during risky hours timeframe or braking harshly while using the mobile phone might excessively affect accident risk. All the above should be further investigated to conclude on their significance to crash risk modelling.

It should be mentioned however that some of the indicators mentioned above such as alcohol use cannot be taken into account in the driving behaviour models of the present analysis as they cannot be captured efficiently yet. Nevertheless, it is very likely for scientists to be able to monitor these factors in an easy and reliable manner in the near future and therefore exploit this information as well.

Travel behaviour-based Insurance (PAYD)

Several studies focus on the correlation between kilometres travelled and driving risk and therefore the determination of the probability of a driver to get involved in an accident. In the initial form of PAYD models, mileage was only included as a travel behaviour characteristic. This implementation was based on the fact that mileage and accident risk are proved to be close related based on past research conducted on the field. There are certainly many studies (Litman, 2005, Bordoff and Noel, 2008) that result to a close relationship between VMT (vehicle miles travelled) reduction and the crash risk reduction. Edlin (2003) finds that the elasticity of the number of crashes occurring with respect to VMT is approximately 1.7, from which it can be inferred that if total mileage were reduced by 1%, this would lead to a 1.7% reduction of the number of crashes. Results from other research indicates that the elasticity of crash risk is around 1.2 (ICBC Research Services Data, 1998) and more specifically, that the 1981-1982 recession led to a 10% VMT and 12% insurance claims reduction in British Columbia. Ferreira and Minikel (2010) revealed that, in support of the above, the statistical significance of mileage and risk's positive correlation is high. It should be highlighted that the above findings are interpreted based on the concept of elasticity, which is defined as the relative importance of an independent variable in terms of its influence on the dependent variable. In other words, it can be explained as the percent change in the dependent variable engendered by a 1% change in the independent variable (Washington et al., 2010).

On the other hand, much research is conducted on the relationship between mileage and crashes with a number of them indicating that there are serious grounds to believe that this relationship is neither linear nor proportional (Janke, 1991, Litman, 2008). The number of road accidents divided by the number of kilometres driven by a group of users should not therefore be expected to remain constant. Recent research (Ferreira and Minikel, 2010) concludes that when all vehicles are considered together with class or territory differentiation, the relationship between risk and mileage is less-than-proportional and when these factors are not taken into consideration the relationship is less-than-linear.

It is also found in literature (Janke, 1991, Langford et al., 2013) that most lower mileage drivers groups (e.g. young and older drivers) tend to have a higher crash rate compared to that of higher mileage drivers. The general trend is that per mile crashes decrease as annual mileage increases (Litman, 2008) which is mainly attributed to factors such as driving more kilometres in congested urban areas where crash risk is higher and less driving experience for low mileage drivers, medical conditions for older drivers etc.

The early stage of the mileage-based insurance scheme that appeared later was presented by some studies in the past as the Pay-at-the-Pump (PATP) approach. PAYD and PATP approaches share many similar characteristics and the same conceptual basis considering that fuel consumption and mileage are somehow correlated. According to literature, PATP is probably the second most influential method of UBI and considers fuel consumption as its main estimating parameter of insurance premiums instead of mileage.

Based on the above it is clear why first studies were focusing mainly on the development of mileage-models considering mileage as the most (and sometimes the only) influential

factor for crash risk. Mostly used models are presented and described below. It is highlighted that risk predictability increases when mileage is incorporated together with other rating factors in the model and does not stand alone (Litman, 1997, Ferreira and Minikel, 2010). As shown (Ferreira and Minikel, 2010), when combined with space and behavioural information of the miles driven, mileage provides a great explanatory power and therefore is deemed to be a powerful supplement for the rest of the traditional insurance rating factors (e.g. experience and territory). This would further increase fairness among motorists as drivers would be charged a flat-rate premium per mile, differentiated based on other driving characteristics as well.

Furthermore, it has been found (Lourens et al., 1999) that the influence of sex and education variables is minimized when annual mileage is taken into consideration for crash prediction. At the same time, a strong positive correlation between traffic violation commitment and crash involvement (independent of the annual mileage driven) is seen in literature (Rajalin, 1994, Massie et al., 1997, Lourens et al., 1999) and a well-documented age influence (young driver's age group) is proved. There are a few researchers though (Ellison et al., 2015b) who dealt with the problem of modelling driving behaviour using other exposure spatiotemporal indicators as independent variables instead of mileage e.g. speed limits, school zones, rain, time of the day/ week, number of passengers, vehicle and driver's demographic characteristics.

Pay-at-the-pump (PATP)

Wenzel (1995) argued that insurance premiums should be estimated based on motor use i.e. mileage. The travel behaviour-based system proposed was actually a per-gallon surcharge for consumers, a method similar to the PATP method, because VMT is a good predictor of insurance claims. It was also suggested that premiums should be the sum of a variable amount on the basis of fuel consumption (per-gallon surcharge), plus a fixed amount on the basis of location, vehicle safety characteristics and driving record, most of which are travel behaviour characteristics.

In other proposed forms of PATP (Sugarman, 1994), the funds gathering should take place at the pump in the form of fuel surcharges collected by a governmental or county organization founded for the specific reason. It was suggested that additional charges should be imposed based on drivers' driving record and experience as well as on vehicle ownership, apart from the fuel surcharge. The prior amount was suggested to be defrayed either as an annual instalment or as a once-off fee. It should be noted that this method would substitute tort liability or lawsuit system not for material damages but only for bodily injuries. The author draws the conclusion that this new system will provide better compensation, fairer funding and most of all greater safety for most users. On top of the benefits presented, it is argued (Litman, 2004, Sugarman, 1994) that the new vehicle injury plan (VIP) would assist in overcoming many problems that appear in today's insurance policy e.g. a large percentage of premiums is attributed to other reasons such as claims administration, duplication of other sources of compensation, pain and suffering rewards or lost to fraud, the enormous number of seriously injured victims that are vastly undercompensated, the unsatisfying claims process, the long delays of many bodily injury claims payment and that safer driving and safer vehicles are insufficiently encouraged.

Some researchers (Khazzoom, 2000) estimated the marginal travel behaviour risk of the average driver to be approximately 2c/mile and suggested a fuel surcharge of 50c/gallon. They also argued for VMT-based over PATP insurance stating that the former removes uninsured motorists from the road and does not encourage them solely to switch to fuel efficient vehicles burdening this way the environment and that it does not have any implementation problems. Research on PATP (Kavalec and Woods, 1999, Khazzoom, 1999, Khazzoom, 2000) indicates that this insurance scheme results to welfare benefits with both a direct and an indirect manner. The average driver might be benefitted either directly by paying less for insurance premiums and having an enhanced road safety system or indirectly by enjoying societal benefits such as reduced external costs e.g. reduced energy consumption, congestion, greenhouse gases, emissions etc.

Nonetheless, because of the drawbacks referred above for the PATP method, it was not extensively implemented. Kavalec and Woods (1999) claimed that introducing a surcharge for gasoline is an incentive for consumers to drive vehicles that are energy efficient in order to reduce their tax exposure and therefore not reduce their annual mileage significantly. Khazzoom (2000) raised the issue that differences in vehicle fuel efficiency are probably leading to a discrepancy between drivers that is yet a fairer policy than today's lump sum policy. PATP might also cause a slight shift to energy efficient vehicles, according to the author, a fact that will increase the above mentioned discrepancy even more. Previously (Khazzoom, 1999), criticism against PATP could be classified into two categories which are the criticism of PATP design such as state bureaucracy, uncertainty of insurers' income and long-distance motorists penalization and the consequences of adopting this new method such as the burden on lower income insurers and the shift to fuel efficient vehicles.

Mileage-based insurance

Because of the PATP method drawbacks, efforts thereafter focused on distance-based methods that are "penalizing" driving in a more direct way. For instance, the potential of paying premiums proportionally to vehicle-kilometre use (PAYD) was examined (Weaver, 1970) as a possible solution for the economic asymmetry that exists in the vehicle insurance market. Research results indicate that the new insurance method has the potential to reduce transaction costs, lead to more cost-efficient consumer behaviour, reduce premiums and benefit insurance companies, allowing for the creation of enhanced insurance policies that will be representative of the actual individual risk of each user. Past research also examined both social benefits and obstacles that are likely to result from the implementation of such a policy e.g. reducing GHG emissions and CO_2, dependence on oil, lowest number of crashes, the reduced need for maintenance of the infrastructure etc.

A Texas mileage research conducted by a US PAYD provider (Progressive Insurance, 2005), was outstanding in terms of the number of observed vehicles and the experiment's duration (36 months and 203,941 vehicles insured; nonetheless, the authors do not provide a detailed description of their sample selection). This study's final report presents the relationship between annual mileages and insurance losses incurred for different coverage types using a methodology that was based on regression analysis. It was

shown that there is a strong impact by the number of vehicle-miles travelled by the user on insurance claims (dependent variable). The basic model tested was a linear regression model featuring a > 0.82 R^2 (goodness-of-fit indicator). No more variables other than annual mileage are tested for correlation with insurance claims in this study.

A mileage-based model (PAYD) development and evaluation (Bordoff and Noel, 2008) resulted that each household can reduce vehicle insurance contributions up to $ 270. It was pointed out that if a per kilometre charge policy was applied, drivers would have an extra incentive to drive less, which would result in a reduction to the total number of crashes. A reduction of the number of vehicles was estimate to be around 8%, a figure equivalent to $ 50-60 million due to reduced harmful effects on driving. The latter would also lead to a reduction of 2% and 4% to carbon dioxide emissions and oil consumption respectively. Regarding projected annual vehicle-kilometres, it is also shown (Nichols and Kockelman, 2014) that the average vehicle will be driven less by 2.7% (a reduction of 237 miles per year), with benefits of only $ 2.00 per vehicle for average consumers, with a premium that is partly fixed and partly based on mileage. Because of the convex relationship between vehicle mileage and accident probability, drivers with lower annual vehicle-kilometres are expected to receive the greatest social benefits. Therefore, PAYD policy can reduce vehicle-kilometres travelled annually and lead to a fairer premiums system, which supports the findings of literature thus far.

Examples of the above mentioned PAYD models in practice are National General (http://www.nationalgeneral.com/auto-insurance/smart-discounts/low-mileage-discount.asp) which is providing a discount of up to 54% and Metromile (https://www.metromile.com/insurance/) Insurance companies which charging 3.2¢ per mile.

Behaviour-based insurance (PHYD)

The main argument of literature thus far (Kantor and Stárek, 2014) against current PAYD systems is that there are several weaknesses and shortcomings since they focus solely on the number of driven kilometres and not on driving behaviour which is more significant. Evaluating a user's driving behaviour is most of the times more crucial to crash risk estimation than the quantity of kilometres he has driven. Modelling the individual driving patterns of drivers in an efficient manner is a matter of significant importance for crash risk modelling, since it allows not only to sufficiently understand differences between driving behaviours but to take them into consideration as well.

Linear modelling approaches are used by most researchers (Iqbal and Lim, 2006, Boquete et al., 2010) to model PHYD insurance. A UBI model that takes into account driving behaviour attributes is implemented for example by Boquete et al. (2010). The on-board system installed in vehicles, transmitted data to the CC using mobile data service. The concept of this research was to build a model for premium estimation based on how much (mileage), where (Zones used), when (Day/night) and how (excessive speeding, harsh accelerations, number of vehicle passengers, mobile phone use) a vehicle is driven. Insurance premiums were calculated as a sum of a linear combination of the above mentioned indicators and their coefficients plus a fixed charge imposed to each

driver. Other recent studies (Iqbal and Lim, 2006) also included driving behavioural attributes in cost calculation incorporating exposure characteristics such as weather and light conditions risk, rush hour risk and road network risk terms as well as speeding risk terms in the form of the percentage of driving over the speed limit after detecting the road network type used. For premium cost computation, this study (Iqbal and Lim, 2006) proposed the product of a base rate for each driver by all risk factors calculated for each indicator used (road network type, excessive speeding etc.).

Other studies (Kantor and Stárek, 2014) proposed alternative methods such as a fuzzy-linguistic approximation apparatus which is proved to be a suitable tool considering the insufficient exact knowledge and the large possible combinations of the parameters used as model's input. The process of driving pattern assessment was successfully incorporated by a concise algorithmic procedure and a projection of that evaluation into the insurance premium was made. The algorithm consisted of six algorithmic steps i.e. data collection, meteorological conditions evaluation, vehicle dynamic qualities determination, manoeuvre type determination, manoeuvre style evaluation and finally number of penalty points assignment and determination of driving style sanctions. The types of manoeuvres taken into consideration were driving straight, turning, overtaking, speeding, aggressive deceleration, non-fluent driving (frequent acceleration and deceleration) but the manoeuvre style is evaluated for driving straight, turning, overtaking and aggressive braking. The parameters visibility, deteriorated road conditions, sufficient vehicle performance, acceleration in x and y axes, speeding, motorways and roads (directions separated or not separated) were used as inputs for the fuzzy model. An algorithmic procedure to form the scoring procedure describing the risk profile based on the figure of merits of actuarial relevance is also followed by Handel et al. (2014) using smartphones as measurement probes. Parameters used included speeding, road network type, risky and rush hours driving, harsh acceleration, harsh braking, harsh cornering, manoeuvre type, trip duration, energy consumption, trip distance and smoothness.

Chowdhury et al. (2014) proved that the statistical analysis and algorithmic approach applied to estimate driving score are able to capture the relationship between jerk energy (first derivative of acceleration in m/s^3) and speed. A scoring mechanism for monitoring a vehicle was successfully established based on this relationship through large-scale data collection of a large number of vehicles that was made possible by OBD devices and smartphones. According to the authors, the algorithm presented can serve either as service analytics or for PHYD insurance model premium computation.

There are also some studies in literature where several methods were tested to find the best fit. The potential of high-resolution travel behaviour data for PAYD insurance pricing was demonstrated by Paefgen et al. (2013) by training and testing the applicability of three different approaches, logistic regression, neural network, and decision tree classifiers and compare their outcomes. Speeding, road network, risky and rush hours, mileage and day of the week were the significant predictor variables in this study with vehicle mileage to be the strongest single predictor variable; it was noted that particularly for logistic regression its predictive power was further improved by applying a logarithmic transformation. PAYD insurance data recorded by in-vehicle data recorders (IVDR) from 1600 vehicles and obtained from an insurance provider were exploited by Paefgen et al.

(2014). The authors developed and validated a variety of models to investigate and explain the differences that exist between vehicles that get involved in crashes and those that do not. Logistic regression modelling techniques are applied to estimate accident probability. As shown, crash risk fluctuates during the day (lower for daytime and higher for nightfall), the week (lower risk on Friday and weekends), road network type (higher risk on urban roads) and velocity range (mid-range (60 - 90 km/hr) velocities are associated with lower risk compared to low-range (0 - 30 km/hr) and higher range(90 – 120 km/h)).

Hultkrantz & Lindberg (2011) tested a variation of PHYD named Pay-As-You-Speed and simulated an insurance scheme based only on speed limit exceedance. The experiment's duration was two months and participants that took part were divided in two groups; those that were receiving a malus/bonus for their speeding behaviour and those who were not. A fixed monthly payment deducted every time the participant was violating speeding traffic rules was received by each participant of the first group. According to findings, severe speeding violations were reduced during the first month but, after participants received their feedback reports with an account of earned payments, those not given a penalty did not change their behaviour in the second month. Research performed so far on PHYD schemes, indicated that it presents many potentials and appears to have many benefits. Despite the fact that PHYD is undoubtedly the best way to evaluate driving behaviour and estimate crash risk, it still remains a sharp shift from today's lump sum policy; an alteration that requires significant effort to be diffused in society. PAYD methodologies implemented so far seems to be very persistent and unilateral in terms of the parameters considered. Regarding travel behaviour-based modelling, mileage is not the only factor influencing crash risk and therefore multivariate travel behaviour-based insurance models should be developed to consider parameters such as road network used, time-of-the-day driving etc. (on the top of mileage driven).

Table 2.6: Usage-Based insurance model literature.

Reference	Model Type	Speeding	Road Network	Risky hours	Rush hours	Harsh acceleration	Mileage	Harsh braking	Manoeuvre	Harsh Cornering	Day of the week	Visibility	Territory	Vehicle type	Passengers	Trip duration	Weather	Smoothness	Trip distance	Jerk Energy	Mobile phone	Energy consumption	Fuel	Gender	Class
Khazzoom (2000)	Linear																						√		
Progressive Insurance (2005)	Linear						√																		
Bordoff & Noel (2008)	Linear						√																		
Ferreira & Minikel (2010)	Non-Linear (Poisson)						√																		
Ferreira & Minikel (2010)	Linear						√																		
Buxbaum (2006)	Linear						√																		
Zantema et al. (2008)	Linear						√																		
Litman (2005)	Linear						√																		
Ferreira & Minikel (2010)	Non-Linear (Poisson)																							√	√
Ferreira & Minikel (2010)	Linear																							√	√
Ferreira & Minikel (2010)	Linear						√																	√	√
Kantor & Stárek (2014)	Non-Linear (algorithm)		√	√	√		√																		
Iqbal & Lim (2006)	Non-Linear	√	√		√								√			√									
Boquete et al. (2010)	Non-Linear	√		√		√	√							√	√					√					
Handel et al. (2014)	Non-Linear (algorithm)	√	√	√	√	√		√	√	√							√	√	√		√				
Paefgen et al. (2013)	Non-Linear	√	√	√	√		√				√														
Paefgen et al. (2014)	Generalized Linear/ Non- Linear	√	√	√	√		√				√														
Chowdhury et al .(2014)	Non-Linear																			√					
Hultkrantz & Lindberg (2011)	Non-Linear	√																							

Overall

The scope of studying UBI is to develop a premium calculation system based on travel and/or driving behavioural characteristics and to ultimately create reliable models that associate driving risk and travel behaviour (PAYD models) and/or driving behaviour (PHYD models) and charge road users based on driving risk. As for PAYD, literature review revealed that premium calculation method is based only on travel behaviour characteristics. Risk is correlated only with the exposure of a vehicle, assuming that the probability of a crash occurrence increases as some indicators referred above (e.g. driven kilometres) increase. Figure 2.3 illustrates that the traditional insurance approach considers neither the exposure of a vehicle nor the behaviour of a user and assigns an "average premium", which corresponds to the "average driver" and consequently to an "average crash probability", to a specific vehicle and driver. PHYD is on the other hand based on the evaluation of user's driving and travel behaviour leading to a more realistic estimation of the corresponding risk. PHYD models incorporate a large number of parameters allowing for the accurate estimation of driving risk. The final outcome of PHYD models is an individual risk indicator that depicts the risk associated with a user's driving behaviour. Since premium calculation in PHYD is based on the evaluation of driving behaviour of a user, it can be concluded that it leads to a more realistic assessment of driving risk than PAYD approach does.

Traditional motor insurance has started to gradually transform into UBI during the last few decades. The question remaining is to what extent UBI will be widely adopted and which indicators are going to be fully incorporated. UBI is expected to play a key role in future motor insurance market and therefore it will significantly influence traffic safety. Figure 2.3 illustrates the types of insurance that currently exist in the marketplace as well as a prediction on the form of future motor insurance. Since motor insurance's trend is to implement innovative schemes that embed travel and behavioural factors it is believed that future models will be in the form of Pay-As-How-You-Drive (PAHYD) including parameters from both PAYD and PHYD models.
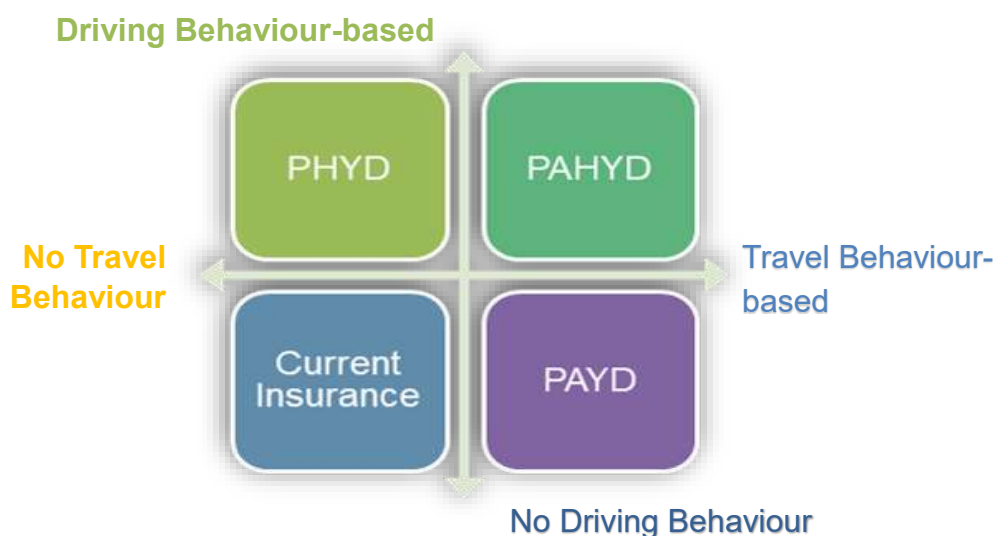
Figure 2.8: UBI and current insurance policies

It can be said that the PAYD model using fewer parameters as risk indicators is a more simplistic approach. Nonetheless, there are significant advantages since (a) implementation is easier, (b) there is a significantly shorter period for developing and verifying the model (less data required and, significant information may be found in literature and reports of relevant organizations), (c) it is targeted to vehicles that are less often used. On the contrary, PHYD is a more sophisticated approach that aims to (a) associate driving risk with a large number of indicators quantifying driving behaviour in a realistic manner (b) raise driving awareness and motivate drivers to evaluate and improve their own driving behaviour and (c) decrease insurance companies' claims via driving improvement.

According to literature (Litman, 1999) on how well insurance pricing schemes represent crash risk, the best performing models are those taking into account time and location information (PHYD), followed by mileage-based models (PAYD), PATP models (PAYD), fixed vehicle charges models (current insurance policy) and external costs (not charged to drivers) models respectively.

Finally, despite the significant contribution of past research on PAYD pricing, a small percentage to date has dealt with PHYD systems. As previously mentioned, this methodology is proposed to cope with the problem of estimating and assigning individual crash risk and therefore estimate personalized insurance premiums. It is deemed to be a more objective approach since it takes into account several significant factors such as sudden braking / acceleration events, driving over the speed limits, mobile usage etc. which makes it a tool that is more reliable for calculating the probability of accident involvement. Future research should mainly focus on this as well as on developing and evaluating PAYD and PHYD models and compare their efficiency.

Finally, all metrics used in UBI modelling are describing and representing a driver's behaviour in an explicit way. Apart from these metrics though, it is worth mentioning that there are factors affecting crash risk and are not yet considered in UBI modelling e.g. mobile phone, seatbelt (recorded from the OBD) usage, alcohol use, reaction time, time to collision (from naturalistic driving experiments), vehicle maintenance condition, vehicle safety rating etc. The combining effect of two different driving characteristics should also be examined such as using the mobile phone and driving over the speed limits. Although some of these factors cannot be currently monitored in an easy and reliable manner, most of them can or will be able to be efficiently captured in the near future.

Literature review conducted above reveals a PAYD schemes trend, which are mainly focusing on the effects, externalities and potentials that UBI offers. Despite the fact that the potential arising from the implementation of PAYD schemes on insurance companies and drivers has been thoroughly examined (Husnjak et al., 2015), PHYD is apparently not exhaustively modelled thus far.

The specific section of the present doctoral dissertation's literature review constitutes a systematic effort to gather, group and present the most scientifically significant studies relevant to UBI approaches which are particularly focused on PAYD and PHYD methodologies. Unlike the past, there is an obvious trend for more personalized motor insurance. Therefore, personal driving characteristics (travel and/or driving behavioural) are gradually incorporated into insurance models instead of estimating insurance

premiums based solely on demographic characteristics such as age, number of years holding a driving licence etc.

Literature review conducted above revealed the extensive effort of analysis and evaluation of PAYD methods. Its small-scale implementation thus far has demonstrated a great influence on all levels, economic, social, environmental, etc. This is a ground-breaking first attempt to alter the conventional insurance charging approach that is currently old-fashioned and unfair to many users and research proves that does not to contribute in crash reduction which is the goal of road safety.

A gradual global transition towards PAYD/PHYD insurance can be envisaged in the near future. Low-risk drivers (low-mileage, less risky drivers etc.) will gain several incentives for opting out of traditional insurance in favour of alternative new insurance policies such as mileage-based insurance (Parry, 2004); this is becoming increasingly feasible while telematics systems are gradually incorporated in modern vehicles. Governments are also likely to encourage this upcoming trend in the future through relative legislation and political decisions such as subsidies, tax waivers for insurance companies offering alternative policies like these aforementioned.

A simplistic approach to calculate annual crash risk is to estimate the product of per-mile crash risk and annual mileage (Litman, 2008). Although imposing drivers to reduce their annual mileage would probably lead to reduced crash risk, this approach does not take into consideration two important factors. First, a driver that is penalized based only on mileage driven is not incentivized at all to improve driving behaviour. Per-mile risk remains an unspecified factor that fluctuates over time and therefore although mileage might be reducing, total crash risk can still be increasing. Second, insurance system remains unfair and the cross-subsidies phenomenon is not eliminated since per-mile accident probability is considered to be the same for all drivers and is not individually estimated. Consequently, behavioural aspects of driving should be embedded in insurance models to contribute towards current trends of personalized vehicle insurance.

Supporting the above mentioned, even if assumed that per-mile crash risk remains constant and annual mileage is reducing throughout the year, total individual crash risk reduction cannot be estimated since it depends on behavioural characteristics that are not currently recorded and therefore not considered in today's UBI. Driving information e.g. number of harsh braking and acceleration events occurred, percentage of excessive speeding, road type etc. should be included in driver's evaluation so as a per-mile risk factor could be estimated for each driver. In other words, risk factor is risk's increase rate which indicates how total individual risk is increased as mileage raises. Estimating this factor is the only way to precisely predict individual crash risk and consequently, charge a fair amount to each driver based on the risk he reflects. Since technological solutions exist nowadays and conditions for recording and managing real-time big data efficiently are met, there is a need for science to move towards that direction.

Based on the review conducted in this sub-chapter, it is concluded that UBI is expected to improve traffic safety as most UBI models (Paefgen et al., 2014) focus on determining the relationship between road safety parameters, such as crash risk, and travel and behavioural indicators such as mileage, risky hours, number of harsh braking/ acceleration events etc. This can be implemented by classifying each driving style on a

continuous scale from low to high risk (Sagberg et al., 2015) and subsequently estimate a safety scoring for each driver. It was also observed that a preferable practice to collect and transmit driving data for further analysis are IVDR such as OBD devices and smartphones. It is strongly believed that smartphones will be mostly used for data acquisition in the future since hardware cost of IVDR and smartphone penetration rate is high.

As for the indicators used in today's UBI models, the predominant among them are mileage, speeding, road network type and risky and rush hours driving. It is anticipated that apart from these, more behavioural parameters e.g. the number of sudden braking/ acceleration/ cornering events, mobile phone use etc. will be used a lot in future models because they represent crash probability better. Most barriers for the wide diffusion of UBI schemes have been overtaken, yet a few still exist namely the relatively medium capability level of cloud computing services for analysis and exploitation of big data, and the problems in data quality originating from the devices mentioned above. Future UBI is likely to be transformed in such a way as to adopt parameters from both PAYD and PHYD insurance schemes establishing a new Pay-As-How-You-Drive (PAHYD) model that embeds both travel behavioural and driving behavioural parameters.

Regarding future directions of driver's individual safety scoring, since it is possible to classify each driving style on a continuum scale from low to high risk, research should focus on the exploitation of actual accident record data. Data analysis techniques (artificial intelligence, big data analysis etc.) of data collected from naturalistic driving experiments are anticipated to play a significant role in the future. In naturalistic driving experiments where the user drives and reacts naturally, it is easier to "capture" those driving habits, styles and indicators exploiting normal driving data recorded and associate them with actual crash data. This is likely to be extremely convenient for researchers especially if datasets including both crash-involved and crash-free data exist that could be directly linked to driving behaviour indicators and/or styles. Review revealed that smartphone exploitation in the framework of naturalistic driving experiments is an innovative cost-effective approach for gathering and transmitting large amounts of travel and behavioural data. This methodology is now diffused and will be gradually used more in future UBI research.

In terms of the indicators usually incorporated in UBI models, there are several influencing most traffic crashes and related insurance claims that are not yet taken into consideration. Past review highlighted (Sagberg et al., 2015) that crash involvement is thus far predicted mainly using factors that indicate aggressive and/or impatient driving such as driving over the speed limits and a high frequency of driving-related violations. Since driver drowsiness and distraction (Kaplan et al., 2015) are two factors that mostly influence traffic crashes and related insurance claims, they should also be taken into account in UBI modelling along with other namely alcohol use, ecological driving and vehicle maintenance condition.

From a road safety perspective, estimating individual crash risk and charge based on that would reward good drivers for driving safely and therefore eliminate the cross-subsidies phenomenon. It would also serve as a strong incentive for more risky drivers to improve their driving behaviour, optimize their travel options and reduce their degree of exposure

by receiving feedback and monitoring their driving performance and preferences and paying less for insurance premiums. As a result, an insurance model incorporating individual driving characteristics would enhance traffic safety on a total and individual level by motivating drivers positively and negatively to alter their travel behaviour and improve their driving behaviour. It is suggested from the above that there are numerous and important challenges emerging on this research field, which will be further investigated in the near future.

2.5) Critical synthesis

Taking into account literature review conducted, it is deemed 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. 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. Within this dissertation, the quantification of the need for driving data collection in driving performance assessments is also investigated based on data collected through smartphone devices.

It can be said from the above that the most significant human factors that 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 study the potential of measuring driving safety efficiency using microscopic driving data collected from smartphone devices. It was 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. For the purposes of this study, drivers will be considered as DMUs, which is deemed to be rational since a driver is a unit that makes

decisions for a given mileage range about the number of events occurring and the time of mobile phone usage and speed limit violation. Driving attributes (metrics and distance recorded) will be considered the inputs and outputs of the DEA program. More details on the structure of the DEA formulation implemented are given below. It is also important to address the problem of the computation time required for a DEA algorithm to run 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 data.

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. It is evident (Litman, 2008) that 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 considered to be 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.

2.6) Research questions

Based on the results of the literature review, the research questions of this doctoral dissertation are formulated as shown below:

1) How well can driving safety efficiency be benchmarked? Can data science techniques and large-scale data provide sufficient answers?

2) What are the temporal evolution characteristics of driving efficiency? What do the drivers' groups formed represent?

3) What is the required amount of driving data that should be collected for each driver?

4) How can the least efficient trips of a database be identified?

Chapter 3: Methodological Approach

3.1) General methodological framework

The purpose of this chapter is to present the main structure of the methodological approach followed, including the processing procedure and the analysis conducted using large-scale data collected from smartphone devices. The description of every statistical, non-parametric, linear programming, non-supervised etc. methodology used in this thesis is given in details in this chapter. The overall methodological approach followed in this research is demonstrated below in order to achieve the objectives specified after the literature review conducted in this thesis.

Overview

Figure 3.1 illustrates the general methodological framework that is applied thereinafter. 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.

Figure 3. 1: General methodological framework of the present doctoral dissertation.

The above graphical illustration of the methodological approach is used to provide a broader and more comprehensive picture of the workflow that takes place and results to the better understanding on how driving efficiency can be analysed, using data science techniques for large-scale data. Further details on the methodological background and implementation of the techniques applied in this thesis are presented in each of the following sections.

3.2) Methodological steps

3.2.1) Smartphone data preparation

This step includes the data collection as well as the processing procedure that took place after data are obtained from the OSeven databases in order to be utilizable in the DEA models.

Data collection

The large-scale data that will be used within this research are mainly data collected from OSeven's smartphone application e.g. distance driven in km, driving duration in seconds, seconds of driving over the speed limit, seconds of mobile usage etc. In this step, the requirements of the data collection procedure are presented. To accomplish the goals set in this doctoral research, a sample of one hundred and seventy one (171) drivers from a 7 months period is acquired from the sophisticated database of OSeven, which contains large-scale driving data recorded from smartphones, constructing thus a database of 49,722 trips. For each individual part of the analysis conducted herein, a different part of this database is exploited because of the different requirements of each analysis.

A proportion of the participants also participated in a survey designed for the purpose of assessing driving behaviour. Answers were collected from 43 and 39 drivers in urban and rural road respectively regarding accident history, tickets received, number of road traffic code infringements, demographics (age, nationality, occupation etc.), driving experience and characteristics, driver's vehicle etc.

All data are properly prepared, as described below, in order to meet the requirements set and they can be imported in the DEA models developed.

3.2.2) Large-scale data investigation

Investigation of metrics-distance ratio evolution

In this step, it is examined whether or not the sum of metrics is proportionally increased to the sum of distances. One of the fundamentals of CRS DEA that cannot be overlooked is that the ratio of the inputs are increasing linearly to outputs. Therefore it is essential to examine how driving metrics are evolved in time compared to distance travelled not only in total but in each moving window examined as well.

Adequate driving data

In conjunction with the previous step, the amount of adequate driving data sample that should be collected for each driver is estimated in this step to ensure the significance of the results arising. As mentioned above, for the analysis of the total driving behaviour as

well as in moving window considered for the temporal analysis of driving efficiency, driving metrics should be linearly increased to distance travelled in order to apply CRS DEA. Therefore, the amount of data collected for each driver should be exceeding the minimum amount of data found that is required in each time step and in total as well.

3.2.3) Safety efficiency index estimation

After literature review revealed the existing knowledge gaps, research questions are set, and data are prepared and examined for adequacy, this leads to the estimation of the safety efficiency index. This index can be acquired using seconds of mobile usage, seconds of driving over the speed limits, and the number of harsh acceleration and braking events as inputs in the DEA model to measure safety efficiency. For the estimation of this index, distance travelled is used as output.

3.2.4) Trip efficiency analysis

As mentioned above, the main scope of the trip efficiency analysis is to a) implement an algorithm that outperforms the standard DEA algorithm in terms of computation time and b) provide a methodology for identifying less efficient trips. It is highlighted that a heuristic or meta-heuristic approach could also be investigated, but this is not within the scope of the present research, which aims to estimate the accurate solution of the DEA efficiency index.

In general, input and output selection is a critical procedure for DEA and great care should be taken to address this major issue. Nonetheless, because of the fact that the objectives of the computation time investigation, it is not deemed necessary to select inputs and outputs based on a specific safety concept. Three outputs and inputs are examined in each problem and more specifically the combinations of distance per road type with the number of harsh acceleration and braking events, seconds driving over the speed limit and seconds of mobile phone usage per road type respectively. These combinations create four different DEA problems but herein only harsh acceleration per road type with distance per road type is presented to avoid chattering. All models provided similar results and therefore conclusions drawn can be generalized regardless of the variables chosen in the model.

In all scenarios tested, results showed that all methods yield the same accurate solution as the standard DEA approach tested in terms of identifying the most efficient DMUs and peers, lamdas and theta values and calculating the efficient level of inputs and outputs for each DMU. This is a weighty outcome because for the first time tests proved the efficacy of the proposed methodology for performing a multiple input and output CH DEA. In the specific experiment, distance per road type travelled is used as output for DEA and convex hull algorithm is applied before applying standard DEA.

Data used in this study are metrics recorded in the form of absolute values i.e. the number of harsh acceleration and braking events, seconds driving over the speed limit and seconds using the mobile phone. All metrics are recorded per road type (urban, rural,

highway) e.g. ha$_{urban}$, ha$_{rural}$, ha$_{highway}$, hb$_{urban}$, hb$_{rural}$, hb$_{highway}$ etc. Convex hull's dimension is determined by the sum of the number of inputs and outputs of the DEA problem. Overall, the three approaches (Standard, RBE and CH DEA) are tested for seven different scenarios, i.e. for 100, 500, 1000, 2500, 5000, 7500 and 10088 trips.

The amount of computational memory required to perform the Convex Hull – DEA (CH DEA) approach is notably high. Quickhull algorithm applied herein does not support medium-sized inputs in 9-D and higher, which is the limitation of the present study. This is the reason why the authors choose to test their models only for three inputs and outputs in order to create a convex hull problem of 6-D which is less than the algorithm's capacity and can be calculated as described in the previous section.

An important note is that trips with zero sum of inputs were excluded from the analysis since this problem cannot be defined in that case. The DEA cannot be solved for a DMU with a sum of inputs equal to zero because this the same as having a business that produces some outputs without using any inputs, which is irrational. This is not deemed to cause any effect on the identification of the less efficient trips since practically only the most efficient trips are omitted from the analysis and trip efficient is relatively estimated and therefore, although the absolute value of efficiency is changing the k^{st} percentile of the least efficient trips will include exactly the same trips.

The methodological steps followed for the computation time investigation are presented below:

Assuming a set E of a number of N trips, four overall models are created for testing the processing time of standard DEA, RBE DEA and convex hull DEA. Total distance travelled in urban, rural and highway roads as outputs (three dimensions) in all models whereas a) total number of harsh acceleration events occurred, b) total number of harsh braking events occurred, c) total seconds of mobile phone use and d) total seconds of driving over the speed limits in urban, rural and highway roads are used as inputs (three dimensions) in model 1, 2, 3 and 4 respectively. One dimension will be created for each input and output i.e. six dimensions of the convex hull. The specifications of the models implemented are shown in Table 3.3.

For this model, convex hull algorithm is ran to estimate the set E_e consisting of the number N_e most efficient trips. Consequently, each trip m of the set $\{E - E_e\}$ of the non-efficient trips will be run with the set E_e creating $(N - N_e)$ DEA linear problems with $N_e + 1$ (trip examined in each iteration) variables each time to calculate the efficiency θ and the slacks λ of the peers of each trip m of the set $\{E - E_e\}$ of the non-efficient trips.

Required running time for each approach is estimated, compared and the optimal solution is determined. Results of this analysis are presented in the Results chapter.

Table 3.1: Inputs and outputs of the DEA models used in the trip efficiency analysis

Model Type	Overall model	
	Set of Inputs used	Set of Outputs used
1	1) ha_{urban} 2) ha_{rural} 3) $ha_{highway}$	1) $distance_{urban}$ 2) $distance_{rural}$ 3) $distance_{highway}$
2	4) hb_{urban} 5) hb_{rural} 6) $hb_{highway}$	1) $distance_{urban}$ 2) $distance_{rural}$ 3) $distance_{highway}$
3	1) $speeding_{urban}$ 2) $speeding_{rural}$ 3) $speeding_{highway}$	1) $distance_{urban}$ 2) $distance_{rural}$ 3) $distance_{highway}$
4	1) $mobile_{urban}$ 2) $mobile_{rural}$ 3) $mobile_{highway}$	1) $distance_{urban}$ 2) $distance_{rural}$ 3) $distance_{highway}$

The methodological steps for the least safety efficient trips identification are:

a) Total distance travelled is used as DEA output and the total number of harsh acceleration and braking events occurred, total seconds of speed limit exceedance, total seconds of mobile usage are used as DEA inputs.

b) A DEA model from table 3.3 is developed (in the road type examined).

c) The least efficient trips among the $n = 100$ trips are identified by the DEA model and the efficiency of each trip is acquired. In order to identify the least efficient trips, the k^{st} percentile of the least efficient trips of the database is taken based on the efficiency index assigned to each trip by DEA LPs. For instance, if $k = 5$ and therefore the target is to determine the list of the 5% least efficient trips of a database of 100 trips, trips are sorted by their efficiency index and the 5 least efficient are selected.

The trip efficiency analysis could not stand alone since there are trips with zero sum of metrics recorded (harsh acceleration/ braking events, seconds of mobile use, seconds of speeding) which cannot be included when applying DEA methodology. For the same reason, the most efficient trips cannot be included in the DEA model development since their efficiency cannot be defined. As a result, the methodology presented here for least efficient trip identification is not estimating the actual efficiency (trips with zero metrics cannot be included); nonetheless the least efficient trips can be relatively identified. Since the scope of this section is to provide the methodology for the least efficient trips identification, results of this analysis are not presented in the Results chapter.

As for the methodology for estimating the efficient level of inputs and outputs of a trip, it is provided in 3.2.1. This is the actual level of metrics (seconds driving over the speed limit etc.) that should have been reached in the specific trip in order to become efficient. The results of this analysis are presented in the Results chapter.

3.2.5) Driver efficiency analysis

Efficiency estimation

As mentioned above, the main scope of the driver efficiency analysis is to a) provide a methodology based on the DEA approach for driving safety efficiency determination and b) investigate the evolution of driving safety performance over time and cluster drivers based on the characteristics identified.

Models representing driving behaviour in all road types are constructed, with multiple inputs and outputs. Input and output selection is a critical procedure for DEA and should be linked to the conceptual specifications of each problem. Dyson et al. (2001) discussed several issues that should be taken into account before applying DEA to a dataset. One of the pitfalls is that the efficiency score might be miscalculated when input and output variables are in the form of percentiles and/or ratios simultaneously with raw data (Cooper et al., 2006).

The large-scale data used in this study are metrics recorded in the form of absolute values i.e. the number of harsh acceleration and braking events, seconds driving over the speed limit and seconds using the mobile phone. All metrics are recorded per road type (urban, rural, highway) e.g. ha_{urban}, ha_{rural}, $ha_{highway}$, hb_{urban}, hb_{rural}, $hb_{highway}$ etc. In this specific experiment, distance per road type travelled is used as output for DEA and the best performing algorithm identified in the trip efficiency analysis (convex hull DEA) is applied. Two different models are developed. The variables' combinations for structuring these models in each road type was based on literature review. These two models include all traffic safety parameters found in literature review and account for the overall safety profile of the driver. The specifications of the models implemented are given below and are illustrated in Table 3.3. The reason why analysis is not conducted in highways is provided below.

As for the driver efficiency analysis, in each model the cumulative metrics monitored during the total period recorded are used as inputs and outputs in the DEA models developed. In other words, the final database used in each model includes the cumulative value (for the period that each driver was recorded) of each variable considered in the DEA models, constructing thus a database with one row per driver.

Table 3.2: Inputs and Outputs of the DEA models used in the driver efficiency analysis

DEA models	Urban	Rural
Set of Inputs used	1) ha_{urban} 2) hb_{urban} 3) $speeding_{urban}$ 4) $mobile_{urban}$	1) ha_{rural} 2) hb_{rural} 3) $speeding_{rural}$ 4) $mobile_{rural}$
Set of Outputs used	1) $distance_{urban}$	1) $distance_{rural}$

where the index x determines the road type of each model.

Evolution of driving efficiency

The temporal evolution of average driving efficiency is investigated using different databases of metrics accumulated over different timeframes. Time series are decomposed to acquire trend, volatility and estimate stationarity.

A sample of 230 and 150 trips is taken into account respectively for urban and rural roads. This results to 100 users of data_sample_1 in both road types and to 43 and 39 users, for which questionnaire data are available, in urban and rural areas respectively. Trips chosen for the analysis are the last trips of each driver that were recorded. This is to ensure that there has been some time since the driver is being recorded so that the fact that is being monitored is not influencing his/ her driving behaviour any more.

Unfortunately, the same analysis cannot be conducted on highways since it is found in previous steps of the analysis that there are very few end-users for whom a time series can be created with several observations. The analyses showed that there should be a moving window of at least 75, 81 and 116 trips in urban, rural and highways respectively in which driving performance is calculated to create the required time series. Table 3.4 summarizes the sample used in this specific analysis for each road type. As it appears, it is not feasible to perform the analysis for highways since there are only 18 of the data_sample_1 drivers and 7 of the data_sample_2 that have the adequate total distance and number of trips. Even if the analysis was performed with these drivers, the length of the time series would not be enough to ensure the significance of the results. The two last columns of table 3.4 represent a) the number of participants that have at least as many trips required in "No of trips" column and b) the number of participants that have at least as many trips required in "No of trips" column and have also responded the questionnaire. More details on the sample choice are given in the data collection chapter.

Finally, it is highlighted that the analyses are performed separately for the samples with and without available questionnaire data to compare the clusters arising and their characteristics. This procedure will evaluate the potential of driving safety efficiency benchmarking without having any knowledge on the personal information (age, gender, accident record etc.) of the user that is being assessed.

Table 3.3: Number of drivers participated in the analysis of the temporal evolution of driving efficiency in each road type

Road type	No of trips	Required moving window (trips)	No of participants of the data_sample_1	No of participants of the data_sample_2
Urban	230	75	100	43
Rural	150	81	100	39
Highway	150	116	18	7

The methodological steps to achieve this are:

For each driver:

1) The last 230 and 150 trips are kept in urban and rural roads respectively

2) The efficiency models are applied

3) The efficiency of each driver is estimated in a moving window of 75 trips in urban roads and 81 in rural

4) The time series of the driver's efficiency evolution is created

5) The volatility of driving efficiency is estimated

6) The Augmented Dickey-Fuller (ADF) and KPSS test for unit root and stationarity respectively, is performed

7) When the null hypothesis of both the ADF and KPSS test are rejected for a time series, then it is considered fractionally integrated and present long memory. An ARFIMA model is applied in this case to estimate the order d of the time series. This variable takes values in the $[-1,1]$ region for fractionally integrated time series, close to 0 for stationary time-series and close to 1 for unit-root time series

8) Time series trend is acquired by estimating the coefficient b of the linear regression model that best fits the time series data (slope)

The efficiency models that will be applied in urban and rural roads are the following:

1) One per road type, developing the following model combinations (models of table 3.3):

    a) Urban road: Output: total distance travelled in urban roads, Inputs: total number of harsh acceleration events occurred in urban roads, total number of harsh braking events occurred in urban roads, total seconds of speed limit exceedance in urban roads, total seconds of mobile usage in urban roads.

    b) Rural road: Output: total distance travelled in rural roads, Inputs: total number of harsh acceleration events occurred in rural roads, total number of harsh braking events occurred in rural roads, total seconds of speed limit exceedance in rural roads, total seconds of mobile usage in rural roads.

As mentioned above, for each model the cumulative metrics monitored during the period examined are used as inputs and outputs in the DEA models developed. As a result, a different database is created in each step of the moving window and a new DEA model is developed respectively to estimate driving efficiency in the specific step and consequently the temporal evolution of total driving efficiency. Thereafter, the cumulative metrics of the total recording period of the **230** and **150 trips** are taken into consideration to estimate a driver's total driving efficiency.

Drivers clustering

Based on the variables created from the analysis conducted above:

1) Total efficiency of the driver during the complete recording period

2) Volatility of the sequence

3) Time series trend

4) Stationarity of the time series

5) Questionnaire data (used only in the data_sample_2)

A K-means algorithm is employed for clustering drivers. The optimal number of clusters is determined using the elbow method. Driving characteristics of each cluster arose are analysed and conclusions drawn are presented.

3.3) Theoretical background

3.3.1) Data envelopment analysis

DEA is a non-parametric approach that does not require any assumptions about the functional form of a production function and a priori information on importance of inputs and outputs. DEA allows each DMU to choose the weights of inputs and outputs which maximize its efficiency. The DMUs that achieve efficiency equal to unit are considered efficient while the other DMUs with efficiency scores between zero and unit are considered as inefficient. The first DEA model proposed by (Charnes et al., 1978) is the CCR model that assumes that production exhibits constant returns to scale i.e. outputs are increased proportionally to inputs. DEA models can also be distinguished based on the objective of a model; that can be either outputs maximization (output-oriented model) or inputs minimization (input-oriented model).

Let us use X and Y to represent the set of inputs and outputs, respectively. Let the subscripts i and j to represent particular inputs and outputs respectively. Thus x_i represents the ith input, and y_j represent the jth output of a DMU. The input-oriented CCR model evaluates the efficiency of DMU_o by solving the following (envelopment form) linear program (Ramanathan, 2003) and its mathematical formulation is formulated as:

$$\min \theta_B$$

Subject to the following constraints:

$$\theta_B * x_o - X * \lambda \geq 0 \qquad\qquad (1)$$

$$Y * \lambda \geq y_o$$

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

where λ_i is the weight coefficient for each DMU_i that is an element of set λ, X is the set of inputs, Y is the set of outputs and θ_B is a scalar representing the efficiency of reference DMU_0. The objective function of this linear programming problem (DEA) is min θ_i i.e. minimize the efficiency of DMU_i. In order to benchmark the efficiency of all DMUs (of each DMU) of the database, this linear programming problem should be solved for each DMU_i. This is radically increasing the processing time of the problem as the number of DMUs and especially the dimensions (every extra input or output added to the problem increases the dimension by one unit) of the problem are increased. This is the reason why this research makes use of other techniques to reduce computation time.

Although a trip and a driver cannot literally behave as a decision-making unit, it can be evaluated as a DMU and therefore, it will be considered as such for the purpose of this research. As mentioned above, this is deemed a correct assumption on a trip/ driver basis since a) all variables used are continuous quantitative variables as those used in previous DEA studies and b) a driver should reduce his mileage and the frequency of some of his driving characteristics. The mathematical formulation of DEA for the driving problem examined here is presented in the next section. It is noted that from now on, DMUs will be referred as either trips or drivers depending on the problem examined each time. For brevity purposes, DMUs are referred only as drivers in the next section but the constraints for solving the trip efficiency problem has exactly the same formulation and constraints.

Table 3.4: Description of the per trip variables recorded

Variable name	Variable short description
ha_x	number of harsh acceleration events in X road type
ha_{urban}	number of harsh acceleration events in urban road
ha_{rural}	number of harsh acceleration events in rural road
$ha_{highway}$	number of harsh acceleration events in highway
hb_x	number of harsh braking events in X road type
hb_{urban}	number of harsh braking events in urban road
hb_{rural}	number of harsh braking events in rural road
$hb_{highway}$	number of harsh braking events in highway
$speeding_x$	total seconds of speed limit violation in X road type
$speeding_{urban}$	total seconds of speed limit violation in urban road
$speeding_{rural}$	total seconds of speed limit violation in rural road
$speeding_{highway}$	total seconds of speed limit violation in highway
$mobile_x$	total seconds of mobile phone usage in X road type
$mobile_{urban}$	total seconds of mobile phone usages in urban road
$mobile_{rural}$	total seconds of mobile phone usage in rural road
$mobile_{highway}$	total seconds of mobile phone usage in highway
$distance_x$	total distance driven in X road type
$distance_{urban}$	total distance driven in urban road
$distance_{rural}$	total distance driven in rural road
$distance_{highway}$	total distance driven in highway

Driving metrics recorded are illustrated in table 3.1. On a driver and trip basis, seconds of mobile use, number of harsh acceleration and braking events occurred and seconds of speed limit exceedance are used as DEA inputs while total distance travelled is used as DEA output for driving efficiency benchmarking. As literature review revealed a) these are the most important driving metrics that affect driving safety efficiency among those recorded from smartphone devices and b) a methodology that is capable of incorporating behavioural risk per unit of exposure should be developed.

It is assumed that this study should adopt an input-oriented (IO) DEA model, since the objective is to minimize the number of harsh acceleration, harsh braking events etc. (inputs) that occur per driving distance unit rather than to maximize driving distance (outputs) for given metrics (output-oriented (OO) DEA model). In terms of road safety, the latter would increase the exposure of a driver (kilometrage) and therefore crash risk (Tselentis et al., 2017). Nonetheless, this is considered a minor issue since it is related to the general notion of the research problem and it only affects the formulation of the problem and not the research outcomes. It is also proved in this research that the sum of all metrics (inputs) recorded (e.g. the number of harsh acceleration and braking events occurred in each $trip_i$) converge to a constant and changes proportionally to the sum of driving distance (output) and therefore the driving efficiency problem is considered a constant-returns-to-scale (CRS) problem and is solved as such.

Mathematical formulation of DEA for the driving problem

Let $\mathbf{x}_i$ and $\mathbf{y}_j$ represent the set of inputs and outputs of a DMU. The input-oriented CCR model evaluates the efficiency of DMU_0 by solving the (envelopment form) linear program (Ramanathan, 2003) presented below. Considering each driver as a DMU and taking into account the principles of DEA (Charnes et al., 1978), the mathematical formulation for the specific driving efficiency problem examined herein is:

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

Subject to the following constraints:

$\text{Driving\_Efficiency}_B * x_o - X * \lambda \geq 0$ (2)

$Y * \lambda \geq y_o$

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

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 etc.), Y is the set of Outputs (distance travelled) and $\text{Driving\_Efficiency}_B$ is a scalar representing the efficiency of reference DMU i.e. $driver_0$. The objective function of DEA is min θ_i i.e. to minimize the efficiency of $driver_i$. To benchmark the efficiency of each and all drivers of the database, this linear programming problem should be solved for each $driver_i$.

Efficient level of inputs and outputs for non-efficient drivers/ trips

After DEA LPs of (1) are solved and the efficiency index $\text{Driving\_Efficiency}_B$ and coefficients λ_i are estimated for each driver the efficient level of inputs and outputs at which each driver could optimally reach can be calculated. The efficient level of inputs for driver i can be calculated as the product sum of the lamdas and the input values of each of the identified peers whereas to find the efficient level of outputs for the same driver, each output value should be divided by theta value. Considering $driver_i$ as the reference DMU and a set of m drivers, where $m \in \mathbb{N}$ is the number of $driver_i$ peers, the efficient level of $Metric_i$ can be estimated using following formula (3):

$$Metric_i = \sum_{j=1}^{m} \lambda_j * Metric_j$$ (3)

More specifically, considering $driver_i$ as the reference DMU and a set of m drivers, where $m \in \mathbb{N}$ is the number of $driver_i's$ peers, the efficient level of e.g. ha_{urban} can be estimated using following formula (4):

$$ha_{urban_i} = \sum\nolimits_{j=1}^{m} \lambda_j * ha_{urban_j} \tag{4}$$

On the other hand, the efficient level of e.g. $dis\tan ce_{urban}$ is calculated from formula (5):

$$distance_{urban} = distance_i / Driving\_Efficiency_i \tag{5}$$

It should be noted that a DMU achieves its efficient level by reaching the efficient level of either its inputs or outputs. Additionally, a DMU is deemed to have achieved the efficient level when it reaches unit efficiency. Based on the above, it can be concluded that the required change of each driving attribute that was taken into consideration in order for a driver to shift either to the efficient frontier or to another driving class can be estimated. This can be achieved by solving the optimization problem for a specific input or output given the target efficiency ($Driving\_Efficiency_B$), which is the upper or the lower limit of the class that the driver is shifting in case of efficiency decrease or increase respectively.

Reduced basis entry (RBE) algorithm

The Reduced Basis Entry (RBE) algorithm for DEA is iteratively solving the DEA LP for all DMUs in the database. The main difference from the standard DEA approach is that if the reference DMU_0 examined is found to be inefficient in an iteration, it is excluded from all the next solutions. Therefore, each time a non-efficient DMU is recognized, variables are reduced by one and as a result, the running time of the next LP will be lower. Thus, total computations are less expensive in terms of time. The pseudocode of RBE algorithm is given below:

```
for every DMUx in DMUset:
    θx,λx=DEA(DMUx,inputset,outputset)
    if θx<1:                                                    (6)
        DMUset.remove(DMUx)
        delete[inputx]
        delete[outputx]
```

where DEA is the function written for solving the DEA LPs given reference DMU name, input matrix and output matrix, θ_x is the estimated efficiency for DMU_x, λ_x is the weight coefficient of DMU_x and input and output are the matrices containing inputs and outputs respectively. This algorithm results in constructing two sets comprising of the thetas and lambdas of all DMUs in the dataset.

3.3.2) Convex hull

The convex hull of a set of points is the smallest convex set that contains all points in the set. Reducing the required computation time for finding the optimal solution of convex hull is a fundamental problem for mathematics and computational geometry. In quickhull algorithm (Barber et al., 1996), that is used herein, it is assumed that points are in a general position, so that their convex hull is a simplicial complex (Preparata & Shamos, 1985). Its vertices and facets represent a d-dimensional convex hull. A point is deemed to be extreme, and, therefore, lies on the hull, if it is a vertex of the convex hull. Each facet comprises of a set of vertices, a set of neighboring facets, and a hyperplane equation. The ridges of the convex hull are the (d - 2) - dimensional faces. The point where the vertices of two neighboring facets intersect, constructs a ridge. Quickhull makes use of two geometric operations (Barber et al., 1996), oriented hyperplane through d points and signed distance to hyperplane. It represents a hyperplane by its outward-pointing unit normal and its offset from the origin. The inner product of the point and normal plus the offset represents the signed distance of a point to a hyperplane. A halfspace of points that have negative distances from the hyperplane is defined by the hyperplane. A point is above the hyperplane if this distance is positive.

Assuming a set E containing N DMUs convex hull algorithm will initially estimate the set E_e consisting of the number N_e of the most efficient DMUs. Consequently, each DMU m of the set $\{E - E_e\}$ of the non-efficient DMUs will be run with the set E_e creating $(N - N_e)$ DEA linear problems with $N_e + 1$ (DMU that is evaluated) variables. This will allow for the calculation of efficiency θ and slacks λ of the peers for each DMU m of the set $\{E - E_e\}$ of the non-efficient DMUs.

It should be mentioned that the DEA - convex hull algorithm consists of three different steps namely convex hull solution, determination of the efficient DMUs, DEA solution for non-efficient DMUs. At the first step, convex hull points are identified creating thus a superset of N_c points that includes all efficient DMUs. Nonetheless, because some of the convex hull points are not efficient DMUs since they do not lie on the efficiency frontier, N_c DEA LPs are solved to find the N_e efficient DMUs. During the third step, $(N - N_e)$ DEA LPs one for each of the inefficient DMUs to estimate parameters θ_i and λ_i.

3.3.3) Driver's behaviour volatility measure

Since it is crucial to observe how each driver alters everyday behaviour and whether or not there is a stability in his/her driving behaviour, the natural logarithm of the ratio of the performance of two consecutive time steps is estimated (Mantouka et al., 2018). 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, $E_t = (1,…,n)$, 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:

$$r_{t,i} = \ln(\frac{E_{t,i}}{E_{t-1,i}})$$
(7)

This indicator has a positive value, when the driver improves his overall driving efficiency in the next time step and negative values indicate a deterioration of overall driving efficiency. 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:

Driver's behaviour volatility = $\sqrt{\dfrac{\sum\limits_{t=1}^{n}(r_{t,i} - \overline{r_i})^2}{n-1}}$
(8)

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

3.3.4) Driving efficiency time series

Stationarity

Stationarity is one of the most important concepts in time series analysis. This is because time series models may apply only to stationary data. Therefore, before proceeding with time series modelling, data must not have a trend. 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 (Tsay 2002).

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 (Shumway and Stoffer, 2000). For $n = 2$, the joint probability distributions of (Y_{t_1}, Y_{t_2}) and (Y_{t_1+L}, Y_{t_2+L}) are the same and their covariances are equal (Washington et al., 2010)

$$COV\left(Y_{t_1}, Y_{t_2}\right) = COV\left(Y_{t_1+L}, \mathrm{Y}_{t_2+L}\right) \tag{9}$$

The previous condition depends only upon the lag L. The covariance between Y_t and Y_{t+L} is called autocovariance (γ_κ) and is a function that gives the covariance of the process with itself at pairs of time points. It is given by

$$\gamma_\kappa = COV\left(Y_t, Y_{t+L}\right) = E\left[\left(Y_t - \mu\right) \cdot \left(Y_{t+L} - m\right)\right] \tag{10}$$

and $\gamma_0 = VAR\left(Y_t\right) = \sigma^2$

When a time series is not stationary, stationarity is usually obtained through first-order differencing in transportation research (Washington et al., 2010) i.e. Y_t is $Z_t = Y_t - Y_{t-1}$. 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 (Granger & Engle, 1987) 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 one differencing to become stationary whereas the alternative is that the time series is already stationary.

There are some cases though where a process does not have a unit root but is near to it. In most cases time series are usually differenced d times to achieve stationarity, where d is the integration order d and the number of existing unit roots. Nonetheless, when imposing erroneous differentiation parameters (Granger & Joyeux, 1980) d is practically forced to be equal to 1 when it is not. This leads to overdifferentiation of the time series and forces an artificial correlation structure on the prediction models. Additionally, $I(0)$ and $I(1)$ model structures cannot account for persistence in a time series commonly referred to as "long memory". A series is said to have "long memory" when significant autocorrelation is observed in wide timeframes. This is easily noticed by examining the autocorrelation function plot of a time series.

Short and long-memory time series are necessary to be fractionally integrated, with the order d taking on any fractional or integer value in the $[-1,1]$ region. Data are generally modelled better using fractional integration because strict $I(0)$ or $I(1)$ processes are avoided and both long-term persistence and short-term correlation are explicitly modelled (Hosking, 1981). The order d denotes the status of a time series in terms of stationarity (Hosking 1981; Odaki 1993) i.e. when $d = 1$ it is a unit-root process, when $d = 0$ it is stationary, when $0 \leq d \leq 1/2$, it is fractionally integrated (Karlaftis and Vlahogianni 2009) and exhibits long memory and when $1/2 \leq d \leq 1$ stationarity in the series cannot be verified.

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 (Zarnowitz & Ozyildirim, 2006, Bianchi et al., 1999, Cameron, 2005, Pranab, 1968, Hodrick & Prescott, 1997, Rotemberg, 1999):

1) Least squares linear regression

Interlacing of the time series points by a straight line with the lowest sum of squared distances (in direction of y-axis) from all the points. This is the most usual choice called a least-squares fit, which minimizes the sum of the squared errors in the data series y.

$$y = ax + b = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} x + \frac{\sum_{i=1}^{n} x_i^2 y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \tag{11}$$

where x_i denotes the time points of measurements as in the previous equations. Given a set of points in time t, and data values Y_t observed for those points in time, values of a and b are chosen so that

$$\sum_t \left[y_t - (at + \hat{b}) \right]^2 \tag{12}$$

is minimized. Here $at + b$ is the trend line, so the sum of squared deviations from the trend line is what is being minimized. This can always be done in closed form since this is a case of simple linear regression. It can be therefore inferred that trend is the slope of the least squares line.

2) Theil-Sen regression

Nonparametric variant of the previous trend statistic computed as a median slope from all of the slopes of pairs of points in the time series (such that, the first point is predecessor of the second one in the pair).

3) Delta

Basic difference of the final and initial time point of the time series:

$$Delta = y_n - y_1 \tag{13}$$

4) The Hodrick-Prescott Trend

Let a time series Y_t be viewed as the sum of a growth (trend) component g_t and a cyclical component $c_t : Y_t = g_t + c_t$ for $t = 1, ..., T$. The growth component should be smooth, so that the procedure recommended by Hodrick and Prescott (1997) is to minimize

$$\sum_t^T c_t^2 + \lambda \sum_t^T \left[(g_t - g_{t-1}) - (g_{t-1} - g_{t-2}) \right]^2 \tag{14}$$

where the parameter λ is positive. The larger λ, the smoother is the result; if λ, which penalizes variability in g_t, is large enough, approaches $g_0 + \beta t$. Hodrick and Prescott favour $\lambda = 1,600$ for quarterly data, but show that the numbers change little if λ is reduced or increased by a factor of four.

5) The Rotemberg Trend

Rotemberg (1999) proposes a heuristic method of time series decomposition which estimates the value of λ given two parameters, k and v. Using the earlier notation, let $Y_t = g_t + c_t$, where g_t and c_t are trend and cycle components of the time series Y_t,

respectively. The parameter k ensures that the estimated trend minimizes the covariance of two values of the cyclical component, c_t and c_{t+k}. The parameter v ensures that the trend and cycle components, g_t and c_t, are orthogonal over the horizon of v periods. Specifically, Rotemberg estimates the trend by minimizing

$$\sum_{t=1+k}^{T} c_t * c_{t-k} + \lambda \sum_{t=2}^{T-1} \left[\left(g_t - g_{t-1} \right) - \left(g_{t-1} - g_{t-2} \right) \right]^2 \tag{15}$$

λ is chosen as the lowest parameter value such that the following constraint holds

$$\sum_{t=k+v}^{T-k-v} c_t * \lambda \left[\left(g_{t+v} - g_t \right) - \left(g_t - g_{t-v} \right) \right] = 0 \tag{16}$$

Rotemberg recommends that k be set to equal 16 quarters on the admittedly somewhat arbitrary ground that historically NBER business cycle troughs for the U.S. have been four years apart on average. (The dispersion around this average is very large.) With large k, the minimization of (5) results in a trend that is quite smooth and not very sensitive to either the cyclical movements of the series nor the choice of v. With low $k - a$ fortiori, with zero k, which is the case in the H-P trend – the effects are opposite. Rotemberg chooses v to equal five quarters.

Literature review revealed that given a set of time series data and the desire to estimate its trend, there are several functions that might be chosen for the fit. 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 (Mills, 2003). 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. 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 thesis.

3.3.5) K - means clustering

K-means clustering is an unsupervised machine learning technique used when unlabelled data (data that are not categorized or grouped) exist. This algorithm 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:

1) The centroids of the K clusters, which can be used to label new data

2) Labels for the training data (each data point is assigned to a single cluster)

Clustering allows for 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.

Algorithm

An iterative refinement is used by the K-means machine learning 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:

1. Data assignment:

Each data point is assigned to the nearest existing centroid in this step, based on the squared Euclidean distance. In other words, if c_i is the collection of centroids in set C, each data point x is assigned to a cluster based on

$$\underset{c_i \in C}{\arg\min} \, dist(c_i, x)^2 \tag{17}$$

where dist($\cdot$) is the standard (L2) Euclidean distance. Let the set of data point assignments for each ith cluster centroid be S_i.

2. Centroid update:

In this step, the centroids are re-estimated by taking the mean of all data points assigned to that centroid's cluster.

$$c_i = \frac{1}{|S_i|} \sum_{x_i \in S_i} x_i$$ (18)

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.

Number of clusters

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. Figure 3.1 shows an elbow method example.
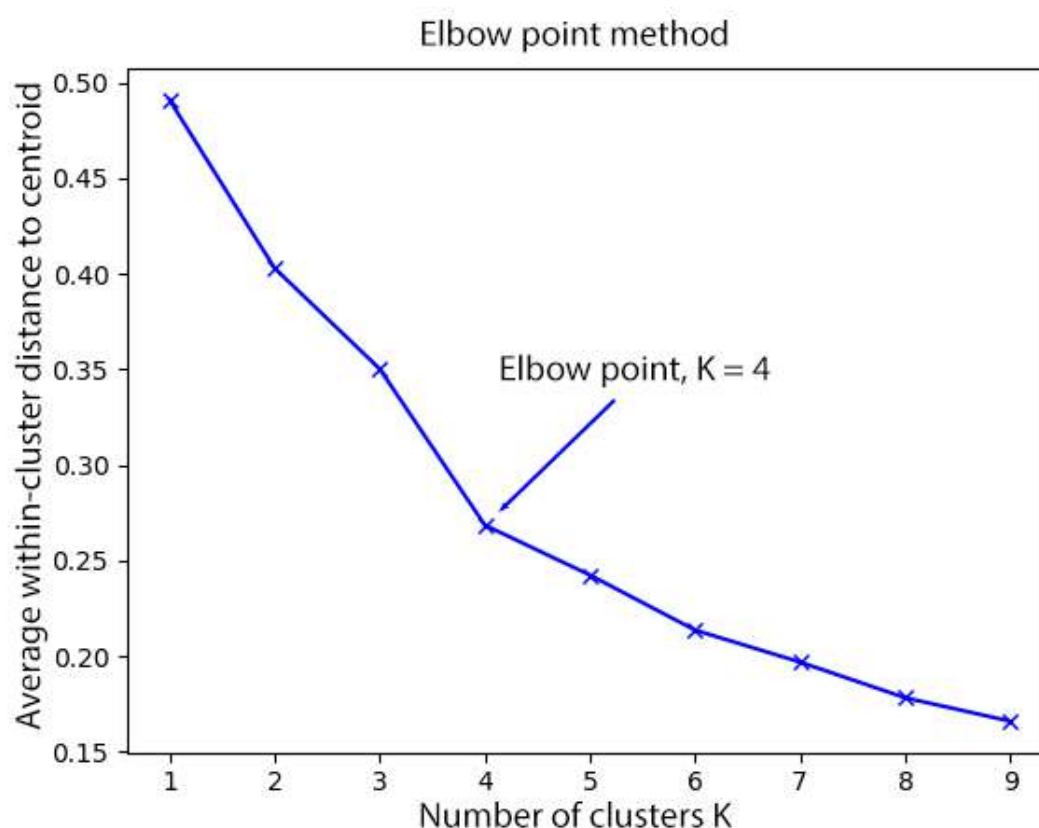
Figure 3.2: Elbow method example

A number of other techniques exist for estimating the optimum number of clusters K, such as information criteria, the information theoretic jump method, cross-validation, the silhouette method, and the G-means algorithm. Additionally, insights into how K-means algorithm is clustering data for each K are provided when monitoring the distribution of data points across groups.

3.3.6) Cumulative event rate convergence

The statistical analysis using the data collected from the smartphone will be conducted to determine the driving distance at which the rate of the driving indicators converges to a stable index and therefore DEA can be applied in each time step and no more data are required to be collected in total. In this research the magnitude of change measurement in a time series is employed which is decreasing over distance (km) as the specific magnitude converges on its average rate. At the same time, this means that the rate at which an event (number of harsh acceleration/ braking events, seconds of mobile phone usage, seconds driving over the speed limits per 100km) occurs also converges to its average rate e.g. the average rate of harsh acceleration events for the specific driver (average number of harsh acceleration events). For each driver and after each trip that took place, the above metrics were calculated by diving the total number of occurred events by the total distance driven thus far, constructing thus a time series of average

events per km. The mathematical formulation for calculating the convergence index of the event rate is given in the following section.

Convergence index

Assuming that we have calculated the time series of the total events per km travelled (CR_i) for n trips, the following formula is used for calculating the convergence rate of events:

$$CI_i = \left|(CR_i - CR_{i-1}) / CR_{i-1}\right| \quad \forall i \in N^* in[2, n]$$ (19)

where,

$$CR_i = \sum_1^i E / \sum_1^i km$$

Chapter 4: Data Collection

4.1) Recording procedure

A mobile App developed by OSeven Telematics is employed for the purposes of this study to record driving behaviour of the participating users, exploiting the hardware sensors of the smartphone device and a variety of APIs to read sensor data and transmit it to a central database.

OSeven has developed an integrated system for the recording, collection, storage, evaluation and visualization 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. The system developed integrates a data collection, transmission, processing and visualization procedure using smartphones, the main features of which are outlined in the next paragraphs.

This subsection comprises an overview of the OSeven Telematics data flow system. Therefore, it should be highlighted that none of the procedures described herein was implemented as part of this research, but the existing system of OSeven was exploited to acquire the necessary information for the analysis conducted afterwards. It is also noted that all ML algorithms developed for raw driving data analysis e.g. harsh event detection is company-owned and therefore all relevant information is kept confidential.

4.1.1) Data recording system

The data recording is initiated automatically in the mobile apps when a driving status is recognized and again it stops automatically when a non-driving status is recognized. Trip recording also continues after the vehicle is idled for five minutes, to consider the case that the driver continues his trip with a few minutes stop. All extra information collected after the "end of driving" are discarded using the machine learning techniques described below. The recorded data come from various smartphone sensors and data fusion algorithms provided by Android (Google) and iOS (Apple). A mobile application is developed to record user's behaviour by exploiting the hardware sensors of the smartphone device and a variety of APIs, which read sensor data and temporarily store it to Smartphone's database before transmitted to the central database. After the transmission, everything is discarded from the mobile phone.

Indicatively, technology sensors that are integrated in mobile phone are:

1) Accelerometer[1]

---

[1]these sensors are recording attributes in x, y, z axes

2) Gyroscope[1]

3) Magnetometer

4) GPS (speed, course, longitude, latitude)

Fusion Data provided by iOS and Android:

1) Yaw, pitch, roll

2) Linear acceleration[1]

3) Gravity[1]

The frequency of the data recording varies depending on the type of the sensor with a minimum value of 1Hz. It is noticeable that a massive dataset can be achieved and a very significant contribution can be made by the use of mobile phone applications to the collection of driving characteristics. The basic operating frame of the data flow is shown in Figure 1.
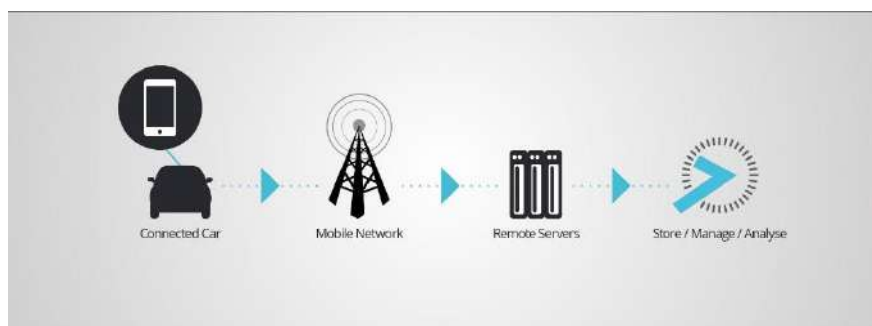
Figure 4.1: Oseven data flow system

4.1.2) Data transmission

After the end of the trip, the application is transmitting all data recorded to the central database of the OSeven backend office via an appropriate communication channel such as a Wi-Fi network or cellular network (upon user's selection) such as a 3G/4G network (online options) based on the user settings.

To achieve the interoperability between those sides, an API is built which is used to deliver data from an online service to another client application. This architecture is used to transfer and receive data between systems, supporting their interoperability between them. Making data accessible over the World Wide Web with an API, empowers third party systems data to be submitted to the database and makes the information easily available.

The total volume of data for an average driver is estimated around 50Mb/ month.

4.1.3) Data storage, security and privacy issues

Large-scale 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 (e.g. GDPR). The API used supports user authentication and encryption to prevent unauthorized data access.

4.1.4) Data processing

After data is stored in the cloud server for central processing and data reduction, it is converted into meaningful behaviour and safety related parameters (i.e. big data handling and processing). This is achieved by using the two big data processing methods which include two families of techniques, big data mining techniques and machine learning (ML) algorithms.
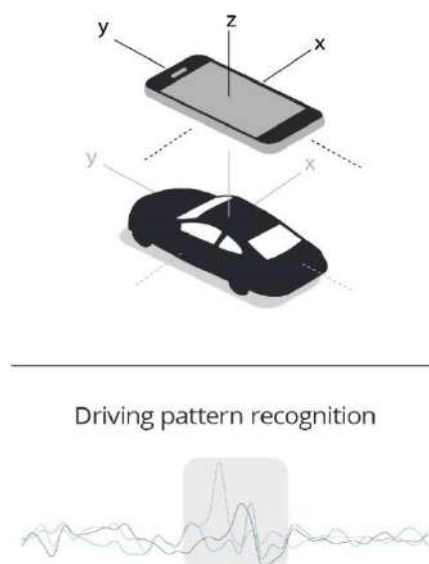
Figure 4.2: Yaw, Pitch, Roll

Machine learning methods (filtering, clustering and classification methods) are used to clean the data from noise and errors, and to identify repeating patterns within the data. Subsequently, these data patterns will be processed by means of big data mining techniques, in order to calculate the necessary parameters and derive behaviour indicators to be used in the analysis. In other words, the highly spatially and time disaggregated data from the Smartphone are processed in order to derive useful road safety indicators. Artificial intelligence methods allow for the detection of aggressive behaviour of the driver in the form of harsh events, the observed distraction of the driver due to the use of mobile phone, the identification of travel modes, the speed limit exceedance as well as where the determination of the time and spatial characteristics of all the above. The procedure of the ML algorithms and big data mining techniques include data filtering and outlier detection, data smoothening, driver clustering and classification,

detection of speeding regions, harsh acceleration/braking/cornering events, mobile usage, risky hours driving, travelling mode and driver or passenger recognition.

The procedure of the machine learning algorithms and big data mining techniques includes the simple steps mentioned below:

1) Data filtering and outlier detection

2) Data smoothening (when needed)

3) Driver clustering and classification

4) Recognition of the speeding regions (duration of speeding, exceedance of speed limit etc.)

5) Detection of harsh acceleration/ braking/ cornering events

6) Detection of mobile usage (talking, texting, surfing)

7) Identification of driving during risky hours (distance in risky hours periods)

8) Determination of the travelling Mode (car, mass transit, bicycle, motorcycle)

9) Driver or Passenger recognition

After the ML process is completed, a variety of different indicators is calculated that is useful to the user and for the evaluation of driving behaviour. These indicators are divided into two distinct categories, risk exposure and driving behaviour indicators. The main risk exposure indicators arising are:

1) Total distance (mileage)

2) Driving duration

3) Type(s) of the road network used (given by GPS position and integration with map providers e.g. Google, OSM)

4) Time of the day driving (rush hours, risky hours)

combined with other data sources (speed limits etc.)

The main driving behaviour indicators arising are:

1) Speeding time (duration of driving over the speed limits, speed limit exceedance etc.)

2) Number and severity of harsh events:

   i) Harsh braking (longitudinal acceleration)

   ii) Harsh acceleration (longitudinal acceleration)

   iii) Harsh cornering (angular speed, lateral acceleration, course)

3) Driving aggressiveness (e.g. average positive and negative acceleration)

4) Time of mobile phone usage
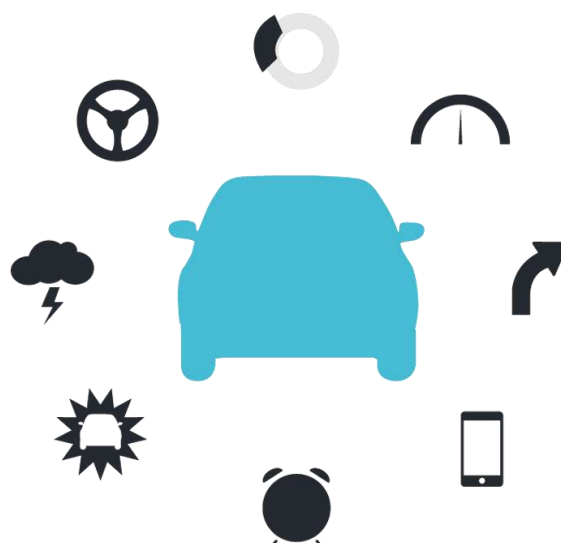
Figure 4.3: Driving risk indicators

These indicators along with other data (e.g. data from maps) are subsequently exploited to implement individual traveller's statistics, on all road networks (urban, highway, etc.) and under various driving conditions, enabling the creation of a large database of driving characteristics.

The final step of the data processing procedure is the development of the driving behaviour model. Aggregated data are analysed and the evaluation system is calibrated based on the whole sample. The driving behaviour model includes several indicators for driving benchmarking and finally aggregates the whole procedure on a trip basis for each driver in the sample in order to produce the final rating per driver. Each trip and therefore each driver is benchmarked based on the characteristics mentioned above. Final evaluation produced comprises both a total and a per indicator rating. Processed data is transferred to the smartphone apps and/or web platform offering user-friendly environments for the users to get their analytics and reports. The data visualization procedure is described in the next section.

4.1.5) Data visualization

The results of all the aforesaid procedure are accessible in the Smartphone app and the web portal, where it is available for the user to see all detected events and their place on the map as well as all scores (overall and per category). Thus, the driver is provided with a user-friendly way to realize the trip sections with risky driving behaviour and avoid similar behaviours in the future. At the same time the insurance companies have access to the data of their clients using the OSeven web portal. The driving scores are used for the determination of the insurance premium and/or the loyalty programs.
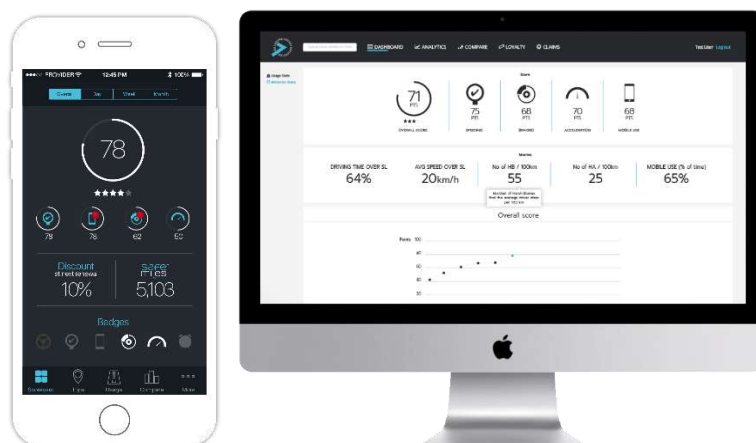
Figure 4.4: Mobile App and Web portal

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. The procedure how inputs and outputs are selected will be described in the next section. Data filtering and DEA improvement algorithms are performed in Python programming language and several scripts are written for this reason. Python packages used include Pandas and Numpy for numeric calculations and transformations, scipy that features quickhull algorithm and pulp for linear programming problem construction. More details on the algorithm implementation are given below. Coding is applied using Pycharm IDE Community edition, for Python & Scientific development. The computer used for the computation time estimation is an Intel® Core™ i7 CPU K 875 @ 2.93GHz × 8 featuring a 2.0 GiB Ram memory running on Ubuntu 16.04 LTS. More details on the algorithmic implementation are given below.

4.2) Data sample

4.2.1) Overview

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.

Figure 4.5: Driving sample used in each part of the analysis

For the purposes of this doctoral research, a sample of one hundred and seventy one (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. More details on the database selection are given below. The selection made is presented in table 4.1 and illustrated in figure 4.5.

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

In all three parts of the research, the indicators exploited are the distance travelled, the number of harsh acceleration events, the number of harsh braking events, the seconds of using the mobile phone, the seconds of speeding per trip travelled. For driver efficiency

analysis, it was necessary to cumulate data per trip as described below. The definition of these indicators is given in table 4.2.

Table 4.2: Description of the variables recorded

Variable name	Variable short description
ha_x	number of harsh acceleration events in X road type
ha_{urban}	number of harsh acceleration events in urban road
ha_{rural}	number of harsh acceleration events in rural road
$ha_{highway}$	number of harsh acceleration events in highway
hb_x	number of harsh braking events in X road type
hb_{urban}	number of harsh braking events in urban road
hb_{rural}	number of harsh braking events in rural road
$hb_{highway}$	number of harsh braking events in highway
$speeding_x$	seconds of speed limit violation in X road type
$speeding_{urban}$	seconds of speed limit violation in urban road
$speeding_{rural}$	seconds of speed limit violation in rural road
$speeding_{highway}$	seconds of speed limit violation in highway
$mobile_x$	seconds of mobile phone usage in X road type
$mobile_{urban}$	seconds of mobile phone usages in urban road
$mobile_{rural}$	seconds of mobile phone usage in rural road
$mobile_{highway}$	seconds of mobile phone usage in highway
$distance_x$	distance driven in X road type
$distance_{urban}$	distance driven in urban road
$distance_{rural}$	distance driven in rural road
$distance_{highway}$	distance driven in highway

Additionally, data have been collected from a questionnaire, which was administered to a proportion of the drivers that were selected for the analysis conducted herein. More details on the questionnaire administered are given below. It is highlighted that two different data samples, data_sample_1 and data_sample_2, were used in the driver efficiency analysis, which included participants that did not answer the questionnaire administered and participants that answered the questionnaire administered, respectively.

4.2.2) Large-scale data investigation

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.

Figures 4.6, 4.7, 4.8, 4.9, 4.10, 4.11, 4.12, 4.13, 4.14, 4.15 illustrate some descriptive statistics concerning the attributes of the driving sample collected from the smartphone devices. The first five figures present the sample collected on a trip basis while the latter five figures present the sample collected on a driver basis. Figure 4.6 and 4.11 presents the average trip duration, average trip driving duration (duration of a trip with no stops

included) and average trip distance travelled respectively. Figure 4.7, 4.8, 4.12 and 4.13 presents the average number of harsh events (acceleration and braking respectively) occurred in urban, rural and highway road network per 100 km distance travelled in each road network. Figure 4.9, 4.10, 4.14, 4.15 illustrate the percentage of time using the mobile phone and the percentage of time driving over the speed limits per trip travelled.

It is evident that most trips have an average trip duration less than 20 minutes, an average trip driving duration less than 15 minutes and that most trips have a length between 0km and 10km. Additionally, it appears that the average number of harsh acceleration events occurred per 100km is higher than the number of harsh braking occurred. Especially for the urban and rural road types, this difference is sharp. As for the mobile usage, there is low or no mobile phone usage for the majority of the trips performed whereas the distribution of the percentage of speed limit exceedance appears to be more balanced. It should be highlighted though that speed limit exceedance takes place for almost two thirds of the trips recorded for urban roads. On the other hand, mobile phone usage takes place in more than the two thirds of the trips recorded for all road types. It is noted that the total number of trips is not equal to the sum of the trips illustrated in each sub-figure because there are several trips that were not performed in all road types.

Figure 4.6: Histogram of the i) average trip duration (minutes), ii) average trip driving duration (minutes) and iii) average trip distance (km) of the driving sample's per trip characteristics (from left to right).

Figure 4.7: Histogram of the i) average ha$_{urban}$/ distance$_{urban}$, ii) average ha$_{rural}$/ distance$_{rural}$, iii) average ha$_{highway}$/ distance$_{highway}$ per 100 km of the driving sample's per trip characteristics (from left to right).

Figure 4.8: Histogram of the i) average hb_{urban}/ $distance_{urban}$, ii) average hb_{rural}/ $distance_{rural}$, iii) average $hb_{highway}$/ $distance_{highway}$ per 100 km of the driving sample's per trip characteristics (from left to right).
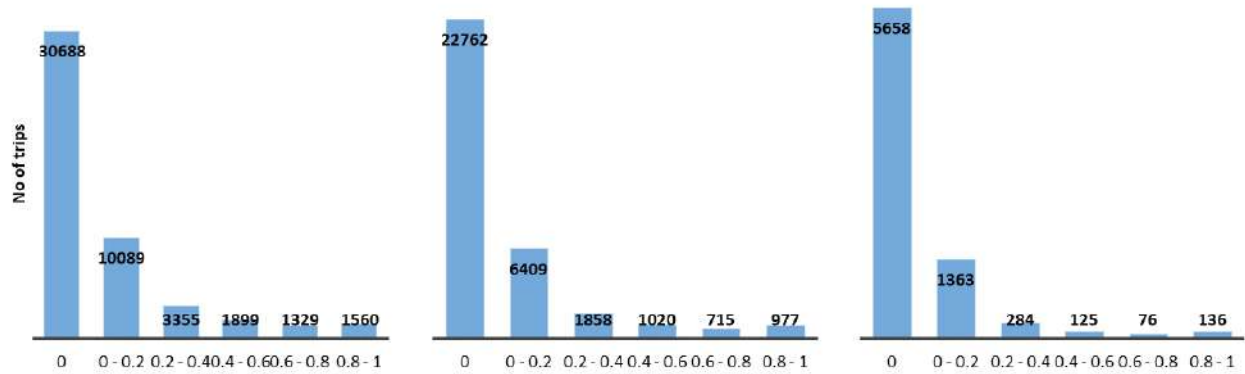
Figure 4.9: Histogram of the percentage of time using the mobile phone per trip driving duration in i) urban, ii) rural iii) highway of the driving sample's per trip characteristics (from left to right).
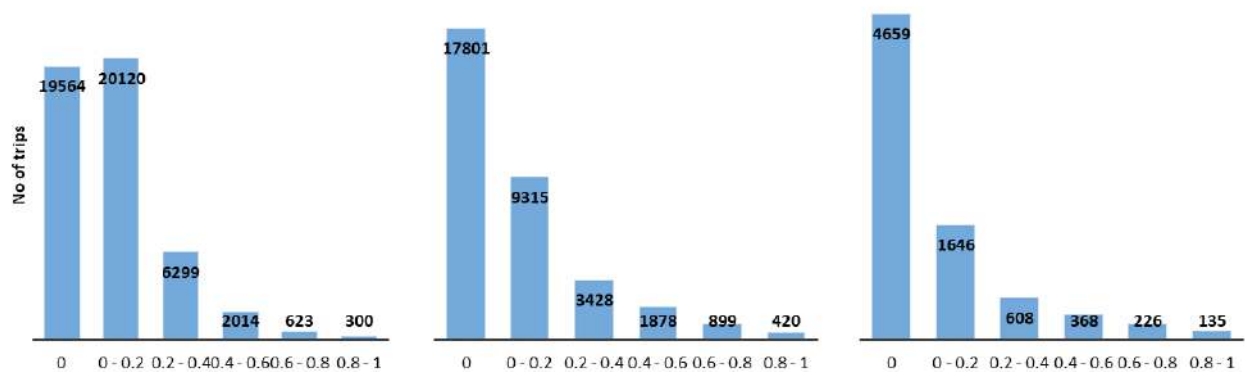
Figure 4.10: Histogram of the percentage of time driving over the speed limits per trip driving duration in i) urban, ii) rural iii) highway of the driving sample's per trip characteristics (from left to right).

Figures 4.11, 4.12, 4.13, 4.14, 4.15 illustrate some descriptive statistics regarding the attributes of the driving sample collected from the smartphone devices. Figure 4.11 presents the average trip duration, average trip driving duration (duration of a trip with no stops included) and average trip distance travelled respectively. Figure 4.12 and 4.13 presents the average number of events (harsh acceleration and braking respectively) occurred in urban, rural and highway road network per 100 km distance travelled in each road network. Figure 4.14 and 4.15 illustrate the percentage of time using the mobile phone and the percentage of time driving over the speed limits per trip travelled. It is evident that most drivers have an average trip duration around 60 minutes, an average trip driving duration of less than 60 minutes and that most drivers were monitored for less than 3000km. Additionally, it appears that the average number of harsh acceleration events occurred per 100km is higher than the number of harsh braking occurred in all road types. Mobile phone usage is limited to less than 5% of driving trip duration for the one third of the drivers in urban roads and for the 2 thirds in highways. As for the speed limit exceedance, the majority of the drivers driver over the speed limit between 10% and 30% of the driving trip duration in urban and rural roads and less than 10% in highways.
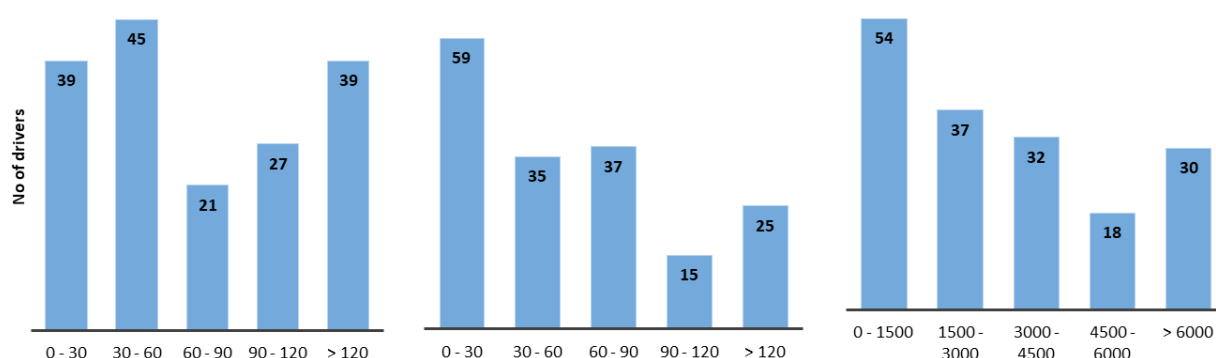
Figure 4.11: Histogram of the i) average trip duration (minutes), ii) average trip driving duration (minutes) and iii) average trip distance (km) of the driving sample's per driver characteristics (from left to right).
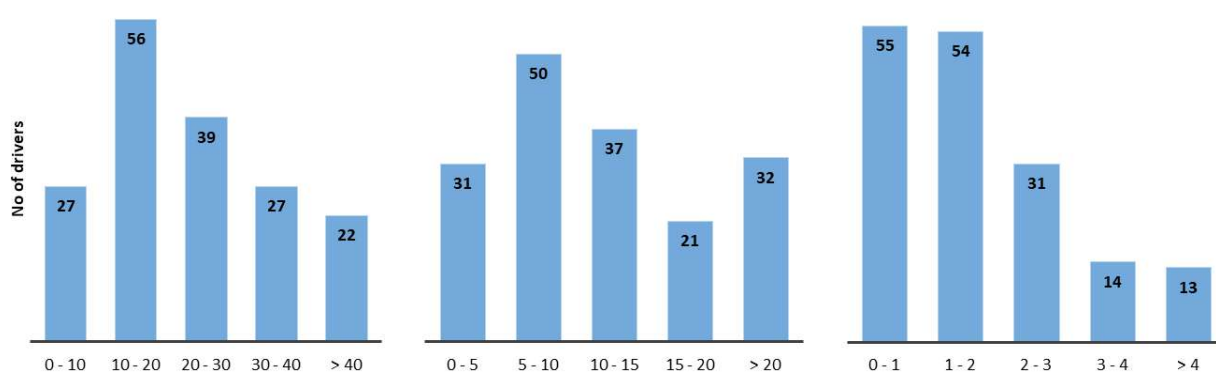
Figure 4.12: Histogram of the i) average ha$_{urban}$/ distance$_{urban}$, ii) average ha$_{rural}$/ distance$_{rural}$, iii) average ha$_{highway}$/ distance$_{highway}$ per 100 km of the driving sample's per driver characteristics (from left to right).
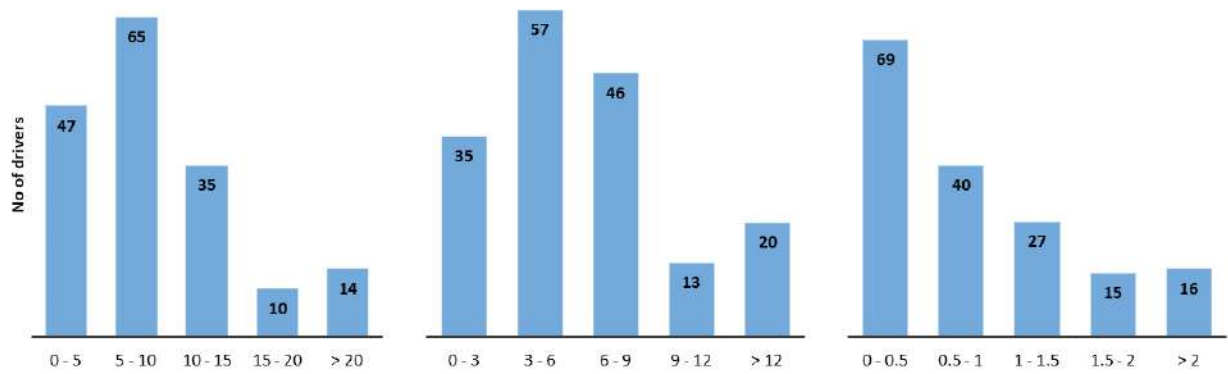
Figure 4.13: Histogram of the i) average hb_{urban}/ $distance_{urban}$, ii) average hb_{rural}/ $distance_{rural}$, iii) average $hb_{highway}$/ $distance_{highway}$ per 100 km of the driving sample's per driver characteristics (from left to right).
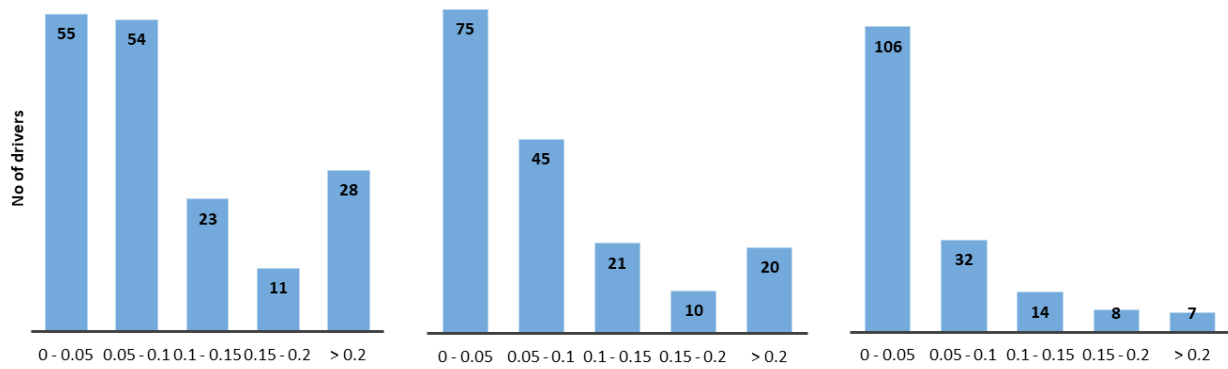
Figure 4.14: Histogram of the percentage of time using the mobile phone per trip driving duration in i) urban, ii) rural iii) highway of the driving sample's per driver characteristics (from left to right).
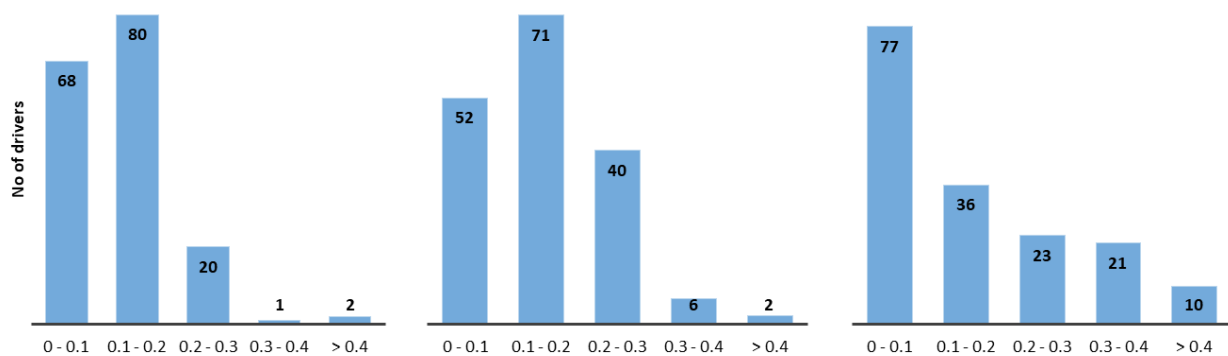
Figure 4.15: Histogram of the percentage of time driving over the speed limits per trip driving duration in i) urban, ii) rural iii) highway of the driving sample's per driver characteristics (from left to right).

It appears that the majority of the drivers (131 out of 171) have driven between 20 and 200 hours, while only a few spent more than 17 hours on road. Regarding distance covered, 155 drivers drove between 500 and 10,000 km. Only two drivers demonstrated a mileage of more than 10,000km.

It should be highlighted that the required amount of driving data will be quantified in driving time and not in driving distance in order for the results to be comparable with other past studies (Shichrur et al. (2014)). The difference between trip duration and trip driving duration is that the first includes possible stops. They are both measured in driving hours.

4.2.3) Trip efficiency analysis

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.

Figures 4.16, 4.17, 4.18, 4.19, 4.20 and table 4.3 illustrate some descriptive statistics concerning the attributes of the driving sample collected from the smartphone devices. Figure 4.16 presents the average trip duration, average trip driving duration (duration of a trip with no stops included) and average trip distance travelled respectively. Figure 4.17 and 4.18 presents the average number of events (harsh acceleration and braking respectively) occurred in urban, rural and highway road network per 100 km distance travelled in each road network. Figure 4.19 and 4.20 illustrate the percentage of time using the mobile phone and the percentage of time driving over the speed limits per trip travelled.

It is evident that most trips have an average trip duration less than 30 minutes, an average trip driving duration less than 15 minutes and that most trips have a length between 0km and 10km. Additionally, it appears that the average number of harsh acceleration events occurred per 100km is higher than the number of harsh braking events occurred. Especially for the rural and highway networks, this difference seems to be sharp. As for the mobile usage, there is low or no mobile phone usage for the majority of the trips performed whereas the distribution of the percentage of speed limit exceedance appears to be more balanced. It should be highlighted though that almost half of the trips recorded show a speed limit exceedance between 20% and 40% of driving time for urban roads. It is noted that the total number of trips is not equal to the sum of the trips illustrated in each sub-figure because there are several trips that were not performed in all road types.
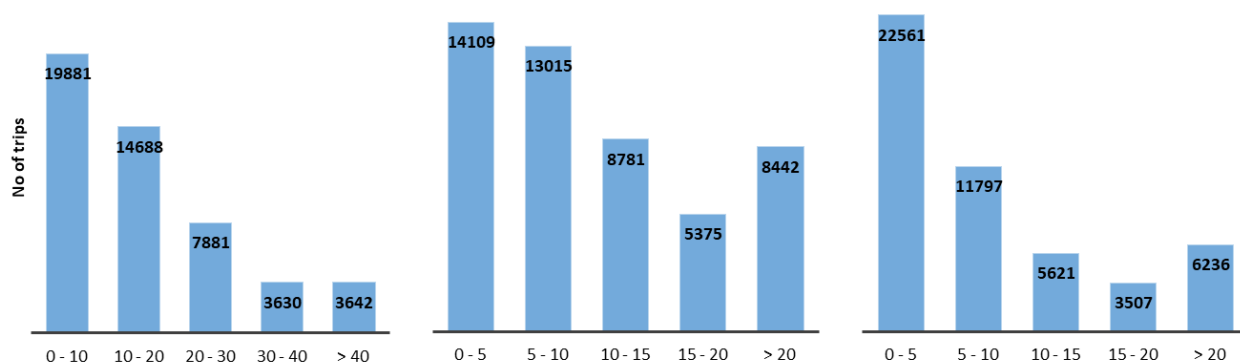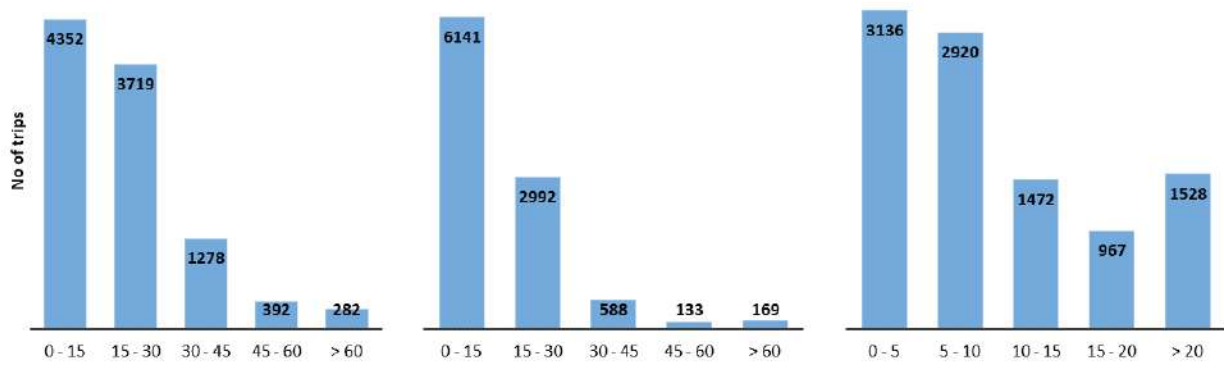
Figure 4.16: Histogram of the i) average trip duration (minutes), ii) average trip driving duration (minutes) and iii) average trip distance (km) of the driving sample (from left to right).
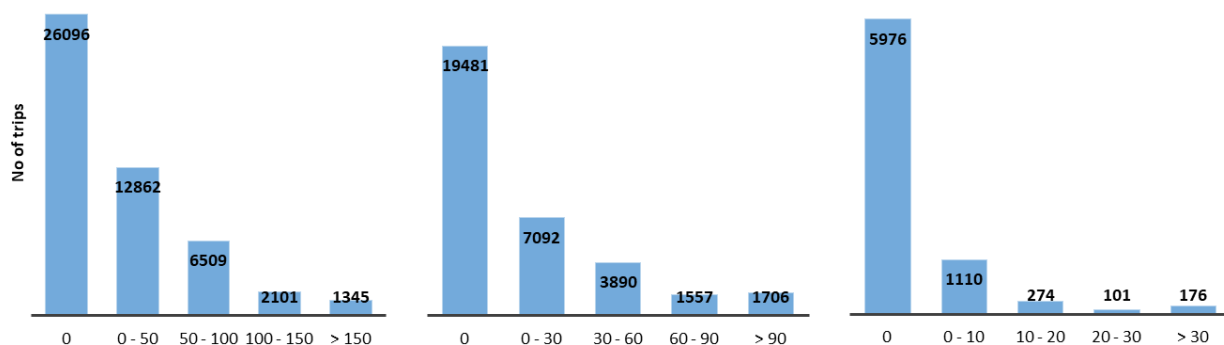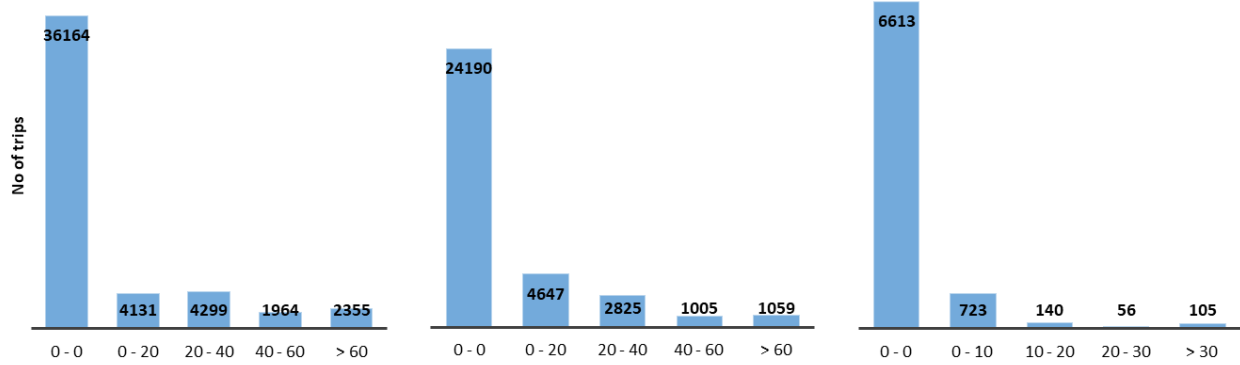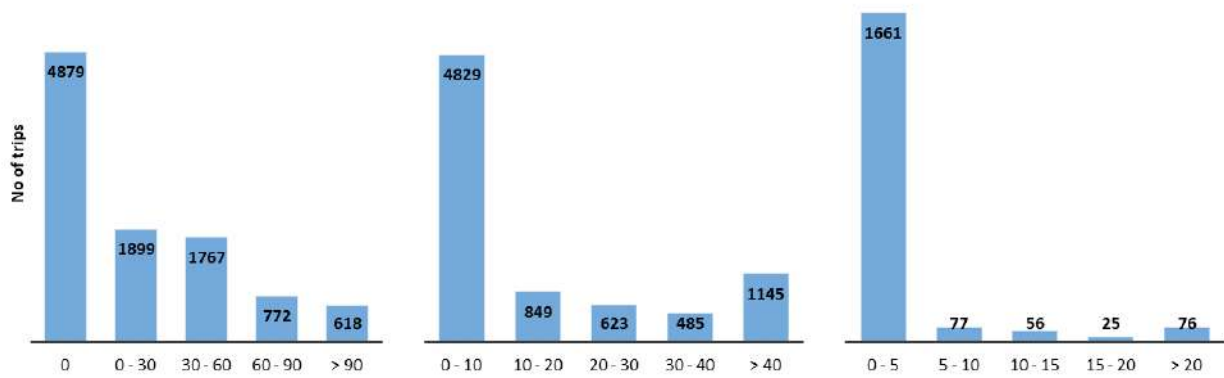

Figure 4.17: Histogram of the i) average ha_{urban}/ $distance_{urban}$, ii) average ha_{rural}/ $distance_{rural}$, iii) average $ha_{highway}$/ $distance_{highway}$ per 100 km of the driving sample (from left to right).
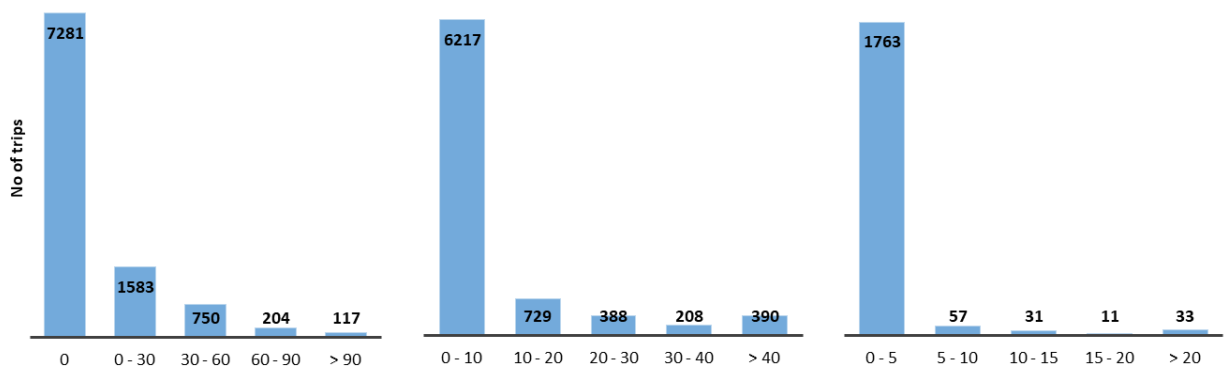
Figure 4.18: Histogram of the i) average hb_{urban}/ $distance_{urban}$, ii) average hb_{rural}/ $distance_{rural}$, iii) average $hb_{highway}$/ $distance_{highway}$ per 100 km of the driving sample (from left to right).
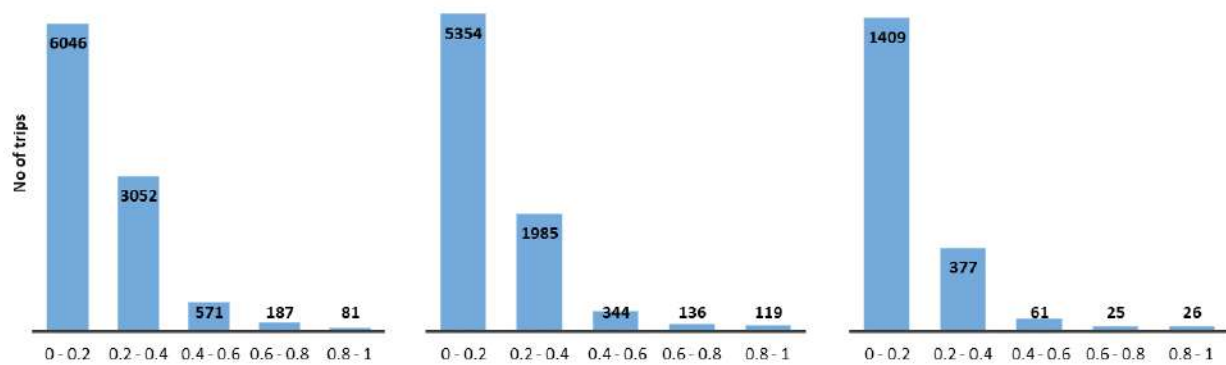
Figure 4.19: Histogram of the percentage of time using the mobile phone per trip driving duration in i) urban, ii) rural iii) highway of the driving sample (from left to right).
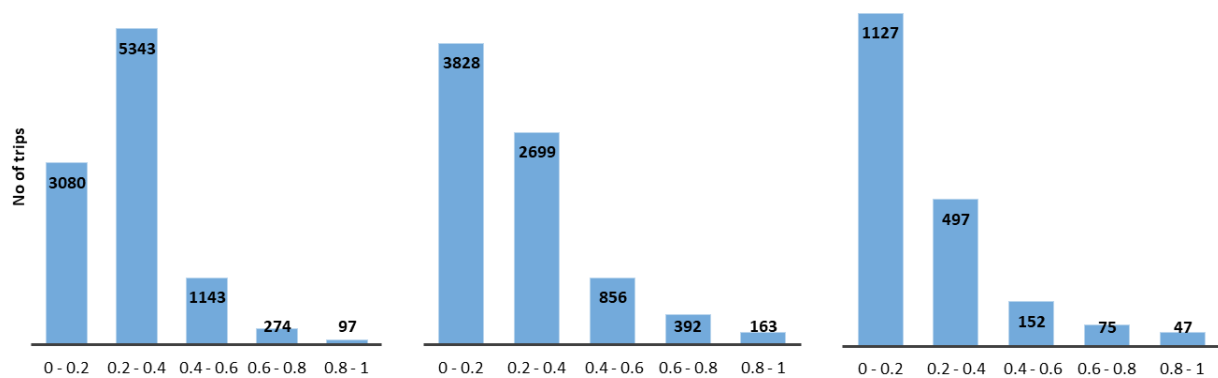
Figure 4.20: Histogram of the percentage of time driving over the speed limits per trip driving duration in i) urban, ii) rural iii) highway of the driving sample (from left to right).

Table 4.3 provides some descriptive statistics of the per trip values of the variables recorded. In other words, the final database includes the value of each variable considered in the DEA models, constructing thus a database with one row per trip, the descriptive statistics of which are presented in table 4.3. Taking into account the skewness of the metrics, the distributions of harsh acceleration and harsh braking appear to be more symmetric compared to the rest of the metrics collected. The only exception is the case of harsh acceleration events recorded in highways that show a higher skewness. The table also confirms that the number of harsh acceleration events is higher than the number of harsh braking events in all road types and that the average percentage of speed limit exceedance is higher than the average percentage of mobile phone usage. Nonetheless, the median of these two values is 0 in almost all road types (except from the speeding$_{urban}$ value) which shows that there is no mobile phone usage and/or speed limit exceedance over the 50% of the trips.

137

Table 4.3: Descriptive statistics of the per trip values of the variables recorded

	Distance (km)	HA	HB	Mobile (sec)	Speeding (sec)
Urban					
Min	0.00	0.00	0.00	0.00	0.00
Max	144.41	452.83	432.76	1.03	1.00
Average	4.95	25.30	8.98	0.06	0.12
Standard Deviation	5.34	38.65	22.00	0.14	0.16
Median	3.68	7.45	0.00	0.00	0.05
Kurtosis	215.90	11.92	44.37	12.80	5.64
Skewness	10.26	2.70	5.03	3.35	2.17
Rural					
Min	0.00	0.00	0.00	0.00	0.00
Max	130.35	1023.08	422.18	1.00	1.00
Average	5.24	18.38	7.43	0.06	0.12
Standard Deviation	8.65	40.22	21.25	0.15	0.20
Median	2.73	0.00	0.00	0.00	0.01
Kurtosis	50.56	84.66	118.36	15.64	4.24
Skewness	5.64	6.66	8.33	3.79	2.11
Highway					
Min	0.00	0.00	0.00	0.00	0.00
Max	292.38	399.16	244.02	1.09	1.00
Average	2.95	3.26	1.52	0.05	0.10
Standard Deviation	15.27	17.86	9.46	0.15	0.20
Median	0.00	0.00	0.00	0.00	0.00
Kurtosis	139.46	288.01	340.86	22.01	5.56
Skewness	10.82	15.00	15.83	4.46	2.42

4.2.4) Driver efficiency analysis

For the purposes of this part of the research, 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.

Figures 4.21, 4.22, 4.23, 4.24, 4.25, 4.26, 4.27, 4.28 and table 4.4 illustrate some descriptive statistics concerning the attributes of the driving sample used for the analysis of the temporal evolution of driving efficiency. Figure 4.21, 4.23, 4.25 and 4.27 presents the total trip duration, total driving distance travelled and the average number of harsh acceleration events per 100 km driven for the urban data_sample_1, the rural data_sample_1, the urban data_sample_2 and the rural data_sample_2 respectively. Figure 4.22, 4.24, 4.26 and 4.28 presents the average number of harsh braking events per 100 km driven, the average percentage of time using the mobile phone and the average percentage of time driving over the speed limits for the urban data_sample_1, the rural data_sample_1, the urban data_sample_2 and the rural data_sample_2 respectively.

As for the data_sample_1, it is evident that most drivers were recorded for less than 100 hours and between 800km and 1200km in urban roads whereas for rural drivers were monitored for less than 60 hours and 800 km. Again, it appears that the average number of harsh acceleration events occurred per 100km is higher than the number of harsh braking events occurred per 100km. Mobile usage detected is relatively low and significantly lower in rural than urban roads. On the other hand, the distribution of the percentage of speed limit exceedance appears to be more balanced and similar in both road types except from the first two ranges of 0-2% and 2-4%. It is highlighted though that a speed limit exceedance of more than 4% of the driving time is showed for the 40% of the drivers in urban roads and 36% in rural roads.
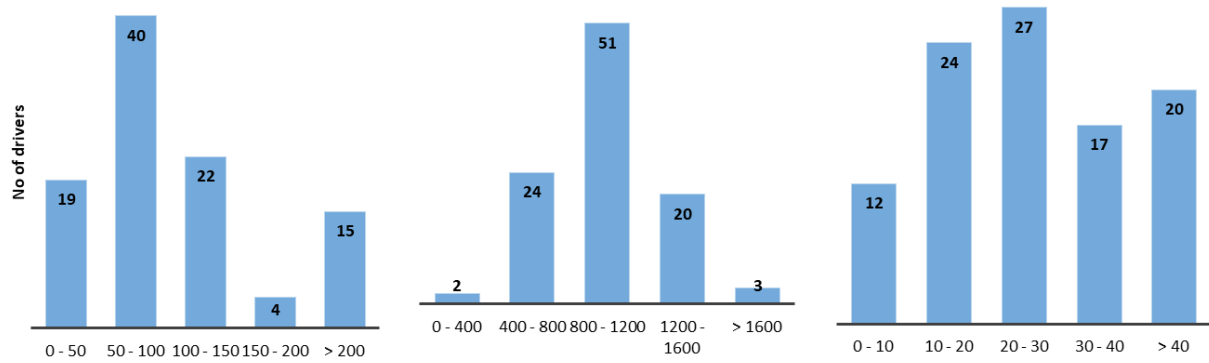
Figure 4.21: Histogram of the i) total trip duration (hours), ii) total driving distance (km) and iii) average number of ha_{urban}/ $distance_{urban}$ per 100 km of the data_sample_1 travelled in urban road type (from left to right).
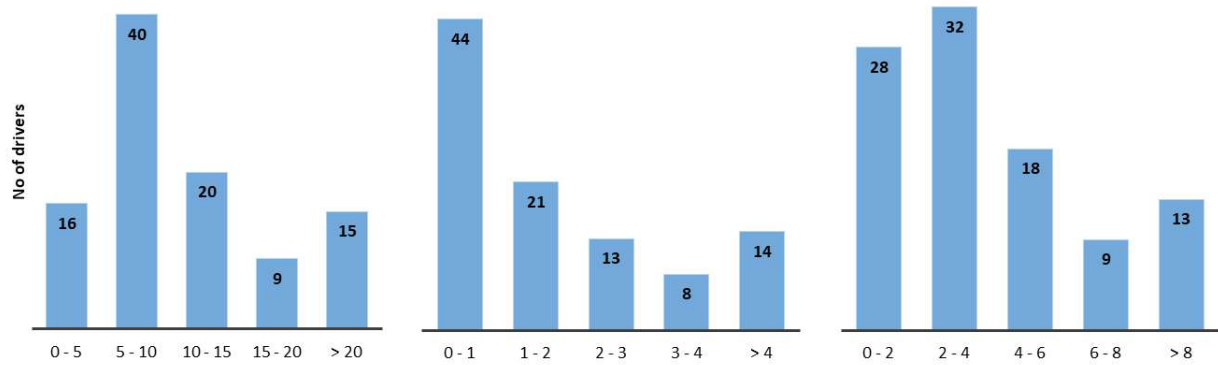
Figure 4.22: Histogram of the i) average number of hb_{urban}/ $distance_{urban}$ of 100 km ii) average percentage of time using the mobile phone and iii) average percentage of time driving over the speed limits of the data_sample_1 travelled in urban road type (from left to right).
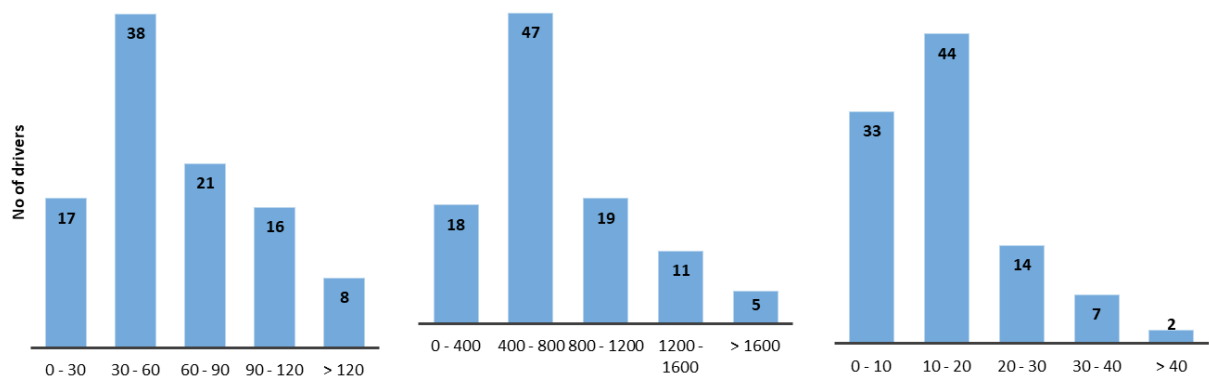
Figure 4.23: Histogram of the i) total trip duration (hours), ii) total driving distance (km) and iii) average number of ha_{rural}/ $distance_{rural}$ per 100 km of the data_sample_1 travelled in rural road (from left to right).

140
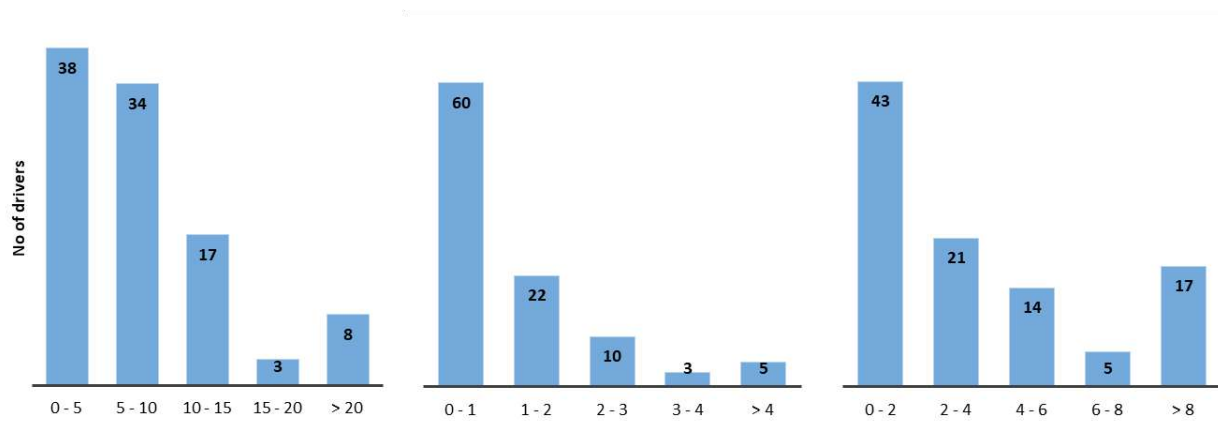
Figure 4.24: Histogram of the i) average number of hb$_{rural}$/ distance$_{rural}$ per 100 km, ii) average percentage of time using the mobile phone and iii) average percentage of time driving over the speed limits of the data_sample_1 travelled in rural road type (from left to right).

As for the data_sample_1, it is evident that most drivers were recorded for less than 100 hours and less than 1200km in urban roads whereas for rural drivers were monitored for more than 60 hours and less than 800 km. Again, it appears that the average number of harsh acceleration events occurred per 100km is higher than the number of harsh braking events occurred per 100 km. More precisely, most drivers performed more than 20 harsh acceleration events in urban roads and more than 10 in rural roads while the number of harsh braking events was less than 10 in both road types. Mobile usage detected is relatively low and significantly lower in rural than urban roads. The percentage of speed limit exceedance shows a similar distribution in both road types except from the last range of above 8%, which shows a higher concentration. It is highlighted though that a speed limit exceedance of more than 4% of the driving time is showed for around a 42% of the drivers in urban roads and 33% in rural roads.
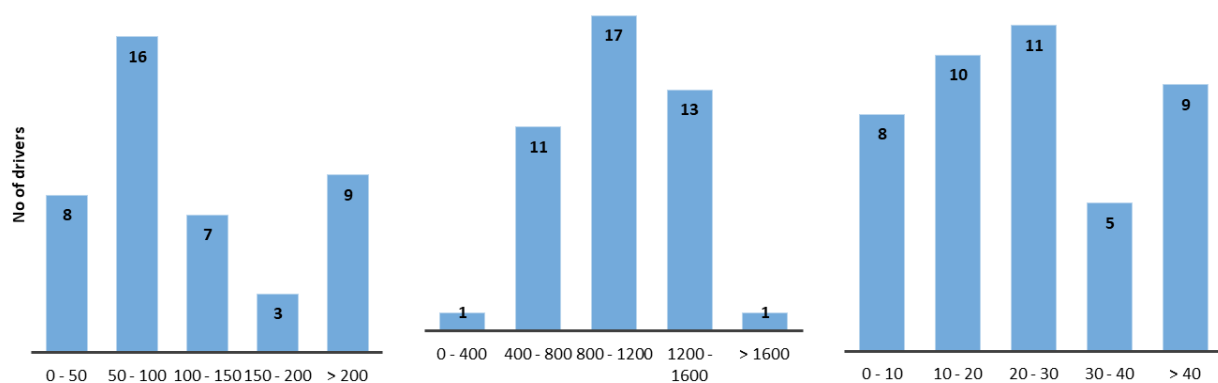
Figure 4.25: Histogram of the i) total trip duration (hours), ii) total driving distance (km) and iii) average number of ha$_{urban}$/ distance$_{urban}$ per 100 km of the data_sample_2 travelled in urban road type (from left to right).
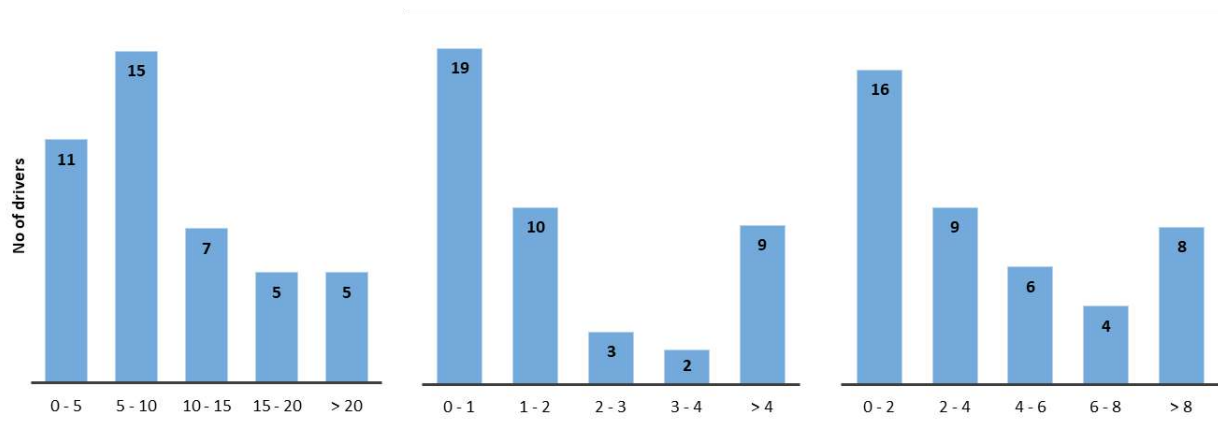
Figure 4.26: Histogram of the i) average number of hb_{urban}/ $distance_{urban}$ 100 km ii) average percentage of time using the mobile phone and iii) average percentage of time driving over the speed limits of the data_sample_2 travelled in urban road type (from left to right).
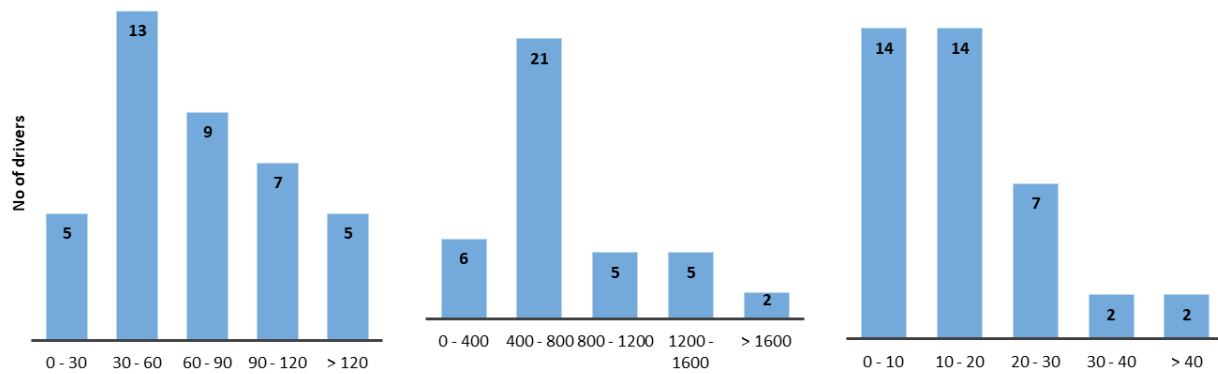
Figure 4.27: Histogram of the i) total trip duration (hours), ii) total driving distance (km) and iii) average number of ha_{rural}/ $distance_{rural}$ per 100 km of the data_sample_2 travelled in rural road type (from left to right).
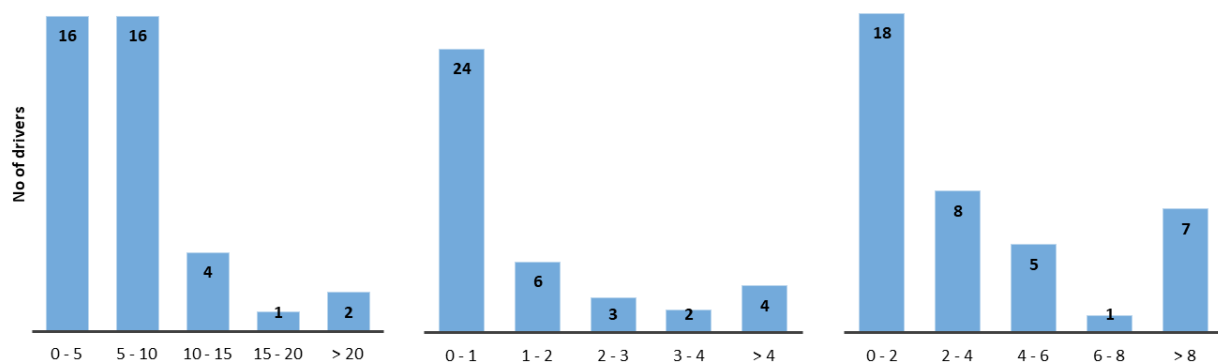
Figure 4.28: Histogram of the i) average number of hb_{rural}/ $distance_{rural}$ per 100 km ii) average percentage of time using the mobile phone and iii) average percentage of time driving over the speed limits of the data_sample_2 travelled in rural road type (from left to right).

142

Table 4.4 provides some descriptive statistics of the cumulative per driver values of the variables recorded. In other words, the final database includes the cumulative value (from the number of trips recorded in each road type) of each variable considered in the DEA models, which equals to the sum of metrics recorded in each trip of the specific driver, constructing thus a database with one row per driver, the descriptive statistics of which are presented in table 4.4. For instance, if driver X has a number of Y recorded trips, the number of harsh acceleration events of each of the Y trips is summed and used in the DEA models. Taking into account the skewness of the metrics, almost all data appear to be skewed right. The only exception is the case of distance in urban road type which is more normally distributed with a light tail in the data of the data_sample_1 and with a slightly negative skewness in the data of the data_sample_2. The kurtosis of the harsh braking and acceleration events is observed to be significantly higher and slightly higher than all the rest of the metrics, respectively. The table also confirms that the number of harsh acceleration events is higher than the number of harsh braking events in all road and data (data_sample_1/ data_sample_2) types and that the average percentage of speed limit exceedance is higher than the average percentage of mobile phone usage.

Table 4.4: Descriptive statistics of the cumulative per driver values of the variables recorded – Analysis of the temporal evolution of driving efficiency

	Distance (km)	HA	HB	Mobile (sec)	Speeding (sec)
Urban - data_sample_1					
Min	362.01	3.05	0.42	0.00	0.32
Max	2138.24	131.77	62.95	12.04	17.18
Average	992.31	27.81	11.73	1.92	4.22
Standard	312.03	18.40	9.71	1.96	3.19
Median	1012.32	24.06	8.51	1.24	3.51
Kurtosis	1.02	9.05	12.48	6.45	2.78
Skewness	0.41	2.15	2.94	2.05	1.54
Rural - data_sample_1					
Min	52.99	1.26	0.85	0.00	0.08
Max	2459.47	68.45	41.49	8.13	26.15
Average	777.86	15.17	8.33	1.18	4.22
Standard	470.13	10.19	6.52	1.33	4.46
Median	702.22	13.26	6.46	0.75	2.73
Kurtosis	2.38	7.34	6.52	6.75	6.55
Skewness	1.34	2.04	2.13	2.12	2.24
Urban - data_sample_2					
Min	362.01	3.05	0.42	0.05	0.32
Max	1804.49	131.77	60.50	7.69	17.18
Average	1010.03	27.89	11.14	2.02	4.51
Standard	317.31	22.88	10.42	1.98	4.08
Median	1012.96	22.05	7.76	1.16	2.77
Kurtosis	-0.29	8.76	11.04	0.57	1.38
Skewness	-0.08	2.43	2.85	1.21	1.39
Rural - data_sample_2					
Min	133.40	1.26	0.85	0.00	0.08
Max	2374.86	68.45	41.49	8.13	19.43
Average	769.80	15.82	7.52	1.35	4.08
Standard	479.81	12.99	7.00	1.73	4.52
Median	628.28	12.35	5.39	0.57	2.40
Kurtosis	2.37	6.58	13.89	4.89	4.19
Skewness	1.42	2.22	3.30	2.07	2.01

4.2.5) Questionnaire

A number of users that participated in the driver efficiency analysis conducted, took part in the online survey, which was administered by OSeven Telematics. The questions that are most related to the present research are:

1) Which year did you obtain your driving license?

2) How many years have you been driving?

4) How many km do you drive per year?

5) Do you own the vehicle you most frequently use?

6) How much is your gas consumption while driving the vehicle?

7) In how many accidents have you been involved as a driver (either with your responsibility or not) to date?

8) In how many accidents have you been involved as a driver (either with your responsibility or not) during the last 3 years?

9) In how many accidents have you been involved as a driver with injury/injuries of any of the drivers/passengers (either with your responsibility or not) to date?

10) In how many accidents have you been involved as a driver with injury/injuries of any of the drivers/passengers (either with your responsibility or not) during the last 3 years?

11) In how many accidents have you been involved as a driver with material damage only (either with your responsibility or not) to date?

12) In how many accidents have you been involved as a driver with material damage only (either with your responsibility or not) during the last 3 years?

13) How many driving fines have you received for road traffic violations during the last 3 years?

14) What is your country of residence?

15) What is your gender?

16) What is your age?

17) What is your education?

18) What is your profession?

Based on the questions 2) and 7) a new variable is created to account for the number of accidents occurred to date per 10 years of driving. To acquire a better picture on the distribution of the drivers participated in the online survey, the following figures 4.29 and 4.30 are provided:
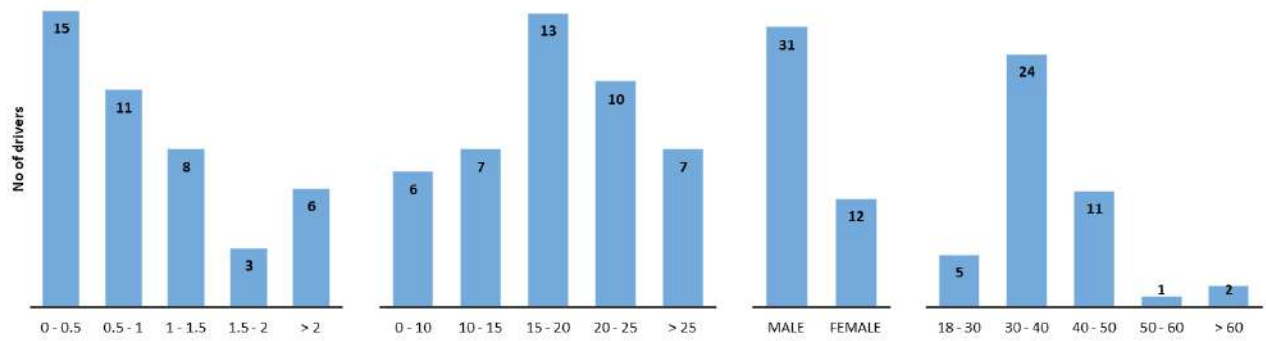
Figure 4.29: Histogram of the i) average number of accidents occurred to date/ year of driving ii) years of driving experience iii) gender distribution and iv) age distribution of the drivers that answered to the questionnaire and travelled in urban road type (from left to right).
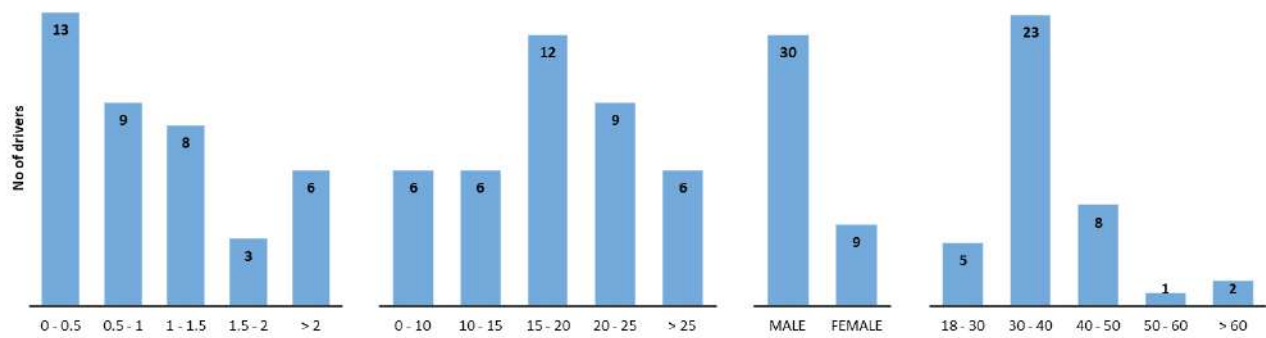
Figure 4.30: Histogram of the i) average number of accidents occurred to date/ 10 years of driving ii) years of driving experience iii) gender distribution and iv) age distribution of the drivers that answered to the questionnaire and travelled in rural road type (from left to right).

As expected, since most drivers in both road type samples are common, both distributions appear to be similar. The majority of the drivers have a driving experience of more than 15 years and less than one accident per 10 years. It also appears that approximately the 25% of participants are females and that 50% of the sample belong to the 30-40 age range.

146

Chapter 5: Implementation and Results

In this chapter, analysis conducted is presented step by step and the results arising are described and explained. For the sake of brevity, the number of harsh acceleration events, the number of harsh braking events, the seconds of mobile usage and the seconds of driving over the speed limit will be referred as HA, HB, MU and SP respectively.

Before proceeding to this stage of the doctoral research, data are filtered and prepared so that they meet the requirements set and they can be imported in the DEA models developed. Data filtering are performed in Python programming language and several scripts are written for this reason. In each step of the analysis, a different database is used and to this end, the initial database obtained should be filtered. For instance, to perform the analysis required for the temporal evolution of driving efficiency, it is necessary to obtain a specific number of each driver's last trips.

DEA improvement algorithms are also performed in Python programming language. Python packages used include pandas and numpy for numeric calculations and transformations, scipy that features quickhull algorithm, pulp for linear programming problem construction and scikit-learn for machine learning k-means clustering. More details on the algorithm implementation are given below. Coding is applied using Pycharm IDE Community edition, for Python & Scientific development. The computer used for the computation time estimation is an Intel® Core™ i7 CPU K 875 @ 2.93GHz × 8 featuring a 2.0 GiB Ram memory running on Ubuntu 16.04 LTS. More details on the algorithmic implementation are given below.

5.1) Large-scale data investigation

As defined in the methodological steps, in this step of the analysis it is examined whether or not the sum of metrics is proportionally increased to the sum of distances.

One of the fundamentals of CRS DEA that cannot be overlooked is that the inputs are increasing linearly to outputs. It is therefore essential to investigate the evolution of driving metrics in time, compared to distance travelled, not only in total but also in each moving window examined. Additionally, the amount of adequate driving data sample that should be collected for each driver is estimated in this step to ensure the significance of the results arising. As mentioned above, for the analysis of the total driving behaviour as well as in moving window considered for the temporal analysis of driving efficiency, driving metrics should be linearly increased to distance travelled in order to apply CRS DEA. Therefore, the amount of data collected for each driver should be exceeding the minimum amount of data found that is required in each time step and in total.

The statistical analysis using the data collected from the smartphone was conducted to determine the driving distance at which the rate of the driving indicators converges to a

stable index and therefore DEA can be applied in each time step and no more data are required to be collected in total. In this research the magnitude of change measurement in a time series is employed which is decreasing over distance (km) as the specific magnitude converges on its average rate. At the same time, this means that the rate at which an event (number of harsh acceleration/ braking events, seconds of mobile phone usage, seconds driving over the speed limits per 100 km) occurs also converges to its average rate e.g. the average rate of harsh acceleration events for the specific driver (average number of harsh acceleration events). For each driver and after each trip that took place, the above metrics were calculated by diving the total number of occurred events by the total distance driven thus far, constructing thus a time series of average events per km. The mathematical formulation for calculating the convergence index (CI_i) of the event rate is given in the methodological approach chapter.

A moving average of the magnitude CI_i is calculated for all participants. The moving window considered is 40 trips and it is deemed to be within acceptable margins when the average change measurement for the moving window of 40 trips and for at least 200 km is less or equal to 5% (0.05). The reason of choosing 40 trips and 200 km is that all drivers included in the analysis should had driven at least 40 trips that approximately equals the typical monthly number of trips for a driver assuming that each driver drives in average 2 trips of 5 km a day for 5 working days a week. Considering approximately one month for monitoring and assessing a driver's behaviour is deemed a period long enough for capturing the short-term changes in driver's behaviour and short enough for ignoring the long-term changes in driver's behaviour that will be captured in the analysis of temporal evolution of driving efficiency.

The value 5% reflects the per mille change in measured change of the respective harsh event rate i.e. the moving average is attempting to capture the time after which the average per mille change is steadily less than 0.05. The reason why this value was selected can be better explained using figure 5.1.
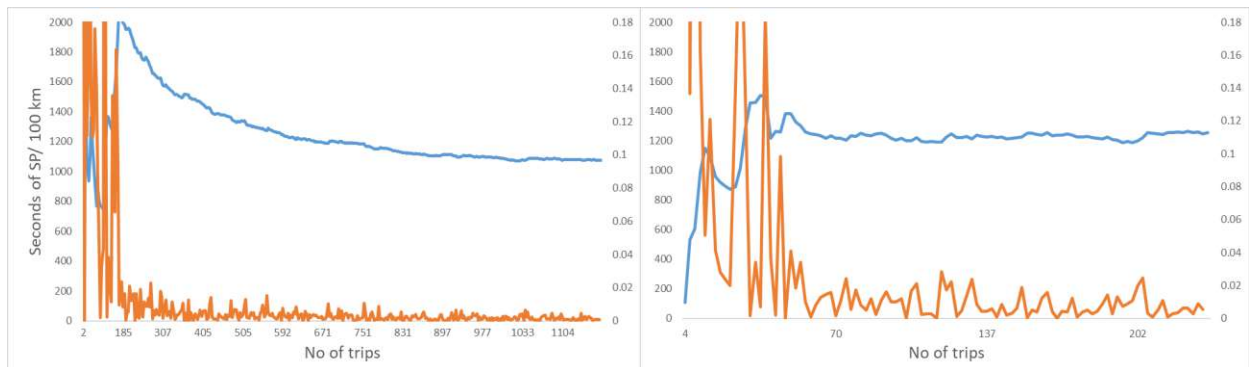
Figure 5.1: The evolution of the average cumulative speeding event rate per 100 km and convergence index over distance (km) for two drivers whose speeding behaviour is (a) converged (b) non-converged

Figure 5.1a and 5.1b demonstrates the evolution of the average cumulative speeding event rate per 100 km and the convergence index of two individual drivers of a 758km

and 239km total driving distance respectively. The blue line refers to the average cumulative speeding event rate per 100 km and the orange line to the convergence index. The blue and the orange line are plotted on the primary Y-axis and the secondary Y-axis, respectively. It is evident that in figure 5.1a the driver's behaviour is gradually converging after approximately 270km of driving as the average cumulative speeding events per 100 km and the average convergence index are not significantly altered from that time and after. On the other hand, it can be said from figure 5.1b that the driver's convergence index and the average cumulative speeding event rate is not converged since the average cumulative speeding event rate is fluctuating and the convergence index is significantly increased over 0. The same analysis was conducted for all drivers and the optimum value for average measurement change was found to be (should probably use median value as measurement) around 5%. In addition to that, this value was chosen because it can be considered a secure low per km change to draw statistically significant conclusions.

For each metric, the procedure described above was implemented to find whether a driver's behaviour is converged or not and what is the required distance to reach that point. For the total sample of converged drivers, the analysis per each metric recorded was conducted for four different value range categories, which were defined by the three percentiles of 25%, 50% and 75% of the converged drivers' sample. This categorization is implemented to enable the investigation of necessary recording time for drivers of different value ranges 1, 2, 3 and 4.

The results of the analysis conducted are presented in the following section. The number of necessary kilometres for the examined metrics to converge are plotted with the respective metric value for each road type in figures 5.2, 5.3 and 5.4 in order to observe any obvious trend or correlation between these two magnitudes. In addition, tables 5.1, 5.2 and 5.3 provide some descriptive statistics of the metric values recorded and the distance travelled to convergence by users per value range category in each road type.

5.1.1) Urban

Figure 5.2 illustrates the distance in km that is required for each driver's behaviour to converge based on the cumulative average of the HA, HB, MU and SP in urban road type. To begin with, it is difficult to draw a clear conclusion from the scatter plots, which showed that there is no apparent data trend for all metrics. Especially for HA and SP per 100km, distance required for convergence appears to be lower than the other two metrics no matter what the value range of the metrics is. On the other hand, MU and HB per 100km are more random in terms of the required distance and scatter points do not tend to concentrate in lower distance values. Nevertheless, as distance required for convergence increases, MU value range decreases.

Figure 5.2: Required distance (km) for each driver's behaviour to converge based on the cumulative average per 100 km of HA, HB, MU and SP in urban road type

Table 5.1 also confirms that there is no correlation between required distance and the range of values for the metrics recorded. Therefore, their cumulative average does not affect the distance required to acquire a clear picture for an individual driver. Nonetheless, there appears to exist a weak positive correlation between HA and required distance, which shows that the more aggressive/ risky a driver is, the more he/she should be monitored in order to be assessed.

Table 5.1: Correlation matrix between HA, HB, MU and SP and required distance for each driver's behaviour to converge in urban road type

	Distance$_{urban}$
ha$_{urban}$	0.210
hb$_{urban}$	-0.017
mobile$_{urban}$	-0.072
speeding$_{urban}$	-0.036

150

Table 5.2: Descriptive statistics of metric values and distance (*100km) per percentile range category in urban road type

Metric	Percentile range	Metric descriptive statistics				Distance to convergence		
		Average	St. Dev	Min	Max	Average	Median	St. Dev
HA	0% – 25%	8.18	2.61	-	11.61	3.89	3.09	1.77
	25% – 50%	15.92	2.8	11.61	20.54	3.99	3.36	2.08
	50% – 75%	25.01	2.48	20.54	30.14	4.15	3.22	2.46
	75% – 100%	43.23	10.92	30.14	-	5.26	5.19	2.93
HB	0% – 25%	3.05	1.26	-	4.95	5.15	4.78	1.79
	25% – 50%	6.17	0.69	4.95	7.32	4.08	3.35	1.55
	50% – 75%	8.6	0.83	7.32	10.61	5.58	4.31	3
	75% – 100%	15.65	3.12	10.61	-	4.69	3.41	2.48
MU	0% – 25%	204	101	-	332	5.59	4.07	3.95
	25% – 50%	495	78	332	606	4.4	3.47	1.74
	50% – 75%	799	111	606	1041	4.43	3.81	2.47
	75% – 100%	2063	994	1041	-	4.93	3.66	3.1
SP	0% – 25%	727	194	-	947	4.89	3.39	3.08
	25% – 50%	1081	67	947	1198	3.43	2.93	1.23
	50% – 75%	1402	119	1198	1594	3.78	3.12	1.88
	75% – 100%	1919	318	1594	-	4.55	3.48	2.87

Table 5.2 illustrates the descriptive statistics of each metric and distance per percentile range category that resulted from the analysis performed regarding the required distance to driving behaviour convergence. It is clear that for almost all metrics, performing the analysis for drivers of higher and lower percentile value range does not provide an extra value since there is no apparent trend as the metric range increases. The only exception might be HA, which presents a slight positive trend in required distance. All these observations were also highlighted in table 5.1 and figure 5.2, which lead to the conclusion that different sampling periods are not required for drivers of different percentile value range. This is probably because drivers of different percentile value ranges present similar standard deviation values as appears in figure 2 and therefore it does not take more distance for their driving characteristics to converge to their average level. Only drivers of the 75-100% percentile value range present a significantly higher standard deviation, which could probably be attributed to the fact that outlier metrics' values exist in this specific range.

As for the characteristics per driving risk level (percentile value range), more risky drivers perform 43.23 and 15.65 harsh acceleration and braking events per 100km whereas less risky drivers perform 8.18 and 3.05 harsh acceleration and braking events per 100km on average respectively. On the other hand, less risky drivers use their mobile phone

approximately for 204 sec/ 100km and drive over the speed limits for 727 sec/ 100km. Finally, risky drivers demonstrated a speed limit violation of 1919 sec/ 100km on average and a 2063 sec/ 100km mobile usage.

It is highlighted that the maximum median value of distance to convergence is selected for the determination of the required sampling distance since compared to the average value the specific statistical measure ignores the outlier values that are likely to exist in the sample. The maximum median distance value of the table is taken into consideration because all metrics should have converged to their cumulative average in order to claim that a) the driving sample acquired is adequate and b) the input/ output ratio is relatively constant to perform DEA analysis. This ensures that enough data are taken into consideration for the analysis even in the case of initially underestimating the required sampling distance due to other factors such as biased sample acquisition etc.

According to table 5.2, 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.

5.1.2) Rural

Figure 5.3 illustrates the distance in km that is required for each driver's behaviour to converge based on the cumulative average of the HA, HB, MU and SP in rural road type. It is difficult to draw a clear conclusion from the scatter plot of SP, which showed that there is no apparent data trend. For the rest of the metrics recorded, there is a weak negative correlation between the metric value and the distance to convergence. This indicates a weak negative data trend, which is confirmed both from figure 5.3 and from table 5.3. The interpretation of this is that the higher the metric value, the lower the distance required to be collected from a driver to acquire a clear picture regarding his/her driving behaviour and that as distance required for convergence increases, metrics values decrease. As for SP per 100km, it appears to be more random in terms of the required distance and scatter points do not tend to concentrate in lower metric values.

As illustrated in the graphs, more risky drivers require to be monitored for less kilometres for their behaviour to converge. This is observed in most graphs, as there are almost no drivers with a high cumulative average of metrics among those that require a higher distance for their behaviour to be converged. This is unclear only in the SP graph where drivers seem to have a more random trend. This is probably because much less distance is required for most drivers to obtain a clearer picture on their SP behaviour.
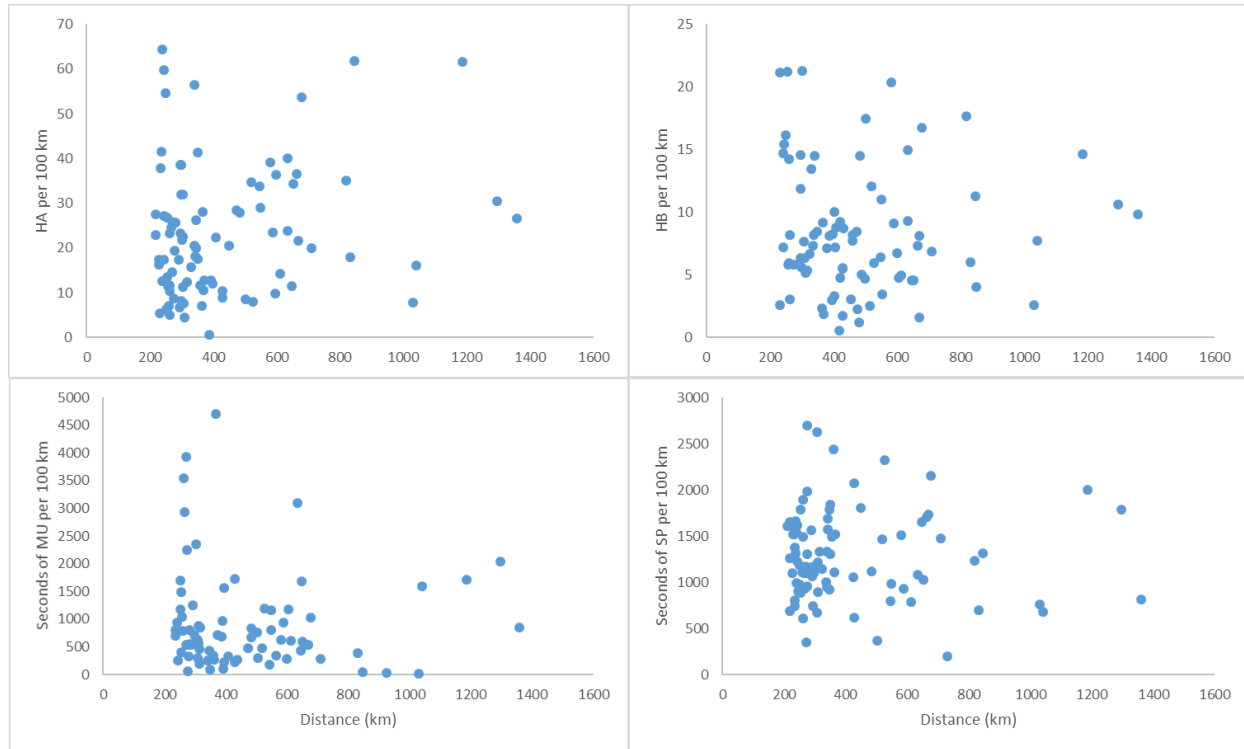
Figure 5.3: Required distance (km) for each driver's behaviour to converge based on the cumulative average per 100 km of the following metrics in rural road type: (a) Number of harsh acceleration events, (b) Number of harsh braking events, (c) Seconds of mobile usage and (d) Seconds of driving over the speed limit

Table 5.3 also confirms that there is a weak negative correlation between required distance and all metric values apart from SP, which shows that the more risky a driver is, the less he/she should be monitored in order to be assessed. Therefore, the value of SP's cumulative average does not affect at all the distance to convergence required to be collected for an individual driver.

Table 5.3: Correlation matrix between HA, HB, MU and SP and required distance for each driver's behaviour to converge in rural road type

	Distance$_{rural}$
ha$_{rural}$	-0.249
hb$_{rural}$	-0.265
mobile$_{rural}$	-0.227
speeding$_{rural}$	-0.089

153

Table 5.4: Descriptive statistics of metric values and distance (*100km) per percentile range category in rural road type

Metric	Percentile range	Metric descriptive statistics				Distance to convergence		
		Average	St. Dev	Min	Max	Average	Median	St. Dev
HA	0% – 25%	3.69	1.7	-	6.42	5.07	4.31	2.19
	25% – 50%	8.8	1.33	6.42	10.79	4.05	3.76	1.34
	50% – 75%	13.58	1.98	10.79	17.02	4.99	4.17	2.7
	75% – 100%	27.27	11.04	17.02	-	3.49	3.36	0.85
HB	0% – 25%	2.05	0.84	-	3.15	5.2	4.39	2.21
	25% – 50%	4.35	0.67	3.15	5.6	4.29	4.03	1.26
	50% – 75%	6.92	0.69	5.6	8.28	4.89	4.40	2.21
	75% – 100%	13.54	7.16	8.28	-	3.89	3.60	1.87
MU	0% – 25%	85	48	-	157	6.3	5.79	2.61
	25% – 50%	263	50	157	371	4.85	4.42	2.09
	50% – 75%	511	92	371	747	5.01	4.48	1.92
	75% – 100%	1334	684	747	-	4.11	3.62	1.66
SP	0% – 25%	454	170	-	745	4.48	3.99	1.89
	25% – 50%	851	74	745	970	4.44	3.97	2.21
	50% – 75%	1142	112	970	1315	4.19	3.88	1.72
	75% – 100%	1526	181	1315	-	4.25	3.56	1.95

Table 5.4 illustrates the descriptive statistics of each metric and distance per percentile range category that resulted from the analysis performed regarding the required distance to driving behaviour convergence. There exists a weak negative trend for all metrics besides SP, but it remains unclear whether or not an extra value is provided to the results arising when performing the analysis separately for less and more risky drivers. Therefore, this should be further investigated in future research. The only exception is SP, which presents no apparent trend in required distance. All these observations were also highlighted in table 5.3 and figure 5.3, which lead to the conclusion that different sampling periods are likely to be required for drivers of different driving risk level. This is probably because behaviour of typical and less risky drivers fluctuates more before it converges and therefore they need to be monitored more so as their driving characteristics have converged to their average level. Apparently, this is not the case in urban road since there is no obvious trend as the level of metrics increases. This is attributed to the fact that drivers are more volatile on the specific road type and as a result, the amount of data required to be collected is not strongly related to the aggressiveness of each driver. On the other hand, more risky drivers are likely to have a

more steady, risky though, behaviour. Again, drivers of the 75-100% percentile value range present a significantly higher standard deviation, which could probably be attributed to the fact that outlier metrics' values exist in this specific range.

As for the characteristics per driving risk level, more risky drivers perform 27.27 and 13.54 harsh acceleration and braking events per 100km whereas less risky drivers perform 3.69 and 2.05 harsh acceleration and braking events per 100km on average respectively. On the other hand, less risky drivers use their mobile phone approximately for 85 sec/ 100km and drive over the speed limits for 454 sec/ 100km. Finally, risky drivers demonstrated a speed limit violation of 1526 sec/ 100km on average and a 1334 sec/ 100km mobile usage.

For the same reasons as in urban roads, the maximum median value of distance to convergence is selected for the determination of the required sampling distance. According to table 5.4, 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.

5.1.3) Highway

Figure 5.4 illustrates the distance in km that is required for each driver's behaviour to converge based on the cumulative average of the HA, HB, MU and SP in rural road type. As in rural road type, it is difficult to draw a clear conclusion from the scatter plot of SP, which shows no apparent data trend. For the rest of the metrics recorded, a weak negative correlation exists between the metric value and the distance to convergence. Specifically for MU, this trend seems to be even weaker. This data trend is also confirmed both from figure 5.4 and from table 5.5. The interpretation of this is that the higher the metric value, the lower the distance required to be collected from a driver to acquire a clear picture regarding his/her driving behaviour and that as distance required for convergence increases, metrics values decrease. As for SP per 100km, it appears to be more random in terms of the required distance and scatter points do not tend to concentrate in lower metric values.

As illustrated in the graphs, more risky drivers require to be monitored for less kilometres for their behaviour to converge. This is observed in most graphs, as there are almost no drivers with a high cumulative average of metrics among those that require a higher distance for their behaviour to be converged. This is unclear only in the HA and the SP graph where drivers seem to have a more random trend. This is noticed in all road types and leads to the conclusion that SP metric is the least critical when estimating the

155

sampling distance required. This is probably because SP behaviour is more random and therefore more monitoring is required for most drivers to obtain a clearer picture.

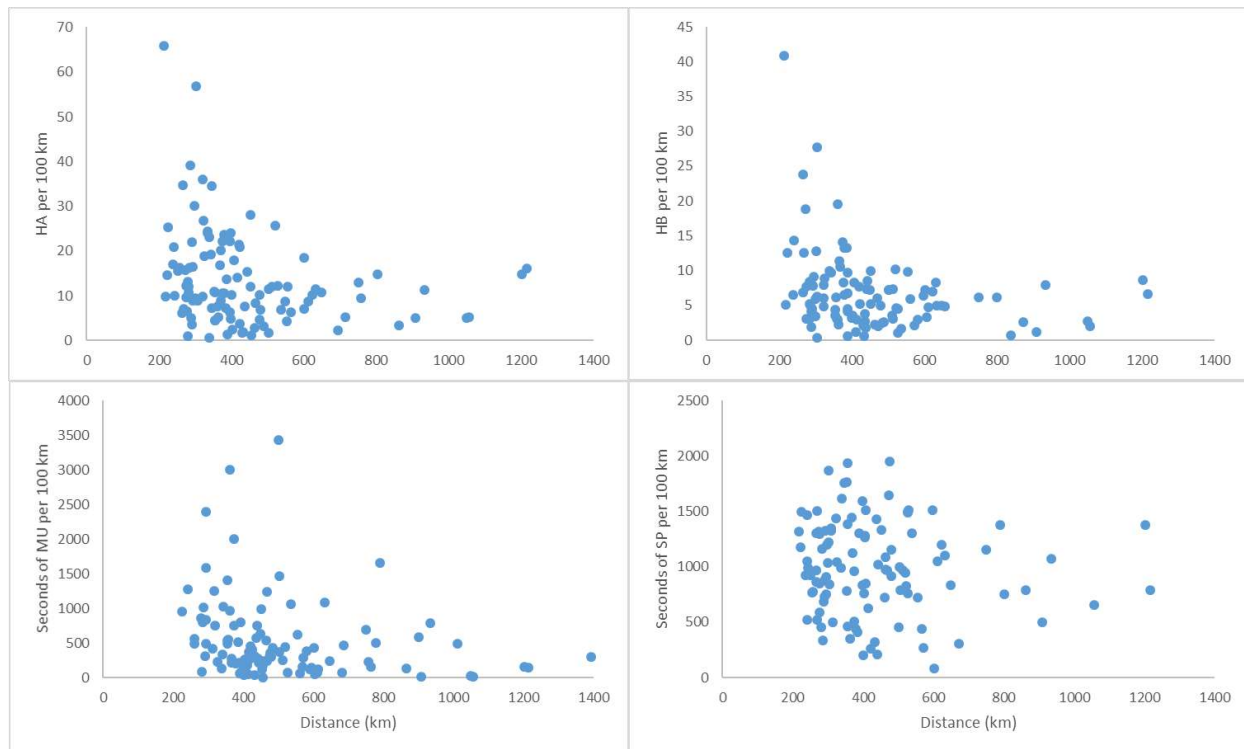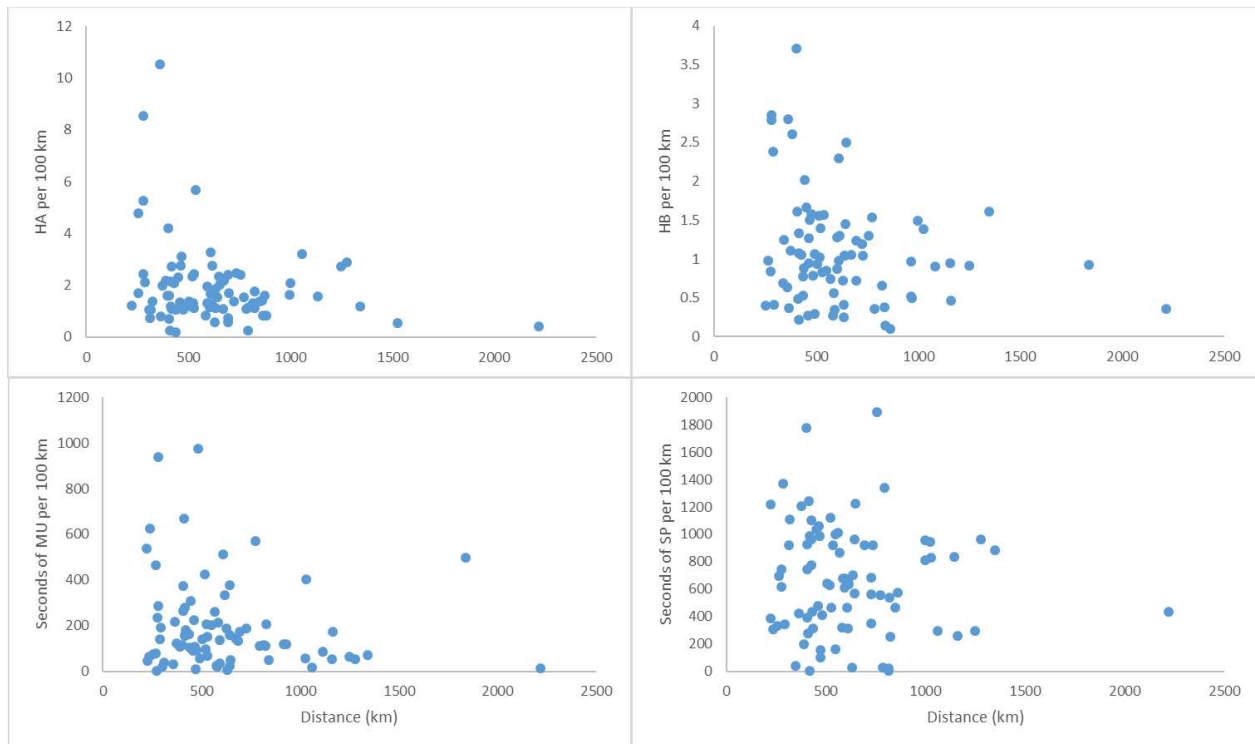*Figure 5.4: Required distance (km) for each driver's behaviour to converge based on the cumulative average per 100 km of the following metrics in highways: (a) Number of harsh acceleration events, (b) Number of harsh braking events, (c) Seconds of mobile usage and (d) Seconds of driving over the speed limit*

Table 5.5 also confirms that there is a weak negative correlation between required distance and all metric values apart from SP, which shows that the more risky a driver is, the less he/she should be monitored in order to be assessed. It appears that this correlation is weaker when it comes to MU. Therefore, the value of SP's cumulative average does not affect at all the distance to convergence required to be collected for an individual driver.

Table 5.5: Correlation matrix between HA, HB, MU and SP and required distance for each driver's behaviour to converge in highways

	Distance$_{highway}$
ha$_{highway}$	-0.223
hb$_{highway}$	-0.227
mobile$_{highway}$	-0.147
speeding$_{highway}$	-0.069

156

Table 5.6: Descriptive statistics of metric values and distance (*100km) per percentile range category in highways

Metric	Percentile range	Metric descriptive statistics				Distance to convergence		
		Average	St. Dev	Min	Max	Average	Median	St. Dev
HA	0% – 25%	0.74	0.29	-	1.1	6.78	5.85	4.4
	25% – 50%	1.26	0.12	1.1	1.54	6.39	6.05	2.34
	50% – 75%	1.87	0.24	1.54	2.3	6.07	6.11	2.42
	75% – 100%	3.77	2.12	2.3	-	5.93	5.29	2.89
HB	0% – 25%	0.36	0.12	-	0.56	7.05	5.92	4.13
	25% – 50%	0.83	0.1	0.56	0.97	7	5.62	3.76
	50% – 75%	1.15	0.13	0.97	1.38	5.72	6.06	1.74
	75% – 100%	2.05	0.64	1.38	-	5.42	4.72	2.51
MU	0% – 25%	35	20	-	66	7.22	5.92	4.68
	25% – 50%	101	18	66	135	6.23	4.98	2.85
	50% – 75%	174	26	135	223	5.64	5.40	1.96
	75% – 100%	455	206	223	-	5.33	4.45	3.49
SP	0% – 25%	193	124	-	346	6.1	5.50	2.85
	25% – 50%	505	91	346	641	6.49	5.92	3.92
	50% – 75%	807	100	641	950	6.66	6.01	2.95
	75% – 100%	1168	249	950	-	5.44	4.64	2.41

Table 5.6 illustrates the descriptive statistics of each metric and distance per percentile range category that resulted from the analysis performed regarding the required distance to driving behaviour convergence. The main picture seems to be similar to that acquired from the analysis on rural road type. There exists a weak negative trend for all metrics besides SP, but it remains unclear whether or not an extra value is provided to the results arising when performing the analysis separately for less and more risky drivers; this should be further investigated in future research. Again, the exception is SP that presents no apparent trend in required distance. All these observations were also highlighted in table 5.5 and figure 5.4, which lead to the conclusion that different sampling periods are likely to be required for drivers of different driving risk level. This is probably because behaviour of medium and less risky drivers fluctuates more before it converges and therefore they need to be monitored more so as their driving characteristics have converged to their average level. As mentioned above, this is apparently not the case in urban road since there is no obvious trend as the level of metrics increases. Again, this is attributed to the fact that drivers are more volatile on the specific road type and as a result, the amount of data required to be collected is not strongly related to the

aggressiveness of each driver. On the other hand, more risky drivers are likely to have a more steady, risky though, behaviour. In contrast to the results found in the other two road types, drivers of the 25-50% and especially the 0-25% percentile value range present a significantly higher standard deviation, which could probably be attributed to the fact that more randomness exists in this specific range.

As for the characteristics per driving risk level, more risky drivers perform 3.77 and 2.05 harsh acceleration and braking events per 100km whereas less risky drivers perform 0.74 and 0.36 harsh acceleration and braking events per 100km on average respectively. On the other hand, less risky drivers use their mobile phone approximately for 35 sec/ 100km and drive over the speed limits for 193 sec/ 100km. Finally, more risky drivers showed a speed limit violation of 1168 sec/ 100km on average and a 455 sec/ 100km mobile usage. Metrics of all percentile value ranges across all different metrics are found to be significantly lower than those found in the other two road types.

For the same reasons as in the other two road types, the maximum median value of distance to convergence is selected for the determination of the required sampling distance. According to table 5.6, 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.

The driving efficiency problem can be therefore 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.

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.

5.2) Trip efficiency analysis

5.2.1) Multiple input-output DEA

Overall, the three approaches (Standard, RBE and CH DEA) are tested for 7 different scenarios, i.e. for 100, 500, 1000, 2500, 5000, 7500 and 10088 DMUs, the results of which are presented in the following section.

The amount of computational memory required to perform the Convex Hull DEA (CH DEA) approach is notably high. Quickhull algorithm applied herein does not support medium-sized inputs in 9-D and higher, which is the limitation of the present study. This is the reason why the authors choose to test their models only for six inputs and three outputs in order to create a convex hull problem of 9-D which can be calculated as described in the previous section (every DEA input is divided by every DEA output). Three outputs and six inputs are examined, instead of two or four for instance, for the results to be easily explained from a transportation engineering perspective. The combinations of the number of harsh acceleration and braking events, seconds driving over the speed limit and seconds used the mobile phone per road type with distance per road type were used to create 6 different DEA models. Nonetheless, only harsh acceleration and braking per road type with distance per road type (DEA model type 1) is chosen to be presented herein to avoid chattering. All models provided similar results and therefore conclusions drawn can be generalized regardless of the variables chosen in the model. The specifications of the models implemented are shown in Table 5.7.

In every scenario tested, results showed that CH DEA method yielded exactly the same results as the other two approaches tested in terms of the most efficient DMUs identification, the lamdas and theta values estimation, the peers' determination and the efficient level of inputs and outputs calculation for each DMU. This is a weighty outcome because for the first time tests proved the efficacy of the proposed methodology for performing a multiple input and output CH DEA.

Table 5.7: Inputs and outputs of DEA models used in trip efficiency analysis

DEA Model type	Set of Inputs used	Set of Outputs used
1	1) number of harsh acceleration events in urban road 2) number of harsh acceleration events in rural road 3) number of harsh acceleration events in highway 4) number of harsh braking events in urban road 5) number of harsh braking events in rural road 6) number of harsh braking events in highway	1) total distance driven in urban road 2) total distance driven in rural road 3) total distance driven in highway
2	1) number of harsh acceleration events in urban road 2) number of harsh acceleration events in rural road 3) number of harsh acceleration events in highway 1) total seconds of mobile phone usage in urban road 2) total seconds of mobile phone usage in rural road 3) total seconds of mobile phone usage in highway	1) total distance driven in urban road 2) total distance driven in rural road 3) total distance driven in highway
3	1) number of harsh acceleration events in urban road 2) number of harsh acceleration events in rural road 3) number of harsh acceleration events in highway 1) total seconds of speed limit violation in urban road 2) total seconds of speed limit violation in rural road 3) total seconds of speed limit violation in highway	1) total distance driven in urban road 2) total distance driven in rural road 3) total distance driven in highway
4	1) number of harsh braking events in urban road 2) number of harsh braking events in rural road 3) number of harsh braking events in highway 1) total seconds of mobile phone usage in urban road 2) total seconds of mobile phone usage in rural road 3) total seconds of mobile phone usage in highway	1) total distance driven in urban road 2) total distance driven in rural road 3) total distance driven in highway
5	1) number of harsh braking events in urban road 2) number of harsh braking events in rural road 3) number of harsh braking events in highway 1) total seconds of speed limit violation in urban road 2) total seconds of speed limit violation in rural road 3) total seconds of speed limit violation in highway	1) total distance driven in urban road 2) total distance driven in rural road 3) total distance driven in highway
6	1) total seconds of mobile phone usage in urban road 2) total seconds of mobile phone usage in rural road 3) total seconds of mobile phone usage in highway 1) total seconds of speed limit violation in urban road 2) total seconds of speed limit violation in rural road 3) total seconds of speed limit violation in highway	1) total distance driven in urban road 2) total distance driven in rural road 3) total distance driven in highway

5.2.2) Computational time reduction

Results illustrated in table 5.8, indicate a superiority of the proposed method over the standard and RBE DEA approaches in terms of computation time. As anticipated, CH

160

DEA approach significantly outperformed the other two especially for samples of a higher scale. Results are presented not only as absolute values but also as percentages of improvement, in order for the results to be representative regardless of a computer's performance.

As anticipated, computation time appears to be approximately linearly increased in CH DEA method as the time required for each LP to be solved depends only on the number of the efficient DMUs found in the first step of the process. The number of used DMU in the LP in each iteration is kept constant (N_e plus the reference DMU in each iteration) and as a result, the total time is proportionally increased to the total number of DMUs. It should be noted at this point that the results arising show that the difference in computation time is increased as the number of DMUs is increased. Therefore, comparing scenarios that include a higher number of DMUs would not add more value in this research since the point is to a) investigate the results arising from the application of a multiple inputs and outputs CH DEA and b) apply it on transport data.

In the specific DEA problem presented in table 5.8, the density of the efficiency DMUs is found to be very low which reduces the computation time considerably since each of the 10073 LPs (a total of 10088 trips minus 15 efficient) that needs to be solved has only 16 (15 efficient plus 1 reference DMU in each LP) DMUs. RBE was also confirmed to perform faster than standard approach especially for larger datasets. Nonetheless, the percentage of running time improvement over the standard DEA approach is kept constant aside from the sample size. On the other hand, RBE is found to be significantly slower than CH DEA; ranging between 33.33% and 99.97% from 100 to 10088 DMUs respectively. It is evident that for small-scale samples of less than 500 DMUs the computational time gain is not worthwhile and probably standard approach should be preferred.

Finally, standard approach and RBE is proved to be a non-feasible option for analysing large-scale data using DEA which need several days (more than 40 and 12 days respectively) of processing on a conventional computer. This implies that alternative solutions such as the one examined in this paper should be further investigated and appraised especially when it comes for analysing large-scale data with DEA. The efficacy of the CH DEA algorithm investigated here, in terms of running time, provides encouraging insights for future enhancements on DEA addressing the issue of reducing its computation time. All solutions examined herein are exact solutions of the problem and it would be interesting to investigate on the potential of improving running time using heuristic and meta-heuristic algorithmic solutions, which provide an approximate solution to the problem. Of course, in this case it would also be important to investigate whether the approximate solution given is satisfactory.

Table 5.8: Computation time for seven scenarios

No of DMUs	Computation time (sec)			CH DEA % computation time improvement over		RBE DEA % computation time improvement over Standard DEA
	Standard DEA Approach	RBE DEA	Convex Hull DEA	Standard DEA Approach	RBE DEA	
100	11	6	4	63.64%	33.33%	45.45%
500	477	169	21	95.60%	87.57%	64.57%
1000	3250	1121	41	98.74%	96.34%	65.51%
2500	44435	15570	94	99.79%	99.40%	64.96%
5000	398485	123986	180	99.95%	99.85%	68.89%
7500	1400909	444498	231	99.98%	99.95%	68.27%
10088	3519372	1089731	314	99.99%	99.97%	69.04%

* Inputs = ['ha_urban', 'ha_rural', 'ha_highway'], Outputs = ['distance_urban', 'distance_rural', 'distance_highway']

Running time results are also illustrated in figure 5.5; convex hull results are plotted in the secondary axis because computation time showed that convex hull significantly outperforms the other two approaches tested and therefore demonstration would not be distinguishable.
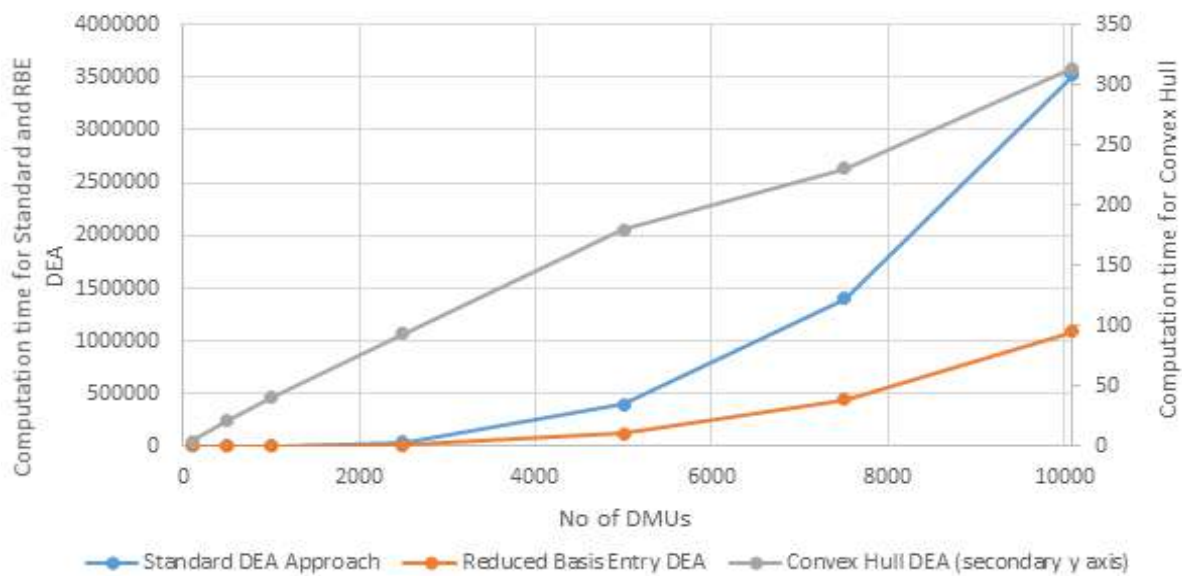
Figure 5.5: Computation time of the three methodologies implemented

5.2.3) Efficient level of DEA inputs and outputs

Table 5.9 shows lambdas and theta for ten (first nine non-efficient trips plus one efficient trip) trips of the DEA model type 1, where L_i stands for the lamda coefficient of the efficient $trip_i$ that acts as a peer for the trip examined each time. For the purpose of brevity, not all lambdas and thetas calculated are presented herein. For instance, in the first row of the table where DEA is solved for $trip_1$, the value of the theta coefficient is 0.008332 (less efficient) and the lamda coefficients L_{745}, L_{4403}, L_{5293} are equal to 0.014510, 0.008332, 2.222700 respectively. The efficient level of inputs for $trip_1$ can be calculated as the product sum of the lamdas and the input values of each of the identified peers whereas to find the efficient level of outputs for the same trip, each output value should be divided by theta value. Again, taking $trip_1$ as example, the efficient level of ha_{urban} can be estimated using formula (4) presented in the methodological approach:

$$Efficient \text{ level of } ha_{urban_1} = \lambda_{745} * ha_{urban_{745}} + \lambda_{4403} * ha_{urban_{4403}} + \lambda_{5293} * ha_{urban_{5293}} \Rightarrow$$

$$ha_{urban_1} = 0.015 * 27 + 0.008 * 0.26 + 2.223 * 0.32 = 1.11$$

On the other hand, the efficient level of e.g. $dis\tan ce_{urban}$ is calculated from formula (5) presented in the methodological approach:

$$Efficient \text{ level of } dis\tan ce_{urban_1} = dis\tan ce_{urban_1} / theta_1 \Rightarrow$$

$$dis\tan ce_{urban_1} = 587 / 0.008 = 70444$$

Table 5.9: Lamdas, thetas, real and efficient level of metrics (distance (km) and ha per road type) for the first 9 non-efficient trips (DMUs) and one efficient trip

Trip No	Real level of metrics						Theta	Lamdas of peers: Trip No				Efficient level of metrics					
	$distance_{urban}$	$distance_{rural}$	$distance_{highway}$	ha_{urban}	ha_{rural}	ha_{higway}		745	4403	5293	9493	$distance_{urban}$	$distance_{rural}$	$distance_{highway}$	ha_{urban}	ha_{rural}	ha_{higway}
1	587	1105	612	133	77	2	0.008332	0.014510	0.008332	2.222700	-	70444	132617	73458	1.11	0.65	0.02
2	428	751	439	128	18	9	0.021543	0.085333	0.096945	1.334662	-	19852	34840	20362	2.76	0.39	0.19
3	142	266	147	69	64	7	0.002526	-	0.008839	0.536261	0.000183	56400	105471	58273	0.17	0.16	0.02
4	567	1059	587	117	78	8	0.007910	0.008943	0.031639	2.120794	-	71624	133841	74185	0.93	0.62	0.06
5	489	917	503	28	65	1	0.021047	-	0.010524	1.841262	0.025249	23224	43566	23902	0.59	1.38	0.02
6	723	1353	745	55	83	4	0.015964	-	0.031928	2.721888	0.016139	45261	84767	46694	0.88	1.33	0.06
7	488	896	508	155	38	6	0.013056	0.054085	0.039169	1.716472	-	37396	68613	38898	2.02	0.50	0.08
8	375	693	389	156	55	8	0.007112	0.024599	0.028447	1.355048	-	52732	97453	54682	1.11	0.39	0.06
9	797	1489	831	168	70	4	0.012218	0.040842	0.024436	2.955081	-	65212	121905	67989	2.06	0.86	0.05
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
745	524	736	632	27	0	0	1.000000	1.000000	-	-	-	524	736	632	27	0	0
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.

*Inputs = ['ha$_{urban}$', 'ha$_{rural}$', 'ha$_{highway}$', 'hb$_{urban}$', 'hb$_{rural}$', 'hb$_{highway}$'], Outputs = ['distance$_{urban}$', 'distance$_{rural}$', 'distance$_{highway}$']

5.3) Driver efficiency analysis

This subchapter describes the analysis performed on a driver level including the temporal evolution of driving efficiency.

As mentioned earlier in the methodological approach chapter, 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.

5.3.1) Models' specification and sample used

As explained above, DEA models representing driving safety efficiency in urban and rural types are developed with multiple inputs and one output. The critical process required for input and output selection as well as the pitfalls that might arise from it are discussed in the methodological approach chapter. The form of the inputs and outputs used in the analysis is described in Chapter 3 and the variables along with their description are given in table 4.2. As shown, metrics used as inputs are the number of harsh braking and accelerations events, seconds driving over the speed limit and seconds used the mobile phone, while metric used as output is distance travelled.

Each driver is deemed a DMU with an aggregate performance for the entire monitoring period. His driving behaviour is considered equivalent to the sum of the driving characteristics that were recorded for the period examined. For instance, the total distance travelled in urban network is equivalent to the sum of the distance travelled in urban 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 (20):

$$indicator_i = \sum_{j=1}^{N_i} indicator_{ij} \qquad (20)$$

recorded $\forall trip_j, j \in (1, N_i)$ that took place by $driver_i$. As described above, each driver is treated as a distinct DMU to be analysed in DEA and therefore the linear program constructed (see (1)) has a number of variables (λ_i, θ_B) that is equal to the number of drivers plus the efficiency for $driver_0$. The number of constraints on the other hand is equal to the sum of a) the number of inputs ($\theta_B * \chi_0 - X * \lambda \geq 0$), b) the number of outputs ($Y * \lambda \geq y_0$) and c) the number of drivers ($\lambda_i \geq 0$). The DEA procedure described by (1) is followed separately for each of the two different road types (urban, rural) as described in table 3.3. As mentioned above, no model was developed for highway road type because there were not enough users to be analysed.

Table 3.3: Inputs and Outputs of the DEA models used in driver efficiency analysis

DEA models	Urban	Rural
Set of Inputs used	1) ha_{urban} 2) hb_{urban} 3) $speeding_{urban}$ 4) $mobile_{urban}$	1) ha_{rural} 2) hb_{rural} 3) $speeding_{rural}$ 4) $mobile_{rural}$
Set of Outputs used	1) $distance_{urban}$	1) $distance_{rural}$

As shown in chapter 3, table 3.4 summarizes the sample used in this specific analysis for urban and rural road type. It is not feasible to perform the analysis for highways since there are only 18 drivers of the data_sample_1 that have the required total distance and number of trips; for only 7 out of which, questionnaire data are available. Even if the analysis was performed with the specific drivers' sample, the time series would not be long enough to ensure the significance of the results. The two last columns of table 3.4 represent a) the number of participants that have at least as many trips required in "No of trips" column and b) the number of participants that have at least as many trips required in "No of trips" column and have also responded to the questionnaire administered.

Table 3.4: Number of drivers participated in the analysis of the temporal evolution of driving efficiency in each road type

Road type	No of trips	Required moving window (trips)	No of participants of the data_sample_1	No of participants of the data_sample_2
Urban	230	75	100	43
Rural	150	82	100	39
Highway	150	116	18	7

5.3.2) DEA model illustration

It is noted that only models incorporating two-inputs/ one output or one-input/ two outputs (the number of dimensions should be equal to 3) can be visualized in 2D figures and therefore, models of table 3.3 cannot be illustrated. Nonetheless, in order to acquire a better picture of the DEA outcomes, two models in urban and rural type are developed that account for drivers' aggressiveness (the number of harsh acceleration and braking events occurred are considered) and their results are presented in figure 5.6. It is evident that there are only two efficient DMUs for urban and rural road, which confirms the results of the DEA LPs. In each subplot of figure 5.6, $dis\tan ce_x / ha_x$ and $dis\tan ce_x / hb_x$ is plotted in axis Y and X respectively along with the envelopment line accounting for the efficiency frontier. Extending the line joining the origin and $driver_i$, it crosses the efficiency frontier at a point where virtual $driver_i'$ is created which represents the optimal performance which the specific $driver_i$ can achieve. The two points that comprise the start and end of the line that the virtual $driver_i'$ crosses are called the peers of the $driver_i$ examined.
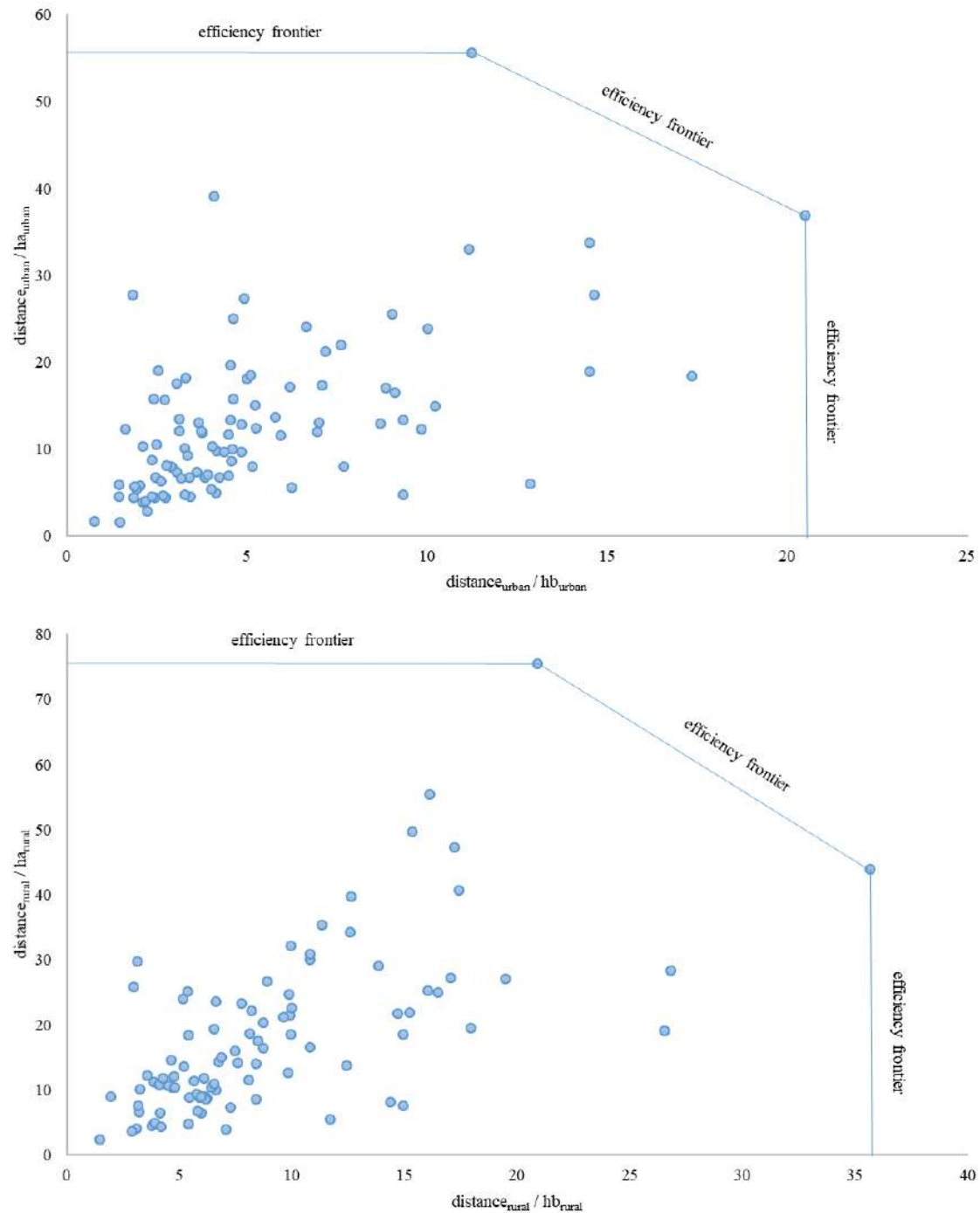
Figure 5.6: Efficiency frontier of drivers' aggressiveness per road type

The closer a $driver_i$ is to the efficiency frontier, the higher his/her efficiency index is. The influence of outliers to the DEA solution is obvious since most drivers appear to be near the origin. Nonetheless, the solution remains reliable, as the efficiency index calculated is comparable to that of the rest of the drivers set. The scope of this paragraph is to explain the DEA results visualized in figure 5.6 and not to present the models' outcome. This is the reason why further details of the models' results illustrated in the figure are not given.

5.3.3) Driving efficiency classification

The results of DEA are the efficiency index θ_i and coefficients λ_i for each driver. This allows for the classification of the whole set of drivers to most efficient, weakly efficient and non-efficient. Since the absolute value of the efficiency index cannot be somehow interpreted unless it is compared to the efficiency index of the rest of the drivers set, the percentiles of the drivers set's θ_i are used to classify drivers. The percentile thresholds specified was 25% and 75%, which separate the subsets of non-efficient and weakly efficient as well as weakly efficient and most efficient DMUs respectively. For each of the data_sample_1 and the 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. From here on, for brevity purposes class 1 drivers will be 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 will be referred to as weakly efficient and non-efficient drivers.

For the better understanding, the 5th column of table 5.10 shows the driving attributes of the 2nd class of the data_sample_1 of 100 drivers who have driven in urban road type and whose efficiency's percentiles are between 25-75%. Results are presented as the driving efficiency, the number of harsh acceleration and braking events occurred, the number of seconds for driving over the speed limits and the number of seconds using the mobile phone per 100 kilometres driven.

Main characteristics of drivers efficiency classes

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.

Another observation for the data_sample_1 is that the number of harsh events occurring in rural road is higher than in urban and that the number of harsh acceleration events occurred in all road types is at least twice as much as the number of harsh braking events. For instance, in urban roads, the number of harsh acceleration events ranges from 8.82 to 21.49 per 100km while the number of harsh braking events from 3.68 to 9.64 for most efficient to non-efficient drivers. This difference becomes larger in rural road. The same

difference between urban and rural road types is noticed for mobile usage and speeding as well except from the most efficient drivers who tend to use the mobile phone in urban as much as in rural road. A similar observation is made for the data_sample_2 as well.

In general, it is concluded that mobile usage in urban road is limited to approximately 5, 3.5 and 2.5 minutes per 100km of driving for non-efficient, weakly efficient and most efficient drivers respectively whereas the respective amount of speed limit violation is 20.5, 14.5 and 6. As for the rural road type, mobile usage road is approximately 9.5, 7 and 3 minutes per 100km of driving for non-efficient, weakly efficient and most efficient drivers respectively whereas the respective amount of speed limit violation is 26, 16.5 and 12.

Observing the shift between efficiency classes in urban road, the move from class 1 to 2 is mainly affected by the number of harsh acceleration and braking events and hardly by mobile phone usage while the time driving over the speed limits plays a more important role when moving from class 2 to 3. As for rural road, speeding mainly affects the shift from class 1 to 2 whereas the number of harsh acceleration events and mobile phone usage mostly influences the move from class 2 to 3.

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

The noticeable difference in the data_sample_2 is that $mobile_{urban}$ in class 1 and 2 of drivers is higher than $mobile_{rural}$ and that $speeding_{urban}$ is higher than $speeding_{rural}$ in class 1 and 3 of drivers and slightly higher in class 2 of drivers. This difference might be due to

the lower number of participants in the data_sample_2 and therefore a larger sample should be exploited to further investigate it.

As for the shift between efficiency classes in urban road, moving from class 1 to 2 is mainly affected by the number of harsh acceleration and braking events and hardly by the time driving over the speed limits while mobile phone usage plays a more important role when moving from class 2 to 3. The number of harsh acceleration and braking events are also the most influencing factors for shifting between classes 1 and 2 in rural road, whereas the move from class 2 to 3 is mainly affected by mobile phone usage; speeding factor seems to equally affect both shifts from class 1 to 2 and from class 2 to 3. It is a matter of great importance to identify these parameters in order to provide significant targeted macroscopic recommendations for further improvement to a certain group of drivers whose safety efficiency is known.

Regarding the median driving efficiency of each efficiency class, the efficiency of class 3 is equal to 1.00 in both road types examined and therefore it is evident that there is a high density of efficient drivers in the data_sample_2, which is probably due to the lower number of drivers. It is mentioned earlier in the literature review that when the ratio of the sum of inputs and outputs to the total number of DMUs is decreased to 1, the number of the efficient DMUs is increased. In the case examined herein, this may lead to the conclusion that a higher number of efficient drivers, than it actually exists, is found to exist in the data_sample_2 and therefore results arising might be slightly biased. This should be further investigated in the future in order to shed more light on it.

Figure 5.7 is a histogram that shows the number of drivers of the data_sample_1 that fall into each efficiency range in urban and rural road types. It is conspicuous that the driving efficiency of most drivers is concentrated in the 0.2-0.6 efficiency range in both road types and that only a small percentage of drivers fall into the upper and lower ranges. This is probably because most efficient drivers have significantly higher efficiency and consequently the rest of the drivers' efficiency is relatively estimated to be much lower. On the other hand, since driving efficiency is relatively estimated, it can be concluded that drivers of the 0.4-0.6 efficiency range are likely to be drivers of average risk. With regards to the 0.2-0.4 efficiency range, this group should be further investigated to distinguish between average and high risk drivers.
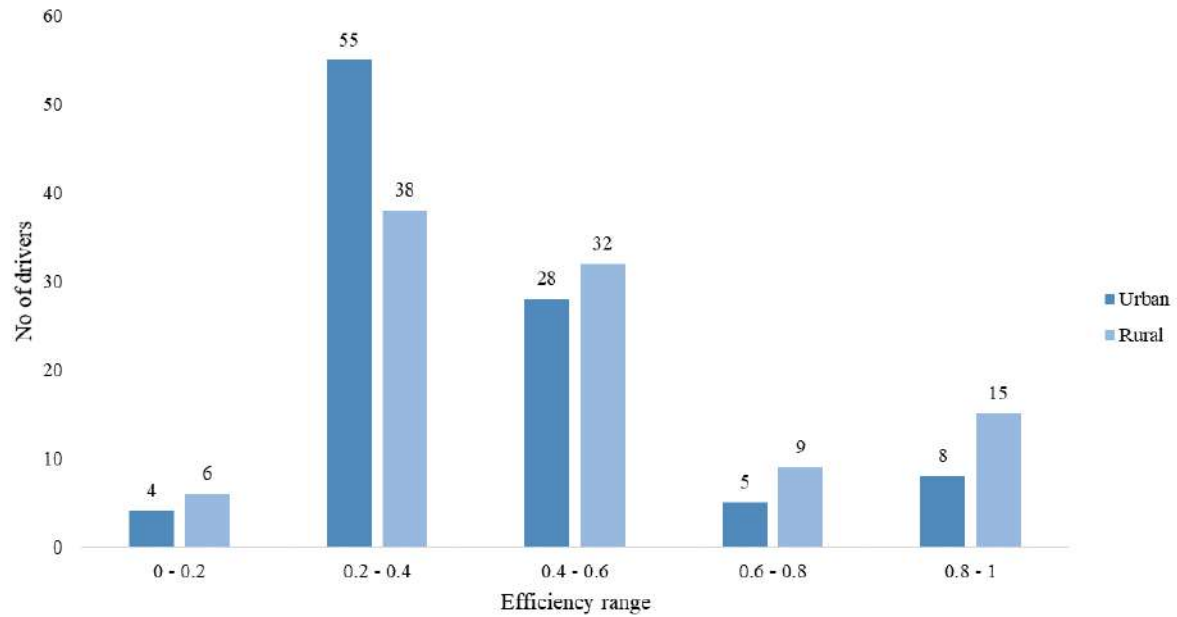
Figure 5.7: Number of drivers of data_sample_1 in each efficiency range for urban and rural road types

Figure 5.8 is a histogram that shows the number of drivers of data_sample_2 that fall into each efficiency range in urban and rural road types. Obviously, figures 5.7 and 5.8 have similar characteristics in terms of the drivers' distribution. Again, most drivers are concentrated in the 0.2-0.6 efficiency range in both road types and only a small percentage of drivers fall into the upper and lower ranges. The difference noticed in figure 5.8 is that there is a higher number of drivers in the 0.8-1 range probably because of the decreased ratio of the sum of inputs and outputs to the number of participants, a parameter that was mentioned earlier in details. In general, the examination of a different class discretization (e.g. class 1: 0-33.3%, class 2: 33.3-66.6%, class 3: 66.6-100%) is proposed for investigation in the future.
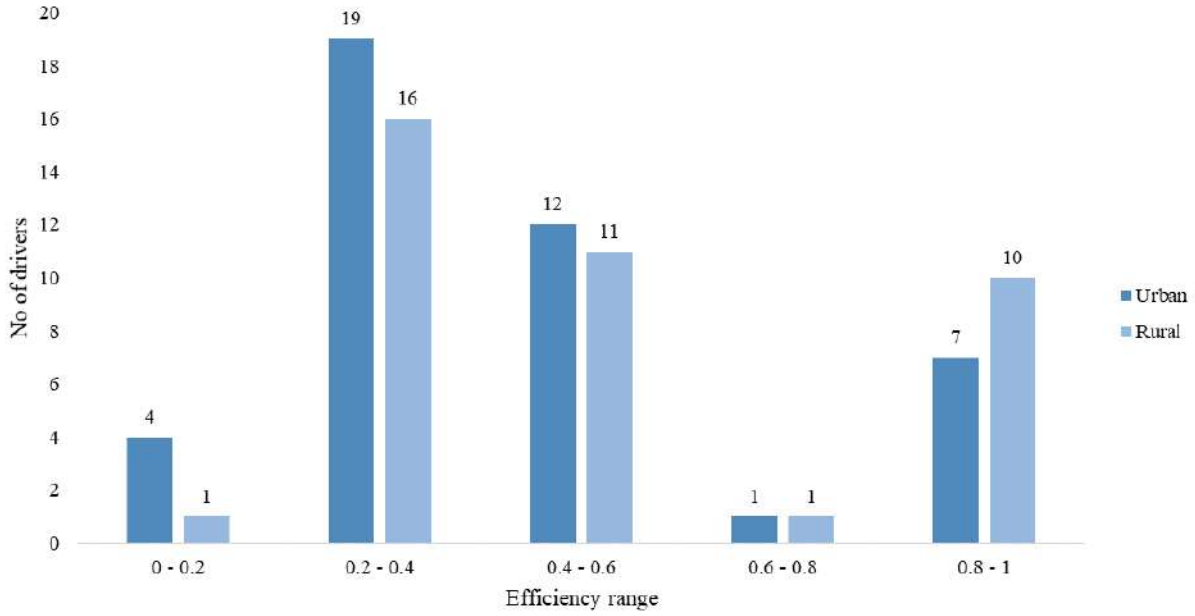
Figure 5.8: Number of drivers of data_sample_2 in each efficiency range for urban and rural road types

5.3.4) Efficient level of inputs and outputs

Table 5.11 shows lambdas and theta for the first twelve (eleven non-efficient drivers plus one efficient driver) drivers of the DEA models of table 3.1 in urban road, where L_i stands for the lamda coefficient of the efficient $driver_i$ that acts as a peer for the driver examined each time. For the purpose of brevity, not all lambdas and thetas calculated are presented herein. For instance, in the first row of the table where DEA is solved for $driver_1$, the value of the theta coefficient is 0.581 (less efficient) and the lamda coefficients L_{34}, L_{40}, L_{42} are equal to 0.52, 0.14, 0.06 respectively. The efficient level of inputs for $driver_1$ can be calculated as the product sum of the lamdas and the input values of each of the identified peers whereas to find the efficient level of outputs for the same driver, each output value should be divided by theta value. Again, taking $driver_1$ as example, the efficient level of ha_{urban} can be estimated using formula (4) presented in the methodological approach:

$$Efficient \text{ level of } ha_{urban_1} = L_{12} * ha_{urban_{12}} + L_{34} * ha_{urban_{34}} + L_{40} * ha_{urban_{40}} + L_{42} * ha_{urban_{42}} =>$$

$$ha_{urban_1} = 0 * 329 + 0.52 * 242 + 0.14 * 86 + 0.06 * 366 = 159.8$$

173

Unfortunately, the exact calculations of ha_{urban_1} are not provided since only the first 12 drivers are shown in the table and therefore the $ha_{urban_{34}}$, $ha_{urban_{40}}$ and $ha_{urban_{42}}$ are not shown. On the other hand, the efficient level of e.g. $dis\tan ce_{urban}$ is calculated from formula (5) presented in the methodological approach:

$$Efficient \text{ level of } dis\tan ce_{urban_1} = dis\tan ce_{urban_1} / theta_1 =>$$

$$dis\tan ce_{urban_1} = 1868 / 0.581 = 3214.3$$

As stated previously, a driver (DMU) is deemed to have achieved the efficient level when it reaches unit efficiency. It should be highlighted though, that a driver should reach either the efficient level of inputs or the efficient level of outputs in order to become efficient and not both at the same time. Of course, if a driver achieves the efficient level of both inputs and outputs it will become the most efficient driver and, therefore, it will define a new efficiency frontier and act as a peer for the rest of the drivers (given that no other driver will achieve the same). It is also obvious from the table that the most efficient drivers of the sample are drivers $driver_{12}$, $driver_{34}$, $driver_{40}$ and $driver_{42}$ who act as peers for the rest of the driving sample. As expected, most peers do not act as peers for all drivers but most drivers have a portion of the most efficient drivers as their peers. It is also expected that all drivers that have unit efficiency such as $driver_{12}$, have a real and efficient level of metrics that is equal for all metrics.

Based on the above, it can be concluded that the required change of each driving attribute that was taken into consideration in order for a driver to shift either to the efficient frontier or to another driving class can be estimated. This can be achieved by solving the optimization problem for a specific input or output given the target efficiency ($Driving \text{ } Efficiency_B$), which is the upper or the lower limit of the class that the driver is shifting in case of efficiency decrease or increase respectively.

Table 5.11: Lamdas, thetas, real and efficient level of metrics (distance (km), ha, hb, speeding (sec), mobile (sec)) in urban road for the first 12 drivers (DMUs)

Driver No	Real level of metrics					Theta	Lamdas of peers: Driver No				Efficient level of metrics				
	$distance_{urban}$	ha_{urban}	hb_{urban}	$speeding_{urban}$	$mobile_{urban}$		12	34	40	42	$distance_{urban}$	ha_{urban}	hb_{urban}	$speeding_{urban}$	$mobile_{urban}$
1	1868	326	134	21712	9954	0.581	-	0.52	0.14	0.06	3214.3	159.8	77.9	12617.9	5784.7
2	2456	574	85	27049	13974	0.696	-	0.19	0.40	0.20	3526.5	154.8	59.2	18838.2	9732.1
3	1634	709	509	15888	42817	0.391	0.79	-	0.20	-	4182.6	277.0	78.1	6206.9	10434.2
4	2219	233	181	27052	12421	0.637	-	0.23	0.31	0.18	3481.0	148.5	59.1	17244.6	7917.9
5	4223	1088	309	37825	29581	0.613	0.30	1.16	0.47	-	6887.3	417.6	189.5	23192.7	18137.9
6	2773	652	251	25829	25887	0.529	0.60	0.51	0.30	-	5245.3	344.7	128.8	13654.8	13685.4
7	2086	265	149	20880	14036	0.619	-	0.52	0.28	0.04	3371.9	163.9	79.2	12917.2	8683.2
8	1789	460	323	17185	5960	0.656	-	0.75	-	0.01	2728.0	184.3	98.1	11269.8	3908.5
9	1630	184	82	30801	8955	0.528	-	-	0.21	0.22	3085.0	97.2	30.8	14784.6	4731.5
10	808	266	95	9985	6562	0.443	0.11	0.24	0.04	-	1824.6	97.2	42.1	4421.6	2905.8
11	3012	913	152	21604	43562	0.585	0.72	-	0.83	-	5149.6	308.3	88.9	12636.1	23107.5
12	1462	329	92	5074	7781	1.000	1.00	-	-	-	1462.0	329.0	92.0	5074.0	7781.0

* Inputs = ['ha_{urban}', 'hb_{urban}', '$speeding_{urban}$', '$mobile_{urban}$'], Output = ['$distance_{urban}$']

5.3.5) Evolution of driving efficiency

After driving efficiency is estimated for the total recording period, the next step is to estimate the characteristics of driving efficiency evolution for each driver. The temporal evolution of average driving efficiency is investigated using different databases of metrics accumulated over different timeframes. Based on the above mentioned DEA input/ output combinations, exactly the same models (models of table 3.3), as in the previous step of the analysis, are developed in each moving window. As mentioned above, for each model the cumulative metrics monitored during the period examined are used as inputs and outputs in the DEA models developed. As a result, a different database is created in each step of the moving window and a new DEA model is developed respectively to estimate driving efficiency in the specific step and consequently the temporal evolution of total driving efficiency.

Time series created from this process are decomposed to acquire volatility, trend and stationarity. The ADF and KPSS tests are performed for unit root and stationarity respectively. When the null hypothesis of both the ADF and KPSS test is rejected for a time series, then it is considered fractionally integrated and presents long memory. In this case, an ARFIMA model is applied to estimate the order d of the time series, which takes values in the $[-1,1]$ region, close to 0 for stationary time-series and close to 1 for unit-root time series. Finally, time series trend is acquired by estimating the coefficient b of the linear regression model that best fits the time series data (slope).

The sample analyses conducted, determined a moving window of 75 and 81 trips long in urban and rural respectively in which driving performance is estimated to create the necessary time series. Table 3.4 summarized the sample used in this specific analysis for each road type. For the reasons explained above, analysis is not performed for highways. Finally, it is highlighted that the analyses are performed separately for the samples with and without available questionnaire data to compare the clusters arising and their characteristics. This procedure will evaluate the potential of driving safety efficiency benchmarking without having any knowledge on the personal information (age, gender, accident record etc.) of the user that is being benchmarked.

Indicatively, figure 5.9, 5.10, 5.11 and 5.12 illustrates seven time series of the urban data_sample_1, the rural data_sample_1, the urban data_sample_2 and the rural data_sample_2 respectively. It is obvious that several different driving patterns exist since the fluctuation of the time series differs between one another. It can be inferred that the drivers who are least efficient in total, also appear to be 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. 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.

It is also evident that there are some common local minimum and maximum for most of them which is attributed to existing efficiency outlier 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 the following graphs.
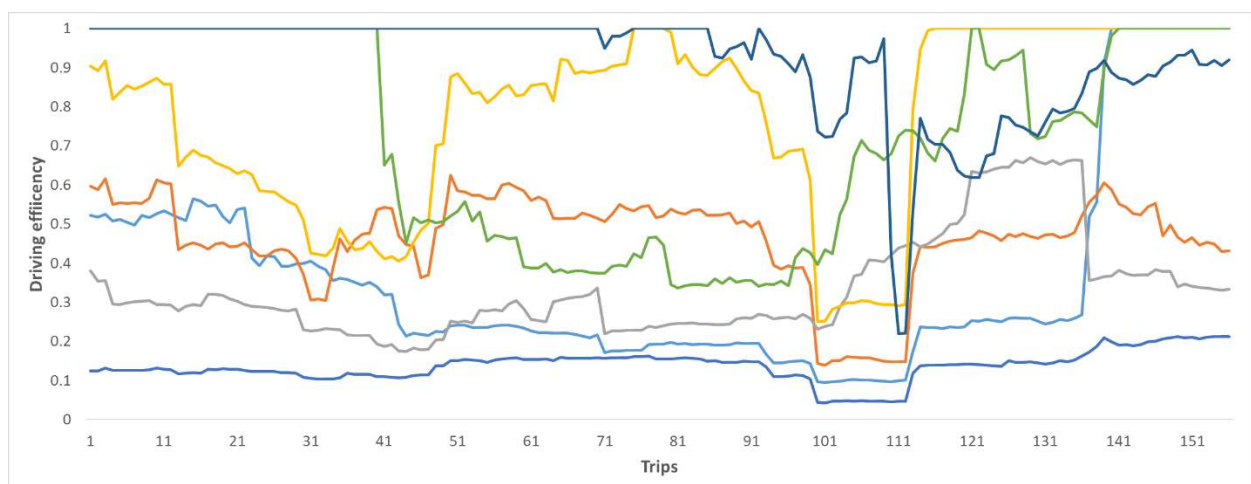
Figure 5.9: Efficiency time series of the urban data_sample_1
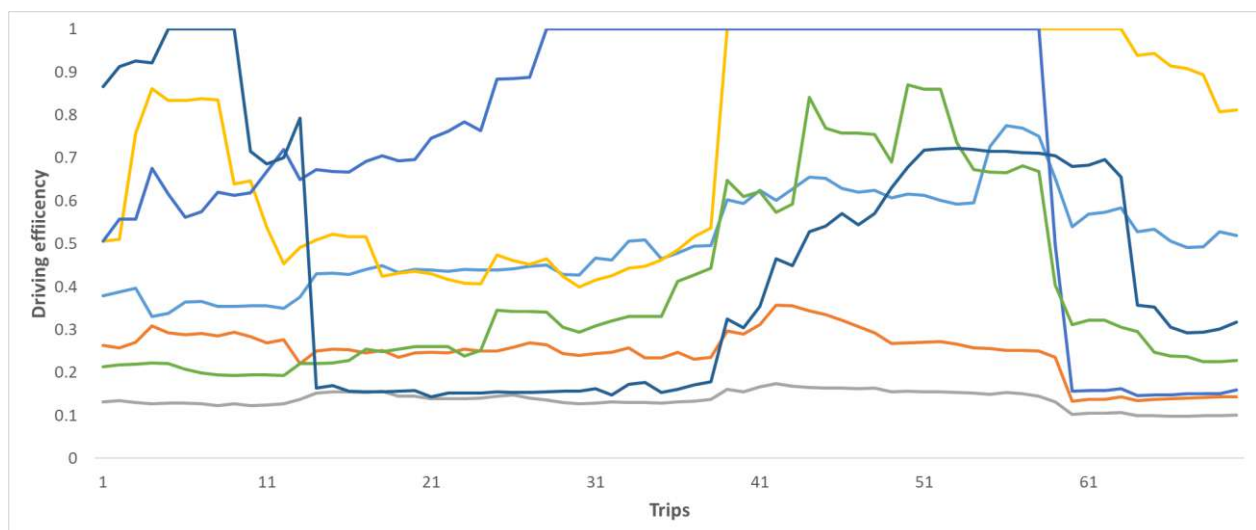
Figure 5.10: Efficiency time series of the rural data_sample_1
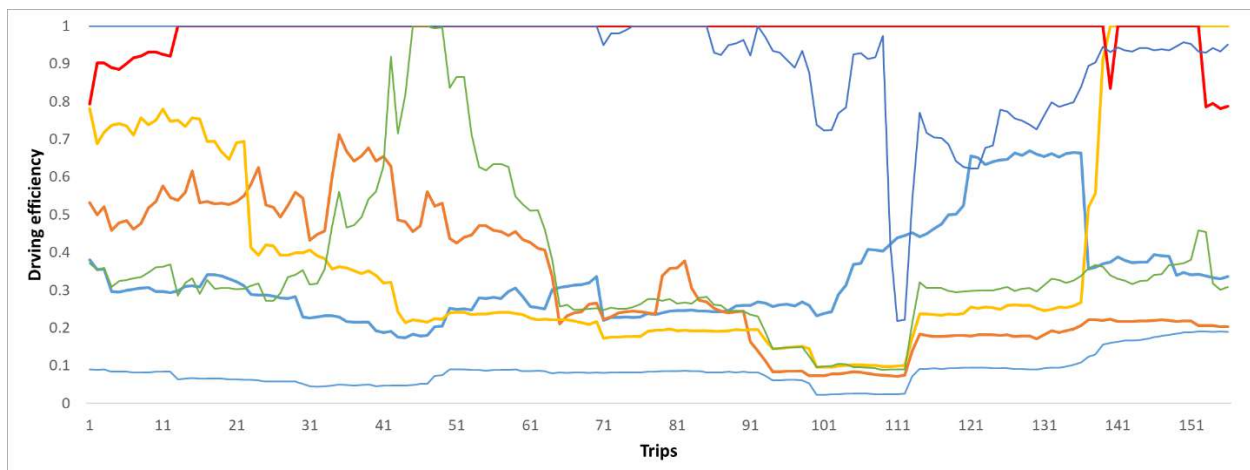
Figure 5.11: Efficiency time series of the urban data_sample_2
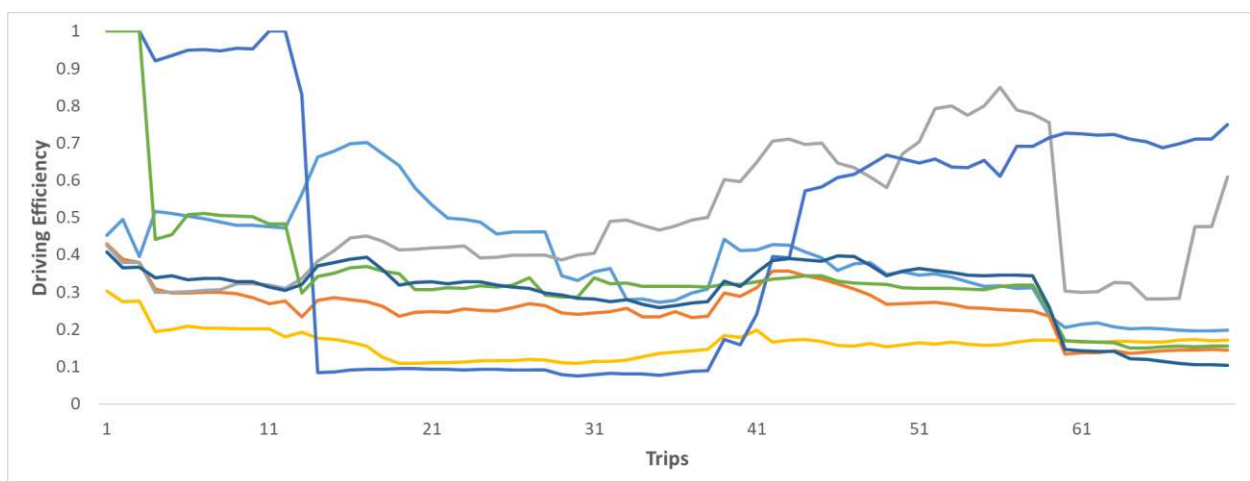
Figure 5.12: Efficiency time series of the rural data_sample_2

Volatility

The methodology how volatility of driving efficiency is estimated is provided in the relative chapter. The statistical characteristics of this measure are illustrated in table 5.12. It appears 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 except from the data_sample_2 of rural road that is slightly higher than the rest. Based on driving volatility's definition, it is inferred that when it is equal to 0, a driver demonstrates a solid performance throughout his/her monitoring. Nonetheless, since estimated efficiencies are rounded in the 10th decimal, this may happen only in the case of unit efficiency. As a result, drivers with steady unit efficiency exist only in rural road since the minimum value of volatility in urban road is higher than 0. On the other hand, rural road sample includes users with a more alterable behaviour, which is evident from maximum volatility that is twice as much as in urban road.

Table 5.12: Descriptive statistics of the driving efficiency volatility of the drivers' sample

Sample type	data_sample_1		data_sample_2	
Road type	Urban	Rural	Urban	Rural
Min	0.022	0.000	0.012	0.000
Max	0.152	0.379	0.144	0.384
Average	0.119	0.111	0.116	0.148
Standard Deviation	0.021	0.055	0.026	0.059
Median	0.123	0.095	0.122	0.144
Kurtosis	7.245	6.393	6.724	7.282
Skewness	-2.388	2.102	-2.402	1.763

Since the kurtosis value is approximately the same for all road and sample types, the distribution's tail is also similar. The fact that it is positive indicates a "heavy-tailed" distribution for all types with outliers. The negative skewness value of the urban sample of both sample types indicates that the distribution's left tail is longer, compared to the right, whereas the positive skewness value of the rural sample testifies the opposite.

Trend

The methodology how trend of driving efficiency is estimated is provided in the relative chapter. The statistical characteristics of this measure are illustrated in table 5.13. 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.

Table 5.13: Descriptive statistics of the driving efficiency trend ($*10^{-3}$) of the drivers' sample

Sample type	data_sample_1		data_sample_2	
Road type	Urban	Rural	Urban	Rural
Min	-4.56	-8.79	-3.23	-5.84
Max	4.09	8.46	4.41	6.80
Average	0.68	0.66	0.74	0.43
Standard Deviation	1.25	2.69	1.42	2.43
Median	0.51	0.80	0.49	0.33
Kurtosis	3.820	3.696	1.629	1.610
Skewness	-0.222	-0.550	0.370	0.02

This is also testified by the high kurtosis value that exist in both road types of the same sample. The high positive kurtosis also indicates a "heavy-tailed" distribution. The negative skewness value of both road types of the data_sample_1 indicates that the distribution's left tail is longer, compared to the right, whereas the positive skewness value of the data_sample_2 testifies the opposite. The trend value range and the

standard deviation of rural road seems to be twice as much as those of the urban road. It is noted that the measure illustrated in table 5.13 is efficiency trend $*10^{-3}$.

Stationarity

The methodology how trend of driving efficiency is estimated is provided in the relative chapter. The statistical characteristics of this measure are illustrated in table 5.14. Observing the number of differences that is required for a time series to become stationary, it is evident there are no urban road users whose driving behaviour is stationary. On the other hand, the relative number in rural roads is low for data_sample_1 but significantly higher for data_sample_2.

Table 5.14: Number of differences required for the driving efficiency time series of the drivers' sample to become stationary

Sample type		data_sample_1		data_sample_2	
Road type		Urban	Rural	Urban	Rural
No of differences	0	0	5	0	15
	1	97	93	41	22
	>1	2	1	2	1
No of fractionally integrated time series		1	1	0	1

For all road and sample types, the number of fractionally integrated time series of driving efficiency is negligible, indicating thus a clear picture concerning time series stationarity. As expected, the number of required differences is higher than 1 only for an insignificantly low number of users.

5.3.6) Drivers clustering

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. Figure 5.13 is representative of the elbow method figures created for each combination of road and sample type in order to find the number of clusters. In most cases, the elbow of the graph appears to exist at k = 3 or 4 indicating that the optimal number of clusters should be chosen between these two values. After several clustering tests performed using both numbers as k, it was found that in average, when k is set to 4, some clusters formed include a significantly the number of users (e.g. 2) and therefore there is a high probability that the results obtained would be biased. As a result, the number k of clusters is set to 3, which is rational considering the sample size used in this study.
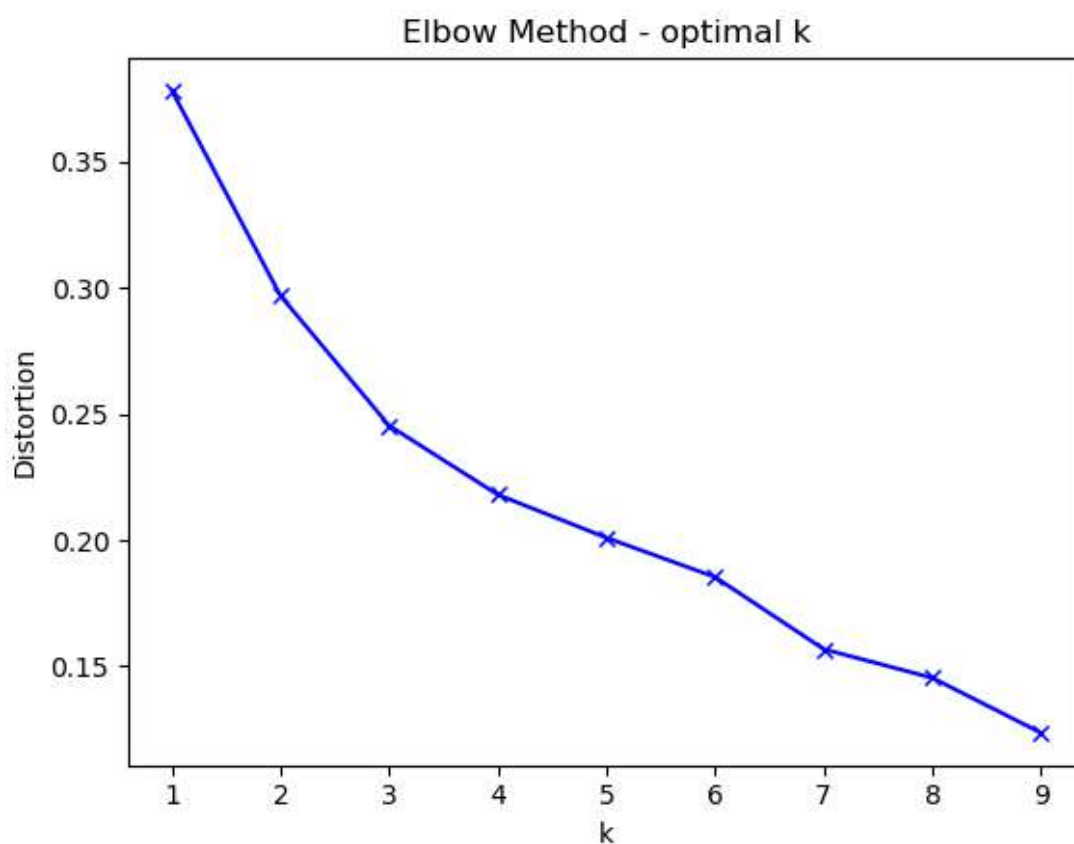
Figure 5.13: Elbow method that determines the optimal number of clusters

It is noted that after the analysis of the time series conducted in the previous step, stationarity is not selected to be included in the final clustering procedure since three out of four driver groups have similar characteristics (number of differences required to become stationary is 1) and therefore it is likely not to play an important role. Nonetheless, clustering was also performed including stationarity and results obtained showed that there was no influence of stationarity on the clusters arising.

Main results of drivers' clusters

Apart from the main attributes based on which clustering is performed, tables 5.15, 5.16, 5.17 and 5.18 present the number of drivers included in each cluster as well as the number of the drivers of data_sample_2 included in data_sample_1. This is necessary since accident data were available only for a portion of the drivers included in the clustering approach and therefore results concerning this type of data are indicatively presented to examine the average driving risk of the cluster resulting. This column is omitted from the data_sample_2 results since corresponding data are available for all drivers of the sample.

Table 5.15: Macroscopic characteristics of the urban data_sample_1

		Trend (*10⁻³)	Volatility	Rating	Accidents/ 10 years of driving experience	Number of drivers	Number of drivers of the data_sample_2
Cluster 1 (typical)	**Min**	-1.045	0.066	0.122	0.000	79	32
	Max	1.686	0.152	0.725	0.300		
	Average	0.516	0.123	0.340	0.109		
	Standard Deviation	0.534	0.013	0.108	0.090		
	Median	0.486	0.124	0.328	0.096		
	Kurtosis	0.303	4.969	0.944	-0.693		
	Skewness	-0.123	-1.438	0.713	0.525		
Cluster 2 (unstable)	**Min**	2.032	0.066	0.448	0.000	13	6
	Max	4.085	0.141	1.000	0.250		
	Average	3.006	0.119	0.673	0.090		
	Standard Deviation	0.628	0.022	0.206	0.078		
	Median	3.067	0.125	0.608	0.077		
	Kurtosis	0.334	-1.815	-2.281	-2.548		
	Skewness	0.209	-1.278	0.732	1.571		
Cluster 3 (cautious)	**Min**	-4.557	0.022	0.367	0.000	8	5
	Max	0.322	0.122	1.000	0.125		
	Average	-1.512	0.080	0.746	0.059		
	Standard Deviation	1.530	0.038	0.263	0.051		
	Median	-0.937	0.090	0.813	0.080		
	Kurtosis	-1.027	0.925	-1.154	3.344		
	Skewness	-1.053	-0.385	-0.237	-0.204		

Table 5.15 reveals the macroscopic characteristics of the urban data_sample_1 clusters that resulted from the model developed. 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. Finally, a low to medium accident frequency is observed in the partial data obtained for some of the cluster's drivers. 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 typical driver. In other words, the behaviour of the users is the typical/ expected and is represented by the average or median values of the driving characteristics.

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 rating which along with the low accident frequency and all the aforementioned demonstrate that this cluster is comprised from 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 behavioural volatility. They also feature a medium to very high average driving efficiency confirmed by the low accident frequency. Consequently, this cluster includes the most safety efficient 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.

Table 5.16 reveals the macroscopic characteristics of the urban data_sample_2 clusters that resulted from the model developed. As in the data_sample_1, cluster 1 also 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 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. Finally, a low accident frequency is observed in the partial data obtained for some of the cluster's drivers. All the above along with the higher number of drivers included in the specific cluster, lead to the conclusion that this cluster is similar to cluster 1 of the data_sample_1 and again, it mainly represent the typical driver.

Drivers of cluster 2 feature a low - medium positive trend indicating an improvement trend in general. This is also confirmed by the medium volatility presented, which testifies the existing positive instability. This cluster's drivers also have a low average rating and a significantly high accident frequency. All the aforementioned demonstrate that this cluster is comprised from drivers with the most risky behaviour of the sample that have a low trend of improvement that can partially be attributed to the fact that there is no room for further deterioration of their behaviour.

Table 5.16: Macroscopic characteristics of the urban data_sample_2

		Trend (*10^{-3})	Volatility	Rating	Accidents/ 10 years of driving experience	Number of drivers
Cluster 1 (typical)	**Min**	-3.230	0.104	0.122	0.000	25
	Max	1.396	0.144	1.000	0.150	
	Average	0.093	0.124	0.368	0.067	
	Standard Deviation	0.998	0.010	0.171	0.052	
	Median	0.382	0.123	0.366	0.069	
	Kurtosis	3.916	-0.694	6.216	-1.360	
	Skewness	-1.780	0.028	1.875	0.045	
Cluster 2 (unstable)	**Min**	0.136	0.081	0.159	0.167	10
	Max	3.168	0.141	0.501	0.300	
	Average	1.151	0.118	0.324	0.230	
	Standard Deviation	0.833	0.017	0.111	0.045	
	Median	1.091	0.122	0.326	0.230	
	Kurtosis	-2.013	-1.146	1.811	-2.224	
	Skewness	1.310	-0.858	0.122	0.008	
Cluster 3 (cautious)	**Min**	-0.219	0.012	0.466	0.000	8
	Max	4.413	0.136	1.000	0.093	
	Average	2.225	0.087	0.877	0.046	
	Standard Deviation	1.772	0.044	0.186	0.038	
	Median	2.941	0.098	1.000	0.057	
	Kurtosis	2.419	0.851	-1.040	-1.365	
	Skewness	-0.398	-0.600	-1.586	-0.240	

As for drivers of cluster 3, they present a medium negative trend and a low behavioural volatility. They also feature a significantly high average driving efficiency combined by a low accident frequency. As also shown in the data_sample_1, this cluster is likely to represent the most safety efficient drivers of the sample.

Table 5.17 shows the macroscopic characteristics of the rural data_sample_1 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 behavioural volatility also appears and as a result, an instability exists in behaviour. Drivers of this cluster also feature an average low total efficiency value, which shows a poor average behaviour. Finally, a low to medium accident frequency is demonstrated in the partial data obtained for some of the cluster's drivers. 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 typical driver. In general, this

cluster is very much alike cluster 1 obtained from the cluster analysis performed in urban roads.

Drivers of cluster 2 present a high negative trend and a high behavioural volatility. Despite the fact that they also feature a medium to high average driving efficiency, a medium to high accident frequency is observed. Consequently, this cluster includes drivers with a medium to high efficiency, present a significant deterioration of their behaviour while being monitored and show a medium to high accidents per year of experience value. This is the most important difference observed between the clustering results obtained from the analyses of the data_sample_1 of the two road types. The latter could probably be attributed to the fact that accident data are available only for 3 out of 12 drivers included in the specific cluster.

Table 5 17: Macroscopic characteristics of the rural data_sample_1

		Trend (*10^{-3})	Volatility	Rating	Accidents/ 10 years of driving experience	Number of drivers	Number of drivers of the data_sample_2
Cluster 1 (typical)	**Min**	-1.987	0.048	0.127	0.000	72	27
	Max	3.375	0.228	0.664	0.300		
	Average	0.764	0.099	0.363	0.117		
	Standard Deviation	1.040	0.035	0.120	0.094		
	Median	0.778	0.091	0.356	0.105		
	Kurtosis	-0.639	2.144	-1.806	-		
	Skewness	-0.252	1.437	0.410	0.455		
Cluster 2 (unstable)	**Min**	-8.785	0.072	0.323	0.050	12	3
	Max	-1.545	0.379	1.000	0.214		
	Average	-4.288	0.155	0.716	0.147		
	Standard Deviation	2.530	0.088	0.246	0.070		
	Median	-3.811	0.125	0.685	0.176		
	Kurtosis	0.412	2.323	-0.250	-0.644		
	Skewness	-0.824	1.490	-0.042	-1.363		
Cluster 3 (cautious)	**Min**	0.000	0.000	0.483	0.000	16	9
	Max	8.455	0.306	1.000	0.143		
	Average	3.904	0.133	0.847	0.064		
	Standard Deviation	2.573	0.072	0.160	0.048		
	Median	4.295	0.115	0.880	0.056		
	Kurtosis	-0.712	1.167	-0.268	-0.962		
	Skewness	0.398	0.789	-0.802	0.256		

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 presented. This cluster's drivers also have a high average rating which along with the low accident frequency and all the aforementioned demonstrate that this cluster is comprised from drivers with less risky behaviour and a constant trend of improvement. This cluster is also very similar to the third cluster that results from the analysis of the data_sample_1 in urban roads.

Table 5.18: Macroscopic characteristics of the rural data_sample_2

		Trend (*10⁻³)	Volatility	Rating	Accidents/ 10 years of driving experience	Number of drivers
Cluster 1 (typical)	Min	-5.837	0.091	0.174	0.000	
	Max	2.298	0.193	0.598	0.143	
	Average	-0.282	0.147	0.382	0.060	
	Standard Deviation	1.911	0.027	0.118	0.049	19
	Median	0.334	0.147	0.374	0.065	
	Kurtosis	2.344	-0.213	-0.684	-1.331	
	Skewness	-1.469	-0.173	0.203	0.059	
Cluster 2 (unstable)	Min	-5.069	0.077	0.205	0.150	
	Max	1.820	0.384	1.000	0.300	
	Average	-0.583	0.148	0.465	0.223	
	Standard Deviation	1.750	0.078	0.212	0.048	11
	Median	-0.240	0.129	0.389	0.222	
	Kurtosis	3.168	8.497	2.510	-1.366	
	Skewness	-1.449	2.757	1.401	0.057	
Cluster 3 (cautious)	Min	0.000	0.000	0.826	0.000	
	Max	6.804	0.308	1.000	0.143	
	Average	3.155	0.150	0.951	0.064	
	Standard Deviation	2.073	0.078	0.061	0.048	9
	Median	2.894	0.144	1.000	0.056	
	Kurtosis	-0.319	2.043	-0.071	-0.962	
	Skewness	0.465	0.220	-1.019	0.256	

Table 5.18 reveals the macroscopic characteristics of the rural data_sample_2 clusters that resulted from the model developed. As in the data_sample_1, cluster 1 has barely no trend compared to the rest of the clusters formed presenting a steady behaviour throughout monitoring. 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. Finally, a low accident frequency is observed in the partial data obtained for some of the cluster's drivers. All the above along with the higher number of drivers included in the specific cluster, lead to the conclusion that this cluster is similar to cluster 1 of the data_sample_1 and once again, it mainly represent the typical driver.

As for drivers of cluster 2, they present a low negative trend and a medium behavioural volatility. They also feature a significantly low average driving efficiency combined by a high accident frequency. All the aforementioned demonstrate that this cluster is comprised of drivers with the most risky behaviour of the sample that have a low trend of behavioural deterioration. The only difference compared to the data_sample_2 of the urban road is that drivers of this cluster retain a steadily low performance throughout the complete recording period.

Drivers of cluster 3 feature a high positive trend indicating an improvement trend in general. This is also confirmed by the medium to high volatility presented, which testifies the existing positive instability. This cluster's drivers also have a high average rating and a significantly low accident frequency. As also shown in the data_sample_1, this cluster is likely to represent the most safety efficient drivers of the sample. The only difference from cluster 3 of the urban road is that it features a positive instead of a negative efficiency trend.

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

As for the necessity of having prior information on driving accident data of the drivers, those data seem to affect only the second cluster of the most unstable drivers, which incorporates drivers that are both less safety efficient and unstable. The forming of the other two clusters is not significantly influenced by the existence of these data. The

qualitative characteristics of the clusters arising for each sample and road type are aggregated and illustrated in table 5.19.

Driving characteristics of the resulting clusters

Table 5.20 demonstrates the driving characteristics of the drivers' clusters As expected, for both sample type models of the urban road, the median of the attributes is reducing while shifting to a class of higher efficiency. The only exception to that is $speeding_{urban}$ of the third cluster of the data_sample_1, which is the highest value of the cluster instead of being the lowest. At this point, it should be highlighted that the cluster analysis performed take into account a variety of factors apart from driving efficiency and therefore, despite the fact that the specific cluster demonstrates the highest median efficiency among the three resulted, it includes some drivers whose safety efficiency is not high. The fact that drivers of this cluster feature a constant behavioural improvement is probably indicative that they are likely to enhance their speeding behaviour in the future. This is probably not very clear for clusters 1 and 2 of the data_sample_2 since the median driving efficiency is similar. Nonetheless, the clusters that present the highest efficiency also feature the lowest driving attributes. As for the rural data_sample_1, this rule applies only for the $mobile_{rural}$ and $speeding_{rural}$.

It is also observed that for all sample and road types the number of harsh acceleration events is higher than the number of harsh braking events occurred. Additionally, the number of harsh events occurring in rural road is lower than in urban for cluster 1 that represents the typical driver. The same difference between urban and rural road types is noticed for mobile usage and speeding of cluster 1 as well.

As for the clusters of the data_sample_2 that include drivers with a high number of accidents per year of experience, they both feature the highest number of harsh acceleration and braking events among all other clusters and the highest number of seconds of mobile usage and speeding at rural road. Nonetheless, the relative urban road cluster shows a level of mobile usage and speeding that is close to the highest one. The fact the urban road cluster of more risky drivers does not present the highest level for these two characteristics could be attributed to the fact that the data_sample_2 used is not very large. Therefore, a larger sample should be exploited to investigate more the specific issue.

The clusters of the cautious drivers of the data_sample_2 demonstrate a significantly lower level of driving characteristics for all metrics considered. This finding is validated by the results of the efficiency analysis conducted in subchapter 5.3.3, which indicates that the specific cluster resulted from the clustering process of the data_sample_2 also represents the cautious drivers. This is also the case for the data_sample_1 except from $speeding_{urban}$, which appears to be the highest among the three clusters and from ha_{rural} and hb_{rural} that are slightly higher than the respective lowest value of the other two clusters. This probably implies the high value of incorporating the accident per year variable into the clustering procedure when trying to form a cluster of safety efficient drivers. To this end, it is suggested to use a larger dataset and investigate further on which variables

should be included in the clustering approach of an data_sample_1 (i.e. without acquiring the accident history of the drivers).

As mentioned above, the second cluster represents the unstable drivers of the sample. The data_sample_2 demonstrates an approximately equal or slightly higher level of driving characteristics for the metrics considered, which along with the fact that the accidents per year variable and the driving efficiency is high and low respectively, leads to the conclusion that the second cluster formed introduces drivers with non-steady and poor driving behaviour. Consequently, it can be inferred that the most risky drivers also tend to present a low trend of efficiency change and a low volatility value.

Table 5.20: Driving characteristics of the drivers' clusters per 100km and per road and sample type

Sample type	Road type	No of drivers	Driving characteristics	Cluster 1 (typical)	Cluster 2 (unstable)	Cluster 3 (cautious)
data_sample_1	Urban	100	efficiency	0.33	0.61	0.81
			ha	26.75	17.20	6.89
			hb	10.05	7.30	3.44
			mu	499	165	60
			sp	1095	619	1240
	Rural	100	efficiency	0.36	0.69	0.88
			ha	14.97	6.36	9.65
			hb	7.59	3.09	3.61
			mu	234	285	86
			sp	988	923	347
data_sample_2	Urban	43	efficiency	0.37	0.33	1.00
			ha	22.05	41.26	14.13
			hb	8.10	15.39	5.28
			mu	653	481	82
			sp	1349	1140	436
	Rural	39	efficiency	0.37	0.39	1.00
			ha	12.35	19.24	5.74
			hb	6.66	6.84	2.46
			mu	316	380	149
			sp	1125	1149	415

Regarding the data_sample_1, it is clear that the second cluster represents highly volatile drivers that introduce a medium positive (urban road) or negative (rural road) trend. Nevertheless, drivers of this cluster show a safety efficiency that is between that of the other two clusters of the typical and cautious drivers or lower. This is also confirmed by the metrics recorded and illustrated in table 5.20 and seems to be a valid observation

since cautious drivers should maintain a low level of metrics to remain at the specific cluster, which requires a behaviour that is less volatile. It is noticeable that the number of drivers included in the first cluster is significantly higher than the rest, especially for the data_sample_1, and accordingly, the driving characteristics of the clusters arising might not be totally representative. This remains to be addressed in the future.

As a general observation, the driving efficiency index of each cluster goes along with the level of its metrics. This validates the methodology proposed in the previous step for the estimation of each driver's safety efficiency. It is also shown here that the driving efficiency and the metrics recorded present a significant difference between the sets of the typical and the cautious drivers. There is a minor divergence from this rule in the case of $speeding_{urban}$, which, as mentioned above, is attributed either to the sample size or to the absence of the accidents per year of driving experience variable.

5.4) Results summary

The most important findings of this research are summarized below:

1) A different sampling kilometrage is required for each a) road type, b) driving metric and c) driving aggressiveness to accumulate enough data to obtain a clear picture of a driver's behaviour and perform DE analysis.

2) There is no significant metric that appears to be critical for the determination of the required amount of data to be recorded in highways.

3) More risky drivers need less monitoring in rural road and highways.

4) A new methodological approach is proposed for estimating the efficient level of inputs and outputs of each trip as well as for the identification of the least efficient trips.

5) The integration of DEA with the convex hull algorithmic approach yielded significantly better results than the rest of the approaches tested.

6) A new methodological approach is provided for driving efficiency evaluation as well as for estimating the efficient level of inputs and outputs of each driver.

7) Mobile usage is not a critical factor when measuring driving efficiency using DEA.

8) The number of harsh acceleration events occurred in all road types is at least twice as much as the number of harsh braking events.

9) The shift between efficiency classes is mainly affected by different driving metrics in urban and rural road.

10) Regarding the analysis of the efficiency time series arising, it appears 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 except from the data_sample_2 of rural road that is slightly higher than the rest.

11) The average trend is observed to be approximately the same between the two road types despite the fact that median trend is diverged.

12) Stationarity is not included in the final clustering procedure since three out of four driver groups have similar characteristics and therefore it would not play an important role.

13) The clustering analysis performed resulted to three driving groups of a) the typical drivers, b) unstable drivers and c) cautious drivers.

14) 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.

Chapter 6: Conclusions

6.1) Overview

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.

The research questions raised are:

1) How well can driving safety efficiency be benchmarked? Can data science techniques and large-scale data provide sufficient answers?

2) What are the temporal evolution characteristics of driving efficiency? What do the drivers' groups formed represent?

3) What is the required amount of driving data that should be collected for each driver?

4) How can the least efficient trips of a database be identified?

The general **methodological** framework applied to answer these questions 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.

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. 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. 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. 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. 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 is found 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.

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.

Literature review revealed that it is significant to study the potential of **benchmarking** driving safety efficiency using **microscopic** driving data collected from smartphone devices. Literature review revealed that it is 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. 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 an 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 as well as a description of the questionnaire administered. 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 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. 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. 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.

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. All drivers chosen to be included in the large-scale data investigation 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 is used, which equals to 10,088 trips. Finally, for the purposes of the driver efficiency analysis, driving data from 100 and 100 drivers were selected for the analysis conducted in urban and rural road respectively. 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.

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.

It is 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. First, the total driver efficiency of the total recorded period is estimated for each driver based on the safety efficiency index defined.

The **evolution** of average driving efficiency over time is also investigated using different databases, accumulated over different timeframes from the beginning of recording time until the end of each timeframe. For the time window of each time step examined, the total driving efficiency of that period is estimated. The **time series** that results is studied and decomposed in its main components, stationarity and trend. 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. The efficiency time-series created is analysed and results are exploited for driver clustering, which lead to the recognition of the main driving profiles. The optimal number of clusters is determined using the elbow method. Driving characteristics of each cluster arose are analysed and conclusions drawn are presented below.

6.2 Main contributions

This doctoral dissertation:

1) Makes use of an **innovative smartphone data collection system** and develops a valuable **methodology** for **large-scale data investigation**.

2) Provides a cutting-edge **methodological framework for evaluating driving safety efficiency** on trip and driver basis based on data science techniques.

3) **Quantifies** the **driving data** that should be collected when **evaluating driving behaviour** in terms of safety.

4) Provides insights on the **main driving** behaviour **profiles** that exist and discusses **their characteristics**.

6.2.1) Large-scale data investigation methodology and the innovative smartphone data collection system

In this research, the **optimization** technique of DEA is applied, which is mainly used so far in operational research. DEA is a linear programming technique and as such, it is performing relatively fast on small-scale databases but much slower when it comes to **large-scale** data. 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.

To this end, a multi-dimensional **convex hull** technique incorporating multiple inputs and outputs is combined with DEA and applied on driving data. To the best of the authors' knowledge, no effort has been made previously to combine the computational geometry procedure of convex hull with DEA for reducing the running time of a large-scale DEA problem that features multiple inputs and outputs. The **scenarios** that were taken into consideration for testing included sets of 100, 500, 1000, 5,000 and 10,088 trips and compared based on running time of each of the algorithms applied. To test the efficiency of the proposed approach, a driving data sample of 10,088 trips, collected using the smartphone devices of 88 drivers, was exploited for the purpose of this study with the ultimate goal of identifying the efficient level of inefficient trips as well as the least efficient trips, based on driving behaviour. This verifies that the convex hull technique **outperforms** the rest in terms of the required processing time.

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. This research succeeds in quantifying the need for driving data collection through smartphone devices when it comes to driving behaviour evaluation using DEA methods. 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. It is essential for researchers nowadays to allocate available resources efficiently due to the emerging necessity of expensive big data computations. As we move on to the **big data** era, it becomes extremely important to address this issue because of the enormous costs that result from the data computations. On the other hand, it is momentous to be capable of exploiting this vast amount of information collected and draw non-biased conclusions from it without having to include more or less information than necessary.

For this purpose, the present research exploits large-scale driving data from a sophisticated data collection system **171** drivers participated to the designed experiment during a 7-months timeframe and a large database of **49,722** trips is acquired. Driving behaviour variables collected include distance travelled, acceleration, braking, speed and smartphone usage. The required monitoring period at which the driving behaviour metrics rate of each driver converges to a stable value is determined by the **statistical methods** used.

On the top of these, this study makes use of data that were collected using an **innovative** approach that is based on a **smartphone** application. This is an approach that is becoming popular nowadays and is considered a fair solution for collecting data in **naturalistic** driving **experiments**. It is also a comparatively less expensive approach, which is very important because it could easily be used in a real case scenario.

These two approaches combined, constitute the **innovation** of this doctoral research in terms of the large-scale **data handling**. The added value of this doctoral research is that it presents how the **dimensionality** of large-scale data can be reduced and valuable conclusions can be drawn from them without putting too much computational effort or losing information during this procedure.

6.2.2) Methodological framework for evaluating driving safety efficiency

Another important contribution of this research is that it suggests a new approach for the **benchmarking** 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. The efficiency **level** of a trip is defined as the maximum value of inputs (e.g. the maximum number of harsh acceleration/ braking events that should occur) or the minimum value of outputs (the minimum distance that should be travelled) that should be reached for a trip to become efficient. 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.

This paper also provides an innovative solid **framework** for **benchmarking** drivers' safety efficiency based on DEA. A database of 100 drivers in urban and rural road was exploited and combinations of several driving analytics collected served as inputs and outputs of DEA models for the estimation of a comparative driving safety efficiency **index** for each driver in the sample. Efficiency is examined in terms of speed limit violation, driving distraction from mobile phone usage, aggressiveness and overall safety on urban,

rural and highway roads and in an overall model. An additional value of the methodology proposed is that it provides the potential to recognize the efficient **peers** of a driver within a driving sample, which enables the estimation of the optimal level of inputs and outputs that each driver should reach to shift to the efficiency frontier or should pass to become even more efficient.

Based on this methodological framework, it is inferred that the change that is required in each driving attribute that was taken into consideration for a driver to **shift** either to the efficiency **frontier** or to another driving **class** can also be estimated. This can be achieved by solving the optimization problem for a specific input or output given the target efficiency ($\text{Driving Efficiency}_\text{B}$), which is the upper or the lower limit of the class that the driver is shifting in case of efficiency decrease or increase respectively.

6.2.3) Driving data quantification for driving behaviour evaluation

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. In this research, the adequate sampling distance to perform DEA using large-scale driving data is investigated and suggestions are made.

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. This is a rational conclusion since the level of metrics of a driver with an average level of metrics may fluctuate a lot within a wide range whereas this is not possible for an aggressive driver whose behaviour has to fluctuate within a relatively high level of metrics in order for his/ her average level of metrics to be high. Nevertheless, this is not strongly indicated from the results arising and therefore it should be further investigated.

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**. This shows that driving behaviour is significantly different in highways from the other two road types and therefore there is not a specific driving metric that requires significantly higher distance for convergence compared to the rest. The most critical driving metrics found to be the number of harsh acceleration events occurred in urban road and highway and the seconds of mobile usage in rural roads.

Additionally, it appears that a **different** data **amount** is necessary to be collected depending on the **road type** examined. Indicatively, the sampling distance suggested by the outcomes of this research are 519km for urban, 579km for rural and 611km for highways. Despite the fact that the above values are of the same order of magnitude, they are likely to be affected by the longer distance travelled, which results from the higher driving speed in each road type. This argument is also reinforced by the lower level of

metrics recorded when moving from an urban to an inter-urban environment. Therefore it would be arbitrary to directly conclude that this difference is due to the longer distance required for the metrics to be converged and should be further examined.

6.2.4) Main driving profiles and their characteristics

Findings pointed towards a potential for classifying driving sample based on the drivers' relative safety efficiency identified. Drivers were divided into **three categories** (non-efficient, weakly efficient and most efficient) based on the 25% and 75% percentile thresholds specified. The highlights of the analysis conducted for each category indicated considerable differences 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. The number of harsh braking events appeared to be 2-3 times less than that of the harsh acceleration events in all road types. The seconds of speeding followed by the number of harsh braking and acceleration events were identified as the key factors for safety efficiency index estimation. On the other hand, **mobile** usage is **not** found to be a **critical** factor in safety efficiency benchmarking 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 investigated in this study, which also specifies the moving time window in which each driver is benchmarked. 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. The analysis of the efficiency time series arising showed 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 except from the data_sample_2 of rural road that is slightly higher than the rest. Apparently, drivers present a more volatile behaviour when driving in rural road, which is likely to imply a higher number of different driving patterns in the specific road type and something that is also confirmed by the higher sampling distance required.

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. This indicates the existence of **higher** outlier trend values in urban road, which probably indicates that the temporal evolution of driving behaviour is more sensitive in the changes made in driving metrics and therefore it is easier and quicker for drivers to become more or less efficient. This is also reinforced by the fact that driving metrics in urban road are higher than those in rural road. 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.

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.

Clustering analysis performed resulted to **three** driving groups, which mainly represent the **typical** 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 typical 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.

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.

6.3) Research innovation and impact

The innovation of this study lies on the several different research questions answered herein. First of all, 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. The way in which it is feasible to perform benchmarking of driving efficiency based on travel and driving behaviour metrics collected from each trip on a driver basis is studied in the present research. Moreover, a new methodological approach is provided for estimating the efficient level of inputs and outputs that each driver should reach to become efficient in terms of safety. To the best of the author's knowledge, this is the first research that contributes towards the efficiency measurement on a driver's level using microscopic large-scale driving data. Furthermore, metrics that reveal information about different driving risk category (aggressiveness, distraction, etc.) are exploited altogether to estimate driving safety efficiency, which constitutes the methodological innovation of this dissertation.

This thesis is also dealing with the problem of data science techniques that can be applied on real transportation problems as the one examined, to deal with the problem driving efficiency benchmarking using the optimization technique of DEA, which is mainly used so far in operational research. Consequently, the potential of applying a DEA methodology of improved performance on large-scale data 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. These two approaches combined, constitute the innovation of this doctoral research in terms of the large-scale data handling. The added value of this doctoral research is that it presents how the dimensionality of large-scale data can be

reduced and valuable conclusions can be drawn from them without putting too much computational effort or losing information during this procedure. On the top of these, this study makes use of data that were collected using an innovative approach that is based on a smartphone application.

Moreover, one of the innovations of the methodology proposed is that there is no need to assign weights to the input variables. This is extremely significant since no weight specification is required, which is very handy when there is no prior knowledge of the weights that should be assigned to each of the attributes used. This is the case in the specific study as well, since the weights that could be assigned to each of the driving metrics considered (e.g. number of harsh events, seconds driving over the speed limits) are under investigation in literature.

It is also very important that this research recognizes 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 dissertation studies the time evolution of driving efficiency and reveals the characteristics of the driving profiles arose. This may assist in the acquisition of a clearer picture regarding the most significant factors that increase accident risk and therefore accident probability as well as towards the provision of more representative and personalized information to drivers for further driving behaviour improvement. To this end, optimization and machine learning techniques are combined.

The impact of this dissertation is that its results can be exploited to provide feedback to drivers on their total driving efficiency and its evolution, whenever an inefficient trip performed, as well as the deviation of their trip efficiency from their overall efficiency. Both are valuable findings since they could potentially be embedded in a platform that provides feedback and recommendations to users regarding their driving behaviour. The results arising could also be used by a smartphone application to advise drivers regarding their driving efficiency either overall or per road type. Drivers could be advised on the driving characteristics that need further improvement to become less risky or every time that an inefficient trip takes place to be further improved. It can also be exploited as an innovative approach to measure the efficiency of a database that includes a vast number of trips, on a trip level. This constitutes a very significant contribution since driving safety efficiency measurement is highly correlated with driving risk, which means that this can potentially affect the accident probability of a driver. As a result, the exploitation of this study's results can lead to a reduction in the total number of accidents.

The fact that this research recognizes the main characteristics of the resulting driving safety efficiency groups from the analysis performed is very important, since more personalized feedback can be provided to users, which could be proved more influential. Nonetheless, results indicated that since the temporal evolution could be steep, 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. Overall, it can be said that the results of this dissertation present a considerable opportunity for road safety, as both trips and drivers could be classified into different efficiency categories (such as efficient, less efficient, non-efficient) and further evaluate their main

characteristics in terms of traffic risk, aggressiveness, eco-driving etc. Finally, all the aforementioned can serve as a recommendation system's service that provides the appropriate stimuli to drivers to improve their behaviour. To this end, gamification policies based on this approach such as competitions, learning goals and awards could contribute to this scope.

Finally, findings could also be useful for developing insurance pricing schemes based on driving usage or characteristics i.e. Pay-How-You-Drive driving insurance schemes, a policy that also conduces to the further enhancement of behaviour and, therefore, driving risk reduction.

6.3) Future challenges

The most important suggestion for future research is probably the exploitation of a larger driving sample. This will assist towards the acquisition of a clearer picture regarding the relationship between the number of harsh acceleration events and seconds of speeding and the total driving duration at which those metrics converge. It is significant to know more about the relationship between the aggressiveness of a driver and the necessary monitoring distance or time and as a result, this should be further investigated. The same applies for the safety efficiency estimation as well, where DEA analysis showed that models' results become more representative of the average characteristics of each class as more trips and drivers are aggregated. Apart from that, as the sample grow bigger the high proportion of efficient DMUs to the total number DMUs is reduced. In this case though, data analysis requires quicker implementations and therefore further research is needed towards improving the implementation of the algorithms to overcome the dimensionality limitation of the quick hull implementation used in this study, so as the algorithm can incorporate input and output matrices of higher dimensions.

Further research should centre to larger samples of trips with a representative sample of drivers population for which more demographic data etc. can be collected. The recent trend in driving data collection and analysis is to collect anonymized data from larger samples in contrast with the classic studies, which used to design driving experiments that collect data from a sample the personal details of which, such as demographics etc. are known. Both approaches have several drawbacks and benefits e.g. the fact that the results of a research that has exploited data that cannot be connected with any personal information cannot be generalized in the population. It is therefore important to somehow bridge the gap between these two approaches and retain the advantages of both. This could potentially be achieved by many means such as obtaining larger samples to respect representativeness, collect data from many countries of which drivers have different driving characteristics etc. This should be the objective of future research as well.

Another important research question raised at this point is whether future research should focus on the investigation of the macroscopic or microscopic behaviour of drivers. Although these two paths are seemingly different, they are likely to be equally useful in determining the variety of driving behaviour patterns. The macroscopic approach would suggest constructing all possible driving profiles and study behavioural shifts among them

over longer time periods. On the other hand, microscopic analysis would focus on the spatial analysis of the time series of the driving metrics collected from each trip and would therefore suggest the opposite i.e. to concentrate on how everyday driving behaviour could be classified as risky or less risky. In any case, future research should focus on the comparison of the results arising from trip and driver efficiency analysis of each driver to evaluate the representativeness of the results. This is probably where the key to this answer lies.

Other DEA's limitations should also be addressed which among others include DEA's sensitivity to outliers and to drivers with zero input attributes that in the present study had to be eliminated from the sample. This is something that has probably affected the time series creation since efficiency is estimated in each time step by comparatively estimating driving efficiency. As a result, when outliers appear, time series might be "pushed" upwards or downwards. It would be very interesting to be capable of comparing driving efficiency even between those drivers that appear to have zero driving metrics recorded. Nonetheless, the solution to this problem could be given by incorporating more driving metrics in the DEA models, which is strongly recommended in the future.

Apparently, not all metrics capturing safety behaviour of a driver have been included in this research and an attempt to do this would probably lead to even stronger models. Finally, a higher number of driving metrics such as headways, eye movement, drowsiness, lane changing etc. that are also significantly influencing accident risk should be used to test whether or not the driving behaviour models are improved. These metrics could be recorded by a variety of different data sources such as cameras, eye-tracking devices, radars, LiDARs, on-board-diagnostic devices (OBD), smartphones etc. It is noted though that most of these data collection methods are relatively expensive to use and this is why smartphone is considered a fair solution for collecting data in naturalistic driving experiments.

The study of the dynamic evolution of driving efficiency raises also the question of how much and how rapidly driving profiles are altering over time. It is a matter of great significance to shed light on this issue since the classification of the drivers based on these characteristics would lead to enhanced recommendations to users and therefore to a further reduction of total driving risk. For this purpose, stationarity should also be studied using a different methodological approach since it is deemed to play an important role in the form of the configuration of the final clustering procedure.

Another recommendation made for future research is study the prediction of drivers' total driving efficiency index or cluster in a set of consecutive periods, applying time-series analysis on the existing time-series of driving efficiency. The way to estimate the level of inputs and outputs that each driver should reach to shift to the cluster of best and most improved drivers that resulted from the analysis of the temporal evolution of driving efficiency.

It is also very important to collect the official accident record of each driver participating in the experiment instead of the stated number of accidents. This would probably lead to results that are more accurate since it is likely that some drivers might have not recalled the correct number of crashes they have been involved in, especially when it comes to older drivers with a large driving experience.

References

Aarts, L., & Van Schagen, I. Driving speed and the risk of road crashes: A review. Accident Analysis & Prevention, Vol. 38, No. 2, 2006, pp. 215-224. https://doi.org/10.1016/j.aap.2005.07.004

Ali, A. I. Streamlined computation for data envelopment analysis. European journal of operational research, Vol. 64, No. 1, 1993, pp. 61-67. https://doi.org/10.1016/0377-2217(93)90008-B

Alm, H., & Nilsson, L. (1994). Changes in driver behaviour as a function of handsfree mobile phones—A simulator study. Accident Analysis & Prevention, 26(4), 441-451.

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

Andersson, Gunnar, and Göran Nilsson. Speed management in Sweden: Speed, speed limits and safety. Swedish National Road and Transport Research Institute, 1997.

Araújo, R., Igreja, Â., de Castro, R., & Araujo, R. E. (2012). Driving coach: A smartphone application to evaluate driving efficient patterns. In Intelligent Vehicles Symposium (IV), 2012 IEEE (pp. 1005-1010). IEEE.

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

Ball, K., Ackerman, M., (2011). The older driver (Training and assessment: Knowledge skills and attitudes. Handbook of Driving Simulation for Engineering, Medicine and Psychology, CRC Press.

Barber, C. B., Dobkin, D. P., & Huhdanpaa, H. The quickhull algorithm for convex hulls. ACM Transactions on Mathematical Software (TOMS), Vol. 22, No. 4, 1996, pp. 469-483. https://doi.org/10.1145/235815.235821

Barr, R.S. and M.L. Durchholz. Parallel and hierarchical decomposition approaches for solving large-scale Data Envelopment Analysis models. Annals of Operations Research, Vol. 73, 1997, pp. 339-372. https://doi.org/10.1023/A:1018941531019

Bianchi, M., Boyle, M., Hollingsworth, D. (1999). "A comparison of methods for trend estimation". Applied Economics Letters. 6 (2): 103–109. doi:10.1080/135048599353726

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

Blana E., Golias J., (2002), "Differences between vehicle lateral displacement on the road and in a fixed-base simulator", Human Factors, Vol. 44, 2.

Bonsall, Peter, Ronghui Liu, and William Young. Modelling safety-related driving behaviour—impact of parameter values. Transportation Research Part A: Policy and Practice Vol. 39, No. 5, 2005, pp. 425-444. https://doi.org/10.1016/j.tra.2005.02.002

Bougnol, M.-L., J.H. Dulá, D. Retzlaff-Roberts, and N.K. Womer. Nonparametric frontier analysis with multiple constituencies. Journal of the Operational Research Society, 2005; 56: 252-266.

Bowers, A.R., Anastasio, R.J., Sheldon, S.S., O'Connor, M.G., Hollis, A.M., Howe, P.D., Horowitz, T.S., 2013. Can we improve clinical prediction of at-risk older drivers? Accident Analysis and Prevention, 59, 537-547.

Brookhuis, K.A., de Vries, G., de Waard, D., 1991. The effects of mobile telephoning on driving performance, Accident Analysis & Prevention 23(4), 309-316.

Burns, P. C., Parkes, A., Burton, S., Smith, R. K., & Burch, D. (2002). How Dangerous is Driving with a Mobile Phone?: Benchmarking the Impairment to Alcohol (Vol. 547). TRL.

Cameron, S. (2005). "Making Regression Analysis More Useful, II". Econometrics. Maidenhead: McGraw Hill Higher Education. pp. 171–198. ISBN 0077104285.

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

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

Cook, W. D., Kazakov, A., & Persaud, B. N. (2001). Prioritising highway accident sites: a data envelopment analysis model. Journal of the Operational Research Society, 52(3), 303-309.

Cook, W. D., Tone, K., & Zhu, J. (2014). Data envelopment analysis: Prior to choosing a model. Omega, 44, 1-4.

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.

Cooper, W. W., Seiford, L. M., & Tone, K. Introduction to data envelopment analysis and its uses: with DEA-solver software and references. Springer Science & Business Media. 2006.

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

Donmez, B., Boyle, L. N., & Lee, J. D. (2006). The impact of distraction mitigation strategies on driving performance. Human factors, 48(4), 785-804.

Dragutinovic, N., & Twisk, D. (2005). Use of mobile phones while driving–effects on road safety. SWOV Institute, Leidschendam.

Dulá, J.H. and B. L. Hickman. Effects of excluding the column being scored from the DEA envelopment LP technology matrix. Journal of the Operational Research Society, Vol. 48, 1997, pp. 1001– 1012. https://doi.org/10.1057/palgrave.jors.2600434

Dulá, J. H., & López, F. J. Algorithms for the frame of a finitely generated unbounded polyhedron. INFORMS Journal on Computing, Vol. 18, No. 1, 2006, pp. 97-110. https://doi.org/%7B0%7D

Dulá, J. H. A computational study of DEA with massive data sets. Computers & Operations Research, Vol. 35, No. 4, 2008, pp. 1191-1203. https://doi.org/10.1016/j.cor.2006.07.011

Dulá, J. H., & López, F. J. Preprocessing DEA. Computers & Operations Research, Vol. 36, No. 4, 2009, pp. 1204-1220. https://doi.org/10.1016/j.cor.2008.01.004

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

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

Elvik, R., Christensen, P., & Amundsen, A. (2004). Speed and road accidents. An evaluation of the Power Model. TØI report, 740, 2004.

Elvik, R. (2009). The Power Model of the relationship between speed and road safety: update and new analyses (No. 1034/2009).

Elvik, R. (2011). Developing an accident modification function for speed enforcement. Safety science, 49(6), 920-925.

Elvik, R. (2014). Rewarding safe and environmentally sustainable driving: systematic review of trials. Transportation Research Record: Journal of the Transportation Research Board, (2465), 1-7.

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

Enev M., Takakuwa A., Koscher K, and Kohno T. (2016) Automobile Driver Fingerprinting. Proceedings on Privacy Enhancing Technologies (1):34-51

Engle, R. F., & Granger, C. W. (1987). Co-integration and error correction: representation, estimation, and testing. Econometrica: journal of the Econometric Society, 251-276.

Gitelman, V., Doveh, E., & Hakkert, S. (2010). Designing a composite indicator for road safety. Safety science, 48(9), 1212-1224.

Global status report on road safety: time for action. Geneva, World Health Organization, 2009 (http://www.who.int/violence_injury_prevention/road_safety_status/2009/en/index.html, accessed 7 February 2013).

Gündüz, G., Yaman, Ç., Peker, A. U., & Acarman, T. (2018). Prediction of Risk Generated by Different Driving Patterns and Their Conflict Redistribution. IEEE Transactions on Intelligent Vehicles, 3(1), 71-80.

Haigney, D., & Westerman, S. J. (2001). Mobile (cellular) phone use and driving: A critical review of research methodology. Ergonomics, 44(2), 132-143.

Handel, Peter, Isaac Skog, Johan Wahlstrom, Farid Bonawiede, Richard Welch, Jens Ohlsson, and Martin Ohlsson. Insurance telematics: Opportunities and challenges with the smartphone solution. IEEE Intelligent Transportation Systems Magazine Vol. 6, No. 4, 2014, pp. 57-70. https://doi.org/10.1109/MITS.2014.2343262

Haque, Md Mazharul, and Simon Washington. The impact of mobile phone distraction on the braking behaviour of young drivers: a hazard-based duration model. Transportation research part C: emerging technologies Vol. 50, 2015, pp. 13-27. https://doi.org/10.1016/j.trc.2014.07.011

Harbluk, J.L., Noy, Y.I., Eizenman, M., 2002. The impact of cognitive distraction on driver visual behaviour and vehicle control, Report No. TP No. 13889 E, Road Safety Directorate and Motor Vehicle Regulation Directorate, Ottawa, Canada.

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.

Hermans, E., Brijs, T., Wets, G., & Vanhoof, K. (2009). Benchmarking road safety: lessons to learn from a data envelopment analysis. Accident Analysis & Prevention, 41(1), 174-182.

Hermans, E., Van den Bossche, F., & Wets, G. (2008). Combining road safety information in a performance index. Accident Analysis & Prevention, 40(4), 1337-1344.

Hill, J., Aldah, M., Talbot, R., Giustiniani, G., Fagerlind, H., Jänsch, M., 2012. Final Report, Deliverable 2.5 of the EC FP7 project DaCoTA.

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.

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

Hong, J. H., Margines, B., & Dey, A. K. A smartphone-based sensing platform to model aggressive driving behaviors. In Proceedings of the 32nd annual ACM conference on Human factors in computing systems, 2014 April, pp. 4047-4056. https://doi.org/10.1145/2556288.2557321

Horberry, T., Anderson, J., Regan, M. A., Triggs, T. J., & Brown, J. (2006). Driver distraction: The effects of concurrent in-vehicle tasks, road environment complexity and age on driving performance. Accident Analysis & Prevention, 38(1), 185-191.

Hosking, J. (1981). Fractional differencing. Biometrika 68(1), 165–176.

Johnson, Derick A., and Mohan M. Trivedi. Driving style recognition using a smartphone as a sensor platform. In Intelligent Transportation Systems (ITSC), 14th International IEEE Conference on, 2011 pp. 1609-1615. https://doi.org/10.1109/ITSC.2011.6083078

Karlaftis, M. and Vlahogianni, E. (2009). Memory properties and fractional integration in transportation time-series. Transportation Research Part C 17, 444–453.

Karlaftis, M.G., Gleason, J.M., Barnum, D.T. '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, 2013

Kelley, K., Clark, B., Brown, V., Sitzia, J., 2003. Good practice in the conduct and reporting of survey research, International Journal for Quality in Health care, Volume 15, N 3, pp. 261-266.

Klauer, S.G., Dingus, T.A., Neale, V.L., Sudweeks, J.D., Ramsey, D.J., (2008). Comparing Real-World Behaviors of Drivers with High vs. Low Rates of Crashes and Near-Crashes. National Highway Traffic Safety Administration, Washington, DC.

Kloeden, C. N., McLean, A. J., Moore, V. M., & Ponte, G. (1997). Travelling Speed and the Risk of Crash Involvement Volume 2-Case and Reconstruction Details. Adelaide: NHMRC Road Accident Research Unit, The University of Adelaide.

Kloeden, C.N., Ponte, G., McLean, A.J., (2001). Travelling Speed and the Risk of Crash Involvement on Rural Roads. Report CR 204. Australian Transport Safety Bureau ATSB, Civic Square, ACT.

Kloeden, C. N., McLean, J., & Glonek, G. F. V. (2002). Reanalysis of travelling speed and the risk of crash involvement in Adelaide South Australia. Australian Transport Safety Bureau.

Kröyer, H. R., Jonsson, T., & Várhelyi, A. (2014). Relative fatality risk curve to describe the effect of change in the impact speed on fatality risk of pedestrians struck by a motor vehicle. Accident Analysis & Prevention, 62, 143-152.

Lamble, D., Rajalin, S., & Summala, H. (2002). Mobile phone use while driving: public opinions on restrictions. Transportation, 29(3), 223-236.

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

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

Mantouka, E. G., Barmpounakis, E. N., & Vlahogianni, E. I. (2018). Mobile Sensing and Machine Learning for Identifying Driving Safety Profiles (No. 18-01416). In: Transportation Research Board 97th Annual Meeting, Washington, DC.

Martić, Milan, Marina Novaković, and Alenka Baggia. Data envelopment analysis-basic models and their Utilization. Organizacija Vol. 42, No. 2, 2009, pp. 37-43. https://doi.org/10.2478/v10051-009-0001-6

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

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

McEvoy, S. P., Stevenson, M. R., & Woodward, M. (2007). The contribution of passengers versus mobile phone use to motor vehicle crashes resulting in hospital attendance by the driver. Accident Analysis & Prevention, 39(6), 1170-1176.

Mills, T. C. (2003). Modelling trends and cycles in economic time series (Vol. 10). Basingstoke: Palgrave Macmillan.

Musicant, O., Lotan, T., Toledo, T. Safety correlation and implication of an In-Vehicle Data Recorder on driver behavior. Presented at 86th Annual Meeting of the Transportation Research Board, Washington, D.C., 2007

Musicant, O., Bar-Gera, H., Schechtman, E., (2011). Individual driver's undesirable driving events – a temporal analysis. 3rd International Conference on Road Safety and Simulation, Indianapolis, 14–16 September 2011.

National Cooperative Highway Research Program (NCHRP), (1999). Evaluation of Traffic Signal Displays for Protected– Permitted Left Turn Control, Traffic Conflict Studies Report, Working Paper 5.

Neale, V.L., Klauer, S.G., Knipling, R.R., Dingus, T.A., Holbrook, G.T., Petersen, A., (2002). The 100 Car Naturalistic Driving Study Phase I: Experimental Design. Report DOT-HS-808-536, Department of Transportation, Washington, DC.

Nilsson G. The effects of speed limits on traffic accidents in Sweden. Sartryck, Swedish National Road and Transport Research Institute, 1982.

Nilsson, G. (2004) Traffic safety dimensions and the power model to describe the effect of speed on safety. Bulletin 221, Lund Institute of Technology, Lund http://lup.lub.lu.se/record/21612

Odaki, M. (1993). On the invertibility of fractionally differenced ARIMA processes. Biometrika 80(13), 703–709.

Odeck, J. (2006). Identifying traffic safety best practice: an application of DEA and Malmquist indices. Omega, 34(1), 28-40.

Okonkwo OC, Griffith HR, Vance DE, Marson DC, Ball KK, Wadley VG (2009) Awareness of functional difficulties in mild cognitive impairment: a multidomain assessment approach. J Am Geriatr Soc 57, 978-984.

Papadimitriou, E., Tselentis, D. I., & Yannis, G. (2018). Analysis of Driving Behaviour Characteristics Based on Smartphone Data. In: Transportation Research Arena, Vienna.

Papantoniou, P., Papadimitriou, E., & Yannis, G. (2015). Assessment of driving simulator studies on driver distraction. Advances in Transportation Studies, (35).

Parker Jr., M.R., Zegeer, C.V., (1988). Traffic Conflict Techniques for Safety and Operations: Observers Manual. Report FHWA-IP-88-027, FHWA, U.S. Department of Transportation, Washington, DC.

Patel, J., Ball, D. J., & Jones, H. (2008). Factors influencing subjective ranking of driver distractions. Accident Analysis & Prevention, 40(1), 392-395.

Patten, Christopher JD, Albert Kircher, Joakim Östlund, and Lena Nilsson. Using mobile telephones: cognitive workload and attention resource allocation. Accident analysis & prevention Vol. 36, No. 3, 2004, pp. 341-350. https://doi.org/10.1016/S0001-4575(03)00014-9

Pavlou, D., Papantoniou, P., Papadimitriou, E., Vardaki, S., Yannis, G., Antoniou, C., & Papageorgiou, S. G. (2016). Which are the effects of driver distraction and brain pathologies on reaction time and accident risk?. Advances in Transportation Studies, (1).

Peden, M. (2004). World report on road traffic injury prevention.

Perkins, S. R., & Harris, J. L. (1968). Traffic conflict characteristics-accident potential at intersections. Highway Research Record, (225).

Philip, P., Sagaspe, P., Taillard, J., Valtat, C., Moore, N., Åkerstedt, T., ... & Bioulac, B. (2005). Fatigue, sleepiness, and performance in simulated versus real driving conditions. Sleep, 28(12), 1511-1516.

Pöysti, L., Rajalin, S., & Summala, H. (2005). Factors influencing the use of cellular (mobile) phone during driving and hazards while using it. Accident Analysis & Prevention, 37(1), 47-51.

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

Preparata, F. P., & Shamos, M. I. Introduction. In Computational Geometry, Springer New York, 1985, pp. 1-35, https://doi.org/10.1007/978-1-4612-1098-6_1

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

Rakauskas, M. E., Gugerty, L. J., & Ward, N. J. (2004). Effects of naturalistic cell phone conversations on driving performance. Journal of safety research, 35(4), 453-464.

Regan, M. A., Lee, J. D., & Young, K. (Eds.). (2008). Driver distraction: Theory, effects, and mitigation. CRC Press.

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.

Robertson, H.D., Hummer, J.E., Nelson, D.C., (1994). Manual of Traffic Engineering Studies. ITE, Prentice Hall, Englewood Cliffs, NJ, pp. 69–87, 219–235.

Rotemberg, J. J. (1999), A Heuristic Method for Extracting Smooth Trends from Economic Time Series, National Bureau of Economic Research Working Paper No. 7439.

Sabey, B. E., & Taylor, H. (1980). The known risks we run: the highway. In Societal risk assessment (pp. 43-70). Springer, Boston, MA.

Sagberg, F. (2001). Accident risk of car drivers during mobile telephone use. International Journal of Vehicle Design, 26(1), 57-69.

Saifuzzaman, Mohammad, Md Mazharul Haque, Zuduo Zheng, and Simon Washington. Impact of mobile phone use on car-following behaviour of young drivers. Accident Analysis & Prevention Vol. 82, 2015, pp. 10-19. https://doi.org/10.1016/j.aap.2015.05.001

Salmon, P. M., Stanton, N. A., Lenné, M., Jenkins, D. P., Rafferty, L., & Walker, G. H. (2017). Human Factors Methods for Accident Analysis. In Human Factors Methods and Accident Analysis (pp. 29-104). CRC Press.

Seiford, Lawrence M. A bibliography for data envelopment analysis (1978-1996). Annals of Operations Research Vol. 73, 1997, pp. 393-438. https://doi.org/10.1023/A:1018949800069

Shen, Y., Hermans, E., Ruan, D., Wets, G., Brijs, T., & Vanhoof, K. (2011). A generalized multiple layer data envelopment analysis model for hierarchical structure assessment: A case study in road safety performance evaluation. Expert systems with applications, 38(12), 15262-15272. https://doi.org/10.1016/j.eswa.2011.05.073

Shen, Y., Hermans, E., Brijs, T., Wets, G., & Vanhoof, K. (2012). Road safety risk evaluation and target setting using data envelopment analysis and its extensions. Accident Analysis & Prevention, 48, 430-441.

Sheridan, T. B. (2004). Driver distraction from a control theory perspective. Human factors, 46(4), 587-599.

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.

Shumway, R. and Stoffer, S. (2000). Time Series Analysis and Its Applications. Springer-Verlag, NY.

Simons-Morton, B. G., Ouimet, M. C., Wang, J., Klauer, S. G., Lee, S. E., & Dingus, T. A. (2009, June). Hard braking events among novice teenage drivers by passenger characteristics. In Proceedings of the... International Driving Symposium on Human

Factors in Driver Assessment, Training, and Vehicle Design (Vol. 2009, p. 236). NIH Public Access.

Speed management. Paris, OECD/ECMT Transport Research Centre (JTRC), 2006

Strayer, D. L., & Drew, F. A. (2004). Profiles in driver distraction: Effects of cell phone conversations on younger and older drivers. Human factors, 46(4), 640-649.

Strayer, D. L., Drews, F. A., & Johnston, W. A. (2003). Cell phone-induced failures of visual attention during simulated driving. Journal of experimental psychology: Applied, 9(1), 23.

Stutts, J. C., Reinfurt, D. W., Staplin, L., & Rodgman, E. A. (2001). The role of driver distraction in traffic crashes.

SWOW, 2010. Naturalistic Driving: observing everyday driving behaviour, SWOW factsheet, Leidschendam, Nederlands

Thanassoulis, Emmanuel. Introduction to the theory and application of data envelopment analysis. Dordrecht: Kluwer Academic Publishers, 2001.

The Traffic Injury Research Foundation's

Theofilatos A., Tselentis. I.D., Yannis. G., Konstantinopoulos M. (2017) "Willingness – to - Pay for Usage-Based Motor Insurance", Proceedings of the 96th Annual meeting of the Transportation Research Board, Washington, D.C, January 8-12, 2017.

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

Törnros, Jan, and Anne Bolling. Mobile phone use–effects of conversation on mental workload and driving speed in rural and urban environments. Transportation Research Part F: Traffic Psychology and Behaviour Vol. 9, No. 4, 2006, pp. 298-306. https://doi.org/10.1016/j.trf.2006.01.008

Tsay, R. (2002). Analysis of Financial Time series. John Willey & Sons, NY.

Tselentis, D. I., Vlahogianni, E. I., & Yannis, G. Comparative Evaluation of Driving Efficiency Using Smartphone Data. Presented at 97th Annual Meeting of the Transportation Research Board, Washington, D.C., 2018

Tselentis, D. I., Yannis, G., & Vlahogianni, E. I. 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

Van Schagen, I., Welsh, R., Backer-Grondahl, A., Hoedemaeker, M., Lotan, T., Morris, A., Sagberg, F., Winkelbauer, M., (2011). Towards a large-scale European Naturalistic Driving study: final report of Prologue, Deliverable D4.2.

Vlahogianni, Eleni I. Computational intelligence and optimization for transportation big data: challenges and opportunities. Engineering and Applied Sciences Optimization.

Springer International Publishing, 2015, pp. 107-128. https://doi.org/10.1007/978-3-319-18320-6_7

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

Wadley VG, Okonkwo O, Crowe M, Vance DE, Elgin JM, Ball KK, Owsley C (2009) Mild cognitive impairment and everyday function: an investigation of driving performance. J Geriatr Psychiatry Neurol 22, 87-94.

Washington, S. P., Karlaftis, M. G., & Mannering, F. (2010). Statistical and econometric methods for transportation data analysis. CRC press.

Wegman, F., & Oppe, S. (2010). Benchmarking road safety performances of countries. Safety science, 48(9), 1203-1211.

WHO Report (2015)

Yang Y., Chen B., Su L., Qin D. Research and Development of Hybrid Electric Vehicles CAN-Bus Data Monitor and Diagnostic System through OBD-II and Android-Based Smartphones. Advances in Mechanical Engineering, Article ID 741240, 2013.

Yannis, G., E. Papadimitriou, and P. Papantoniou. Distracted driving and mobile phone use: Overview of impacts and countermeasures. NTUA Road Safety Observatory. In Proceedings of the Communication Technologies and Road Safety Conference, Abu Dhabi. 2014.

Yannis, G., Tselentis, D. I., Vlahogianni, E. I., & Argyropoulou, A. (2017). Monitoring distraction through smartphone naturalistic driving experiment. In: 6th International Naturalistic Driving Research Symposium, The Hague, Netherlands, 7-9 June 2017.

Young, K. & Regan, M. (2007). Driver distraction: A review of the literature. In: I.J. Faulks, M. Regan, M. Stevenson, J. Brown, A. Porter & J.D. Irwin (Eds.). Distracted driving. Sydney, NSW: Australasian College of Road Safety. Pages 379-405.

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

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

Zarnowitz, V., & Ozyildirim, A. (2006). Time series decomposition and measurement of business cycles, trends and growth cycles. Journal of Monetary Economics, 53(7), 1717-1739.

Appendix I

The papers produced from this doctoral dissertation as well as the chapter of the dissertation they are related to are illustrated in figure 1.
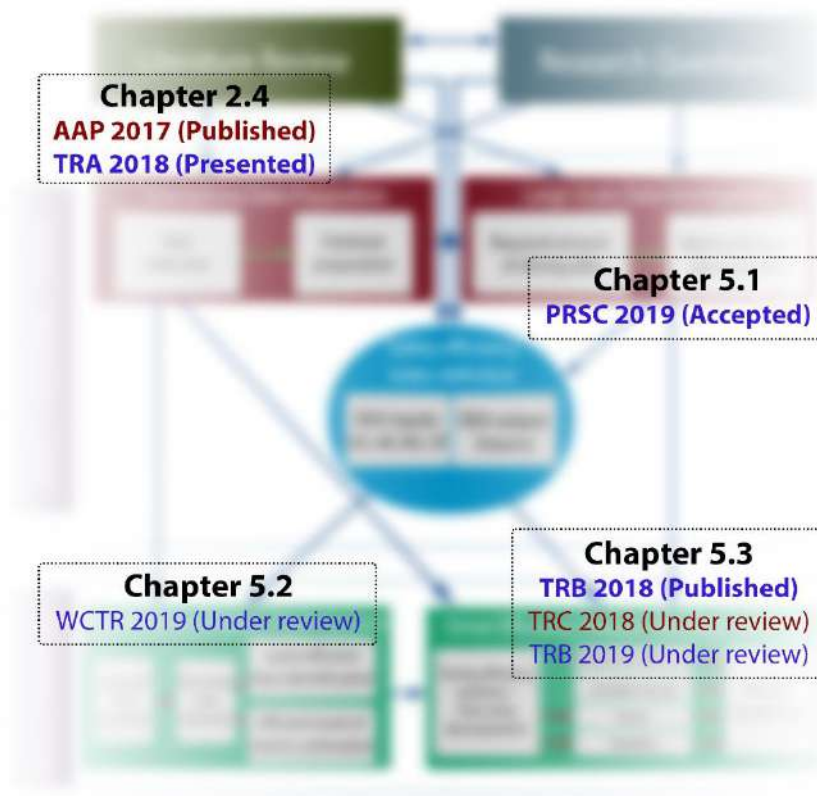

Figure 1: PhD related papers

Tables 1, 2 and 3 illustrate the publications produced thus far regardless of the relation to the specific PhD to scientific journals, international conferences and Greek conferences, respectively. Table 4 illustrates all conference presentations made without full paper review.

Table 1: List of scientific journal publications

Year	Role	Name	Journal	Status	Chapter
2018	Author	"Public opinion on usage-based motor insurance schemes: A stated preference approach"	Travel Behaviour and Society	Published	-
2017	Author	"Innovative motor insurance schemes: a review of current practices and emerging challenges."	Accident Analysis & Prevention	Published	2.2
2017	Co-author	"Road, Traffic, and Human Factors of Pedestrian Crossing Behavior: Integrated Choice and Latent Variables Models."	Transportation Research Record	Published	-
2014	Co-author	"Factors Influencing Freeway Traffic Upstream of an Incident"	Advances in Transportation Studies	Published	-
2014	Author	"Improving short-term traffic forecasts: to combine models or not to combine?"	IET Intelligent Transport Systems	Published	-
2018	Co-author	"Detecting road safety offenders through smartphone data: The case of mobile phone use while driving"	Safety Science	Under Review	-
2018	Author	"Driving Efficiency Benchmarking Using Smartphone Data"	Transportation Research Part C: Emerging Technologies	Under Review	5.3

Table 2: List of international conference publications

Year	Role	Name	Conference	Status	Chapter
2018	Co-author	"Exploring Weather Effects on Powered-Two-Wheeler Safety on Urban Arterials in Athens".	hEART	Published	-
2018	Co-author	"Analysis of Driving Behaviour Characteristics on the basis of Smartphone Data."	Transport Research Arena	Published	-
2018	Author	"Comparative Evaluation of Driving Efficiency Using Smartphone Data."	Transportation Research Board	Published	5.3
2017	Co-author	"Monitoring distraction through smartphone naturalistic driving experiment"	International Naturalistic Driving Research Symposium	Published	-
2017	Co-author	"Willingness - to - Pay for Usage-Based Motor Insurance"	Transportation Research Board	Published	-
2017	Co-author	"Συσχέτιση Χαρακτηριστικών και Επιδόσεων Ασφάλειας του Οδηγού"	International Congress on Transportation Research in Greece	Published	-
2017	Co-author	"Συσχέτιση Δεδηλωμένης και Αποκαλυφθείσας Συμπεριφοράς του Οδηγού με Χρήση των Διαγνωστικών Στοιχείων του Οχήματος"	International Congress on Transportation Research in Greece	Published	-
2016	Author	"About new innovative insurance schemes: pay as/ how you drive." Transportation Research Procedia, 14, 362-371.	Transport Research Arena	Published	2.2
2016	Co-author	"Star rating driver traffic and safety behavior through OBD and smartphone data collection."	International Conference on Artificial Intelligence and Intelligent Transport Systems	Published	-
2016	Co-author	"Road, Traffic, and Human Factors of Pedestria3n Crossing Behavior: Integrated Choice and Latent Variable Models."	Transportation Research Board	Published	-
2019	Author	"Investigating the Temporal Evolution of Driving Safety Efficiency Using Data Collected from Smartphone Sensors"	Transportation Research Board	Published	5.3
2019	Co-author	"Quantifying the Necessary Amount of Driving Data for Driving Behavior Assessment"	Transportation Research Board	Published	-

2019	Author	"Hybrid Data Envelopment Analysis for Large-Scale Smartphone Data Modeling"	World Conference on Transport Research	Under review	5.2
2019	Author	"Investigation of the correlation between stated and revealed driving behaviour using data collected from on-board diagnostics (OBD) devices"	World Conference on Transport Research	Under review	-
2019	Author	"Investigating the Correlation between Driver's Characteristics and Safety Performance"	World Conference on Transport Research	Under review	-
2019	Author	"Driving speed model development using driving data obtained from smartphone sensors"	World Conference on Transport Research	Under review	-

Table 3: List of Greek conference publications

Year	Role	Name	Conference	Status	Chapter
2019	Author	"Quantifying the Need for Driving Data Collection in Driving Behaviour Assessment Using Smartphone Data"	PanHellenic Road Safety Conference	Accepted	5.1
2019	Author	"Harsh manoeuvers investigation using naturalistic driving experiment data collected from smartphone devices"	PanHellenic Road Safety Conference	Accepted	-

Table 4: List of international conference presentations

Year	Role	Name	Published in	Chapter
2016	Co-author	"Willingness to pay for innovative vehicle insurance schemes."	World Conference on Injury Prevention and Safety Promotion	-
2015	Co-author	"Bicycle Traffic Rules Survey."	Meeting of the international traffic safety data and analysis group (IRTAD)	-
2015	Co-author	"Star rating driver traffic and safety behaviour through OBD and smartphone data collection."	International Symposium on Road Safety Behaviour Measurements and Indicators	-

Table 5 illustrates the list of journals at which the author is a reviewer.

Table 5: List of reviewing journals

Since	Role	Journal
2016	Reviewer	**PROMET – Traffic & Transportation** Scientific Journal on Traffic and Transportation Research
2018	Reviewer	**Transportation Research Part C: Emerging Technologies** An International Journal
2018	Reviewer	**IET Intelligent Transport Systems**
2018	Reviewer	**Transportation Research Record:** Journal of the Transportation Research Board

Appendix II

The questionnaire administered to a proportion of the participants of this research is illustrated below. It is highlighted that the questionnaires were collected by OSeven, they were connected with user-IDs and the final questionnaire data were provided for this research in an anonymized format.

oseven

Driving Behaviour Questionnaire

Vehicle

* 6. The vehicle you normally use:

- ○ Belongs to you
- ○ Belongs to another member of your family
- ○ Is leased by you
- ○ Is a corporate vehicle
- ○ Other

* 7. What is the engine capacity of the vehicle?

- ○ < 1000cc
- ○ 1001 - 1200cc
- ○ 1201 - 1400cc
- ○ 1401 - 1600cc
- ○ 1601 - 1800cc
- ○ 1801 - 2000cc
- ○ > 2000cc
- ○ I don't know

* 8. What is the age of the vehicle?

- ○ < 5 years
- ○ 5 - 10 years
- ○ 10 - 15 years
- ○ > 15 years
- ○ I don't know

* 9. The insurance of the vehicle is:

- ○ Third-party insurance (covers only third-party injuries/damages)
- ○ Comprehensive insurance (covers injuries/damages of all parties)
- ○ I don't know

* 10. How do you usually choose your insurance company?

- ○ Agent
- ○ Online
- ○ Other
- ○ I don't know

220

* 11. How much is your gas consumption while driving the vehicle?

○ < 5 lt/100km

○ 5 - 7 lt/100km

○ 7 - 9 lt/100km

○ 9 - 12 lt/100km

○ 12 - 15 lt/100km

○ > 15 lt/100km

○ I don't know

Prev   **Next**

oseven

Driving Behaviour Questionnaire

Driving Behaviour

* 12. The following questions concern your accident history

	to date	during the last 3 years
In how many accidents have you been involved as a driver (either with your responsibility or not)?	[ ]	[ ]
In how many accidents have you been involved as a driver with injury/injuries of any of the drivers/passengers (either with your responsibility or not)?	[ ]	[ ]
In how many accidents have you been involved as a driver with material damage only (either with your responsibility or not)?	[ ]	[ ]

* 13. How many driving fines have you received for road traffic violations during the last 3 years?

- ○ 0
- ○ 1
- ○ 2
- ○ 3
- ○ > 3

* 14. Please characterize your driving behaviour based on the following statements:

	Never	Rarely	Sometimes	Often	Always
I drive over the speed limits	●	●	●	●	●
I brake harshly	○	○	○	○	○
I accelerate harshly	●	●	●	●	●
I turn harshly	○	○	○	○	○
I use my mobile phone while driving	●	●	●	●	●

* 15. How much do you respect speed limits while driving on a:

	1 = Not at all	2	3	4	5 = Very much
Highway	●	●	●	●	●
Rural road	○	○	○	○	○
Urban road	●	●	●	●	●

223

* 16. Please characterize yourself as a driver:

	1 = Not at all	2	3	4	5 = Very much
How skillful driver do you think you are?	●	●	●	●	●
How cautious driver do you think you are?	○	○	○	○	○
How aggressive driver do you think you are?	●	●	●	●	●

* 17. Are the OSeven scores representative of your driving behaviour?

	1 = Not at all	2	3	4	5 = Very much
Overall score	●	●	●	●	●
Speeding	○	○	○	○	○
Mobile use	●	●	●	●	●
Braking	○	○	○	○	○
Accelerating	●	●	●	●	●

* 18. How much has your driving behaviour been improved by using the OSeven app?

	1 = Not at all	2	3	4	5 = Very much
Overall	●	●	●	●	●
Speeding	○	○	○	○	○
Mobile use	●	●	●	●	●
Braking	○	○	○	○	○
Accelerating	●	●	●	●	●

Prev   Next

Powered by
SurveyMonkey
See how easy it is to create a survey.

oseven
DRIVING FORWARD>

Driving Behaviour Questionnaire

Demographics

* 19. Country of residence:

[   ]

* 20. Gender:
○ Male
○ Female
○ I prefer not to answer

* 21. Age:

[   ]

* 22. Marital status:
○ Single/Never married
○ Married
○ Widowed
○ Divorced
○ Separated
○ Other

* 23. Number of children:
○ 0
○ 1
○ 2
○ > 2

225

* 24. Education:

○ Primary School
○ Secondary School
● University Degree
○ Technical Degree
○ Postgraduate Degree (MSc, MEng etc.)
○ PhD
○ Other

* 25. Profession:

○ Private Sector employee
○ Civil Servant
○ Freelancer
○ Entrepreneur
○ Technician
○ Retired
○ Student
○ Unemployed
○ Housework
○ Other

* 26. How familiar are you with SmartPhone Apps?

1 = Not at all	2	3	4	5 = Very much
●	●	●	●	●

* 27. How familiar are you with the Internet?

1 = Not at all	2	3	4	5 = Very much
●	●	●	●	●

Prev   Done

Powered by
SurveyMonkey
See how easy it is to create a survey.

225